Lecture Notes in Networks and Systems 235

Xin-She Yang
Simon Sherratt
Nilanjan Dey
Amit Joshi *Editors*

Proceedings of Sixth International Congress on Information and Communication Technology

ICICT 2021, London, Volume 1

Springer

Lecture Notes in Networks and Systems

Volume 235

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/15179

Xin-She Yang · Simon Sherratt · Nilanjan Dey · Amit Joshi
Editors

Proceedings of Sixth International Congress on Information and Communication Technology

ICICT 2021, London, Volume 1

Springer

Editors
Xin-She Yang
Middlesex University
London, UK

Simon Sherratt
University of Reading
Reading, UK

Nilanjan Dey
JIS University
Kolkata, India

Amit Joshi
Global Knowledge Research Foundation
Ahmedabad, India

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-16-2376-9 ISBN 978-981-16-2377-6 (eBook)
https://doi.org/10.1007/978-981-16-2377-6

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The Sixth International Congress on Information and Communication Technology was held on February 25–26, 2021, digitally on ZOOM and was organized by Global Knowledge Research Foundation. The associated partners were Springer, SPRINGER NATURE, and InterYIT IFIP. The conference provided a useful and wide platform both for display of the latest research and for exchange of research results and thoughts. The participants of the conference were from almost every part of the world (around 85 countries), with background of either academia or industry, allowing a real multinational multicultural exchange of experiences and ideas.

A total of 1150 papers were received for this conference from across 83 countries, among which around 350 papers were accepted and were presented on the digital platform. Due to overwhelming response, we had to drop many papers in hierarchy of the quality. Totally, 51 technical sessions were organized in parallel in 2 days and talks were given on both the days. The conference involved deep discussion and issues which are intended to be solved at global levels. New technologies were proposed, experiences were shared, and future solutions for design infrastructure for ICT were also discussed. The total papers will be published in 4 volumes of proceedings among which this is one.

The conference consisted of several distinguished authors, scholars, and speakers from all over the world. Amit Joshi, organizing Secretary, ICICT 2021, Sean Holmes, Vice Dean International, College of Business, Arts and Social Sciences, Brunel University London, UK, Mike Hinchey, Immd. Past Chair –IEEE UK and Ireland section & Director of Lero and Professor - Software Engineering, University of Limerick, Ireland, Aninda Bose, Sr. Publishing Editor, Springer Nature, Germany, Xin-She Yang, Professor, Middlesex University, Prof. Jyoti Choudri, Professor, University of Hertfordshire, and many were a part of the Inaugural Session and the conference.

The conference was organized and conceptualized with collective efforts of a large number of individuals. We would like to thank our committee members and the reviewers for their excellent work in reviewing the papers Grateful acknowledgements are extended to the team of Global Knowledge Research Foundation for their

valuable efforts and support. We are also thankful to the sponsors, press, print, and electronic media for their excellent coverage of this conference.

London, UK — Xin-She Yang
Reading, UK — Simon Sherratt
Kolkata, India — Nilanjan Dey
Ahmedabad, India — Amit Joshi

Contents

Editors and Contributors

About the Editors

Xin-She Yang obtained his D.Phil. in Applied Mathematics from the University of Oxford and subsequently worked at the Cambridge University and the National Physical Laboratory (UK) as Senior Research Scientist. He is currently Reader in Modelling and Optimization at Middlesex University London and Adjunct Professor at Reykjavik University (Iceland). He is also Elected Bye-Fellow at the Cambridge University and IEEE CIS Chair for the Task Force on Business Intelligence and Knowledge Management. He was included in the "2016 Thomson Reuters Highly Cited Researchers" list.

Simon Sherratt was born near Liverpool, England, in 1969. He is currently Professor of Biosensors at the Department of Biomedical Engineering, University of Reading, UK. His main research area is signal processing and personal communications in consumer devices, focusing on wearable devices and health care. Professor Sherratt received the 1st place IEEE Chester Sall Memorial Award in 2006, the 2nd place in 2016 and the 3rd place in 2017.

Nilanjan Dey is an Associate Professor in the Department of Computer Science and Engineering, JIS University, Kolkata, India. He has authored/edited more than 75 books with Springer, Elsevier, Wiley and CRC Press and published more than 300 peer-reviewed research papers. Dr. Dey is Editor-in-Chief of the International Journal of Ambient Computing and Intelligence; Series Co-Editor of Springer Tracts in Nature-Inspired Computing (STNIC); and Series Co-Editor of Advances in Ubiquitous Sensing Applications for Healthcare, Elsevier.

Amit Joshi is Director of the Global Knowledge Research Foundation and the International Chair of InterYIT at the International Federation of Information Processing (IFIP, Austria). He has edited more than 40 books for Springer, ACM and other reputed publishers. He has also organized more than 50 national and international conferences and workshops in association with the ACM, Springer and IEEE in, e.g. India, Thailand and the UK.

Contributors

Anthony Abdelmassih University St. Joseph (USJ), Beirut, Lebanon

Noreddine Abghour Faculty of Science, University Hassan II, Casablanca, Morocco

Abdelhafid Abouaissa University of Haute-Alsace, Mulhouse, France

Pradeep Achan Amrita Technologies, Amritapuri, India

Ryohei Adachi Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Department of Oral Heal, Kobe Tokiwa College, Kobe, Japan

Fatima Adilova Institute of Mathematics, Tashkent, Uzbekistan

Francis Afedzie Kwofie Fujian Key Laboratory of Automotive Electronics and Electric Drive, Fujian University of Technology, Fujian, China

Alyaa Abdulameer Ahmed Statistics Department, College of Administration & Economic, University of Baghdad, Baghdad, Iraq

Diane Ahrens Technology Campus Grafenau, Deggendorf Institute of Technology, Grafenau, Germany

Mohammed Aissaoui National School of Applied Sciences, Mohammed Premier University, Oujda, Morocco

Sara Ait-Lamallam School of Geomatics and Surveying Engineering, Hassan II Institute of Agriculture and Veterinary Medicine, Rabat, Morocco

Hans Akkermans Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Iman Amer Hameed Al-dahhan Continuing Education Center/ University of Baghdad, Baghdad, Iraq

Nadia Al-Ghreimil College of Computer and Information Sciences, King Saud University, Riyadh, KSA, Saudi Arabia

A. Al-Khowarizmi Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia

Arnulfo Alanís Tecnológico Nacional de México, Instituto Tecnológico de Tijuana, Mexico City, Mexico

Bushra Alhijawi Princess Sumaya University for Technology, Amman, Jordan

Adis Alihodzic University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Aljawharah Almajyul College of Computer Science and Information Technology, Shaqra University, Shaqra, KSA, Saudi Arabia

Sagaya Amalathas University of Southampton, Nusajaya, Johore, Malaysia

Ana Carla Amaro Department of Communication and Art, University of Aveiro, Aveiro, Portugal

Khikmat Anvarov Republican Scientific Centre for Emergency Medicine, Tashkent, Uzbekistan

Irina Arhipova Latvia University of Life Sciences and Technologies, Jelgava, LV, Latvia

Marwan Abdul Hameed Ashour Statistics Department, College of Administration & Economic, University of Baghdad, Baghdad, Iraq

Gerber Aurona Department of Informatics, University of Pretoria, Pretoria, South Africa;
Centre for AI Research (CAIR), Pretoria, South Africa

Maya Avikhana Intelligent Decisions Technologies Systems, Ashdod, Israel

Arafat Awajan Princess Sumaya University for Technology, Amman, Jordan;
Mutah University, Karak, Jordan

Georgia Ayfantopoulou Hellenic Institute of Transport/Centre for Research and Technology Hellas (CERTH/HIT), Thessaloniki, Greece

André Baart Utrecht, The Netherlands

Vitalii Babenko Igor Sikorsky Kyiv Polytechnic Institute, National Technical University of Ukraine, Kyiv, Ukraine

Aliou Badji LI3, Assane Seck University of Ziguinchor, Ziguinchor, Senegal

Rommel Barbosa Federal University of Goiás - UFG, Goiânia, Brazil

Taqadus Bashir Bahria University, Islamabad, Pakistan

Fatma Zohra Belkredim Hassiba Benbouali University of Chlef, Chlef, Algeria

Reda Benkhouya Ibn Tofail University, Kenitra, Morocco

Gundars Berzins University of Latvia, Riga, LV, Latvia

Rajashekhar C. Biradar REVA University, Yelahanka, Bengaluru, Karnataka, India

Fabian Blanco Universidad Cooperativa de Colombia, Bogotá, Colombia

Aleksey Bocharov Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia;
Siberian Federal University, Krasnoyarsk, Russia

Rodrigo Bogarin Costa Rica Institute of Technology, Cartago, Costa Rica

Anna Bon Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Nazanin Borhan Victoria University, Sydney, Australia

A. Botha School of Computing, Science Campus, University of South Africa, Roodepoort, South Africa

Abdelkrim Bouramoul MISC Laboratory, Constantine 2 University, Constantine, Algeria

Mohamed-el-Amine Brahmia LINEACT, CESI Engineering School, Strasbourg, France

W. Bvuma Department of Informatics, University of Pretoria, Pretoria, South Africa

Jorge Ramón Fonseca Cacho Department of Computer Science, University of Nevada Las Vegas, Las Vegas, NV, USA

Thi-Minh-Hue Cai TranData - FPT Software, Quy Nhon, Binh Dinh, Vietnam

Wei Cai Department of Enterprise Development, Inspur Group, Jinan, People's Republic of China

Fiona Carroll Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, UK

Alma Casas-Frausto Tecnológico Nacional de México, Instituto Tecnológico de Tijuana, Mexico City, Mexico

Idriss Chana Moulay Ismail University, Meknes, Morocco

Binghuai Chen Guandong Meteorological Data Center, Guangzhou, China

Mohammed Amine Chenouf National School of Applied Sciences, Mohammed Premier University, Oujda, Morocco

Alick Chingapi University of Pretoria, Pretoria, South Africa

Eun-Sun Cho Chungnam National University, Daejeon, South Korea

Jayanti Choudhary National Institute of Technology Patna, Patna, India

Hassan Chreim Dassault Systèmes, Vélizy-Villacoublay, France

Papa Alioune Cisse LI3, Assane Seck University of Ziguinchor, Ziguinchor, Senegal

António Leça Coelho Laboratório Nacional de Engenharia Civil, Lisboa, Portugal

Ronaldo Costa Federal University of Goiás - UFG, Goiânia, Brazil

Jeffrey Cummings Department of Brain Health, University of Nevada Las Vegas, Las Vegas, NV, USA

Farhad Daneshgar Victoria University, Sydney, Australia

Minh-Hung Dang TranData - FPT Software, Quy Nhon, Binh Dinh, Vietnam

Quang-Vinh Dang Industrial University of Ho Chi Minh city, Ho Chi Minh City, Vietnam

Tran Hoang Chau Dao School of Computer Science and Engineering, International University, VNU-HCMC, Ho Chi Minh City, Vietnam

Victor de Boer Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Diego de Castro Rodrigues Federal University of Goiás - UFG, Goiânia, Brazil

Ruben D'Hauwers IMEC-SMIT-VUB, Brussel, Belgium

Sead Delalić Info Studio d.o.o. Sarajevo, Sarajevo, Bosnia and Herzegovina;
Faculty of Science, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Youssou Dieng LI3, Assane Seck University of Ziguinchor, Ziguinchor, Senegal

Ibrahima Diop LI3, Assane Seck University of Ziguinchor, Ziguinchor, Senegal

Boubacar Diouf Enda-Santé Ziguinchor, Ziguinchor, Senegal

Francis Dittoh University for Development Studies, Tamale, Ghana;
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Jean-François Dollinger LINEACT, CESI Engineering School, Strasbourg, France

Dženana Đonko Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Omar Doukari CESI Engineers School of Nanterre, Paris, France

Kohei Dozono Taylor's University, Subang Jaya, Malaysia

Qiang Duan Inspur Technology Center, Inspur Group, Jinan, People's Republic of China;
Shandong Institute of New Generation Information Industry Technology Co., Ltd., Jinan, People's Republic of China

Xiaohui Duan Department of Enterprise Development, Inspur Group, Jinan, People's Republic of China

Martin Eichler National Center for Tumor Diseases, University Hospital Dresden, Dresden, Germany

Georges Zakka El Nashef Saint-Joseph University of Beirut, Mar Roukoz, Lebanon

Amina Elomri Faculty of Science, University Hassan II, Casablanca, Morocco

Daniel C. Elton Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD, USA

Marischa Elveny Universitas Sumatera Utara, Medan, Indonesia

Lina Fabri Intelligent Decisions Technologies Systems, Ashdod, Israel

Rafic Faddoul University St. Joseph (USJ), Beirut, Lebanon

Luoluo Feng Inspur Technology Center, Inspur Group, Jinan, People's Republic of China;
Shandong Institute of New Generation Information Industry Technology Co., Ltd., Jinan, People's Republic of China

Noussair Fikri Faculty of Science, University Hassan II, Casablanca, Morocco

Matthes Fleck Institute of Communication and Marketing, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland

Ingo Gächter Institute of Communication and Marketing, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland

Sabrina Gado Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany

Ming Gao Baidu Knows Business Department, Baidu Science and Technology Park, Jinan, People's Republic of China

Richard García Universidad Cooperativa de Colombia, Bogotá, Colombia

Fadi Geara University St. Joseph (USJ), Beirut, Lebanon

D. D. Geetha REVA University, Yelahanka, Bengaluru, Karnataka, India

Jürgen Geiser The Institute of Theoretical Electrical Engineering, Ruhr University of Bochum, Bochum, Germany;
Ruhr University of Bochum, The Institute of Theoretical Electrical Engineering, Bochum, Germany

Christoph Glauser IFAAR, Bern, Switzerland

Ramiro Gonçalves UTAD (Universidade de Trás-Os-Montes e Alto Douro), Vila Real, Portugal

Lingqi Guo State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

Mohammad Hajiketabi Department of Applied Mathematics, Imam Khomeini International University, Qazvin, Iran

Damir Hasanspahic University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Tokutaka Hasegawa Nagoya Institute of Technology, Nagoya, Japan

Najva Hassan Department of Electrical and Electronics Engineering, Government Engineering College (Affiliated to APJ Abdul Kalam Technological University), Thrissur, Kerala, India

M. J. Hattingh Department of Informatics, University of Pretoria, Pretoria, South Africa

Andreas Hein Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

Leopold Hentschel National Center for Tumor Diseases, University Hospital Dresden, Dresden, Germany

Yessica Herran Universidad Cooperativa de Colombia, Bogotá, Colombia

Chaminda Hewage Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, UK

Naisi Hua School of Architecture, Harbin Institute of Technology, Harbin, China

Guiye Huang Guandong Meteorological Data Center, Guangzhou, China

Viktoria Iakymchuk Igor Sikorsky Kyiv Polytechnic Institute, National Technical University of Ukraine, Kyiv, Ukraine

Lhassane Idoumghar University of Haute-Alsace, Mulhouse, France

Alisher Ikramov Institute of Mathematics, Tashkent, Uzbekistan

Mauro Sebastián Innocente Autonomous Vehicles & Artificial Intelligence Laboratory (AVAILab), Coventry University, Coventry, UK

Ngomane Issah Department of Computer Science, University of Limpopo, Limpopo, South Africa

Dietmar Jakob Technology Campus Grafenau, Deggendorf Institute of Technology, Grafenau, Germany

Liga Jankovska WeAreDots Ltd., Riga, LV, Latvia

Doris Janssen Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany

Samantha Jiménez Tecnológico Nacional de México, Instituto Tecnológico de Tijuana, Mexico City, Mexico

Rize Jin Tiangong University, Tianjin, China

Wilmar Junior University of Western São Paulo - UNOESTE, São Paulo, Brazil

Mina Esmail Zadeh Nojoo Kambar Department of Computer Science, University of Nevada Las Vegas, Las Vegas, NV, USA

Ajanta Kancharla Mahindra Ecole Centrale, Hyderabad, India

Mohamed Raouf Kanfoud MISC Laboratory, Constantine 2 University, Constantine, Algeria

Seoyeon Kang Chungnam National University, Daejeon, South Korea

Theodoros Karakasidis Physics Department of Civil Engineering, University of Thessaly, Lamia, Greece

Jakob Kasbauer Technology Campus Grafenau, Deggendorf Institute of Technology, Grafenau, Germany

Bannaka Kenya Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Organization for the Advancement of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Department of Oral Heal, Kobe Tokiwa College, Kobe, Japan

Mikhail Vladimirovich Khachaturyan Bauman Moscow State Technical University, Moscow, Russia

Abduhakim Khadjibaev Republican Scientific Centre for Emergency Medicine, Tashkent, Uzbekistan

Imran khan Department of Communication and Art, University of Aveiro, Aveiro, Portugal

Ben Khayut Intelligent Decisions Technologies Systems, Ashdod, Israel

Driss Khebbou Moulay Ismail University, Meknes, Morocco

GeunYong Kim Chungnam National University, Daejeon, South Korea

Masashi Kimura Convergence Lab. Inc., 1-9-5 #806 Kanda-awajicho, Chiyoda-ku, Tokyo, Japan

Evgeniia Valeryevna Klicheva Plekhanov Russian University of Economics, Moscow, Russia

Dennis Knopf Seracom GmbH, Stuttgart, Germany

Hideyuki Kobayashi National Institute of Technology, Miyagi, Japan

Yasuhiro Kozaki Faculty of Education, Osaka Kyoiku University, Osaka, Japan

Martin Kubli aiaibot, Zurich, Switzerland

Ravi Ranjan Kumar National Institute of Technology Patna, Patna, India

Sergei Kurashkin Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

Lyudmila Kushnina Perm National Research Polytechnic University, Perm, Russia

Daniel Lai Multimedia University, Melaka, Malaysia

Peter Lambert IDLab, Department of Electronics and Information Systems, Ghent University – imec, Ghent, Belgium

Garam Lee Department of Brain Health, University of Nevada Las Vegas, Las Vegas, NV, USA

Sook Ling Lew Multimedia University, Melaka, Malaysia

Xiaoxu Liang Polytechnic University of Turin, Turin, Italy

Lyuchao Liao Fujian Key Laboratory of Automotive Electronics and Electric Drive, Fujian University of Technology, Fujian, China

Jianyong Li Guandong Meteorological Data Center, Guangzhou, China

Nan Li Roche Diagnostics, Rotkreuz, Switzerland

Rui Li Inspur Technology Center, Inspur Group, Jinan, People's Republic of China; Shandong Institute of New Generation Information Industry Technology Co., Ltd., Jinan, People's Republic of China

Xue Li Inspur Technology Center, Inspur Group, Jinan, People's Republic of China; Shandong Institute of New Generation Information Industry Technology Co., Ltd., Jinan, People's Republic of China

Márcio Lima Federal University of Goiás - UFG, Goiânia, Brazil

Katharina Lingelbach Institute of Human Factors and Technology Management (IAT), University of Stuttgart, Stuttgart, Germany

Tarmo Lipping Tampere University, Pori, Finland

Mourad Loukam Hassiba Benbouali University of Chlef, Chlef, Algeria

Milen Loukantchevsky University of Ruse, Ruse, Bulgaria

Guang Lu Institute of Communication and Marketing, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland

Khurram Mahmood Bahria University, Islamabad, Pakistan

N. R. Makama School of Computing, Science Campus, University of South Africa, Roodepoort, South Africa

Asmaa Mansour Khoudja Hassiba Benbouali University of Chlef, Chlef, Algeria

Hannes Mareen IDLab, Department of Electronics and Information Systems, Ghent University – imec, Ghent, Belgium

Flavia Marku Computer Engineering, Faculty of Information Technology Polytechnic University of Tirana, Tirane, Albania

Bogart Yail Márquez Tecnológico Nacional de México, Instituto Tecnológico de Tijuana, Mexico City, Mexico

S. W. Mason Leeds, UK

Mikhail O. Matveev Moscow State Institute of International Relations (MGIMO University), Moscow, Russia;
I.M. Sechenov First Moscow State Medical University, Trubetskaya, Moscow, Russia

Oleksandr Matviichuk Igor Sikorsky Kyiv Polytechnic Institute, National Technical University of Ukraine, Kyiv, Ukraine

Owais Mehmood Omnicom Balfour Beatty, York, UK

Usha Mehta Institute of Technology, Nirma University, Ahmedabad, Gujarat, India

Ana Melro Department of Communication and Art, University of Aveiro, Aveiro, Portugal

Stephan Menzel Department of Molecular Haematology, King's College, Strand, London, UK

Rayan Mina Saint-Joseph University of Beirut, Mar Roukoz, Lebanon

Kenichiro Mitsunari Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Faculty of Education, Kobe Tokiwa University, Kobe, Japan

Richard Moist Institute of Communication and Marketing, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland

A. Jorge Morais Universidade Aberta, Lisboa, Portugal;
LIAAD - INESC TEC, Porto, Portugal

Maria Morfoulaki Hellenic Institute of Transport/Centre for Research and Technology Hellas (CERTH/HIT), Thessaloniki, Greece

Khalid Moussaid Faculty of Science, University Hassan II, Casablanca, Morocco

Velempini Mthulisi Department of Computer Science, University of Limpopo, Limpopo, South Africa

Alexander Muenzberg Carl von Ossietzky University of Oldenburg, Oldenburg, Germany;
University of Applied Sciences Kaiserslautern, Kaiserslautern, Germany

Abdullah Mujahid Computational Engineering, Ruhr University of Bochum, Bochum, Germany

Braulio Murillo Pontifica Universidad Católica del Perú, Lima, Peru

Aleksandr Murygin Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

Glykeria Myrovali Hellenic Institute of Transport/Centre for Research and Technology Hellas (CERTH/HIT), Thessaloniki, Greece

Pouyan Nahed Department of Computer Science, University of Nevada Las Vegas, Las Vegas, NV, USA

Namrata Nair Amrita Technologies, Amritapuri, India;
Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Tomoe Nakagawa Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Academic Development Division, Kobe Tokiwa University, Kobe, Japan

Yasuo Nakata Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Organization for the Advancement of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Faculty of Health Sciences, Kobe Tokiwa University, Kobe, Japan

Sankaran Narayanan Amrita Technologies, Amritapuri, India;
Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Ievgen Nastenko Igor Sikorsky Kyiv Polytechnic Institute, National Technical University of Ukraine, Kyiv, Ukraine

Liqaa F. Nawaf Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, UK

Zainab Nayyar Bahria University, Islamabad, Pakistan

Konstantin Nechval Transport and Telecommunication Institute, Riga, Latvia

Nicholas Nechval University of Latvia, Riga, LV, Latvia

Joaquim Neto Universidade Aberta, Lisboa, Portugal;
Laboratório Nacional de Engenharia Civil, Lisboa, Portugal

Huu-Nghia Nguyen TranData - FPT Software, Quy Nhon, Binh Dinh, Vietnam

Thi Thanh Sang Nguyen School of Computer Science and Engineering, International University, VNU-HCMC, Ho Chi Minh City, Vietnam

Ikuhiro Noda Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Organization for the Advancement of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Academic Development Division, Kobe Tokiwa University, Kobe, Japan

Olena Nosovets Igor Sikorsky Kyiv Polytechnic Institute, National Technical University of Ukraine, Kyiv, Ukraine

Konstantinos Oikonomou Department of Informatics, King's College, Strand, London, UK

Petri Oksa Tampere University, Pori, Finland

Frederico Oliveira Federal University of Goiás - UFG, Goiânia, Brazil

Lídia Oliveira Department of Communication and Art, University of Aveiro, Aveiro, Portugal

Masato Omori Faculty of Education, Kobe Tokiwa University, Kobe, Japan

Shih Yin Ooi Multimedia University, Melaka, Malaysia

Lucas Osco University of Western São Paulo - UNOESTE, São Paulo, Brazil

Joon-Young Paik Tiangong University, Tianjin, China

Rosanna Palma Pontifica Universidad Católica del Perú, Lima, Peru

Liga Paura Latvia University of Life Sciences and Technologies, Jelgava, LV, Latvia

Volodymyr Pavlov Igor Sikorsky Kyiv Polytechnic Institute, National Technical University of Ukraine, Kyiv, Ukraine

Shrikant Pawar School of Medicine, Genetics Department, Yale University, New Haven, CT, USA

René Peinl Institute of Information Systems, Hof University of Applied Sciences, Hof, Germany

Matthias Peissner Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany

Irina Perlova Department of Foreign Languages, Linguistics and Translation, Perm National Research Polytechnic University, Perm, Russia

Kristina Permiakova Perm National Research Polytechnic University, Perm, Russia

Thanh-Duyen Phan TranData - FPT Software, Quy Nhon, Binh Dinh, Vietnam

Daniela Piechnik Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany

Jayesh Popat Institute of Technology, Nirma University, Ahmedabad, Gujarat, India

José Antonio Pow-Sang Pontifica Universidad Católica del Perú, Lima, Peru

Qi Qi State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

N. Raghu Kishore Mahindra Ecole Centrale, Hyderabad, India

Rajip Raj Victoria University, Sydney, Australia

Franz Reisewitz University of Applied Sciences Kaiserslautern, Kaiserslautern, Germany

Mohamed Rida Faculty of Science, University Hassan II, Casablanca, Morocco

Norbert Roesch University of Applied Sciences Kaiserslautern, Kaiserslautern, Germany

Abdul Saleem Department of Electrical and Electronics Engineering, Government Engineering College (Affiliated to APJ Abdul Kalam Technological University), Thrissur, Kerala, India

Tero Salminen Tampere University, Pori, Finland

Ravan Saravanan RC Buminiaga Sdn Bhd, Petaling Jaya, Malaysia

Janina Sauer Carl von Ossietzky University of Oldenburg, Oldenburg, Germany; University of Applied Sciences Kaiserslautern, Kaiserslautern, Germany

Jacques Savoy University of Neuchatel, Neuchatel, Switzerland

Loris Schmid IFAAR, Bern, Switzerland;
University of Neuchatel, Neuchatel, Switzerland

Markus Schuler National Center for Tumor Diseases, University Hospital Dresden, Dresden, Germany

Imane Sebari School of Geomatics and Surveying Engineering, Hassan II Institute of Agriculture and Veterinary Medicine, Rabat, Morocco

Mapunya Sekgoari Semaka Department of Computer Science, University of Limpopo, Limpopo, South Africa

Yuriy Seregin Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

Daniel Sernatinger Seracom GmbH, Stuttgart, Germany

Ilya Shilov ITMO University, Saint Petersburg, Russia

Shun Shiramatsu Nagoya Institute of Technology, Nagoya, Japan

Fredys Simanca Universidad Cooperativa de Colombia, Bogotá, Colombia

Yana Ivanovna Sipovskaya Institute of Psychology of the Russian Academy of Sciences, Moscow, Russia

Vilson Siqueira Federal Institute of Tocantins, Palmas, Brazil;
Federal University of Goiás - UFG, Goiânia, Brazil

K. P. Soman Amrita Technologies, Amritapuri, India;
Amrita Center for Computational Engineering and Networking (CEN), Amrita Vishwa Vidyapeetham, Coimbatore, India

Aditya Stanam College of Public Health, The University of Iowa, Iowa City, Iowa, USA

Kathleel Steinhöfel Department of Informatics, King's College, Strand, London, UK

Adriana A. Steyn University of Pretoria, Pretoria, South Africa

Haris Šupić Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Haifeng Sun State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

Maksym Suvorov National University of Pharmacy, Kharkiv, Ukraine

Rahmad Syah Universitas Medan Area, Medan, Indonesia

Abdusy Syarif Universitas Nasional, Jakarta, Indonesia

Nikhath Tabassum REVA University, Yelahanka, Bengaluru, Karnataka, India

Igli Tafa Computer Engineering, Faculty of Information Technology Polytechnic University of Tirana, Tirane, Albania

Kazem Taghva Department of Computer Science, University of Nevada Las Vegas, Las Vegas, NV, USA

Jun Jet Tai Autonomous Vehicles & Artificial Intelligence Laboratory (AVAILab), Coventry University, Coventry, UK

Kunihiko Takamatsu Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Organization for the Advancement of Higher Education, Kobe Tokiwa University, Kobe, Japan;
Life Science Center, Kobe Tokiwa University, Kobe, Japan;
Faculty of Health Sciences, Kobe Tokiwa University, Kobe, Japan

Kazumoto Tanaka Kindai University, Hiroshima, Japan

Praowpan Tansitpong NIDA Business School, National Institute of Development Administration (NIDA), Bangkok, Thailand

Fabiano Tavares Federal Institute of Tocantins, Palmas, Brazil

Yijin Teng Department of Program Marketing, Inspur Group, Jinan, People's Republic of China

Duc Duy Tran School of Computer Science and Engineering, International University, VNU-HCMC, Ho Chi Minh City, Vietnam

My-Linh Tran Quy Nhon University, Binh Dinh, Vietnam

Thi-Minh-Trang Tran TranData - FPT Software, Quy Nhon, Binh Dinh, Vietnam

Jan Treur Social AI Group, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Marimuthu Trishan Department of Informatics, , University of Pretoria, Pretoria, South Africa

Eva Tuba Singidunum University, Belgrade, Serbia

Milan Tuba State University of Novi Pazar, Novi Pazar, Serbia

Wendelien Tuyp Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Vadim Tynchenko Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia;
Siberian Federal University, Krasnoyarsk, Russia

Makoto Uehara National Institute of Technology, Miyagi, Japan

Merwe Van der Alta Department of Informatics, , University of Pretoria, Pretoria, South Africa

C. J. van Staden School of Computing, Science Campus, University of South Africa, Roodepoort, South Africa

Glenn Van Wallendael IDLab, Department of Electronics and Information Systems, Ghent University – imec, Ghent, Belgium

Denis Vasiliev Turiba University, Riga, Latvia

Gatis Vitols Latvia University of Life Sciences and Technologies, Jelgava, LV, Latvia

Nils Walravens IMEC-SMIT-VUB, Brussel, Belgium

Jianhua Wang Inspur Technology Center, Inspur Group, Jinan, People's Republic of China;
Shandong Institute of New Generation Information Industry Technology Co., Ltd., Jinan, People's Republic of China

Jingyu Wang State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

Yongqiang Wang Fujian Key Laboratory of Automotive Electronics and Electric Drive, Fujian University of Technology, Fujian, China

Sebastian Wilhelm Technology Campus Grafenau, Deggendorf Institute of Technology, Grafenau, Germany

Tobias Wirth Institute of Information Systems, Hof University of Applied Sciences, Hof, Germany

Timotius Witono Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia;

Faculty of Information Technology, Maranatha Christian University, Bandung, Indonesia

Thomas Wozniak Institute of Communication and Marketing, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland

Chenwei Wu Fujian Key Laboratory of Automotive Electronics and Electric Drive, Fujian University of Technology, Fujian, China

Reda Yaagoubi School of Geomatics and Surveying Engineering, Hassan II Institute of Agriculture and Veterinary Medicine, Rabat, Morocco

Setiadi Yazid Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Denis D. Yershov Moscow State Linguistic University, St. Moscow, Russia

Qingshan Yin Inspur Technology Center, Inspur Group, Jinan, People's Republic of China;
Shandong Institute of New Generation Information Industry Technology Co., Ltd., Jinan, People's Republic of China

Enkli Ylli Computer Engineering, Faculty of Information Technology Polytechnic University of Tirana, Tirane, Albania

Danil Zakoldaev ITMO University, Saint Petersburg, Russia

Maolin Zhang Fujian Key Laboratory of Automotive Electronics and Electric Drive, Fujian University of Technology, Fujian, China

Xiaoxiao Zhang School of Law, Humanities and Sociology, Wuhan University of Technology, Wuhan, China

Yu Zhang School of Architecture, Harbin Institute of Technology, Harbin, China

Jing Zhao Shandong Yingxin Computer Technology Co., Ltd., Jinan, People's Republic of China

Yanhan Zhao Department of Enterprise Development, Inspur Group, Jinan, People's Republic of China

Qinqiang Zhou Guandong Meteorological Data Center, Guangzhou, China

A. Zhuravishkin Institute of Linguistics of the Russian Academy of Sciences, Bolshoy Kislovsky Ln, Moscow, Russia

Fumin Zou Fujian Key Laboratory of Automotive Electronics and Electric Drive, Fujian University of Technology, Fujian, China

Emir Žunić Info Studio d.o.o. Sarajevo, Sarajevo, Bosnia and Herzegovina;
Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Techniques of Time Series Modeling in Complex Systems

Shrikant Pawar and Aditya Stanam

Abstract The autoregressive moving average (ARMA) and the autoregressive integrated moving average (ARIMA) are important techniques for time point analysis. An ARMA is a stationary model and by taking a series of differences, it is necessary to apply it to stationarity. Data can be classified as serially correlated using three basic criteria: mean, variance and covariance. Detrending is a strategy of removing the trending components from the time series. Detrending data may be used to see subtrends when the data shows an overall increase. Differencing, SARIMA, Croston's intermittent demand forecasting approach and bagging are other statistical methods developed by unstable decision rules to improve the accuracy of forecasts. Time point forecasting is abbreviated as ETS models or exponential smoothing state-space models considering error, trend and seasonality. Here we try to address time point analysis techniques and parameters in complex systems.

1 Introduction

Analysis and simulation of the time series are critical forecasting methods in complex systems mostly based on robust statistical theory. The autoregressive moving average (ARMA) and the autoregressive integrated moving average (ARIMA) are important techniques for time point analysis [1]. A combined general autoregressive model AR(p) and general moving average model MA(q) have several similarities. AR(p) predicts using dependent value while MA(q) predicts series mean [2]. An ARMA is a stationary model and by taking a series of differences it is necessary to apply it to stationarity. The "I" in the ARIMA model calculates variations to reach a static phase. If the model fails to differentiate, it becomes an ARMA [1, 2]. These models

S. Pawar (✉)
School of Medicine, Genetics Department, Yale University, New Haven, CT, USA
e-mail: Shrikant.pawar@yale.edu

A. Stanam
College of Public Health, The University of Iowa, Iowa City, Iowa, USA
e-mail: aditya-stanam@uiowa.edu

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_1

can be applied to data that is classified as serially correlated using three basic criteria: mean, variance and covariance. The correlated mean needs to be constant [3], and the variance should not be homoscedasticity. We can formulate this series using the formula [4]:

$$P(t) = P(t-1) + Pr(t)$$

where error $Pr(t)$ for time t further can be recursively fit in all the Ps with equation [4]:

$$P(t) = P(0) + Sum(Pr(1), Pr(2), Pr(3), \ldots, Pr(t))$$

Here we try to address time point analysis techniques and parameters in complex systems.

2 Discussion

1. Rho Coefficient

Spearman's rank correlation coefficient (Rho) varies in stationary series. It can be used with equation [5]:

$$P(t) = Rho * P(t-1) + Pr(t)$$

The Rho is usually 0, with a perfectly stationary series, while the current $P(t)$ relies on the last $P(t-1)$ cycles. $P(t)$ is usually calculated where the hypothesis tests the interval between two parameters. Another interpretation of which would be with an AR(1) equation as follows [6]:

$$P(t) = alpha * \; P(t{-}1) + error(t)$$

where $(t-1)$ is utilized to minimize the error function.

2. Detrending

Detrending is a strategy of removing the trending components from the time series. It can be represented by the following equation [7]:

$$x(t) = (mean + trend * t) + error$$

A pattern usually refers to a shift of mean over time to reduce distortion basically. Detrending data may be used to see subtrends when the data shows an overall increase [7]. The only difference between quadratic and linear detrending is that the data follows exponential patterns [8].

3. **Differencing**

This method removes unstable stationarity in data. It can be represented with the following equation [3]:

$$P(t)-P(t-1) = ARMA(s, t)$$

In AR(I)MA returning AIC or BIC value distinction is usually considered as an integration component. The equation conducts a search in the order of constraints given over the possible model. It is highly effective on non-seasonal time series with unintended noise [1]. From the above equation, s is a rank of AR, t is a rank of MA and d is a difference.

4. **Linear and Nonlinear Models**

SARIMA or seasonal ARIMA works best with seasonal pattern time points. Seasonal autoregressive integrated moving average (SARIMA) is an extension of ARIMA that supports univariate component [9]. Three new parameters, autoregression (AR), differencing (I) and moving average (MA) with the seasonality period are added to this model [9]. A typical seasonal pattern in time series utilizing SARIMA is shown in Fig. 1a (Left). Data for plotting plots in this paper were used from the R "forecast" package.

Croston's intermittent demand forecasting approach is also a popular technique involving the use of simple exponential smoothing (SES) on time series. It is assumed that the smoothing parameters of both SES applications are equal. Croston's method is mostly used for refined stochastic analysis. It can predict time using simple exponential smoothing (SES) [10]. Croston's method can be summarized with the following equation, where Aj and Pj are the forecasts with $(j + 1)$th size and demand j [10].

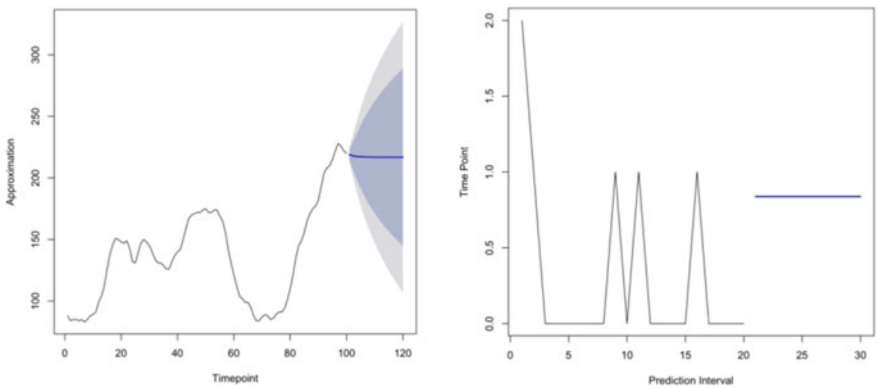


Fig. 1 (Left) **a** Seasonal pattern in time series utilizing SARIMA; (Right) **b** Simple exponential smoothing (SES) using Croston's method

$$Aj = (1 - \alpha)Pj - 1 + \alpha Y * j$$

An example of simple exponential smoothing with prediction interval using Croston's is shown in Fig. 1b (right).

Bagging is another statistical method developed by unstable decision rules to improve the accuracy of forecasts. It is also called bootstrap aggregation, which includes fitting the unregulated model with data approximation producing a large number of bootstrap resamples (Fig. 2a (Left)) and application of decision rule [11]. Removing the instability and decreasing variance requires averaging across resamples. Upon variable selection, bagging can also be applied to the regression model. It is basically an ensemble learning meta-algorithm used to improve the accuracy of predictions and to resolve the overfitting. It starts with a sampling of a training dataset and replacement with some given sample ratio followed by training on a new train set with N boot predictions repeated and averaged [11, 12].

Exponential smoothing forecasts are a tool for the estimation of univariate data in the time series. It's a model where observations sum or lags are predicted. These methods have past observations sum and quantile reduced weights [13]. It is a simple method that does not assume any predefined sequence and also supports seasonal patterns Fig. 2b (right).

Fitting a linear model for time points with trends and seasonality is another important technique (Fig. 3a (left)). It can be easily implemented in R with the function "tslm" [13].

TBATS model or exponential smoothing state-space model with box-cox change (Fig. 3b (right)) utilizes equal handling to accelerate the calculations. The conventional accessor capacities will extricate helpful residuals attributes, where omega is the case cox boundary and phi is the damping boundary, m1–mJ records the occasional cycles utilized in the model and k1–kJ is the quantity of four words utilized for every irregularity [14]. There are a lot more methodologies used to gauge time

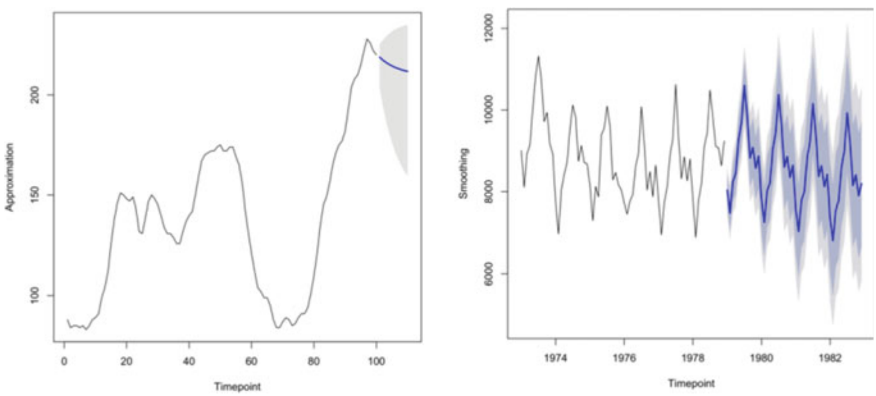


Fig. 2 (Left) **a** Bootstrap resamples with data approximation; (Right) **b** Exponential smoothing forecasts

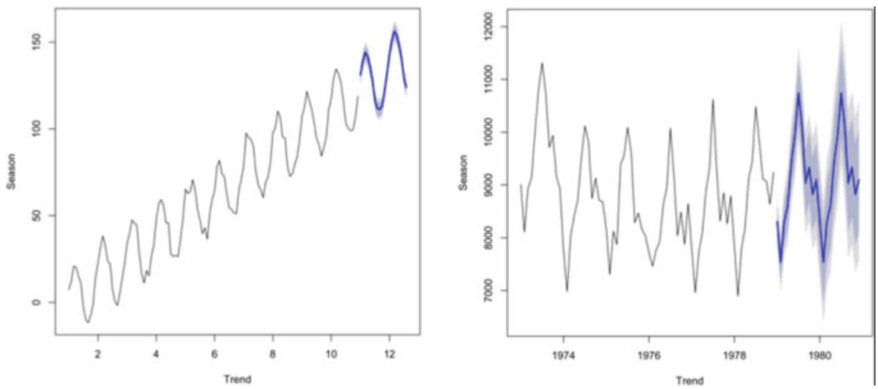


Fig. 3 (Left) **a** Linear fit model for time series; (Right) **b** Exponential smoothing with TBATS model

arrangement results, and they fluctuate broadly in intricacy. As a rule, such models are abridged as ETS models or exponential smoothing state-space models considering error, pattern and irregularity. The ETS models are utilized to demonstrate how a solitary variable will change after some time by recognizing its fundamental patterns and by gauging the impact of past focuses on future time, based on the time between the two focuses [14]. Picking which model to use for information can be a testing approach and requires instinct, just as taking a gander at test measurements. Some significant boundaries to consider while actualizing the above models are the mean mistake or the mean contrast among demonstrated and watched values, root mean squared error, the distinction between the model and the watched qualities and the positive qualities error.

References

1. Swain S et al (2018) Development of an ARIMA model for monthly rainfall forecasting over Khordha District, Odisha, India. In: Recent findings in intelligent computing techniques. Advances in intelligent systems and computing, vol 708, pp 325–331. https://doi.org/10.1007/978-981-10-8636-6_34
2. Hannan EJ (1970) Multiple time series. Wiley series in probability and mathematical statistics. Wiley, New York
3. Shumway S (2011) Time series analysis and its applications. Springer text in statistics. https://doi.org/10.1007/978-1-4419-7865-3
4. Woodward WA, Gray HL, Elliott AC (2017) Applied time series analysis with R. CRC Press. ISBN: 1498734278, 9781498734271
5. Choi SC (1977) Tests of equality of dependent correlation coefficients. Biometrika 64(3):645–647. https://doi.org/10.1093/biomet/64.3.645
6. Altman N, Krzywinski M (2015) Simple linear regression. Nat Methods 12(11):999–1000. ISSN 1548-7091. OCLC 5912005539

7. Wu Z, Huang NE, Long SR, Peng CK (2007) On the trend, detrending, and variability of nonlinear and nonstationary time series. Proc Natl Acad Sci 104(38):14889–14894. https://doi.org/10.1073/pnas.0701020104
8. Baxter M, King RG (1999) Measuring business cycles: approximate band-pass filters for economic time series. Rev Econ Stat 81(4):575–593
9. Brockwell PJ, Davis RA (1991) Time series: theory and methods. Springer Series in statistics. Springer, Berlin
10. Xu Q, Wang N (2012) A review of Croston's method for intermittent demand forecasting. In: Proceedings of 9th international conference on fuzzy systems and knowledge discovery
11. Awajan MA, Ismail MT, Al Wadi S (2018) Improving forecasting accuracy for stock market data using EMD-HW bagging. PLoS ONE 13(7):e0199582. https://doi.org/10.1371/journal.pone.0199582
12. Bergmeir C, Hyndman RJ, Benitez JM (2016) Bagging exponential smoothing methods using STL decomposition and Box-Cox transformation. Int J Forecast 32:303–312
13. Abrami A, Aravkin AY, Kim Y (2017) Time series using exponential smoothing cells. arXiv:1706.02829
14. Naima I, Mahara T, Idrisi AR (2018) Effective short-term forecasting for daily time series with complex seasonal patterns. Procedia Comput Sci 132(2018):1832–1841

Examining the Server's Load Average When Trying to Attack the Firewall

Enkli Ylli, Igli Tafa, and Flavia Marku

Abstract What we have presented in this article has to do with security data. How the firewall provides protection to servers having specific IPs allowed to access a specific port. We have used the Fail2Ban application which applies some rules in the firewall so that it blocks requests from unknown IPs. We create a bash script that sends requests to port 22 also known as the ssh. The communication between the central server and the other ones is made with public keys and there are no passwords. Below we will see how we measure the load average of the server and how the multiple requests affect its work.

Keywords Firewall · Performance · Denial of service

1 Introduction

In our day-to-day life maybe we do not feel the need and importance of securing our information and sensitive data, but that need is increasing more, taking into account the fact that hackers are all around us. Now more than ever, security is everything we should worry about. It is crucial that we can prevent unauthorized access to our online and offline data. This way information remains confidential and secure from hackers. Some security-related terms are firewalls, antivirus, forensics, access control, denial of service, encryption etc.

Firewall definition and what does it do?

Firewall was first used as a term to describe those walls that were more prone to catching fire than the rest of the building. Accordingly, in this stream of thought, a firewall is meant to prevent attacks from known/unknown sources that could be harmful. The primary function of the device is filtering and examining packets.

E. Ylli (✉) · I. Tafa · F. Marku
Computer Engineering, Faculty of Information Technology Polytechnic University of Tirana, Sheshi Nene Tereza, No. 1, Tirane, Albania

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_2

2 Related Works

Reference [1] This article discusses the fact that even though there are major new technologies used to protect data and security has advanced in every aspect known to human, the firewall is still one of our go-to security appliances.

Reference [2] Firewall has mechanisms that allow some IPs to access it and some others that are unauthorized cannot have access. Packet filtering is one of those mechanisms where it analyzes packets depending on their IP address. This is what we do in our experiment as well. We send requests and see the denial of service from a firewall that does not allow these IPs to access the ssh service.

Reference [3] This paper discusses the fact that there are people that try to learn new things that are harmful to society like hacking. Hackers will come into someone's company and steal sensitive information for their own earnings. This is where the firewall comes into play and is important to know what rules a certain server has to access it in order that information does not go into the wrong hands. Some classifications down below help get a better picture of how there are different firewalls offering different ways of protection.

Classification of firewall:

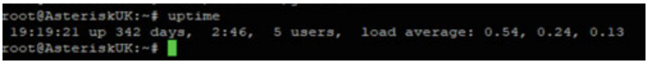
1. Based on features
    a. Packet filtering
    b. Circuit-level gateway
    c. Stateful inspection firewall
    d. Application-level gateway (proxies)
    e. Multilayer inspection firewall
    f. Dynamic firewall
2. Based on usage
    a. Software firewall
    b. Hardware firewall
3. Based on budgets
    a. Commercial or paid firewall
    b. Free or open-source firewall.

Reference [4] We have mentioned in our "Theoretical Explanation" section that Linux load averages are not just for runnable tasks but also for uninterruptible tasks. This article examines the reason behind that. Load averages are for processes or threads that are running or in a waiting state. When averages are 0 the system is not being used, and when the average of 1 min is greater than the 5 and 15 min it means the load is getting greater in value and the opposite means load is getting smaller. The uninterruptible state means that the load average is bigger because of disk or network file system input/output workload and not just CPU demand.

Reference [5] In this paper, we took out something regarding authentication and why that is important for the protection of our data and sensitive information. For example, when logging into some platform what would be the cases of unsecure practices. The need of using passwords that are not weak, and why is it wrong to not use a password at all.

3 Theoretical Explanation

Fail2Ban uses files (scripts, logs, etc.) and denies service to IPs that want to access information when they are not authorized to do such things. This app is used to apply some rules to the firewall and provide protection. In this way, suspicious IPs are dismissed after a specified number of requests. First of all, we should see what's the server's load average in normal conditions.



This can be done with a simple command:

The command uptime is used to show the current time in which we're simulating our scenario, the time for how long the system has been working, the number of users and the load average for 1, 5 and 15 min.

What are we measuring?

Load average is the average number of processes that are running or in an uninterruptable state. Running means using or waiting for CPU and in uninterruptible it means when it is waiting for I/O actions. We explained some of this in the "Related Works" section as well.

4 Experimental Environment

Our experiment was done in the environment of a company testing security with the Fail2Ban app. All servers are in the same INTRANET.

Data:

192.168.10.205—The server where we measure the load average (ubuntu 12.04LTS 1 GB RAM)

192.168.10.6—Central server (ubuntu 16.04LTS 12 GB RAM).

192.168.10. 43 (ubuntu 12.04LTS 8 GB RAM).

192.168.10. 233 (ubuntu 18.04LTS 8 GB RAM).

192.168.10.56 (ubuntu 16.04LTS 8GB RAM).

192.168.10. 225 (ubuntu 16.04LTS 8 GB RAM).

We create a bash script that sends requests to port 22, also known as service ssh (Secure Shell).

The script is executed on a central server. This communicates with public keys with some other servers from which we will send requests to that server we want to measure the load average.

Taking into account that the communication between servers is through public keys, then the connection of this server with the others is done with no password needed. This helps us send requests from different IPs toward the server in which the simulation takes place.

In the script below we have the central server 192.168.10.6 which communicates with keys (no password) with servers “192.168.10.43”, “192.168.10.233”, “192.168.10.56”, “192.168.10.225” from which we will be sending requests to server “192.168.10.205” (where we will measure the load average).

The script is first connected to the respective server with an outside loop (emphasis that the communication is done without the password). Then we are going to send an infinite number of requests toward “192.168.10.205” until the server we’re sending requests from is banned.

```
# Declare variables
currentdate=`date "+%Y-%m-%d"`
destination_project="$1"
declare -a servers_IP=("192.168.10.43" "192.168.10.233" "192.168.10.56" "192.168.10.225")
# get length of an array
arraylength=${#servers_IP[@]}
$k = 0

echo "sending request to server .. "


for (( i=1; i<${arraylength}+1; i++ ));
do

    $connect_from_main_server = ssh root@servers_IP[$i]
    if [$connect_from_main_server == "connection refused" ]
        $alert_msg="There is a problem while connecting from the main server to the host \n"
        echo $alert_msg
        echo "Host is : \n"
        echo servers_IP[$i]
        exit 1


    for (( ; ; ))
    do

        $result_connect = ssh root@192.168.10.205 "cd /var/www"
        if [$result_connect == "connection refused" ]
            break

    done
done
```

5 Results and Interpretation

Case 1: In server 192.168.10.205 we configure Fail2Ban [6] so that it drops requests after a total of 10 requests per ssh service.

```
[ssh]

enabled  = true
port     = ssh
filter   = sshd
logpath  = /var/log/auth.log
maxretry = 10
action   = 3600s
```

We execute the script taking into consideration the configuration we just opted for (that after 10 requests server 192.168.10.205 drops requests). Results can be seen below:

```
root@AsteriskUK:~# uptime
         342 days,  4:08,  5 users,  load average: 1.05, 1.09, 0.10
root@AsteriskUK:~#
```

Case 2: In server 192.168.10.205 we configure Fail2Ban so that it drops requests after a total of 30 requests per ssh service.

```
[ssh]

enabled  = true
port     = ssh
filter   = sshd
logpath  = /var/log/auth.log
maxretry = 30
action   = 3600s
```

We execute the script taking into consideration the configuration we just opted for (that after 30 requests server 192.168.10.205 drops requests).

Results can be seen below:

```
root@AsteriskUK:~# uptime
         342 days,  4:08,  5 users,  load average: 3.05, 2.09, 1.10
root@AsteriskUK:~#
```

Case 3: In server 192.168.10.205 we configure Fail2Ban so that it drops requests after a total of 60 requests per ssh service.

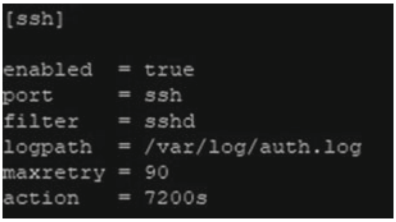
```
[ssh]

enabled  = true
port     = ssh
filter   = sshd
logpath  = /var/log/auth.log
maxretry = 60
action   = 3600s
```

We execute the script taking into consideration the configuration we just opted for (that after 60 requests server 192.168.10.205 drops requests).

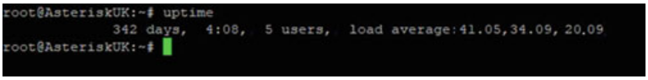
Results can be seen below:

Case 4: In server 192.168.10.205 we configure Fail2Ban so that it drops requests after a total of 90 requests per ssh service.

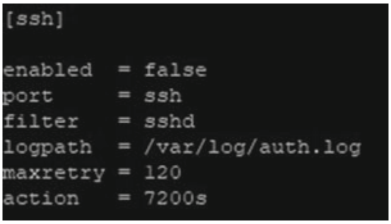


We execute the script taking into consideration the configuration we just opted for (that after 90 requests server 192.168.10.205 drops requests).

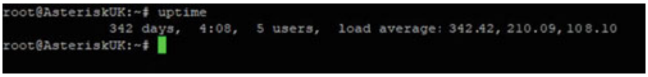
Results can be seen below:



Case 5: In server 192.168.10.205 we configure Fail2Ban so that it does not drop requests.



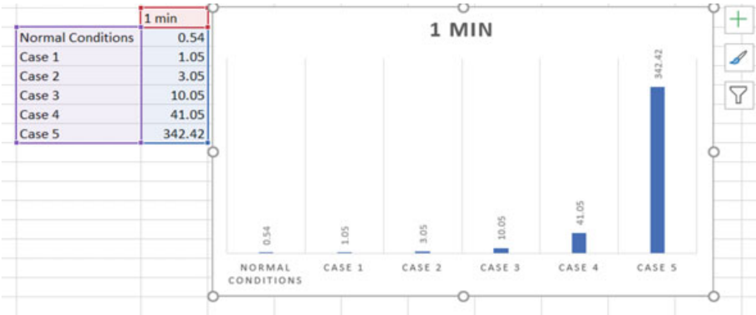
After execution:



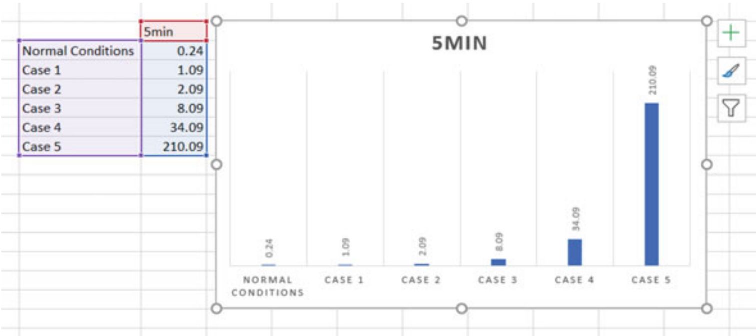
In this particular case, we can see an increase in the load average value until there comes a moment when we cannot access the server anymore. It does not respond and we just restart the machine to return it in working condition.

In the graphics below we can see how the load average changes for all five cases in comparison with the normal condition of the server.

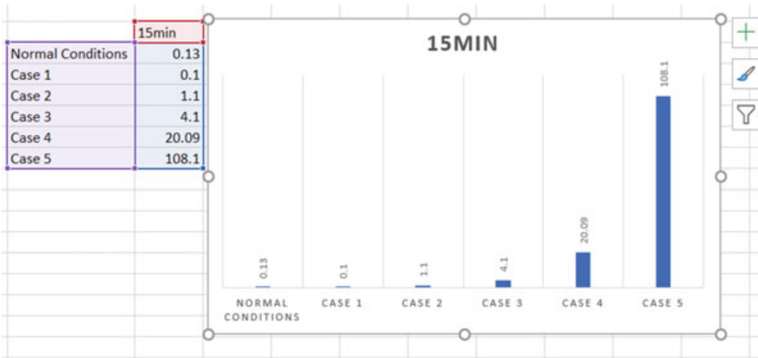
1 minute:



5 minute:



15 minute:



We basically have the same pattern for all three results from the "uptime" command. The load average of the server keeps on increasing causing it to go to a blocked state from which only restart helps it go out of. The increase in the load average means that there are more and more processes crowding the server and the server is not allowing them to perform whatever they want to do in that particular port of his. This means that the firewall is protecting the secure shell even though multiple requests are sent to it.

6 Conclusions

In this simulation, we tried to test how the firewall protects the server from IPs that are not allowed to access it. In this case, the simulation was successful and the port could not be attacked or hacked. No information or operation was allowed to go through it. This talks highly about firewalls that are able to protect information even without proper authentication requirements such as the use of a password. We test the firewall protection by adding some rules in its interface from the Fail2Ban app and we manage to keep the port ssh safe and no foreign IP can access it, so it will be available only to the accepted ones and the performance not deteriorated. Also, it is important to note the change in the value of the load average in the server that we are simulating. It keeps increasing because the server is crowded with requests which it cannot respond to and this causes the machine to block, without giving access. This though in this case, we do not know what would happen in other situations. It is best if there are other forms of authentication used like passwords or public and private keys. This can be done in future works, with the same experiment but requiring some kind of authentication; for example, to see how that affects the simulation. Also, we could try a case when we do not want to access the ssh port but some other port like ftp or tcp service, we could measure the performance of the server to which we are sending requests by evaluating the ARP, PRT, CPU utilization etc.

References

1. https://www.csoonline.com/article/3208344/the-evolution-of-the-firewall.html
2. http://www.ijsrp.org/research-paper-0416/ijsrp-p5278.pdf
3. https://www.scribd.com/document/444237742/3412-320-1-PB
4. http://www.brendangregg.com/blog/2017-08-08/linux-load-averages.html
5. http://interactions.acm.org/archive/view/may-june-2019/dont-blame-the-user
6. http://www.fail2ban.org/wiki/index.php/Main_Page

Blockchain as a Technology Contributor: A Survey

Khurram Mahmood, Zainab Nayyar, and Taqadus Bashir

Abstract Blockchain is a technology that is initially used for cryptocurrencies like bitcoin. The biggest advantage of this technology is that it is immutable data stored in blocks that cannot be changed, rather a new block is added in the chain and a new record can be inserted into it. The previous record shall remain the same and the record of every transaction can be tracked. Many studies and surveys have been done on blockchain mentioning the implications and contributions of blockchain. This paper presents a detailed survey of blockchain mentioning its features, advantages and protocols. In this paper, the authors have presented the working, features and usage of blockchain technology. In the future, based on the characteristics of blockchain technology, the authors will implement blockchain with the cloud of things.

Keywords Blockchain · Technology · Implications and benefits

1 Introduction

Blockchain emerged as a new technology at the beginning of this millennium and gained significant importance during the past few years. It replaced many current digital platforms [1]. Its architecture was used when the virtual currency was launched. It is the fifth-biggest innovation in the field of computing, also known as a distributed ledger of records that cannot be changed and replicated but is verifiable [2]. Though blockchain technology was used initially in developing bitcoin, its usage is not limited to currency and it can also be used in many other fields, like neuro finance, management, health care, vehicular systems, law etc. [3].

Blockchain is based on distributed databases that are highly encrypted in order to prevent fraud in entered records [4]. Financial institutions maintain financial records, hospitals maintain the medical record, universities are concerned with educational record etc., but mostly these records are maintained and stored by third-party entities which are vulnerable to certain security attacks, hacking and data theft. If these

K. Mahmood (✉) · Z. Nayyar · T. Bashir
Bahria University, Islamabad, Pakistan

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_3

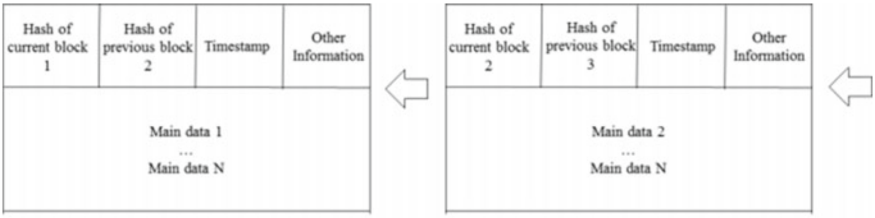


Fig. 1 Structure of blockchain

records are migrated to blockchain-enabled systems then these can be secured from all types of cyber threats regarding data harm and human mischief.

Blockchain technologies are composed of six key characteristics [1]. These are decentralized, transparent, open-source, autonomy, immutable and anonymity. Decentralization is the major feature of blockchain, which means blockchain does not rely on a centralized architecture. All data can be recorded, stored and updated in a distributed way. The data record on the blockchain cannot be fraudulently changed, thus it ensures transparency. It is an open-source technology that people can use freely to develop new applications, can check their records and can easily store their data. Blockchain works with full autonomy in a way that every node can independently transfer or update data without any intervention. No record can be altered in blockchain nodes once it is saved unless someone can take control of more than 51% of that node, so it is immutable (Fig. 1). Blockchain technology provides full trust in data transfer between nodes; even the transactions can be anonymous and one needs to know the persons' blockchain address [5].

2 Literature Review

In [6], multiple proofs-of-concept were reviewed related to blockchain and various data analytics on real-life events are drawn to utilize blockchain technology other than cryptocurrency. Along with this, several research topics on blockchain were also discussed, which included enhancement of security and privacy of blockchain, blockchain-based intelligent systems, data analytics based on blockchain etc. In [7], the literature on the implementation of blockchain was summarized, to find out its implementation beyond cryptocurrency and to enhance its security by reducing the challenges that blockchain is facing.

In [3, 4], the positive implications of blockchain by examining various research results of blockchain for modern organizations that are especially related to the finance or production industry were investigated. Moreover, it was also found that blockchain was not restricted to only these industries but also has the potential to serve in the pharmaceutical industry, event management and other businesses. In [8], a vehicular ad hoc network architecture based on blockchain was proposed.

The new proposed system was totally based on the distributed transport system and named Block-VN. The vehicles were networked and the information was stored using blockchain.

In [9], Medrec, a decentralized medical record management system using blockchain to convert the traditional health records management in a new way to engage patients to view and maintain their health records was proposed. The system provided patients with an unchangeable log and easy access to records. In [2], the key developments of blockchain in various domains and their expected impacts and implications were explored.

2.1 Advantages of Blockchain

In [5], four big advantages of blockchain were identified; they are reliability, security, efficiency and trust. The applications developed on the blockchain are highly reliable because a single point of failure does not affect the whole application. It converts whole decentralized records into a centralized way without affecting its architecture. Only the trusted institutions can access the decentralized ledgers. Blockchain is trusted because its ledgers are distributed among tempered proof nodes. Its network uses a one-way mathematical hash function as input that has variable length and its output has no reflection of input, so it is a highly secured technology. This technology is highly efficient because it runs through the pre-set procedures so that it reduces the cost of labor, increases efficiency and promotes automation.

2.2 Challenges of Blockchain

In [10] the authors believed that blockchain can face three major challenges. The first challenge is the selection of an appropriate and strong cryptographic hash function that is robust and fault-tolerant so that no one can break it. The second is the correct design and implementation of code. The third is the selection of robust and efficient algorithm.

3 Blockchain Protocols

Blockchain ensures the existence consensus mechanism [9] in the network. The greater the consensus mechanism applies in a network, the greater the strength of an application shall be considered. The consensus mechanism proves the sanity and coherence of data. There are three protocols of blockchain. First is proof of work, second is proof of stake and third is practical byzantine fault tolerance. Proof of work is a protocol that is involved in performing extensive data mining calculations and transactions so that a new block can be formed in the blockchain. Proof of stake protocol allows users to mine and validate the block that is created through proof of work. Byzantine fault-tolerance protocol ensures secure communication among the two nodes even if they have the same data [9].

4 Future Discussion

In [7] Gartner's hype cycle of technologies in which blockchain is at its peak was presented. Because of its wide range of usage in society in different spheres, the authors of this paper are aiming to implement this technology in the cloud of things. Currently, there are many security issues that cloud computing is facing like security, confidentiality, transparency etc. When cloud computing is integrated with the internet of things it becomes a cloud of things and the security concerns increased. In order to enhance security and ensure confidentiality and transparency, the author will implement blockchain with the cloud of things. This not only will enhance the security of the system but also ensure transparency in every transaction. For the implementation of this framework, no new protocol will be developed because the already defined protocols of blockchain are fulfilling the need of its implementation (Fig. 2).

5 Conclusion

The concept of blockchain has moved beyond its usage in bitcoin. It can be adopted in many other areas as well. Blockchain transactions are used in both communications whether they are human to human or machine to machine. Blockchain is especially meant to be applied in developing countries where trust and security is a major issue. In this paper emergence, benefits and challenges related to blockchain are highlighted along with the future dimensions which can be helpful for its further futuristic implementations. This paper is well defined that how uniquely we will use blockchain as a technology contributor.

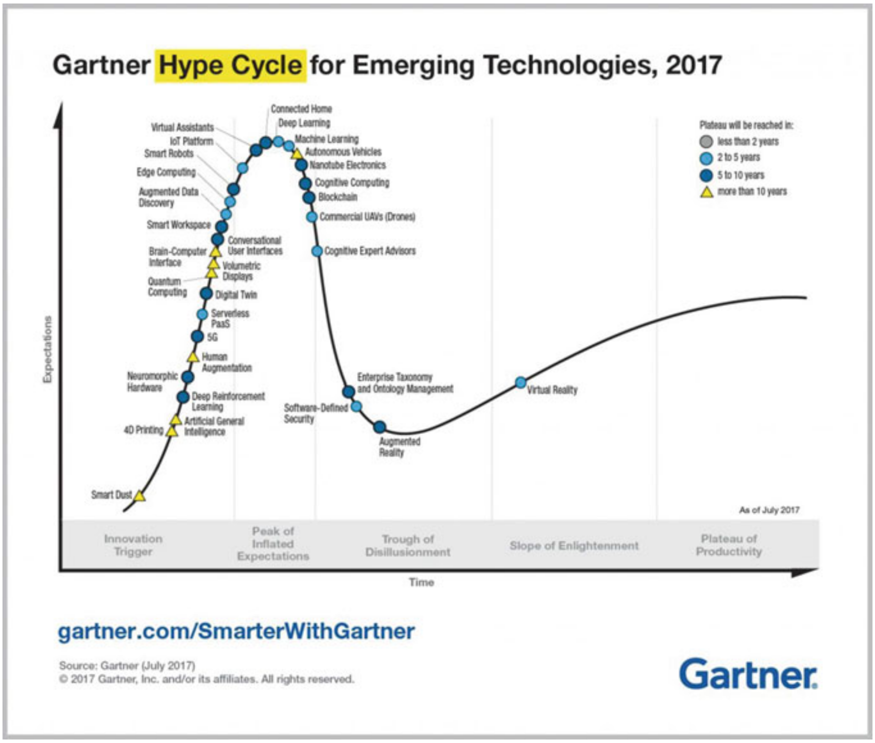


Fig. 2 Gartner hype cycle

References

1. Semantic Scholar (2018) Case study for blockchain in healthcare: "MedRec" prototype for electronic health records and medical research data. /paper/A-Case-Study-for-Blockchain-in-Healthcare. Accessed 13 July 2018
2. Semantic Scholar (2018) A survey of blockchain security issues and challenges. /paper/A-Survey-of-Blockchain-Security-Issues-and-Lin-Liao. Accessed 13 July 2018
3. Aras ST, Kulkarni V (2017) Blockchain and its applications–a detailed survey. Int J Comput Appl 180(3):29–35
4. Beck R, Avital M, Rossi M, Thatcher JB (2017) Blockchain technology in business and information systems research. Bus Inf Syst Eng 59(6):381–384
5. Chen G, Xu B, Lu M, Chen NS (2018) Exploring blockchain technology and its potential applications for education. Smart Learn Environ 5(1):1
6. de Leon C, Stalick A, Jillepalli A, Haney MA, Sheldon FT (2017) Blockchain: properties and misconceptions. Asia Pac J Innov Entrep 11(3):286–300
7. De la Rosa J, Torres-Padrosa V, El-Fakdi A, Gibovic D, Hornyak O, Maicher L, Miralles F (2017) A survey of blockchain technologies for open innovation
8. Miraz MH, Ali M (2018) Applications of blockchain technology beyond cryptocurrency

9. Sharma PK, Moon SY, Park JH (2017) A distributed blockchain based vehicular network architecture in smart city. J Inf Process Syst 13(1):84
10. Vo HT, Kundu A, Mohania M (2018) Research directions in blockchain data management and analytics
11. Zhao JL, Fan S, Yan J (2016) Overview of business innovations and research opportunities in blockchain and introduction to the special issue. Financ Innov 2(1):28

GeLaP: German Labeled Dataset for Power Consumption

Sebastian Wilhelm, Dietmar Jakob, Jakob Kasbauer, and Diane Ahrens

Abstract Due to the increasing spread of smart meters, numerous researchers are currently working on disaggregating the power consumption data. This procedure is commonly known as Non-Intrusive Load Monitoring (NILM). However, most approaches to energy disaggregation first require a labeled dataset to train these algorithms. In this paper, we present a new labeled power consumption dataset that was collected in 20 private households in Germany between September 2019 and July 2020. For this purpose, the total power consumption of each household was measured with a commercial available smart meter and the individual consumption data of 10 selected household appliances were collected.

Keywords Smart meter · Power meter · Dataset · Load disaggregation · Single-measurement · Energy conservation · Scientific data

1 Introduction

On August 29, 2016 the German Federal Government passed the '*Gesetz zur Digitalisierung der Energiewende*', which states, among other things, that in Germany existing analog electricity meters must be replaced by smart meters by 2032 at the latest [1]. However, this will primarily support the network operators, as the smart meter roll-out enables them to better coordinate electricity consumption and elec-

S. Wilhelm (✉) · D. Jakob · J. Kasbauer · D. Ahrens
Technology Campus Grafenau, Deggendorf Institute of Technology, Hauptstraße 3, 94481 Grafenau, Germany
e-mail: Sebastian.Wilhelm@th-deg.de

D. Jakob
e-mail: Dietmar.Jakob@th-deg.de

J. Kasbauer
e-mail: Jakob.Kasbauer@th-deg.de

D. Ahrens
e-mail: Diane.Ahrens@th-deg.de

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_5

tricity generation and thus keep the power grid stable. The consumers usually do not benefit directly from the conversion to smart meters.

Numerous investigations are currently conducted on the re-identification of specific devices from aggregated power consumption data [6, 23, 24]. The aim of most researches is to give the residents the opportunity to monitor their own power consumption and to identify energy guzzlers and thus help to save energy [10].

To perform the disaggregation of the power consumption data—which is commonly known as Non-Intrusive Load Monitoring (NILM)—it is first required to develop intelligent algorithms, which are trained by a labeled dataset [25].

Although numerous public datasets of power consumption data exist (see Sect. 7), there is still a gap in high-frequency data collected by commercial smart meters. For example, for the most frequently cited dataset, REDD, the aggregated consumption data were collected using peripheral measuring devices installed in a so-called 'REDDBox' near the circuit breaker box [9]. The use of smart meter data has the advantage that these are currently being installed throughout Germany (and also in other countries) and have built-in data interface. Developed NILM algorithms can access the power consumption data via these interfaces without additional hardware.

In our project *BLADL* a dataset of load data in 20 German private households was collected between September 2019 and July 2020. This dataset includes the total power consumption of the households measured by a commercial available smart meter and the individual consumption data for 10 appliances in the household.

The main contribution of this paper is to outline the data collection process and present the full collected data.

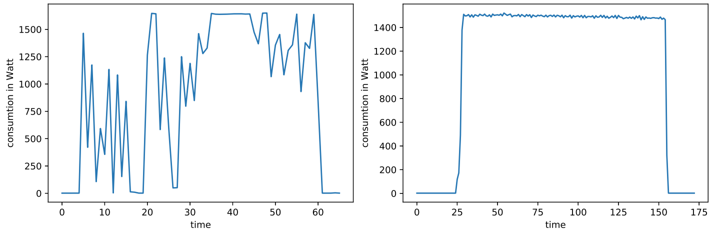
The remaining of the paper is structured as follows: the use case of data collection in the scope of the project *BLADL* as well as an example method of disaggregation is introduced in Sect. 2. Section 3 describes the equipment used for data collection, the methodology of data transmission and storage as well as concerns on ethics and privacy. A detailed description of the generated dataset is given in Sect. 4. The tested households in which the data were collected are presented in Sect. 5. In Sect. 6 the quality of the measurements is discussed. Section 7 provides a survey of related datasets. The paper is concluded in Sect. 8.

2 Use Case and Disaggregation—BLADL

As part of the research project *BLADL*, re-identified power consumption data are investigated to detect potential emergency situations in a household.

The information of the activity state of individual household devices can be used to draw conclusions about the activity of the resident [18, 19]. This allows creating individual activity profiles of the residents, which in turn can indicate possible emergency situation in the case of anomalies.

For data disaggregation Motif Clustering algorithms [15] were used to extract common consumption patterns—called *motifs*—from the label data recordings of individual devices. Figure 1 illustrates as an example two extracted motifs of a coffee



(a) coffee machine (b) microwave

Fig. 1 Examples for extracted motifs from a coffee machine and a microwave

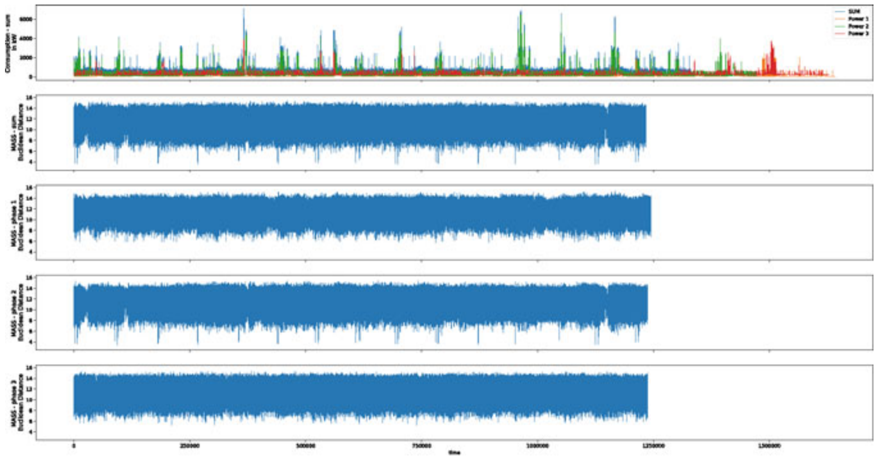


Fig. 2 Time series for the measured values (total power consumption and the phases) and the corresponding distance profiles for each stream calculated with *MASS* for the coffee machine example motif (Fig. 1a)

machine and a microwave. In a second step the previously extracted motifs were identified in the total power consumption (or in the streams of the phases) of a household using the *MASS*—Motif Search [16] algorithm. The *MASS* algorithm allows searching for time series sub-sequences and results in an array of distances. This array of distances enables to identify similarities between sub-sequences [13, 16]. A distance in the distance profile close to 0 is most similar to the sub-sequence.

Figure 2 shows an example for a distance profile, calculated using the *MASS* for the coffee machine motif (Fig. 1a). Downward deflections in the distance profile indicate a matching sub-sequence—i.e., the activity of the device.

3 Methodology of Data Collection

In the following, we present our general methodology for data collection. An extended documentation on organizational, technical, ethical, and legal requirements was published by Wilhelm et al. in [22].

3.1 Measuring Instruments

For data acquisition we measured on the one hand the total power consumption of the households using a smart meter and on the other hand the individual power consumption of 10 selected devices using single power measure tools.

As smart meters, the commercially available *Q3M* devices from *EasyMeter GmbH*, Bielefeld were used with the smart meter gateway *Discovergy Meteorit 2.0* from *Discovergy GmbH*, Heidelberg. This smart meter gateway allows the sampling of the measured values in a resolution of 1 Hz via a REST-API.

In order to record the individual consumption of the selected devices, *ALL3075v3* single power measure tool from *ALLNET GmbH*, Germering was used. These devices allow measurements in a resolution of up 7 Hz, which can be accessed via a local network interface (API).

3.2 Data Transmission and Storage

All measured values were stored centrally and structured on a PostgreSQL database. For this purpose, the smart meter data was periodically fetched and stored using the REST-API provided by the manufacturer.

Since the individual electricity meters do not have integrated buffer storage for the measured values and were configured for data security reasons in a way that the measured values can only be requested in the local (home) network, a single-board computer was used in each of the test households. This computer was responsible for continuously fetching the measured values of the individual electricity meters, synchronizing them in time using NTP and transmitting them to the central database.

Figure 3 shows the software architecture for data acquisition and transmission.

3.3 Ethics and Privacy Considerations

Gaining special importance in this project are the aspects of ethics and privacy since sensitive data concerning the highly personal sphere of life of the test persons are collected and processed. In order to comply with ethical norms and values as well as

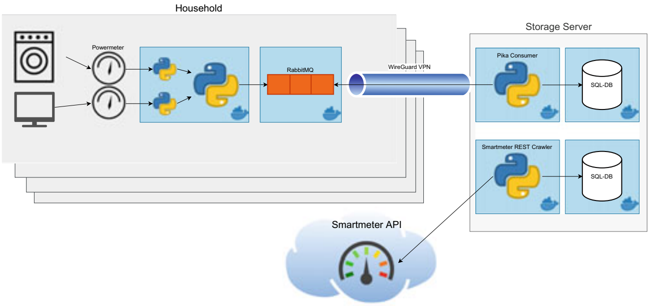


Fig. 3 Software architecture for data acquisition and transmission [22]

a maximum of process transparency for the test participants, an ethical reflection on the project—especially for the data collection and further processing—was created in advance based on a guideline of the *Ethics Commission DGP e.V.* [4].

To ensure the highest level of data security and data protection, all data collected, processed, and analyzed within our project are stored exclusively on the internal servers of the *Deggendorf Institute of Technology*. We stored all measured data by default pseudonymized. Further we implemented a Data Privacy Management according to Jakob et al. [7].

4 Dataset Description

In the period from September 2019 to July 2020, the following data were collected from a total of 20 private households in Germany:

$$
\begin{aligned}
household =& (location, \#residents, space)\\
measurement_{total} =& (household, timestamp, power1, power2, power3)\\
device =& (type, location_{ih}, description)\\
measurement_{device} =& (household, device, timestamp_{requ}, timestamp_{repl}, power)
\end{aligned}
$$

where:

$$
\begin{aligned}
location &= \text{Postcode of the location of the household}\\
\#residents &= \text{Number of residents in the household}\\
space &= \text{Floor space of the household (in m}^3\text{)}\\
timestamp &= \text{NTP-Synchronized UNIX Timestamp (in ms)}\\
power1 &= \text{Power over Line Conductor 1 (in W)}
\end{aligned}
$$

$$
\begin{aligned}
power2 &= \text{Power over Line Conductor 2 (in W)}\\
power3 &= \text{Power over Line Conductor 3 (in W)}\\
type &= \text{Type of the device}\\
location_{ih} &= \text{Location of the device in the household}\\
description &= \text{Extended device designation}\\
power &= \text{Electrical Power (in W)}\\
timestamp_{requ} &= \text{NTP-Synchronized UNIX Timestamp}\\
&\quad \text{(in ms) of the data request}\\
timestamp_{repl} &= \text{NTP-Synchronized UNIX Timestamp}\\
&\quad \text{(in ms) of the data reply}
\end{aligned}
$$

The parameters #*residents* and *space* are based on information we have received from the residents. Especially the information on flat space can therefore only be considered as a tendency value.

The parameters $timestamp_{requ}$ and $timestamp_{repl}$ are caused by the architecture of the data collection of the label data (Sect. 3). Since the individual single power measure tools are not NTP-synchronized, the timestamp of the label data was generated on the single-board computer. The single-board computer used for this purpose requested the current measured value at regular time intervals via a local network interface of the individual electricity meters. The $timestamp_{requ}$ and $timestamp_{repl}$ timestamps represent the time at which a request for the current measurement value was sent to the meter ($timestamp_{requ}$) and the time when the measurement value was received on the single-board computer ($timestamp_{repl}$). The difference between the two timestamps thus can be read as the network run time and the processing time on the single power measure tool.

To ensure that no household can be de-identified by the *location*, #*residents* and *space* parameters, we k-anonymized this parameter [21]. Therefore, we have limited the postcode parameter to the first two characters (e.g., 94***). The information about the number of residents living in the household as well as the flat space we divided up in pairs of two groups (#*residents*: $<=2$ and >2; *space*: $<=150\,\text{m}^3$ and $>150\,\text{m}^3$). This results in a 4-anonymous dataset.

The root-folder contains the documents `readme.md`, `main-info.txt` and `main-info.json` as well as the 20 folders for the individual households (Fig. 4). The `readme.md` gives some basic information on the dataset and its use. The documents `main-info.txt` and `main-info.json` contain the metadata for the measurements of all households in both human and machine-readable format.

For each test household there is a separate folder with the documents `device-info.txt` and `device-info.json`. These containing the household-specific metadata including the device assignment. A separate CSV document with the raw measurements was created for each measured device (e.g., `label-001.csv`). The total household electricity measurements are stored in `smartmeter.csv`.

Fig. 4 Directory structure of the dataset

```
/
├── readme.md
├── main-info.txt
├── main-info.json
├── hh-01
│   ├── device-info.txt
│   ├── device-info.json
│   ├── label-001.csv
│   ├── label-002.csv
│   ├── [...]
│   └── label-010.csv
├── hh-02
├── [...]
└── hh-20
```

For deployment, the folders of the individual household measurements were packed into a `tar` archive and compressed using `xz`. This reduced the total file size from 80 to 6.9 GB without loss of information.

5 Households Overview

We collected the data from September 2019 to July 2020 in 20 German households. A total of four complete setups of the submetering equipment were used in rotation and stayed in the households for 28 days on average. The total power consumption was measured using a smart meter, which remained in the households. Table 1 lists additional information on the individual households in an anonymized form. Furthermore, it is described for each household in detail how long the data recording took place for both—individual consumers and total power consumption.

In addition to the total power consumption of the household, the individual consumption data for 10 appliances was measured in each household. When selecting the appliances for individual measurements, we primarily tried to use devices where the power consumption indicates a direct human interaction. Standalone devices such as refrigerators were only considered in a few cases. Table 2 gives an overview of the individual devices measured per household.

Where possible, the exact model code of the devices was recorded within the metadata.

Table 1 Additional information on the individual households

Household	#residents	Flat space	Duration submetered	Duration aggregated
hh-01	<=2	$>150\,m^3$	25 days 23 h	90 days 19 h
hh-02	<=2	$<=150\,m^3$	6 days 0 h	90 days 19 h
hh-03	>2	$<=150\,m^3$	28 days 18 h	168 days 18 h
hh-04	>2	$>150\,m^3$	23 days 20 h	143 days 4 h
hh-05	>2	$<=150\,m^3$	44 days 20 h	272 days 15 h
hh-06	>2	$<=150\,m^3$	34 days 17 h	210 days 19 h
hh-07	<=2	$>150\,m^3$	34 days 20 h	210 days 19 h
hh-08	>2	$>150\,m^3$	31 days 21 h	162 days 11 h
hh-09	>2	$>150\,m^3$	44 days 0 h	272 days 15 h
hh-10	>2	$>150\,m^3$	3 days 10 h	90 days 19 h
hh-11	>2	$<=150\,m^3$	30 days 18 h	161 days 18 h
hh-12	<=2	$<=150\,m^3$	14 days 0 h	158 days 17 h
hh-13	<=2	$>150\,m^3$	23 days 19 h	159 days 18 h
hh-14	>2	$>150\,m^3$	34 days 21 h	90 days 19 h
hh-15	>2	$>150\,m^3$	18 days 19 h	161 days 18 h
hh-16	<=2	$>150\,m^3$	34 days 22 h	210 days 19 h
hh-17	<=2	$>150\,m^3$	28 days 23 h	161 days 18 h
hh-18	<=2	$<=150\,m^3$	20 days 17 h	91 days 11 h
hh-19	>2	$>150\,m^3$	43 days 0 h	70 days 16 h
hh-20	<=2	$<=150\,m^3$	23 days 20 h	161 days 18 h

6 Data Quality Reflection

To examine the data quality on total power consumption, an example period of 20 days from one of the measured households was considered. We have observed that the smart meter gateways do not constantly deliver data with a resolution 1 Hz. We found that the time intervals between two readings in our data extract ranged from 818 to 117,006 ms as shown in Fig. 5.

At a closer look at the time stamps of the measurement data, a continuous time offset could be observed in the data sections without gaps. It can be seen that after about 20 samples the time stamp increases by 1 ms. Consequently, this means that the resolution of the data is only about 0.99 Hz. Figure 6 shows this time shift.

Even if both, smart meter and single-board computer, were time synchronized using NTP, we could detect a minor time shift of the measured values. This offset is approximately 3 s.

Figure 7 shows the measurements of one phase of the smart meter to which the coffee machine is connected.

Table 2 Individual monitored devices per household

Household	Devices
hh-01	Coffee machine, computer, electric kettle, extractor fan, floor lamp, microwave, radio, television (2x), washing machine
hh-02	Battery vacuum cleaner, charger, computer, dryer, monitor, printer (2x), stereo system, television, washing machine
hh-03	LED lamp, bedside lamp, bread slicer, charger, coffee machine, dryer, lamp, stereo system, television, washing machine
hh-04	Coffee machine, electric kettle, extractor fan, lamp, microwave, radio, television, television-receiver, washing machine, water pump
hh-05	Battery vacuum cleaner, floor lamp, fully automatic coffee machine, garage door opener (2x), television (3x), washing machine, water pump
hh-06	Coffee machine (2x), floor lamp + phone, hair straightener, hairdryer, radio, television (2x), washing machine, water baller
hh-07	Computer, electric kettle, floor lamp (2x), fully automatic coffee machine, microwave, printer, radio, television, vacuum cleaner
hh-08	Thermomix, battery vacuum cleaner, coffee machine, electric kettle, hairdryer (2x), television (2x), television-receiver, toaster
hh-09	Coffee machine, dryer, extractor fan, garage door opener (2x), microwave, television (2x), vacuum cleaner, washing machine
hh-10	Thermomix, electric kettle, extractor fan, fully automatic coffee machine, printer, radio, sound-system (2x), television (2x)
hh-11	Coffee machine, computer, dryer, electric kettle, floor lamp, juicer, television (2x), television-receiver, toaster
hh-12	DVD player, bedside lamp, bread slicer, coffee machine, dryer, floor lamp, fully automatic coffee machine, microwave, television, washing machine
hh-13	Thermomix, charger, coffee machine, floor lamp, garage door opener, radio (2x), television, washing machine (2x)
hh-14	Thermomix, charger, coffee machine, electric kettle, microwave, radio, television, toaster, vacuum cleaner robot, washing machine
hh-15	Battery vacuum cleaner, circulation pump, coffee machine, dryer, flat iron, floor lamp, language assistance system, television, toaster, washing machine
hh-16	Battery vacuum cleaner, coffee machine, computer, garage door opener, hairdryer, light alarm clock, printer, telephone, television, toaster
hh-17	Coffee machine, dryer, electric kettle, floor lamp (2x), garage door opener, radio, stereo system, television, toaster
hh-18	Bedside lamp, coffee machine, dryer, electric kettle, floor lamp, hairdryer, printer, television, toaster, washing machine
hh-19	Thermomix, baking machine, dryer, floor lamp, fully automatic coffee machine, printer (2x), television (2x), washing machine
hh-20	Advent crib, coffee machine, electric kettle, floor lamp (2x), fridge, television (2x), toaster, washing machine

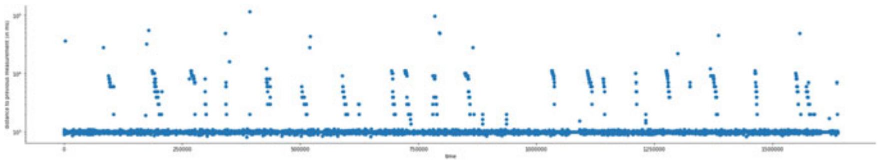


Fig. 5 Time intervals between two consecutive samples in a example period of 20 days

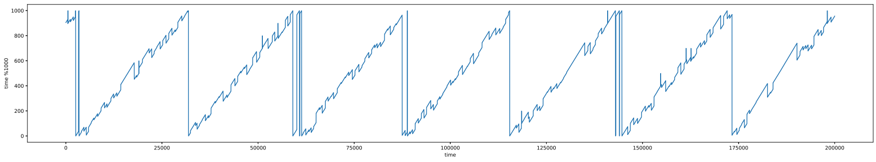


Fig. 6 Time shift of smart meter measurements over 20,000 samples

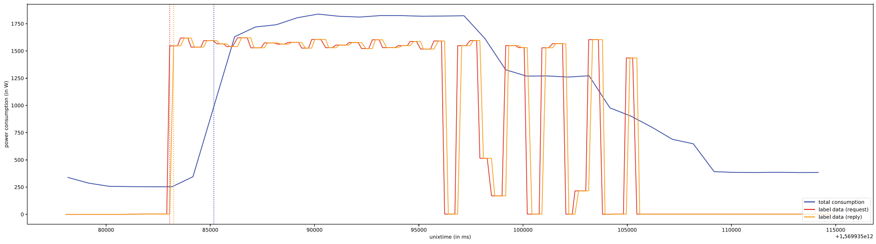


Fig. 7 Measured values of one phase of a smart meter and the individual measured values of a coffee machine simultaneously collected while a coffee was brewed. The dotted lines show the first measured value, which indicates a deflection

In addition, the measurements of the individual electricity meter on the coffee machine, which were taken at the same time, are shown. The time offset between the deflections is clearly visible.

7 Related Datasets Survey

Due to the increasing popularity of smart meters, the interest in device re-identification and energy disaggregation rises. By August 22nd, 2019 Liu [11] identified 617 relevant scientific document in the field of NILM. The fact that a lot of research is currently being carried out in the NILM field is also indicated by the large number of publicly available datasets with load profiles which are shown in Table 3. The datasets differ basically in the number of households investigated, the number and type of sensors used, the resolution of the data, the duration of the test phases in the individual households, and the measured features [14].

Table 3 Existing datasets for energy consumption in households. Extended reprint form [14]

Dataset	Location	Duration	#Households	#Sensors (per house)	Features	Resolution
ACS-F1 [17]	Switzerland	1 h session (2 sessions)	N/A	100 devices in total (10 types)	I, V, Q, f, φ	0,1 Hz
AMPds [12]	Greater Vancouver	1 year	1	19	I, V, pf, F, P, Q, S	0.016 Hz
BLUED [5]	Pittsburg, PA	8 days	1	Aggregated	I, V, switch events	12 Khz
GREEND [14]	Austria, Italy	1 year (2–5 month compl.)	9	9	P	1 Hz
HES	UK	1 month (255 houses) 1 year (26 houses)	251	13–51	P	0.0083 Hz
iAWE [3]	India	73 days	1	33 sensors (10 appliance level)	V, I, f, P, S, E, φ	1 Hz
IHEPCDS[a]	France	4 years	1	3 circuits	I, V, P, Q	0.016 Hz
OCTES[b]	Finland, Iceland, Scottland	4–13 month	33	Aggregated	P, Energy price	0.14 Hz
REDD [9]	Boston, MA	3–19 years	6	9–24	Aggregate: V, P Submetered: P	15 Khz (aggr.), 0.33 Hz (sub)
Sample dataset[c]	Austin, TX	7 days	10	12	S	0.016 Hz
Smart* [2]	Western Massachussets	3 months	1 Submetered +2 (Agg + Sub)	25 circuits, 29 appliance monitors	P, S (circuits), P (submetered)	1 Hz
Tracebase [20]	Germany	N/A	15	158 devices in total (43 types)	P	0.1–1 Hz
UK-DALE [8]	UK	499 days	4	5 (house 3) 53 (house 1)	Aggregated P, Sub P, switch status	16 Khz (aggr.), 0.16 Hz (sub)
GeLaP	Germany	Aggr. 157 days (average) sub 28 days (average)	20	13	Aggregated P, Sub P	1 Hz (aggr.), 5 Hz (sub)

[a] https://tinyurl.com/IHEPCDS
[b] http://octes.oamk.fi/final/
[c] https://www.pecanstreet.org/project/consortium/

Apart of our dataset, only *BLUED* [5], *GREEND* [14], *iAWE* [3], *REDD* [9], *Smart** [2] and *UK-DALE* [8] provides high-frequency measurements with a resolution of at 1 Hz. Thereby, our dataset contains measurements from 20 households, the other datasets only consider between 1 and 9 households.

Due to the high-resolution data, fine granular analyses like motif clustering [15] can be performed.

8 Conclusion and Future Work

Between September 2019 and July 2020, we collected the total power consumption data of 20 households in Germany using conventionally available smart meters with a resolution of 1 Hz. In addition, we measured the individual power consumption of 10 devices in each household where the electrical load indicates direct human interaction. This paper outlines the data collection process and presents the structure and meta-information of the published dataset. In addition, the quality of the data collected was reflected in Sect. 6. This dataset is now publicly available for further research.

GeLaP can be used in future research to develop new NILM algorithms or to evaluate existing approaches. In addition, general statistics on power consumption in private households can be generated, for which the dataset provides some relevant metadata.

Due to the extensive documentation of the data collection process, future researchers will also have the opportunity to adopt the methodology of data collection and expand the dataset.

The dataset collected within this work can be downloaded under:
https://mygit.th-deg.de/tcg/gelap

Acknowledgements This work was funded by the *Bavarian State Ministry of Family Affairs, Labor, and Social Affairs*.

References

1. Bmwi - häufig gestellte fragen rund um das messstellenbetriebsgesetz (msbg) und intelligente messsysteme (smart meter). https://www.bmwi.de/Redaktion/DE/FAQ/Intelligente-Messsysteme-Zaehler/faq-intelligente-netze-intelligente-zaehler.html. Accessed 21 June 2019
2. Barker S, Mishra A, Irwin D, Cecchet E, Shenoy P, Albrecht J, et al (2012) Smart*: an open data set and tools for enabling research in sustainable homes. SustKDD 111(112):108
3. Batra N, Gulati M, Singh A, Srivastava MB (2013) It's different: insights into home energy consumption in India. In: Proceedings of the 5th ACM workshop on embedded systems for energy-efficient buildings. ACM, pp 1–8

4. Ethikkommission DGP e.V. Fragen zur ethischen Reflexion. Last accessed 23 August 2019
5. Filip A (2011) Blued: a fully labeled public dataset for event-based nonintrusive load monitoring research. In: 2nd workshop on data mining applications in sustainability (SustKDD), p 2012
6. Herrero JR, Murciego ÁL, Barriuso AL, de la Iglesia DH, González GV, Rodríguez JMC, Carreira R (2017) Non intrusive load monitoring (NILM): a state of the art. In: Advances in intelligent systems and computing. Springer International Publishing, pp 125–138
7. Jakob D, Wilhelm S, Gerl A (2020) Data privacy management (DPM): a private household smart metering use case
8. Jack K, William K (2015) The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Sci Data 2(1)
9. Kolter JZ, Johnson MJ. Redd: a public data set for energy disaggregation research. In: Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA, vol 25, pp 59–62
10. Liu H (2019) Introduction. In: Non-intrusive load monitoring. Springer, Singapore, pp 1–21
11. Liu H (2020) Non-intrusive load monitoring. Springer, Singapore
12. Makonin S, Popowich F, Bartram L, Gill B, Bajić IV (2013) AMPds: a public dataset for load disaggregation and eco-feedback research. In: 2013 IEEE electrical power and energy conference. IEEE, pp 1–6
13. Marrs T (2019) mass-ts
14. Monacchi A, Egarter D, Elmenreich W, D'Alessandro S, Tonello AM (2014) GREEND: an energy consumption dataset of households in Italy and Austria. In: 2014 IEEE international conference on smart grid communications (SmartGridComm). IEEE, pp 511–516
15. Mueen A, Keogh E (2017) Time series data mining using the matrix profile: a unifying view of motif discovery, anomaly detection, segmentation, classification, clustering and similarity joins
16. Mueen A, Zhu Y, Yeh M, Kamgar K, Viswanathan K, Gupta C, Keogh E (2017) The fastest similarity search algorithm for time series subsequences under euclidean distance. http://www.cs.unm.edu/~mueen/FastestSimilaritySearch.html
17. Paganelli F, Paradiso F, Turchi S, Luchetta A, Castrogiovanni P, Giuli D (2015) Appliance recognition in an OSGi-based home energy management gateway. Int J Distrib Sens Netw 11(2)
18. Perkowitz M, Philipose M, Fishkin K, Patterson DJ (2004) Mining models of human activities from the web. In: Proceedings of the 13th conference on world wide web, WWW'04. ACM Press
19. Philipose M, Fishkin KP, Perkowitz M, Patterson DJ, Fox D, Kautz H, Hahnel D (October 2004) Inferring activities from interactions with objects. IEEE Pervasive Computing 3(4):50–57
20. Reinhardt A, Baumann P, Burgstahler D, Hollick M, Chonov H, Werner M, Steinmetz R (2012) On the accuracy of appliance identification based on distributed load metering data. In: 2012 sustainable internet and ICT for sustainability (SustainIT). IEEE, pp 1–9
21. Sweeney L (2002) k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst 10(05):557–570
22. Wilhelm S, Jakob D, Kasbauer J, Dietmeier M, Gerl A, Elser B, Ahrens D (2021) Organizational, technical, ethical and legal requirements of capturing household electricity data for use as an AAL system. In: Proceedings of the international congress on information and communication technology. Springer
23. Wong YF, Ahmet Sekercioglu Y, Drummond T, Wong VS (2013) Recent approaches to non-intrusive load monitoring techniques in residential settings. In: 2013 IEEE computational intelligence applications in smart grid (CIASG). IEEE
24. Zhuang M, Shahidehpour M, Li Z (2018) An overview of non-intrusive load monitoring: approaches, business applications, and challenges. In: 2018 international conference on power system technology (POWERCON), pp 4291–4299, Nov 2018
25. Zoha A, Gluhak A, Ali Imran M, Rajasegarar S (2012) Non-intrusive load monitoring approaches for disaggregated energy sensing: a survey. Sensors 12(12):16838–16866

Smart Platform Designed to Improve Poultry Productivity and Reduce Greenhouse Gas Emissions

Irina Arhipova, Gatis Vitols, Liga Paura, and Liga Jankovska

Abstract Poultry is the world's most widespread bird species and the major contributor to ammonia and greenhouse gases emission. In order to improve poultry, productivity enterprises used new technological innovations to manage and monitor poultry farms. The objective of the study is to analyse the existing smart poultry management systems that allow taking a decision for the most appropriate feeding process and the lowest level of CO_2 and NH_3 emission. A smart poultry management system is a crucial component of modern poultry farm. The research directions and technologies related to smart poultry management systems are analysed, which are used for poultry productivity, health and welfare prediction. There is not much research on the use of a multi-criteria decision model in poultry farming, where productivity and welfare improvement and greenhouse gases emission reduction are taking into account at the same time. The necessary challenges of the smart platform implementation are designed for the improvement of poultry productivity and reduction of GHG emissions.

Keywords Poultry farm management · Machine learning · GHG reduction

I. Arhipova (✉) · G. Vitols · L. Paura
Latvia University of Life Sciences and Technologies, Jelgava 3001, LV, Latvia
e-mail: irina.arhipova@llu.lv

G. Vitols
e-mail: gatis.vitols@llu.lv

L. Paura
e-mail: liga.paura@llu.lv

L. Jankovska
WeAreDots Ltd., Riga 1050, LV, Latvia
e-mail: Liga.Jankovska@wearedots.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_6

1 Introduction

Agriculture is one of the leading sectors of most of the developing countries and is one of the major tools to improve the standard of living of rural populations and provide food for urban populations [1]. In 2018, 5% of the agricultural sector in Europe was based on poultry and 2.4% on eggs farming [2]. Chickens are today the world's most widespread bird species. More than 50 billion chickens are used annually for meat and egg production [3]. Laying hens are specially harvested for egg production nowadays, and each laying hen lays about 300 eggs a year [4].

It is known that even though many technical advancements take place across all agricultural and farming industries, there is a lot of space for improvement. Smaller enterprises still heavily rely on manual labour and know-how type of the acquired knowledge. A lot is done based on the "sense and feeling" of a professional poultry specialist, and new technological innovations to manage and monitor poultry farms are not used on small farms. Nowadays, only the most advanced and large enterprises are using smart IT solutions that allow better management of their enterprise, and even then, small changes (e.g. inadequate amount of nutrients in the feeding recipe) spotted too late can lead to increased costs or loss of revenue.

The greatest part of poultry farming costs is expenditure on feed. There is a lack of tools to determine which feed recipe is the most appropriate. It is possible to determine the effect on laying productivity, impact on falling and diseases of the birds etc., by analysing different recipes. Simultaneously more and more ethical concerns regarding the well-being of the animals of farming, including birds and chickens, lead to increased competition and put pressure on businesses to increase production efficiency and decrease costs. Egg production is a relatively specific business, and unlike other industries, it is not possible to increase production volumes quickly. Increasing them takes considerable capital investment and time. The poultry sector is actively considering an issue of gas emissions, which increasingly pollute nature. There are a limited amount of solutions, algorithms or processes for analysing the accumulated data from a perspective that solves the issue of choosing the optimal feed recipe with the purpose to reduce gas emissions.

Poultry is an important contributor to ammonia (NH_3) emission and also contribute to greenhouse gases (GHG) emission [5, 7]. Compared to the ruminants, poultry produces a slightly high amount of GHG and represents 8% of the livestock sector's emissions [6], but due to an increase in chicken meat and egg production, poultry is the major source of environmental pollution [7]. Many factors and their combination affect the intensity of GHG, and the main factors can be mentioned as species, feeding and housing system, and region. Some of the key factors are summarized as follows:

- Diet composition and feed conversion ratio (FCR)—efficiency with which chickens convert kg of feed into kg of meat or eggs influence on GHG emission [6]. Diet with high protein level improves production in poultry farms. However, a nutritionally balanced low-protein diet is a way to reduce the environmental burden [8] and decrease GHG. However, the dietary protein should provide the physiological and other requirements of the birds and not influence birds' performance, profit, and the environment [7].

- Manure production and manure management: manure production differs among species due to differences in diets and metabolism processes and within species due to differences in management at the farm. Compared to other species of animals' broilers take the second and laying hens fifth rank in terms of manure production per animal unit [7]. Methane (CH_4) level at the farm or around depends on the ambient temperature and a manure management system. Higher ambient temperatures combined with anaerobic conditions are the cause of high CH_4 level at the farm [6].
- Energy use: around 35–40% of emissions from eggs and broiler production are from energy use, what types of fuels are used and how effective is energy conversion. As well, there are differences between regions, which can be explained by animal feeding and productivity and manure management [6].

It is known that the poultry industry, especially in the Baltic region, has a shortage of poultry specialists. Therefore, the development of an IT solution for the poultry industry is viewed as the highest priority in order to ensure that the knowledge can be sustained and continuously applied across the wider sector and more enterprises. With the introduction of IT solutions, these challenges can be successfully addressed. As mentioned above, optimizing the process of feeding and reducing the costs of production will not only have a significant socio-economic impact on the industry as a whole but also will make a significant contribution to nature protection by reducing harmful carbon dioxide (CO_2) and NH_3 emissions. According to the EU target, CO_2 emissions should drop up to 40% by 2030, compared to 1990 (EU target).

The objective of this research is to analyse the existing smart poultry management systems that allow taking a decision for the most appropriate feeding process and the lowest level of CO_2 and NH_3 emission for every particular egg and broiler producer to enhance productivity, reduce disease, mortality and to take care of the environment. The following proposed tasks are defined:

- To analyse the typical data sources in poultry farms and factors affecting poultry productivity,
- To analyse the research directions and technologies related to smart poultry management systems,
- To identify the necessary challenges of the smart platform implementation designed for improvement poultry productivity and reduction of GHG emissions.

2 Solutions for Poultry Sector

Multiple IT management systems are proposed by companies and organizations for poultry producers. Typically, management systems have multiple components and are built using three-layer architecture: sensing layer, transmission layer and application layer [9]. A majority of systems are commercial, however, there are certain proposals by authors using open-source and low-cost hardware [10]. Typically, the sensing layer involves multiple internet of things (IoT) components, such as video cameras, temperature sensors, air humidity sensors, light intensity sensors, carbon dioxide sensor and others [11]. The transmission layer provides functionality to transmit the data collected by sensors to certain management software for decision support which is located in the application layer.

Another research direction related to the poultry sector is to product traceability solutions [12] using RFID, QR codes, sensors and data transmission layer, for example, mobile data network. Cloud computing has been proposed as a viable solution to overcome certain limitations that appear using sensing, transmission and application layers in larger farms, for example, mobility issues, coverage of signal, energy access, ubiquitous access, etc. [13]. Agriculture robotics is another field of research that incorporates all three layers mentioned above. A review of 24 universities research on agriculture robotics revealed that at the present moment robotics focuses on monitoring the environmental conditions in farms (using various sensors), evaluation of chicken health (using audio and video processing), egg picking (using computer vision and machine learning) and encouragement of chicken movement (using moving robots and other methods) [14].

2.1 Smart Poultry Management Systems and Role of Computer Vision and Machine Learning

A smart poultry management system is a crucial component of modern poultry farm. Such systems can help decision-making processes to reduce costs and use of resources in production, addressing multiple other aspects, animal well-being, reduction of gas emissions, smart management of energy consumption and others. Basically, smart management systems include components from a more general concept-precision livestock farming. Implementation of such systems typically include technologies such as [15] sensors using, processes automatization and data-driven decision-making platforms.

Technologies can address raising issues of precision feeding, detection of diseases, detection of abnormal events in farms and others. Sensors allow the collecting of data in real time from poultry farms which lead to larger data sets over time. Data must be processed and valuable data must be extracted by an information system that is implemented on the farm. If sensors are connected to the internet, then the internet of things network can be created and the possibilities to further connect other devices and equipment raise, thus allowing automation of many farm processes. The collected data can be analysed and data-driven decisions based on various criteria can be performed. Typical data sources in poultry farms are field sensors such as humidity sensor, temperature sensor, weight sensor, etc., CCTV cameras, infrared cameras and sensors to control light in the farm. All the sensors can be wired or wireless, with wireless sensors being the main implementations nowadays. Wearable sensors are applied for animal welfare. Non-invasive wearable sensors such as leg accelerometers can record behavioural changes. There are also invasive wearable sensors such as vaginal boluses that are used to identify animal health and the introduction of new wearable sensors that help to identify particular diseases of animals [16].

There are multiple solutions available as products that farmers can buy and implement on farms. A large portion of available solutions offers accounting and financial planning functionality. Inventory (amount of birds, vaccines, etc.) functionality is also present in various solutions on the market. Management systems [17] offer to enter data manually (feed amount, vaccinations, medications, etc.) and provide reports such as medical record report or others. Some software [18], besides basic reporting, offer forecasting options using machine learning techniques, but without adjustments of criteria or the use of computer vision and machine learning to provide egg counting and water consumption monitoring [19]. A majority of solutions focus on historical data recordings and reporting, thus providing more monitoring functionality [20, 21] or hardware, including sensors [22–25]. Only a few available solutions as products have integrated functionality offered by advances in machine learning and computer vision. It is related to the fact that computational powers and availability of data sets and frameworks only recently were provided for broader use, and application of these technologies for precision agriculture is an emerging field of research.

Published research shows that computer vision and machine learning in poultry farming is applied to solve tasks of detection, identification, classification and prediction. As demands for production are increasing, one of the typical tasks where machine learning algorithms and computer vision are used is in the prediction of broiler weight [26, 27]. Animal health detection is another task; for example, there are studies that applied support vector machine (SVM) to classify sick broilers and computer vision system to identify sick broilers [28]. In another research, the hock burns in broiler chicken were identified by SVM using routinely collected farm management and slaughterhouse data [29]. Disease identification in the early stages is an important task performed manually and recently supported by machine learning [28, 30, 31].

Since the introduction of convolutional neural networks and various modifications of these networks, video and image process application in precision agriculture is rapidly growing. Typically, video analysis is performed where frames from video are extracted and processed; that is extracting, for example, low, mid and high-level features from the image and producing a trainable classifier. After that, the probability is generated, for example, hen sick/healthy or capturing certain characteristics of hen behaviour analysis with MI and CVS have been studied more and there are publications on this topic [32–34] usually experimenting with changing environmental factors, recording animal behaviour and applying machine learning algorithms to classify the behaviour.

Historical data is also combined with SVM, for example, to forecast drops in egg production [35] achieving an accuracy of 0.9854 in forecasting a day before and keeping high forecasting accuracy for 3 days in advance. A systematic review [36] of machine learning technology to record animal behaviour concluded that at present the stated research field is new and there is a broad diversity of approaches. Communication of researchers and common approaches to solving the issue is lacking at the moment which is usual for the emerging field of research. The authors also concluded that for poultry farms the machine learning challenge is the number of animals kept in a small volume of spaces which raise technological problems to detect a single animal with abnormal behaviour. Another systematic review [37] on the application of IT solutions in broiler farms concluded that solutions mainly address good health principles of broilers, mainly focusing on locomotor problems. Video and images are analysed in farms, but still, sound analysis is less used.

2.2 Multi-criteria Decision-Making in the Poultry Production and Importance of GHG Reduction

According to the Food and Agriculture Organization (FAO) of the United Nations report, during the 25 years period (2005–2030), global demand for chicken meat will grow by 61% and chicken eggs by 39% [6]. In 2019, the FAO estimates that 14.5% of all anthropogenic GHG emissions come from the livestock industry [38]. While beef produces most GHG, the pork and poultry industry are also main contributors [6]. Providing balanced chicken nutrition and improving feed conversion of the individual chicken are the solutions for reducing the GHG emissions, as well the following factors and their combination have a significant effect on the emission intensity in poultry production in different regions: feed conversion ratio, chicken productivity and manure management. At the same time, the other factors should be considered, for example, indoor air quality and humidity are the causes of emissions [5].

The previous results show examples of IT solutions to detect in the early stage the problems in poultry farming for its productivity increasing. But there is a lack of information on the multi-criteria decision-making (MCDM) approach used in poultry production. Usually, the decision process consists of the following stages:

- Identification and structuring problem to be solved,
- Model development, criteria selection and its weighting,
- The best solution selection among the alternatives.

Ribeiro [39] stresses that poultry farmers in future can be successful to apply precision agriculture techniques, control and monitor factors that affect poultry growth, including poultry feed ratio, humidity level at farm and others. Ribeiro also introduces the method to generate action plans by expert and then apply them in a monitoring process that is automatized using the sensor network. The important efforts for data analysis are continuous data collection, data collection technologies and poultry farm real-time automatic monitoring. The minimum records which must be recorded by the poultry producer are determined by FAO [40]. For example, Sallabi et al.'s research describes the process of implementation of real poultry production electronic mobile documentation system which collects, stores and processes the data that are related to daily activities and environmental conditions (temperature, humidity and ammonia concentration) at the poultry farms [41].

Based on the previously collected data at the poultry farm, a learning model (the machine learning techniques) for intelligent agents has been applied to classify patterns that improve feed conversion using feed data and indoor condition data such as light, ventilation, temperature and relative humidity. Experimental results have shown the feed conversion improvement by 6% and the broilers weight gain has increased by 7%. The suggested model improves the poultry farming process compared to the poultry specialist's decision [42]. Not only it is necessary to develop the most appropriate feeding and keeping process for the lowest level of CO_2 and NH_3 emission for every particular egg and broilers producer to enhance productivity, but as well reduce the disease, mortality and take care of the environment. It is necessary to develop the multi-criteria decision model for the best solution selection with respect to the following criteria:

- to maximize productivity or reduce disease and mortality,
- to minimize the level of CO_2 and NH_3 emission.

One of the existing solutions of the multi-criteria decision-making (MCDM) approach is developed for poultry production [43]. Investigation in laying hens feeding shows that MCDA can be used to optimize laying hens fed ingredients with the carbon footprint reducing and insignificant increase in expenses [44]. Good monitoring and collecting of the poultry feeding and the farm indoor environmental data at a technologically developed farm with data processing and mathematical data modelling are viable to improve poultry productivity.

3 Smart Platform for Improvement of Poultry Productivity and Reduction of Greenhouse Gas Emissions

The proposed platform is designed with the main task in mind: to allow precise manipulations to feed recipe by analysing an existing data to identify the effects of different feed recipes on the productivity and identification of additional factors which affect productivity such as seasonality and diseases. Estimate GHG production based on manipulations with feed and other factors, seasonality aspects, early detection of diseases, etc. Further, there is a need to introduce an algorithm that will allow migrating aggregated data sets in the system to define the most appropriate feed recipe by retrieving the best productivity and the lowest level of emissions at the same time. As a result, every particular egg producer will be able to optimize the feeding process, enhance productivity, and reduce disease, mortality and carbon and ammonia emissions.

3.1 Challenges of Implementation

There are certain challenges that need to be addressed in the real-world development of the platform. The identified challenges are:

- The system must be adaptable and customizable to add only the functionality that is needed. It can be seen that there is a wide spectrum of functions and purposes of systems for poultry management that can be complicated and often include unnecessary functions for a particular farm.
- System must be a cloud-based solution (also deployable on the premises servers), which allows farms to have little IT investments to be made at the enterprise level while using a sophisticated IT system and data centre, means cost reduction, better and safer data retention and an eco-friendlier approach.
- Usability and visualisations of data are critical for the system as main users will be the farmers and farm managers who might have less experience in the use of IT solutions.
- The system must provide support of file formats that are used in the field for data exchange and possibilities to add manual entries. It is possible that certain farms still do not use any IT solution for keeping records.
- As society is becoming aware of animal welfare issues and are demanding transparency and improvements in poultry production, the system needs to record data about the well-being of animals, recordings of environmental factors, diseases, etc. Thus supporting customers with the ability to provide full traceability of process "farm to table" and contribution to the introduction of better farming practices.
- There should be an implementation of data visualisations and estimation module to reduce greenhouse gas emissions.

3.2 Available Preliminary Data and Experimental Object

The eggs production farm was incorporated in a preliminary study, and the inputs and the outputs of production system data are collected at the farm. The eggs are produced intensively at the farm—over 50 million eggs a year. The hens at the farm are housed in battery cages. Eggs production system started when 16 weeks of age hens are placed in battery cages and begin laying eggs. The environmental conditions (temperature, humidity), amount of feed and water, mortality rate, hens age and productivity are controlled at the farm. Usually, after approximately 80 weeks of eggs production hens are slaughtered, but 5% die before this.

At the farm, 19 hens are put in one cage in a space of 14,490 cm^2 or 762 cm^2 per one hen. The number and age of hens in days and weeks are recorded. Mortality and mortality rate are recorded daily and cumulative mortality is calculated. The battery is provided by automatic feeders and waterers. The feed is given ad libitum, and the chemical composition of the feed, feed amount and rest are controlled at the farm. Feed consumption is calculated. During the productive cycle inside (in two different levels of cages) and outside temperature and inside humidity are recorded daily. The building is equipped with a conditioning system. Feed and environmental conditions are the significant factors that affect GHG emissions [5, 8, 45]. The number of eggs, weight and category are recorded daily at the layer farm. Feed consumption per kg egg mass and per dozen eggs can be calculated using the recorded traits at the farm. Egg production is related to feed intake and feed consumption ratio that affects GHG emissions [46].

4 Conclusions

Smart poultry management is a subcomponent of precision livestock farming and the whole field experience changes because of the rapid availability of technologies (machine learning, computer vision) and pressure from the government and consumer demands. Poultry farms existing solutions typically provide hardware (sensors, controllers, etc.) and software solutions that allow entering of data manually to generate reports with certain exceptions, where automation such as egg counting is implemented.

Machine learning in poultry farms is applied for the identification of animal health, detection of behaviour and forecasting tasks. Compared to previous study results, the chickens are today the world's most widespread bird species and they are the major contributors of NH_3 and GHG emission, hence the poultry farms produce 8% of the livestock sector's emissions. In order to reduce ammonia and GHG emission, as well to develop the most appropriate poultry feeding and keeping and to increase poultry productivity and at the same time to reduce disease and mortality, for the best solution, it is necessary to develop and apply the multi-criteria decision model.

Further research is finding an algorithm that will allow migrating aggregated data sets in the system and the proposed solution will bring optimal poultry productivity and will have a significant contribution to nature protection by reducing harmful CO_2 and NH_3 emission. The main task of proposed platform is the analysis of data to identify the effects of different feed recipes on the productivity and identification of additional factors which affect productivity and lower level of emissions at the same time.

Acknowledgements The research leading to these results has received funding from the Specific Objective 1.1.1 "Improve research and innovation capacity and the ability of Latvian research institutions to attract external funding, by investing in human capital and infrastructure" 1.1.1.1. measure "Support for applied research" project No. 1.1.1.1/19/A/145 "HENCO2: Cloud based IT platform designed to improve poultry productivity and reduce greenhouse gas emissions".

References

1. Westenbrink G, Beers G, Joosten F, Dijkxhoorn Y, Broek J, Lee J, Groot N (2017) Agricultural sector development. Guidelines for aid & trade support in Dutch PSD-partner countries. Ministries of agriculture, nature & food quality and foreign affairs in the Hague, the Netherlands
2. Augère-Granier ML (2019) The EU poultry meat and egg sector: main features, challenges and prospects: in-depth analysis and European Parliament. Directorate-general for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament, Brussels
3. Thornton A (2019) This is how many animals we eat each year. World Economic Forum homepage. https://www.weforum.org/agenda/2019/02/chart-of-the-day-this-is-how-many-animals-we-eat-each-year. Accessed 06 July 2020
4. Cheng HW (2010) Breeding of tomorrow's chickens to improve well-being. Poult Sci 89(4):805–813
5. Brouček J, Čermák B (2015) Emission of harmful gases from poultry farms and possibilities of their reduction. Ekologia Bratislava 34(1):89–100
6. MacLeod M, Gerber P, Mottet A, Tempio G, Falcucci A, Opio C, Vellinga T, Henderson B, Steinfeld H (2013) Greenhouse gas emissions from pig and chicken supply chains—a global life cycle assessment. Food and Agriculture Organization of the United Nations (FAO), Rome
7. Malomo GA, Bolu SA, Madugu SA, Usman ZS (2018) Nitrogen emissions and mitigation strategies in chicken production. Animal Husbandry and Nutrition. IntechOpen
8. Belloir P, Méda B, Lambert W, Corrent E, Juin H, Lessire M, Tesseraud S (2017) Reducing the CP content in broiler feeds: impact on animal performance, meat quality and nitrogen utilization. Animal 11(11):1881–1889
9. Yu Q, Yan Z, Xiu-li W, Bao-quan L (2018) Intelligent poultry environment control system based on internet of things. In: Sun X, Pan Z, Bertino E (eds) Cloud computing and security. ICCCS 2018. Lecture notes in computer science, vol 11067. Springer, Cham, pp 407–417
10. Raghudathesh GP, Deepak DJ, Prasad GK, Arun AB (2017) IoT based intelligent poultry management system using Linux embedded system. In: Balekai R, Yatnalli VC, Lata SH, Kumar BS (eds) 2017 International conference on advances in computing communications and informatics, ICACCI 2017. Institute of Electrical and Electronics Engineers Inc., Udupi, pp 449–454
11. Balachandar S, Chinnaiyan R (2020) Internet of things based reliable real-time disease monitoring of poultry farming imagery analytics. In: Pandian A, Senjyu T, Islam S, Wang H (eds) Proceeding of the international conference on computer networks, Big Data and IoT (ICCBI-2018). ICCBI 2018. Lecture notes on data engineering and communications technologies, vol 31. Springer, Cham, pp 615–620
12. Yu X, Liu P, Ren W, Zhang C, Wang J, Zheng Y (2018) Safety traceability system of livestock and poultry industrial chain. In: Sun X, Pan Z, Bertino E (eds) Cloud computing and security. ICCCS 2018. Lecture notes in computer science, vol 11068. Springer, Cham, pp 3–12
13. So-In C, Poolsanguan S, Rujirakul K (2014) A hybrid mobile environmental and population density management system for smart poultry farms. Comput Electron Agric 109:287–301
14. Ren G, Lin T, Ying Y, Chowdhary G, Ting KC (2020) Agricultural robotics research applicable to poultry production: a review. Comput Electron Agric 169:105216
15. Astill J, Dara RA, Fraser EDG, Roberts B, Sharif S (2020) Smart poultry management: smart sensors, big data, and the internet of things. Comput Electron Agric 170:105291
16. Yoshioka K, Higaki S, Ozawa T, Mikami O (2019) Early detection of livestock diseases by using wearable wireless sensors. FFTC Agricultural Policy Platform (FFTC-AP)
17. TULASI Homepage. https://www.tulassi.com/products/poultry-management/layer-management. Accessed 06 July 2020
18. UNITAS Software Homepage. https://www.poultrymanager.com/sectors/pullets-commercial-egg. Accessed 06 July 2020
19. FANCOM toward thinking Homepage. https://www.fancom.com/poultry/total-automation-for-layers. Accessed 06 July 2020

20. DACS innovation in ventilation Homepage. https://dacs.dk/about-dacs/climatization-in-a-poultry-house. Accessed 06 July 2020
21. FarmAlarm Monitoring Systems Homepage. https://farmalarm.com. Accessed 06 July 2020
22. Smart Barn Wireless Monitoring and Alarm System Homepage. https://smartbarn.io. Accessed 06 July 2020
23. StienenBE Homepage. https://www.stienenbe.com/en/poulty-farming. Accessed 06 July 2020
24. Schulz Systemtechnik Homepage. https://www.schulz.st/en/industrysector/agriculture. Accessed 06 July 2020
25. BigDuchman Homepage. https://www.bigdutchman.com/en/poultry-growing/products/broiler-breeder-management. Accessed 06 July 2020
26. Mortensen AK, Lisouski P, Ahrendt P (2016) Weight prediction of broiler chickens using 3D computer vision. Comput Electron Agric 123:319–326
27. Amraei S, Mehdizadeh SA, Sallary S (2017) Application of computer vision and support vector regression for weight prediction of live broiler chicken. Eng Agric Environ Food 10(4):266–271
28. Zhuang X, Bi M, Guo J, Wu S, Zhang T (2018) Development of an early warning algorithm to detect sick broilers. Comput Electron Agric 144:102–113
29. Hepworth PJ, Nefedov AV, Muchnik IB, Morgan KL (2012) Broiler chickens can benefit from machine learning: support vector machine analysis of observational epidemiological data. J R Soc Interface R Soc 9(73):1934–1942
30. Rizwan M, Carroll BT, Anderson DV, Daley W, Harbert S, Britton DF, Jackwood MW (2016) Identifying rale sounds in chickens using audio signals for early disease detection in poultry. In: 2016 IEEE global conference on signal and information processing, GlobalSIP 2016-proceedings. Institute of Electrical and Electronics Engineers Inc., Washington, pp 55–59
31. Okinda C, Lu M, Liu L, Nyalala I, Muneri C, Wang J, Zhang H, Shen M (2019) A machine vision system for early detection and prediction of sick birds: a broiler chicken model. Biosyst Eng 188:229–242
32. Pu H, Lian J, Fan M (2018) Automatic recognition of flock behavior of chickens with convolutional neural network and kinect sensor. Int J Pattern Recognit Artif Intell 32(7):1850023
33. Abdoli A, Murillo AC, Yeh CCM, Gerry AC, Keogh EJ (2019) Time series classification to improve poultry welfare. In: Proceedings-17th IEEE international conference on machine learning and applications, ICMLA 2018. Institute of Electrical and Electronics Engineers Inc., pp 635–642
34. Abreu LHP, Yanagi T Jr, Bahuti M, Hernandez-Julio YF, Ferraz PFP (2020) Artificial neural networks for prediction of physiological and productive variables of broilers. Eng Agrícola. Associação Brasileira de Engenharia Agrícola 40(1):1–9
35. Morales IR, Cebrian DR, Blanco EF, Sierra AP (2019) Early warning in egg production curves from commercial hens: a SVM approach. Comput Electron Agric 121:169–179
36. Wurtz K, Camerlink I, D'Eath RB, Fernández AP, Norton T, Steibel J, Siegford J (2019) Recording behaviour of indoor-housed farm animals automatically using machine vision technology: a systematic review. PLoS One 14(12):e0226669
37. Rios HV, Waquil PD, de Carvalho PS, Norton T (2020) How are information technologies addressing broiler welfare? A systematic review based on the welfare quality® assessment. Sustainability (Switzerland) 12(4):1413
38. Food and Agriculture Organization of the United Nations (FAO) Homepage. http://www.fao.org/news/story/en/item/197623/icode. Accessed 07 July 2020
39. Ribeiro R, Casanova D, Teixeira M, Wirth A, Gomes HM, Borges AP, Enembreck F (2019) Generating action plans for poultry management using artificial neural networks. Comput Electron Agric 161:131–140
40. Traceability Manual Poultry Industry, FAO Homepage. http://www.fao.org/tempref/GI/Reserved/FTP_FaoRlc/old/prior/segalim/animal/pdf/traeng.pdf. Accessed 07 July 2020
41. Sallabi F, Fadel M, Hussein A, Jaffar A, Khatib HE (2011) Design and implementation of an electronic mobile poultry production documentation system. Comput Electron Agric 76(1):28–37

42. Ribeiro R, Teixeira M, Wirth AL, Borges AP, Enembreck F (2015) A learning model for intelligent agents applied to poultry farming. In: Proceedings of the 17th international conference on enterprise information systems, ICEIS 2015, vol 1. SCITEPRESS-Science and Technology Publications, pp 495–503
43. Ma Q, Paudel KP, Cui L (2018) A multi-objective optimization problem for using poultry litter in electricity production. Appl Energy 228:1220–1242
44. Moe A, Koehler K, Bryan R, Goddard TW, Kryzanowski L (2014) Multi-criteria decision analysis of feed formulation for laying hens. In: Proceedings of the 9th international conference on life cycle assessment in the agri-food sector, LCA Food. San Francisco, USA, pp 835–844
45. Kocaman B, Esenbuga N, Yildiz A, Laçin E, Macit M (2006) Effect of environmental conditions in poultry houses on the performance of laying hens. Int J Poult Sci 5(1):26–30
46. Willems OW, Miller SP, Wood BJ (2013) Aspects of selection for feed efficiency in meat producing poultry. World's Poult Sci J 69:77–88

RPL-OC: Extension of RPL Protocol for LLN Networks Based on the Operator Calculus Approach

Abdusy Syarif, Mohamed-el-Amine Brahmia, Jean-François Dollinger, Abdelhafid Abouaissa, and Lhassane Idoumghar

Abstract Technical constraints imposed by low-power and lossy networks (LLNs) require to defer complexity to routing protocols in order to efficiently and reliably transmit packets. However, despite these constraints, the deployment of this type of network has increased considerably over the last years, particularly in smart cities area with focus on sensing applications. In order to effectively address this challenge, we propose a new mechanism for IPV6 Routing Protocol for LLNs (RPL) based on the Operator Calculus (OC) approach. In this multi-constrained path optimization problem, OC is applied to extract the feasible end-to-end paths while assigning a rank to each network node. Unlike the standard RPL and its variants which adopt a full distributed strategy, the aim is to provide RPL with a tuple containing the most efficient paths from a source node to the sink by considering multiple routing metrics. The definitive choice of the route is then delegated to RPL in order to take dynamic topology changes into account. The solution thus combines a multi-objective and semi-distributed routing algorithm with the RPL. Furthermore, to benchmark our proposal, we perform a comprehensive evaluation and compare it with other state of the art works. Performance evaluation results show that RPL-OC allows a great improvement compared to OFFL and the standard RPL, mainly in terms of end-to-end delay and energy consumption.

A. Syarif (✉)
Universitas Nasional, Jakarta, Indonesia
e-mail: abdusy.syarif@civitas.unas.ac.id

M.-A. Brahmia · J.-F. Dollinger
LINEACT, CESI Engineering School, Strasbourg, France
e-mail: abrahmia@cesi.fr

J.-F. Dollinger
e-mail: jfdollinger@cesi.fr

A. Abouaissa · L. Idoumghar
University of Haute-Alsace, Mulhouse, France
e-mail: abdelhafid.abouaissa@uha.fr

L. Idoumghar
e-mail: lhassane.idoumghar@uha.fr

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_7

Keywords IoT · LLN · Routing · RPL · Multi-constrained path · OC

1 Introduction

The Internet of Things (IoT) has a considerable impact and plays a major role as a new paradigm in the field of wireless communications. IoT networks are built upon massive interconnection of low cost embedded devices [1]. These tiny devices operate with constraints on low computing power, low memory, low energy, and their interconnects are characterized by high packet losses, low data rates and instability [2]. Therefore, routing protocols play a crucial role in transporting data efficiently and the proposal of an adapted solution becomes a challenge for many researchers. RPL essentially builds a tree topology called a Destination Oriented Directed Acyclic Graph (DODAG) over a mesh network. The DODAG construction is initiated by the root of the tree. To construct the DODAG, RPL uses a set of metrics and constraints through a specific Objective Functions (OFs) [3]. There are two OFs integrated in the standard specifications, the first one is Objective Function zero (OF0) which is a basic OF that has only one metric. It uses the hop count as a routing metric to select the best parents with the lowest rank. The second one is the Minimum Rank with Hysteresis Objective Function (MRHOF), which is based on a dynamic link metric such as expected transmission count (ETX) to recover rank stability [4, 5].

The specificity of RPL is that OFs are separated from the core protocol specification. This advantage allows researchers to improve RPL performance through implementing new OFs adapted to domain application and LLNs requirements. However, in several research works, the authors are limited to combine two or more routing metrics to enhance the routes quality based on parent-to-child relation. Indeed, the OF permits to select the preferred parent from the set of neighbors to be the next hop to the root but based only on the local vision of the nodes, in this case only neighbor nodes. To address this problem, we focus in this paper on optimizing RPL routing protocol in LLNs by using the Operator Calculus approach, called RPL-OC. The objective is to determine the end-to-end optimal paths from source to destination, or in other words from each node to the base station, by taking into account multiple metrics such as energy, latency, RSSI (Received Signal Strength Indicator), and hop count.

Our proposed scheme has interesting highlight: it provides optimal path which respect the total cost of defined constraints in an end-to-end manner. We strengthen the RPL routing protocol to adapt in a flexible way to the needs or demands of IoT applications. We demonstrate that this approach is effective to existing RPL protocol. The performance results show several insight into the behavior of protocol while we evaluate and compare with other related protocols in IoT network scenarios.

2 Related Work

Several enhancements of the OF-based RPL are proposed in the last years. Like in the standard OFs case, a part of these research studies suggest a new objective function based on only single metrics. For example, the authors in [6] use the connection status called Best-Friend ETX as a metric to offer a better quality of service. This new routing metric improves LLN network performances like packet delivery ratio and energy consumption. However, this solution does not take into account metrics related to the node and the impact of repetitive ETX updates in the choice of the parent. Another approaches consist on the composition of many node or link routing metrics to define objective function. There are several solutions proposed in the literature based on lexical metric, additive metric or fuzzy logic composition [7]. The authors in [8] proposed a new OF called ETX, Energy, and Queue length (EEQ) to protect the node which has already consumed excessively its energy during data transfer. The proposed OF combines three metrics ETX, active queue length, and energy consumption to compute nodes rank. EEQ reduces control messages overhead compared to the standard MRHOF and OF0 especially for high intensity traffic. Nevertheless, the performance of EEQ is not compared with OFs using metrics composition while taking into account more parameters like network lifetime and its convergence.

Gaddour et al. [9] proposed a new objective function called OFFL that integrates a set of different routing metrics based on the fuzzy logic approach. Their method uses a set of fuzzy parameters considered for the configuration of a routing decision using hop count, end-to-end delay, residual energy and link quality as metrics. OFFL offers the advantage of considering the application requirements with the objective to decide the best paths to the destination. Each metric depicts an important property of the neighbor node which is suitable for acting as a parent, thereby contributing to building a route toward the root. In the proposed OF, OFFL determines the qualities of a good path.

However, the optimization of multiple metrics is known as a NP-hard problem and is supported by some centralized path computation engine in order to have a global vision of the network. Among these solutions, the Operator Calculus was developed and used by Schott and Staples [10] for the last decade. This algebraic approach has been fitted to find the optimal paths using a combination of routing metrics for different application domains. The main concept underlying the OC approach is the association of graphs with algebraic forms whose properties show information about the associated graphs.

In [11], the authors proposed a routing protocol based on the OC approach to optimize and enhance the lifetime of WSN. The constraints considered to determine the optimal paths were residual energy and bit error rate of node. Their objective was to define the main route that contains high centrality rank in a WSN deployment. The evaluation of centrality was employing the OC approach based on network topology which offers optimal paths for all source node to the sink node. In addition, a number of studies have also been conducted between 2012 and 2020 to improve the RPL routing protocol in different approaches such as CORPL [12], DualMOP [13], QU-RPL [14], ORPL-DT [15], and BPRL [16].

3 RPL-OC Approach

The proposed scheme RPL-OC takes several routing metrics into account; i.e., energy, RSSI, delay and hop count. The combination of these routing metrics generates a multi-constrained path problem. In this work, we are focusing on a static environment where we consider a node whose route to the sink has the lowest hop count and delay, but at the same time has the highest RSSI and residual energy in order to obtain the optimal end-to-end paths. Thus, the multi-constrained path optimization problem in this case is to find routes or paths from a source node to the root whose costs are respecting the constraints given. An in-depth development of OC is exposed in the work of Schott et al. [17]. The aforementioned algorithm is primarily dedicated to solve precomputed routing problems. OC is based on the properties of matrix computation and the definition of algebraic structures. Applied in the context of routing, OC is implemented to solve the *multi-constrained optimal path problem* (MCOP) [11, 18, 19].

A weighted connected simple adjacency graph $G = (V, E)$, formed by a set of nodes V linked by directed or simple edges specified in set E, is represented by a nilpotent square matrix denoted Ψ of order $card(V)$. The OC underlying algorithm is parameterized by Ψ whose components associate algebraic structures. A path is denoted by a unique combination of node labels, constructed through concatenation. Let an alphabet $\sum_n := \{\omega_i : \forall i \in \{1, \dots, n\}\}$ along with $\omega_\emptyset$ the empty symbol. For fixed k, a sequence of symbols forms a *word* denoted as $\omega_u := \omega_{u_1}\omega_{u_2} \dots \omega_{u_k}$. The noncommutative semigroup Ω_n represents the set of unique words, including empty word 0, formed with a given alphabet $\sum_n$. We define the concatenation operation on words $w_u, w_v \in \Omega_n$ as follows:

$$\omega_u \cdot \omega_v = \begin{cases} \omega_{u.v} & \text{if } u \cap v = \emptyset \\ 0 & \text{otherwise} \end{cases}$$

For convenience, a word will be expressed through multi-index notation, such that $u := (u_1, u_2, \dots, u_k)$. Each path is coupled to a weight, quantitatively describing the quality of the routing. A path $u := (u_1, u_2, \dots, u_k)$ of aggregated multi-weight $\mathbf{x} \in \mathbb{R}^m$ will be described in $\mathbb{R}\Omega_n$. With $*$ an m-tuple of associative binary operators on $\mathbb{R}$, for path $\mathbf{p}$ we define the weight as follows:

$$\mathrm{wt}(\mathbf{p}) := (w_{i1} *_1 w_{j1}, \dots, w_{im} *_k w_{jm}), \forall (v_i, v_j) \in \mathbf{p} \tag{1}$$

Given a first vertex v_0 and a last vertex v_∞ in a weighted graph, the set of *feasible paths* from v_0 to v_∞, also expressed $v_0 \sim v_\infty$, relates to each path whose associated overall costs respect some predetermined constraints. The vector $\mathfrak{C}$ is named the *constraint vector* relative to the *constraints relation* $\mathfrak{R} = (R_1, R_2, \ldots, R_m)$, a m-tuple of reflexive, transitive, antisymmetric relations on $\mathbb{R}$. A vector $\mathbf{x} \in \mathbb{R}^m$ is stated to *satisfy* $\mathfrak{C}$, denoted $\mathbf{x} \preceq \mathfrak{C} \Leftrightarrow \mathbf{x}_i \preceq_i \mathfrak{C}_i, \forall i \in [1, m]$, if and only if $(\mathbf{x}, \mathfrak{C}) \in \mathfrak{R}$. The *constraints algebra*, indicated $\mathcal{A}_{\mathfrak{C}}$, is the real associative unital algebra produced by $\{\nu^{\mathbf{x}} : \mathbf{x} \in \mathbb{R}^m\}$ with neutral element $\nu^{\mathbf{0}}$ having multiplication defined as follows:

$$\nu^{w_x}\nu^{w_y} = \begin{cases} \nu^{w_x * w_y} & \text{if } x * y \preceq \mathfrak{C} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

A path $\mathbf{p}$ weight $\mathrm{wt}(\mathbf{p})$ will be expressed via multi-exponent notation. Given a finite graph G wherein every edge is weighted with an m-tuple of positive integers and a constraint vector $\mathfrak{C} = (c_1, \ldots, c_m) \in \mathbb{R}^m$, the multi-constrained path problem requires the computation of the weight $\mathrm{wt}(\mathbf{p})$ of path $\mathbf{p}$ from source vertex v_0 to destination vertex v_∞ in the graph G as follows:

$$\mathrm{wt}(\mathbf{p}) \preceq (c_1, \ldots, c_m) = \mathfrak{C}. \tag{3}$$

The partial order $\preceq$ on the multi-exponents showing in the canonical expansion of any component $u \in \mathcal{A}_{\mathfrak{C}}$ can be totally ordered by sorting. With $\wedge$ and $\vee$ the logical conjunction (AND) and disjunction (OR) and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, this sorting is made lexicographically such that

$$\begin{aligned} \mathbf{x} \preccurlyeq \mathbf{y} \Leftrightarrow (x_1 \prec_1 y_1) \vee [(x_1 = y_1) \wedge (x_2 \prec_2 y_2)] \\ \vee \cdots \vee [(\forall i < m, \quad x_i = y_i) \wedge (x_m \prec_m y_m)] \end{aligned} \tag{4}$$

Within these feasible paths, a preferred or an optimal path may then be selected. For practical purpose, $\mathcal{P}$ will indicate the set of *all paths* in the graph G, and $\mathfrak{P}$ will designate the set of *all possible paths* in G, i.e., $\mathfrak{P} := \{\mathbf{p} \in \mathcal{P} | \mathrm{wt}(\mathbf{p}) \preceq \mathfrak{C}\}$. A path $\mathbf{p} \in X \subset \mathfrak{P}$ is stated to be *optimal* or *preferred* for X if the following relation holds

$$\mathrm{wt}(\mathbf{p}) \preccurlyeq \mathrm{wt}(\mathbf{q}), \ \forall \mathbf{q} \in X \tag{5}$$

Although the $\preceq$ operator imposes an order on the monomials $\{\nu^{\mathrm{wt(p)}}\omega_p : p \in \mathfrak{P}\}$ one may introduce a choice function to extract a specific set of paths out of multiple equivalent choices as expressed in relation (4):

$$\daleth(\sum_{p \in X} \nu^{\mathrm{wt(p)}}\omega_p) = \{\nu^{\mathrm{wt(p)}}\omega_p | \mathrm{wt}(\mathbf{p}) \preccurlyeq \mathrm{wt}(\mathbf{q}), \ \forall \mathbf{q} \in X\} \tag{6}$$

From the previous definitions follows the specification of the *nilpotent multi-weight adjacency matrix* Ψ of $n \times n$ elements modeling graph G with entries in $\mathcal{A}_{\mathfrak{C}} \otimes \Omega_n$. Each matrix component couples a weight to a path, with nonexistent links denoted by 0, such that

$$\Psi_{ij} = \begin{cases} \nu^{w_{ij}} \omega_j & \text{if } (\nu_i, \nu_j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The $\langle v_i|\Psi$ expression designates the ith line of matrix Ψ. Conversely, $\Psi|v_j\rangle$ designates the jth column of matrix Ψ. An entry Ψ_{ij} of matrix Ψ is thus entirely designated by $\langle v_i|\Psi|v_j\rangle$. Extraction of multiple paths is based on computing powers of the Ψ matrix. As elements of Ψ are noncommutative, Ψ^n is defined by the following recurrence relation:

$$\Psi^n = \begin{cases} u_n = u_{n-1} \times \Psi \\ u_1 = \Psi \end{cases} \tag{8}$$

Each new computed power adds up one new hop to each feasible path of the matrix; computing Ψ^l generates paths of length l. Matrix entries resembling multiple alternative paths are represented by summation. Classical distributivity applies when multiplying a summation of $\mathcal{A}_{\mathfrak{C}} \otimes \Omega_n$ entries by another. One may thus compute a path $v_i \sim v_\infty$ by applying the following operations:

$$\nu^0_{\omega(i)} \sum_{l=1}^{n} \langle v_i|\Psi^l|v_\infty\rangle = \nu^0_{\omega(i)} \sum_{\textbf{paths } p:v_i \sim v_\infty} \nu^{w_p} \omega_p \tag{9}$$

Extracted paths are cycle-free guaranteed, as specified by the algebra itself. Paths which do not satisfying constraints are automatically pruned as the algorithm progresses. The optimal preferred paths are outputted after selection handled by the $\daleth$ function.

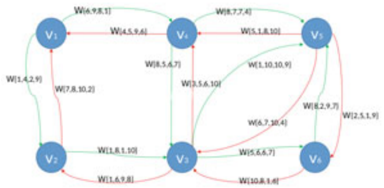
Algorithm 1 RPL-OC approach for paths $v_i \sim v_\infty$

Input: $*, \preceq, \Gamma$
Output: End-to-end hop minimal optimal path for destination $v_i \sim v_\infty$
Collecting info of $G = (V, E)$
Building constraint vector $\mathfrak{C} = buildconstraints(\Psi)$
Building adjacency matrix $\Psi_{ij} = \begin{cases} v^{w_{ij}} \omega_j & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$.
$u = getrow(\Psi, i)$
while $u \neq \overrightarrow{0}$ and $u_{v_\infty} = 0$ **do**
    $u = vecmat(u, \Psi, \mathfrak{C}, *, \preceq)$
end while
return $\daleth(u_{v_\infty})$

The RPL-OC routing algorithm is conceptually summarized in Algorithm 1 through classical algebraic constructions. The algorithm generates a combination of feasible optimal paths for route $v_i \sim v_\infty$ given an adjacency matrix Ψ. It is followed by the construction of the adjacency matrix which entries associate a link and its multi-weight vector. Subsequently, successive vector-matrix operations can be computed as stated in Algorithm 2, in order to generate the feasible paths of interest. Figure 1a depicts a toy example of a 6-node multi-weighted graph. Each edge

Algorithm 2 OC vector-matrix multiplication

function vecmat(u, Ψ, $\mathfrak{C}$, $*$, $\preceq$)
    $n = cols(v)$
    for j in $1 : n$ **do**
        $u_j = \sum_i^n \Big(\sum_{\textbf{paths}\ p : u_j} v^{w_p} \omega_p * v^{w_{ij}} \omega_j \Big)$
    end for
    return u
end function



(a) Six nodes multi-weighted graph.

$$\Psi = \begin{bmatrix} 0 & \nu^{(1,4,2,9)}_{\omega(2)} & 0 & \nu^{(6,9,8,1)}_{\omega(4)} & 0 & 0 \\ \nu^{(7,8,10,2)}_{\omega(1)} & 0 & \nu^{(1,8,1,10)}_{\omega(3)} & 0 & 0 & 0 \\ 0 & \nu^{(1,6,9,8)}_{\omega(2)} & 0 & \nu^{(3,5,6,10)}_{\omega(4)} & \nu^{(1,10,10,9)}_{\omega(5)} & \nu^{(5,6,6,7)}_{\omega(6)} \\ \nu^{(4,5,9,6)}_{\omega(1)} & 0 & \nu^{(8,5,6,7)}_{\omega(3)} & 0 & \nu^{(8,7,7,4)}_{\omega(5)} & 0 \\ 0 & 0 & \nu^{(6,7,10,4)}_{\omega(3)} & \nu^{(5,1,8,10)}_{\omega(4)} & 0 & \nu^{(2,5,1,9)}_{\omega(6)} \\ 0 & 0 & \nu^{(10,8,1,6)}_{\omega(3)} & 0 & \nu^{(8,2,9,7)}_{\omega(5)} & 0 \end{bmatrix}$$

(b) Graph associated adjacency matrix Ψ.

Fig. 1 Sample network graph and associated adjacency matrix

is weighted by a vector of positive integers. The task at hand is to find all feasible optimum paths from vertex v_1 to v_5. Each edge is tied a multi-weight vector subject to the constraint vector $\mathfrak{C} = (20, 17, 20, 18)$, associated to the binary relation vector $\mathfrak{R} = (\leq, \leq, \leq, \leq)$. Thus a path $\mathbf{p}$ of multi-weight $\mathrm{wt}(\mathbf{p})$ is feasible exclusively if $\mathrm{wt}(\mathbf{p}) \preceq (20, 17, 20, 18)$. A path multi-weight vector is computed step-by-step with the help of a vector of binary operators $* = (+, +, +, +)$. Responding to the graph, the associated constrained path-identifying nilpotent adjacency matrix is defined in Fig. 1b. An intermediate result of the operations stated in Eq. 9 applied to the example shown in Fig. 1 is depicted hereafter in column vector ψ:

$$\psi = \begin{bmatrix} \nu^{(14,16,15,5)}_{\omega_{(1,4,5)}} \\ 0 \\ \nu^{(1,10,10,9)}_{\omega_{(3,5)}} + \nu^{(11,12,13,14)}_{\omega_{(3,4,5)}} + \nu^{(13,8,15,14)}_{\omega_{(3,6,5)}} \\ \nu^{(8,7,7,4)}_{\omega_{(4,5)}} + \nu^{(9,15,16,16)}_{\omega_{(4,3,5)}} \\ 0 \\ \nu^{(8,2,9,7)}_{\omega_{(6,5)}} \end{bmatrix}$$

More specifically, $\psi = \nu^0_{w_{(i)}} \sum_{l=1}^6 |\Psi^l|v_5\rangle$ generates the feasible paths to destination v_5. Finally, the constraint satisfying feasible paths from vertex v_1 to v_5 are extracted by inspecting element ψ_1 of vector ψ.

Table 1 Simulation parameters

Parameter	Value
No. of node	25
No. of sink	1
Transmission range	20 m
Interference range	40 m
Area of deployment	$80 \times 110\ m^2$
Radio duty cycling protocol	ContikiMAC
Protocols	RPL-OC, RPL, OFFL
Data type	UDP
Data rate	[1, 10, 20, 30, 40, 50, 60] packets/min
Duration	100 min

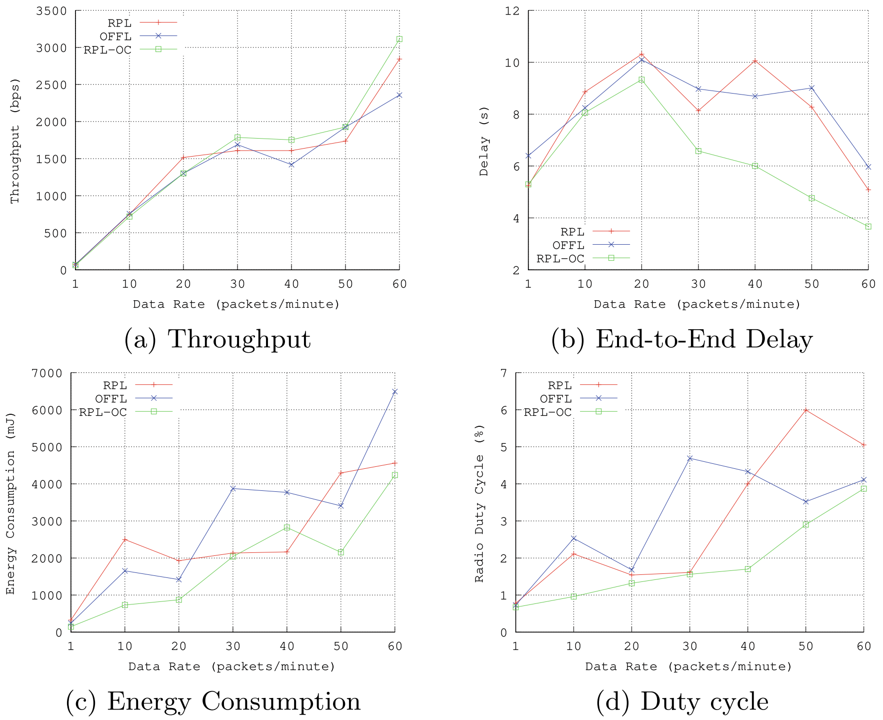
4 Performance Evaluation

In this section, energy consumption, delay, duty cycle, and throughput are investigated in order to evaluate the performance of RPL-OC. Simulations are conducted on Cooja simulator which uses the ContikiRPL. The objective of this simulation study is to investigate the network behavior by comparing RPL-OC with the standard RPL and OFFL [9]. A complementary statistical analysis of several RPL variants is presented in another study [20]. The parameters that we used in our simulation scenarios are given in Table 1.

Scenario I. To assess the throughput of the network, we increased the data traffic from 1 to 60 packets/min and we used UDP traffic to evaluate the round-trip time between source and destination (the root). RPL-OC still has a good performance rather than the others under different data rates. For example, as seen in Fig. 2a, when the data rate is set to 60 packets/min, RPL-OC provides an average throughput of 3112.5 bps, RPL achieves 2842.3 bps, and OFFL obtains 2356.7 bps. This phenomenon is caused by our method that provides the optimum end-to-end paths with respect to constraints in order to deliver the packet to the DAG root.

In terms of delay, RPL-OC obtains lower delay for all scenarios compared to RPL and OFFL, as seen in Fig. 2b. This is due to fact that RPL-OC selects the path according to delay metric thresholds while constructing the path. In general, RPL-OC has lower average end-to-end delay than OFFL and RPL. For data rate of 60 packets/min, OFFL and RPL obtained 5.96 s and 5.08 s respectively. While the data rate increased from 1 to 60 packets/min, we can see that the average end-to-end delay decreased from 5.29 to 3.66 s for RPL-OC.

Figure 2c shows the average energy consumption of our proposed scheme and the comparison with RPL standard and OFFL in relation with data rate. However, RPL-OC has the lowest consumption of energy followed by RPL and OFFL during each 100-min simulation time. At slow data rate, less than 10 packets/min, our proposed RPL-OC and OFFL are better than RPL. Also at higher data transmission, 50 and 60



(a) Throughput

(b) End-to-End Delay

(c) Energy Consumption

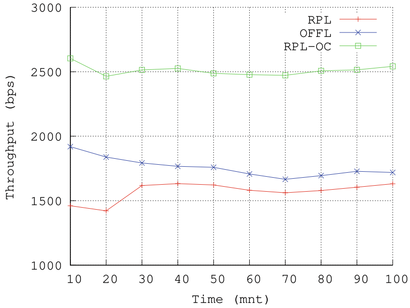
(d) Duty cycle

Fig. 2 Comparison of throughput, end-to-end delay, energy consumption, and duty cycle in relation to data rate in scenario I

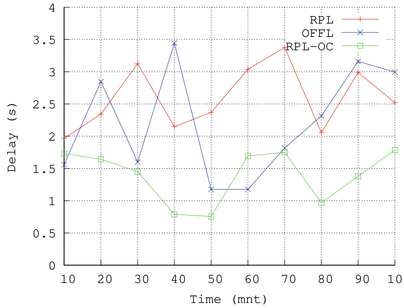
packets/min, RPL-OC obtains lower energy consumption than the others. Through our simulation, we observed that the effect of packet transmission rate on the energy consumption level is high. It shows that RPL and OFFL consume more energy than RPL-OC with the increase in data rate. This phenomenon is due to additional number of data packets transmitted in the network during the simulation. We note for example with 1 packet/minute after 100 min of simulation, RPL-OC reduces the energy consumption by nearly 23% less than RPL.

Figure 2d shows the duty cycle comparison between our proposed algorithm and other protocols during 100-min simulations in relation to data rate. Both RPL-OC and OFFL perform almost the same at the data rate of 1 and 60 packets/min. RPL achieved 0.78% of the average duty cycle which is a little bit higher than RPL-OC and OFFL during 100-min simulation for data rate 1 packet/min. However, for data rate of 50 and 60 packets/min, RPL suffers from the radio duty cycle of more than 5% due to topology maintaining scheme. On the other hand, our proposed mechanism, RPL-OC, attained the lowest average duty cycle.

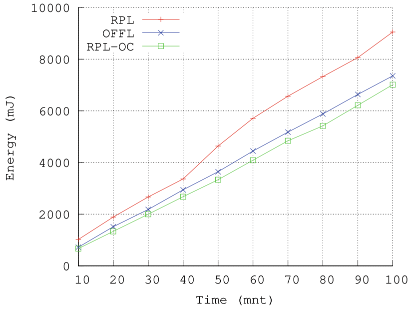
Scenario II. In the second scenario, we conducted tests where some nodes drop one by one during 100 min of simulation with data rate 60 packets/min in order to see the performance of RPL-OC and other protocols. A first node drops after 5 min



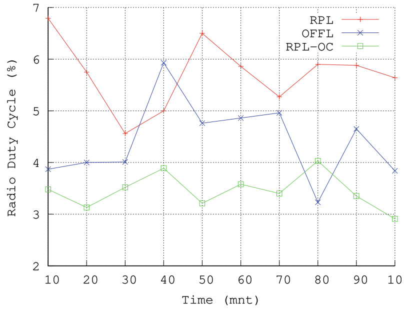
(a) Throughput



(b) End-to-End Delay



(c) Energy Consumption



(d) Duty cycle

Fig. 3 Comparison of throughput, end-to-end delay, energy consumption, and duty cycle in Scenario II for 100-min simulation time

and then every 10 min. Thus, on the overall, 10 nodes drop. In terms of throughput, as depicted in Fig. 3a, RPL-OC obtains highest throughput than the other protocols. This is clear as when there is a change in topology, RPL-OC is able to switch paths which maintain the throughput. It can be seen that RPL-OC experiences about 59% higher throughput when compared to RPL. Moreover, RPL-OC efficiently maximize the stability of network since RPL-OC provides optimal multipath to reach the root. Another reason is that higher link connectivity might reduce the possibility of parent shifting in RPL and OFFL networks. This is shown that RPL is lack of responsiveness and inefficient for this scenario.

Figure 3b demonstrates the average end-to-end delay for scenario II. We notice that the performances of RPL and OFFL were higher than our proposed algorithm. Indeed, by using RPL-OC, the average end-to-end delay is reduced to 1.2 s or 46.2% of RPL. This is due to RPL and OFFL protocols which perform a sequence number of process to update their rank, in order to allow the nodes to select a new preferred parent. Otherwise, RPL-OC has identified a set of optimal parents and has a quick parent shifting process which leading to lower end-to-end delay. While dropped nodes occurred, as seen in Fig. 3c, RPL-OC has a different energy tradeoff from the others. RPL-OC consumes 25.3% less energy than RPL and, in general, it expends lesser energy than the others which means that our scheme is the most energy efficient. This is due to fact that RPL-OC has end-to-end paths with the shorter delay, the better RSSI, and the minimum energy consumption. Even when dead nodes appeared, RPL-OC can provide an alternative path with respect to the constraints.

Whereas we performed dropped nodes scenario on the network during 100-min simulation time, RPL-OC achieved 3.45% of average radio duty cycle, followed by OFFL which obtained 4.41 and 5.71% for RPL, as seen in Fig. 3d. It means that the proposed scheme agreed to reduce the energy consumption and while the source sends the packet, the nodes were continually alive. In addition, it is clear that the RPL network keeps updating its new metrics, especially at higher data rate of 50 packets/min and 60 packets/min.

5 Conclusions

This paper presents our proposed algorithm on routing protocol RPL based on the Operator Calculus approach, called RPL-OC. We have been widely evaluating some RPL variants and performed several simulations with a static network in IoT application domains, such as smart cities. The key to this algebraic approach is that when passing from one path to the next path, all the four metrics, i.e., energy, RSSI, hop count, and delay, are taken into account to estimate the end-to-end optimum paths. We remarked that with a global network topology view, there are clear advantages to RPL-OC to determine the optimal end-to-end paths in LLN networks. In addition, RPL-OC has capability to find the optimal paths and communicates them to the nodes during the construction of the DODAG in a semi-distributed manner. Thus, having recovery routes and fewer signaling messages is made possible. As shown in the result section, our proposed scheme allows to solve the issue of finding the optimal path by regarding maximal RSSI and energy level efficiently but with the minimal end-to-end delay and hop count at once. Future works will address the heterogeneous nodes, mobility, and scalability to provide support for dynamic adaptation. Other metrics such as throughput and latency could be considered in order to improve the reliability and robustness of the proposed schemes.

References

1. Zhao M, Kumar A et al (2017) A comprehensive study of RPL and P2P-RPL routing protocols: implementation, challenges and opportunities. Peer-to-Peer Netw Appl 10(5):1232–1256
2. Gaddour O, Koubâa A (2012) RPL in a nutshell: a survey. Comput Netw 56(14):3163–3178
3. Lamaazi H, Benamar N et al (2016) Study of the impact of designed objective function on the RPL-based routing protocol. In: International symposium on ubiquitous networking. Springer, pp 67–80
4. Aljarrah E, Yassein M et al (2016) Routing protocol of low-power and lossy network: survey and open issues. In: International conference on engineering and MIS. IEEE, pp 1–6

5. Kharrufa H, Al-Kashoash H et al (2019) RPL-based routing protocols in IoT applications: a review. IEEE Sens J 19(15):5952–5967
6. Sanmartin P, Jabba D et al (2018) Objective function BF-ETX for RPL routing protocol. IEEE Lat Am Trans 16(8):2275–2281 Aug
7. Lamaazi H, Benamar N (2020) A comprehensive survey on enhancements and limitations of the RPL protocol: a focus on the objective function. Ad Hoc Netw 96:102001
8. Sarwar S, Rauf S et al (2019) Energy-aware routing in internet of things (IoT). In: International conference on communication, computing and digital systems, March 2019, pp 81–86
9. Gaddour O, Koubâa A et al (2015) Quality-of-service aware routing for static and mobile IPv6-based low-power and lossy sensor networks using RPL. Ad Hoc Netw 33:233–256
10. Feinsilver P, Schott R (2012) Algebraic structures and operator calculus: volume I: representations and probability theory, vol 241. Springer Science & Business Media
11. Syarif A, Abouaissa A et al (2019) New path centrality based on operator calculus approach for wireless sensor network deployment. IEEE Trans Emerg Top Comput 7(1):162–173 Jan
12. Aijaz A, Su H et al (2015) CORPL: a routing protocol for cognitive radio enabled AMI networks. IEEE Trans Smart Grid 6(1):477–485
13. Ko J, Jeong J et al (2015) DualMOP-RPL: supporting multiple modes of downward routing in a single RPL network. ACM Trans Sens Netw 11(2)
14. Kim H, Paek J et al (2015) QU-RPL: queue utilization based RPL for load balancing in large scale industrial applications. In: 12th annual IEEE international conference on sensing, communication, and networking (SECON), pp 265–273
15. Kang D, Kim H et al (2017) ORPL-DT: opportunistic routing for diverse traffic in multihop IoT networks. In: 2017 IEEE global communications conference on GLOBECOM 2017, pp 1–6
16. Tahir Y, Yang S et al (2018) BRPL: backpressure RPL for high-throughput and mobile IoTs. IEEE Trans Mob Comput 17(1):29–43
17. Schott R et al (2017) Generalized Zeon algebras: theory and application to multi-constrained path problems. Adv Appl Clifford Algebras 27(1):45–57
18. Benslimane J, Schott R et al (2015) Operator calculus algorithms for multi-constrained paths. Int J Math Comput Sci 10(1)
19. Cruz-Sanchez H et al (2007) Precomputed routing in a store and forward satellite constellation. In: 66th vehicular technology conference, Sep 2007, pp 240–243
20. Azman N, Syarif A et al (2020) Performance analysis of RPL protocols in LLN network using Friedman's test. In: 7th international conference on internet of things: systems, management and security (IoTSMS), Dec 2020

Evaluating Challenges in Using Big Data in Healthcare

Rajip Raj, Farhad Daneshgar, and Nazanin Borhan

Abstract Data in healthcare refers to large medical data collected from various sources comprising electronic health records (EHR), medical imaging, genomic ordering, pharmaceutical testing, wearables, and medical equipment. The way data is handled, evaluated, and leveraged in healthcare applications has already been affected by big data. The effective use of big data in the healthcare sector can prevail the ability to decrease medical expenses, forecast the outburst of an epidemic, circumvent preventable illnesses, and enhance the overall quality of life. Despite the enormous benefits, applying big data in the healthcare sector poses numerous challenges in terms of the nature of human interactions, healthcare applications, and large and complex data. This qualitative research is the first step of a larger research project that investigates various challenges associated with using big data in healthcare applications. The results of the present study will inform the next project phase in a future study to design and develop effective human interactions in healthcare big data applications. Data are collected from relevant existing literature and then analyzed through a thematic analysis. The coding criteria for such data analysis are those generic codes that are commonly used by similar studies and include technical challenges, administrative challenges, and security and privacy challenges.

Keywords Big data challenges · Healthcare applications · Human interactions

R. Raj (✉) · F. Daneshgar · N. Borhan
Victoria University, Sydney, Australia
e-mail: rajip.raj@student.vu.edu.au

F. Daneshgar
e-mail: farhad.daneshgar@vu.edu.au

N. Borhan
e-mail: nazanin.borhan@vu.edu.au

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_8

1 Introduction

Healthcare sectors are producing extensive information by keeping records of patient care, acquiescence, and regulation requirements. According to [1] the major sources of big data in healthcare are:

- Data generated by machines that include information from digital sensors, wearable devices, intelligent meters, and other machines measuring essential indications.
- Biometric data acquired from physical features of people such as fingerprints, genes, signature, scans of retina, heart rate, blood pressure, X-ray, and other clinical images.
- Human-produced data that include unorganized and semi-organized clinical information such as case documents, laboratory outcomes, hospital admittance, overviews of discharge, and email.
- Transactional data that include health claims information and accounting data.
- Epidemiological data that include important statistical data, data from health surveys, and registrations of diseases.
- Publication data that include information from medical studies and clinical reference data.

Big data in healthcare is defined by its infamous five characteristics [2]: (i) Volume: Data is persistently produced in huge amount from various sources such as electronic health records, real-time health tracking system, clinical trials, labs, research, etc.; (ii) Velocity: It is the requirement for real-time analysis of information from streaming data such as remote patient tracking, sensor data, telemedicine, etc.; (iii) Variety: Data is collected from various sources and can be organized, semi-organized or unorganized; (iv) Veracity: It defines the quality of data that are being collected; and (v) Value: This deals with obtaining value from data.

Although most information is stored in a form of hard copy, the present tendency is to quickly digitize these big quantities of information. Such tendency is due to the mandatory requirements and the ability to escalate the delivery of healthcare while lowering expenses. It is expected that these vast quantities of information known as big data enable a broad variety of medical choice making and healthcare tasks, incorporating assistance for clinical decision-making, disease observational, and administration of population health. Big data in healthcare is immense not only due to its quantity but also due to the variety of kinds of data and the velocity at which it needs to be managed [3]. Through digitizing, mixing, and applying big data efficiently, healthcare organizations can transform from being small clinics to multi-provider networks of big hospitals. Furthermore, responsible care organizations will have a major advantage in perceiving and managing individual and mass health and discovering fraud in healthcare more promptly and effectively. Overall, a successful big data implementation can bring many positive administration and life-saving consequences. Examples are real-time tracking, patient management, utilizing health data for deliberate planning, and improve security.

Despite many benefits, there are some obstacles to the application of big data in the healthcare sector. This research investigates the application of big data in the healthcare industry and identifies challenges that the industry is tackling, through a critical review of the related literature and the authors' analysis of the review. The ultimate goal of our two-phased research project is to use the results of the present study to design and develop effective human interactions in healthcare big data applications. To partially achieve the above goal in the current study, the remaining parts of the paper has been organized as follows: In the next section focused reviews are performed in three related areas, including (i) overall challenges of big data, (ii) application of big data in healthcare, and (iii) challenges for the healthcare industry; the former represents the bulk of findings of the study. To add rigor to the study, the last section of the paper provides the method used to arrive at the review results.

2 Literature Review

The current work-in-progress is the first stage of a two-phased research project where various challenges of big data that would affect healthcare applications are identified. These challenges are synthesized from the current studies and are grouped and presented in the following three sections.

2.1 Overall Challenges of Big Data

Big data challenges reported in the current literature generally fall into two categories: (i) challenges related to the storage and system architecture and (ii) the analytics dimension. This study provides a focused review of both dimensions.

Challenges related to the nature of the data in big data: Such data are extremely large and complex that conventional applications are unable to process them. Both humans and machines can produce big data. It can be produced anywhere in both organized and standard formats as well as unorganized formats. It can be classified broadly based on the sources: machine-produced, human-produced, and organization-produced data, although machine-generated data constitute the bulk of big data. Machine-produced data indicate data produced from live time sensors in industrial machinery or transportation. These data are generated from different detectors, cameras, satellites, log documents, bioinformatics, action trackers, private health trackers, and various other sensory information resources.

Human-generated data is a vast quantity of social media information such as status, pictures, videos, and text messages. Normally, the data produced in this classification are unstructured. On the other hand, information produced by the organization is extremely organized and reliable in essence and contains a specified field or file in the type of records, and commonly stored using a relational database. Several studies currently exist that address the data challenges in big data including [2, 4–10].

The second dimension is related to big data analytics. Currently, there is a greater need to analyze large amounts of data, particularly in healthcare applications. Terabyte and petabyte of data volume are quickly becoming prevalent, and hidden with significant value, waiting for the right computational resources to unlock them. Big data's heterogeneity, size, timeliness, sophistication, and confidentiality issues hinder progression at all stages of the procedure which can generate meaning from the data. Some studies propose specific methods for overcoming underlying challenges in big data analysis with MapReduce system over Hadoop distributed file system [11, 12]. Others provide generic methods for machine learning, data mining, and computational intelligence (e.g. [13]). These studies outline the main features of big data, and how they create a fundamental shift in statistical and numerical approaches as well as computing infrastructure.

2.2 Application of Big Data in Healthcare

Big data in healthcare systems can be utilized for forecasting disease and infection outbreaks, intensify treatment and standard of living, and avoid untimely demise and the spread of diseases. It gives insights on the treatment of disease and caution signs for medication. It will not only assist reduce the number of patients and death rates but will also help the government to decrease medical care costs.

The healthcare industry in Australia is administered and financed by the Australian government and governments of the state and territories, as well as private and non-profit organizations. Governments use Medicare and other programs to fund the majority of spending, which is around 67% of the spending [14]. Personally Controlled Electronic Health Record (PCEHR) is a summary of digital health data that are created by the Australian government. PCEHR aims to provide a safe electronic description of the medical history of individuals that will subsequently contain facts such as present medicines, harmful drug responses, sensitivity, and immunization background in a readily available format. PCEHR is recorded in a group of linked devices with the capacity to enhance data exchange among healthcare operators in order to enhance patient results regardless of the location of the patient throughout Australia [15].

Using big data is now common practice for discovering and managing high-risk and high-cost patients in the United States [16]. There is a dramatic rise in the volume of healthcare data that are accessible electronically as the healthcare institutions in the United States are swiftly embracing the EHR. Concurrently, there have been many achievements in healthcare data analytics for analyzing the vast amount of data and obtain new information through big data. This study discovered six major use cases of big data in healthcare in the United States alone [16]. These are briefly explained below.

In relation to the high-cost patient, another recent study showed roughly 5% of patients account for approximately 50% of the entire expenditure of US healthcare [16]. Predictive analysis can be applied to identify and manage such patients.

Predictive analysis may involve connecting data from various sources, such as healthcare-related, genetics, administrative, reports, social data, and insurance. Healthcare resources are becoming alarmingly restricted, requiring a higher emphasis on meaning. It will therefore be essential to explore analytical methods that discover not only individuals at high risk but also the people at low risk. The second use case is for re-admitting patients that helps reduce the frequency of admitting patients. For instance, the patient with smartphones can be asked to give permission to ingress data from their mobiles for recognizing the patients who are unable to manage a chronic situation better or to keep an eye on those patients who have been recently removed from healthcare centers. Those patients unable to make phone calls or sending emails at their regular rate might be going through depression or other health issues.

The next big data application is triage. It would be helpful to predict the possibilities of complications during the first visit of the patient to healthcare centers for numerous reasons. Such predictions may aid in the allocation of staff and bed, the possibility of transferring to the suitable unit and building, and so on.

The other use of big data analytics is for compensation. There is often a period prior to decompensating—the deterioration of a patient's situation—in which physiological information can be utilized to dictate whether the patient is at risk of decompensating. With the help of big data analytics, it is feasible to track patients in basic care units, nursing care centers, or even at house, to determine some type of decompensating.

Another use of big data analytics is for forecasting which patients are at risk of multiple adverse events such as kidney failure, infection, and unfavorable drug events. Chronic circumstances may stretch over multiple organs, or are systemic in essence; and normally, few of the most expensive conditions to handle can be predicted by big data analysis for possible prevention and risk control [16].

Another study investigated the use of big data for personalized healthcare [17]. The study proposed combining big data analytics with virtual physiological human (VPH) mechanisms for generating powerful and productive results in the field of silicon medicine. Current big data methods inspect a huge collection of information, from numerous clients, in a limited time, recognize collection and interrelation, and creates predictive systems using statistical or machine-based learning simulation methods. Applying statistical processing or artificial intelligence methods gives excellent indicators that are acceptable for the variety of the evaluated sets of data. If a database includes noticeable results for a group of patients, the analyst can instantly calculate the precision of that predictor.

The integration of recommendation systems with big data can appear as a successful solution in the medical industry [18]. The application of the recommendation system in medical care comprises a broad area where patients are recommended about the condition of their health. With the huge volume of data existing in the healthcare field, such information can be evaluated to produce a constructive estimate for the patients. The predictions and suggestions are more precise as we involve a large amount of information. It also notifies the patient of the disease and helps to take the required measures prior to its occurrence.

Another study [19] considered management and delivery as important applications of big data in healthcare. It assists with budgeting of medical facilities provided on the treatment and the insurance covered. It entails information on EHR, data of the patient, treatment trials, and observation based on evidence. It can be used individually to analyze data of administration and delivery.

The next use case category relates to the data that assist in making clinical decisions and judgments. Information such as precise projection of health, electronic health records, genomics, research and development, and medical organizations can assist in clinical decisions [20].

The next application is raised by [20] and is related to the clinical information. This study addresses the datasets accessible from different data sources and clinical data systems that can be utilized for analytics. The information involves external health data from social media, biomedicine sensor data, and treatment data of illness like genomics. The behavior of consumers also needs to be considered. Big data methods are used to analyze the demographics of consumers, and also provides some more varieties of information from them, including information about advertising, customer marketing, and conduct of public and individual. The last aspect addressed by [19] is called support information. It includes efficiency of treatment, laboratory and research documents, public health treatments, medication, error detection, etc.

One study that overlaps with some of the above reviews focuses on specific areas of healthcare where applying big data can create value [21]. The first application is to use for clinical decision support. Big data analytics techniques can be utilized for forecasting results or suggest substitute treatments at the point of care for medical practitioners and patients. The next usage is for personalized care. Probabilistic data mining or analytical approaches can provide early diagnosis and detection prior to the patient developing symptoms of any disease. Detection of patterns by means of real-time wearable sensors for old or paralyzed patients to warn doctors if their crucial variables alter or post-market surveillance of drug efficacy can be achieved. The next application is for public and population health. Big data analytics solutions can extract data from the web and social media to forecast flu and estimate future patterns. Big data can be applied for fraud detection. Different predictive models can be used to forecast and stop transaction fraud and medical claims fraud. Healthcare data can also be used for secondary purposes that relate to the accumulation of clinical information from accounting, patient care, and regulatory documents, to discover useful ideas such as identification of rare disease patients, decisions of treatment, measurement of clinical efficiency, and so on. The last application according to this study [21] is for evidence-based medicine. It includes the application of statistical analyses and quantified research by medical practitioners for detection. This implementation allows physicians to make choices on the basis of their individual opinions as well as best accessible proof.

2.3 Challenges for Healthcare Industry

Reviews in the previous part provide ground for extracting and thematizing challenges that big data healthcare systems are currently facing. Below is a summary of recurring themes extracted. The method of creating these themes is explained in Sect. 3:

1. Capturing and storing a large amount of information [22] are the first challenges in healthcare. The exponential growth of data collected from digital devices, aerial cognitive technologies, remote sensors, software records, cameras, recorders, frequency detector, wireless signal networks, and so on create a major challenge in this area.
2. The next challenge is related to the security and privacy issues of using big data in healthcare. Trustworthiness of data, confidentiality, accuracy, and consistency [23], privacy preservation, and security of data, including network security, authentication of users, and preventing malicious hackers are some aspects of this challenge. In [23–27] major security and privacy-related challenges in the literature are reported.
3. Maintaining the quality of health data and lacking a health-specific standard for handling data are other important challenges in the healthcare industry [19, 21, 28–30]. It is claimed that predictive modeling could have negative impacts on patients if a faulty data model is used [23].
4. Organizational group of challenges is another class of problems that hinders the application of big data in the healthcare domain. Some of these problems are reported as increasing organizational expenditure on big data infrastructures, boosting the analytical and application capacity of the organization, and additional managerial tasks on data [29]. Also, healthcare organizations usually resist change. Unlike other industries such as banking and petroleum and gas, which welcome opportunities in big data by accepting high risks, health organizations tend to avoid the negative risks of big data regardless of the potential opportunities that big data can provide [1]. This is mainly due to insufficient administrative assistance, lack of confidence, judicial problems, and the absence of required and adequate abilities to operate the ICT instruments [1].
5. Medical images are the significant source of information often used for detection, evaluation, and planning of treatment. They vary in size ranging from few megabytes to hundreds of megabytes. The image data needs big storage ability if they have to be stored for a long time. It also requires quick and precise algorithms if any choice helping to automate utilizing the data was to be made. Some of the challenges in image processing are pre-processing, compression of data, real-time implementation, registration, and association, securing data, maintaining privacy, segmentation, mining of data, and validating data [3].
6. Integration of heterogeneous data sources is another challenge in this field [21, 31]. Established healthcare systems have always suffered from compatibility

issues among various hardware and software technologies That in turn will affect interpretation and correlation of results [32].

7. Scalability is another problem in this field [33]. Datastream can conveniently hit up to millions of tuples per second sometimes in real time. Considering the convergence of streams arriving from every healthcare support devices, the issue of centralized databases becomes hard to handle in real time.
8. Another problem related to the above issue is the time-criticality of healthcare applications. Speed is essential in both input and output [25, 33]. Time delay in handling complicated datasets may degrade the quality of patient care [21, 34].
9. Data science training and variation of the workforce is another challenge [35]. In order to better perceive data and turn unstructured or organized numbers into systematic information, data scientists will need to collaborate with doctors, researchers, and patients. Thus, there is diversity in the workforce. At this stage, few opportunities exist for training and education and this has caused a lack of workforce in this field. Miscommunication between users and data analysts is one of the primary issues. It is difficult for the user to understand the data provided by data scientists and such gaps can adversely influence the decision outcomes of healthcare professionals [25].
10. The final group of the challenge is the support. Problems such as finance [20], appropriate laws and rules, ethics, government strategies, coordination inside the ecosystem, and professionals having specialization in healthcare are reported quite frequently in the literature [28].

3 Thematic Analysis

To analyze collected textual data and develop a framework for big data challenges in healthcare applications, this paper adopts a six-step thematic analysis. These steps are expected to lead the study to the identification of categories of challenges that healthcare organizations are facing in relation to big data. This section provides a general overview of the method applied in this study.

In the first and second steps, previous works are reviewed and coded. The idea here is to explore an initial set of codes that describe the phenomena. As a result of this step, a list of codes was identified progressively as the study proceeds through the next five steps, and a final list emerged. In data mining terminology codes identify and mark data features. In step 3 the themes are created. Themes represent contextually meaningful categories of identified classes. Step 4 deals with reviewing and refining themes. Step 5 finalizes definitions of themes and name them. In the final step, the results of the thematic analysis are produced to validate the study and justify the results. Table 1 shows the initial list of extracted codes along with the references.

Table 2 shows the classification scheme of the codes in the previous table. Steps 3 to 6 of the thematic analysis apply here. We found seven themes for this study and show them in the first column of Table 2, and corresponding codes for each theme are shown in the second column of this table.

Table 1 List of codes generated from different papers

Papers	Codes
[28]	Data standard, data sharing, data integration, data storage, data security and privacy, laws and rules
[24]	Data security
[23]	Volume, velocity, veracity
[20]	Cost, data integration, data quality, security and privacy, data fragmentation
[19]	Data standard, heterogeneity, infrastructure of hospitals, data quality, privacy
[22]	Capturing and storing data, data quality, privacy and security, lack of IT professionals
[21]	Data standard, data integration, lack of skill, privacy and security, real-time processing, data analysis, quality of data
[26]	Privacy and security
[36]	Infrastructure issue, data quality, data integration, lack of experts
[1]	Persistence to change, fragmentation, data standard, security, and privacy
[33]	Data quality including scalability, heterogeneity, timeliness
[27]	Data security and privacy
[35]	Data storage, heterogeneity, data quality, timeliness, privacy, scalability, visualization
[35]	Ethical issue, data quality, data access and sharing, workforce
[37]	Data storing, heterogeneity of data transferring, security, and privacy
[28]	Data standard, interoperability, security and privacy, and workforce
[30]	Data quality, data analysis, user training
[32]	Heterogeneity, capturing data, data analysis
[25]	Security and privacy, classifying data, workforce, lack of technology
[31]	Heterogeneity, data quality, cost, lack of experts

Table 2 Categorization of codes into themes

Themes	Codes
Data characteristics challenges	Volume, velocity, veracity
Data quality challenges	Heterogeneity, scalability, visualization
Analysis challenges	Data capturing, data storage, data processing, data sharing, real-time data analysis, interoperability, data integration
Ethical and security challenges	Data privacy and ownership, data security
Workforce challenge	Lack of data scientist and analyst, training of users
Financial challenges	Cost of technology, cost of workforce
Infrastructure issue	No proper infrastructure for technology incorporation

4 Conclusion and Future Work

As part of a larger research project, this work-in-progress research investigated various existing challenges associated with using big data in healthcare applications. Through a critical review of the related literature and the authors' analysis and interpretation of the reviews, data are collected from relevant existing literature and then analyzed through a thematic analysis. The coding criteria for data analysis are those generic codes that are commonly used by similar studies and include technical challenges, administrative challenges, and security and privacy challenges. Results of our thematic analysis resulted in the identification of the following themes: data characteristics challenges, data quality challenges, analysis challenges, ethical and security challenges, workforce-related challenges, financial challenges, and infrastructure challenges.

The ultimate goal of our future study is to develop a set of design directives for effective human interactions in healthcare big data applications.

Acknowledgments The authors would like to thank Victoria University for supporting this publication, and Mr. Abhinav Sharma from Victoria University Sydney for his valuable assistance in this research project.

References

1. Olaronke I, Oluwaseun O (2016) Big data in healthcare: prospects, challenges and resolutions. In: Future technologies conference (FTC), pp 1152–1157
2. Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19(2):171–209
3. Belle A, Thiagarajan R, Soroushmehr SM, Navidi F, Beard DA, Najarian K (2015) Big data analytics in healthcare. In: Biomed research international
4. Ghotkar M, Rokde P (2016) Big data: how it is generated and its importance. IOSR J Comput Eng
5. Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of big data challenges and analytical methods. J Bus Res 70:263–286
6. Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, Shahabi C (2014) Big data and its technical challenges. Commun ACM 57(7):86–94
7. Kadadi A, Agrawal R, Nyamful C, Atiq R (2014) Challenges of data integration and interoperability in big data. In: International conference on big data (big data). IEEE, pp 38–40
8. Katal A, Wazid M, Goudar RH (2013) Big data: issues, challenges, tools and good practices. In: Sixth international conference on contemporary computing (IC3), pp 404–409
9. Khan N, Yaqoob I, Hashem IAT, Inayat Z, M Ali WK, Alam M, Shiraz M, Gani A (2014) Big data: survey, technologies, opportunities, and challenges. Sci World J
10. Kaisler S, Armour F, Espinosa JA, Money W (2013) Big data: Issues and challenges moving forward. In: 46th Hawaii international conference on system sciences. IEEE, pp 995–1004
11. Duggal PS, Paul S (2013) Big data analysis: challenges and solutions. In: International conference on cloud, big data and trust, vol 15, pp 269–276
12. Zhou Z, Chawla N, Jin Y, Willaims GJ (2014) Big data opportunities and challenges: discussions from data analytics perspectives. Comput Intell Mag IEEE 9(4):62–74
13. Fan J, Han F, Liu H (2014) Challenges of big data analysis. Natl Sci Rev 1(2):293–314

14. Australian Institute of Health and Welfare (AIHW) (2019) Online. https://www.aihw.gov.au. Accessed 30 June 2020
15. Australian Digital Health Agency (ADHA), Online. https://www.digitalhealth.gov.au. Accessed 30 June 2020
16. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G (2014) Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff 33(7):1123–1131
17. Viceconti M, Hunter P, Hose R (2015) Big data, big knowledge: big data for personalized healthcare. IEEE J Biomed Health Inf 19(4):1209–1215
18. Shobana V, Kumar N (2017) A personalized recommendation engine for prediction of disorders using big data analytics. In: International conference on innovations in green energy and healthcare technologies (IGEHT). IEEE, pp 1–4
19. Thara DK, Premasudha BG, Ram VR, Suma R (2016) Impact of big data in healthcare: a survey. In: 2nd international conference on contemporary computing and informatics. IEEE, pp 729–735
20. Schaeffer C, Booton L, Halleck J, Studeny J, Coustasse A (2017) Big data management in US hospitals: benefits and barriers. Health Care Manag 36(1):87–95
21. Mathew PS, Pillai AS (2015) Big data solutions in healthcare: problems and perspectives. In: International conference on innovations in information, embedded and communication systems (ICIIECS). IEEE, pp 1–6
22. Sarkar BK (2017) Big data for secure healthcare system: a conceptual design. Complex Intell Syst 3(2):133–151
23. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2(1):3–5
24. Archenaa J, Anita EM (2015) A survey of big data analytics in healthcare and government. Procedia Comput Sci 50:408–413
25. Fatt QK, Ramadas A (2018) The usefulness and challenges of big data in healthcare. Health Commun 3(2):21
26. Martin EM, Skotnes T, Feldman B (2012) Big data in healthcare hype and hope. Dr Bonnie 360:122–125
27. Patil HK, Seshadri R (2014) Big data security and privacy issues in healthcare. International congress on big data, IEEE, pp 762–765
28. Kim MK, Park JH (2017) Identifying and prioritizing critical factors for promoting the implementation and usage of big data in healthcare. Inf Dev 33(3):257–269
29. Kruse CS, Goswamy R, Raval YJ, Marawi S (2016) Challenges and opportunities of big data in health care: a systematic review. JMIR Med Inform 4(4):e38
30. Stylianou A, Talias MA (2017) Big data in healthcare: a discussion on the big challenges. Health Technol 7(1):97–107
31. Asri H, Mousannif H, AI Moatassime H, Noel T (2015) Big data in healthcare: challenges and opportunities. In: International conference on cloud technologies and applications. IEEE, pp 1–7
32. Alexandru A, Coardos D (2016) Big data in healthcare and medical applications in Romania. In: Proceeding of the AQTR
33. Wang L, Alexander CA (2018) Big data and data-driven healthcare systems. J Bus Manag Sci 6(3):104–111
34. Costa FF (2014) Big data in biomedicine. Drug Discov Today 19(4):433–440
35. Zhang X, Pérez-Stable EJ, Bourne PE, Peprah E, Duru OK, Breen N, Berrigan D et al (2017) Big data science: opportunities and challenges to address minority health and health disparities in the 21st century. Ethinicity Dis 27(2):95
36. Noonpakdee W, Phothichai A, Khunkornsiri T (2019) Challenges of big data implementation in a public hospital. In: 28th wireless and optical communications conference. IEEE, pp 1–5
37. Desai PV (2018) A survey on big data applications and challenges. In: Second international conference on inventive communication and computational technologies (ICICCT). IEEE, pp 737–74

Hardware Control of the Electron Beam Energy Density by the Heating Spot

Sergei Kurashkin, Vadim Tynchenko, Yuriy Seregin, Aleksandr Murygin, and Aleksey Bocharov

Abstract Various types of scanning can enhance the quality of welding joints, which allows us to control the energy distribution of the electron beam. In the study conducted by the authors, it was experimentally revealed that the best quality of welded joints is obtained by scanning an electron beam in the form of a sinusoidal raster. The research was conducted for thin-walled aerospace structures. Equipment was developed that realizes scanning in the form of different rasters, considering the frequency of the sweep, thereby removing the high frequency, which will allow getting a more even "saw". As the different types of scanning shapes at which the optimization was carried out, the next types were used: a classical raster, a sinusoidal raster and a truncated raster. The obtained results allow a more meaningful approach to the process of electron beam welding of different types of materials and thin-walled structures and obtaining a consistent seam quality.

Keywords Electron beam welding · Mathematical modeling · Technological parameters · Electron beam oscillation · Modeling of the thermal welding process · Normal distribution law · Welding speed · Control automation

1 Introduction

Today, electron beam welding is a promising area in the aerospace industry. At present day, welding technology provides ample opportunities for the manufacture of assemblies with increased requirements for the quality of the welded joints. To ensure a consistently high quality of the obtained weld, monitoring of the specified process conditions is required. Due to non-compliance with technological parameters, the root defects occurred. Currently, to prevent defects defocusing and different types of

S. Kurashkin (✉) · V. Tynchenko · Y. Seregin · A. Murygin · A. Bocharov
Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
e-mail: scorpion_ser@mail.ru

V. Tynchenko · A. Bocharov
Siberian Federal University, Krasnoyarsk, Russia

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_9

scans are used. The actual method to affect the creature of the penetration channel is using different types of scan paths for the electron beam, like scanning by sinusoidal trajectory [1].

Many experiments and research projects were carried out, but so far, the best method in choosing a sweep has not been suggested. There are some scanning trajectories most used in practice [2]:

In [3, 4] lengthwise trajectory was developed;
In [5–7] x-shaped trajectory was developed;
In [8–10] circular and elliptical trajectory was developed;
In [11, 12] arc and clamp trajectory was developed.

In [13], the authors conducted a research work in which the characteristics of the electron beam distribution density were received by scanning in various forms: classical, sinusoidal and truncated. Sinusoidal and truncated rasters allow you to get a two-hump energy distribution along the coordinate across the joint, and it is necessary to exclude the behavior of root diverts [13].

Earlier at the conferences [14–16], materials were published on the research of scanning paths and optimization of their amplitude using the normalized standard deviation function. In the course of work, various technological modes were used and experiments were carried out on real equipment. As a result, it was found that the best type of scan in terms of the quality of the weld formation is a raster in which the beam is scanned along the joint, according to a unilateral sawtooth law. When using the recommendations obtained in [16] on equipment, during welding of products, the result was a beam smearing effect, which leads to defects. Most beam scanners contain digital-to-analog converters; therefore, it is not difficult to transform the scanning form of the electron beam in the form of displacements along the coordinate grid. In [14–17], the main trajectory was chosen as the scanning raster containing 256 positioning points.

The disadvantage of the previous studies is the high-frequency component that creates interference. In [14–16], the scanning trajectory was implemented at frequencies above 1 kHz.

In earlier research works [18–20], the authors developed a new beam trajectory, an example of such a trajectory is shown in Fig. 1.

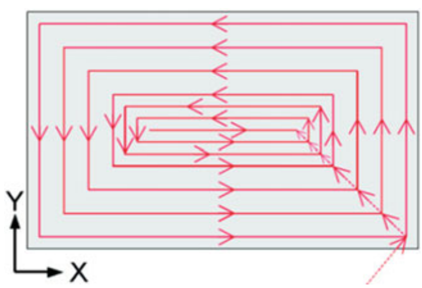


Fig. 1 Renewed paths of beam

In [20, 21], it was determined that overheating occurs due to the difference in the conductivity of the molten and compacted powder. The authors proposed an energy input modulation strategy to consider this difference.

To get this new trajectory of scanning, a new trajectory generation algorism was created. To comply with the limitations of the modulation function, the trajectory was defined as the internal contours of the region that would be melted, moved from the outside to the inside.

In this paper, we propose to implement the hardware, considering the frequency of the sweep, thereby removing the high frequency, which will allow for a more even "saw".

2 Experimental Study

Figure 2 shows a scheme of the hardware for controlling the scanning of an electron beam.

In this research, we used microcontroller unit (MCU)—ATSAM3S4BA and digital-to-analog converter (DAC)—1108PA1.

Figure 3 shows the two scan paths that were chosen by the authors.

Figure 4 shows a macrosection with penetrations during scanning in the form of a classical raster, obtained with a high-frequency component.

This form allows you to get the shape of the welds close to rectangular with almost parallel walls while reducing stress and strain in the welded joints. The radius of the fillet is 0.1–0.2 cm. In the welds, root defects are almost completely absent, porosity is reduced, but the results are still far from the required ones. Lengthwise areas of welded seams presented that the processes of seam formation are stable.

The instability of the penetration depth, in the form of penetration fluctuations in the root of the welded joint, decreased by 2–3 times compared with traditional technology. In [14–16], it was revealed that the raster was the best type of scan; however, there was a high-frequency component, which led to defects. In research [13], it was revealed that the best scanning beam path is sinusoidal. The energy distribution by this scan is shown in Fig. 5.

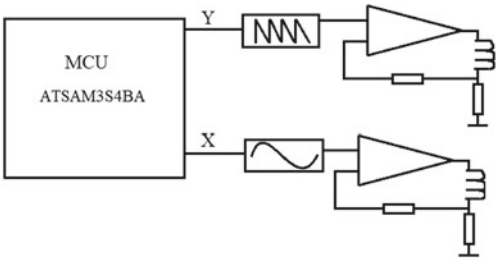


Fig. 2 Functional control unit

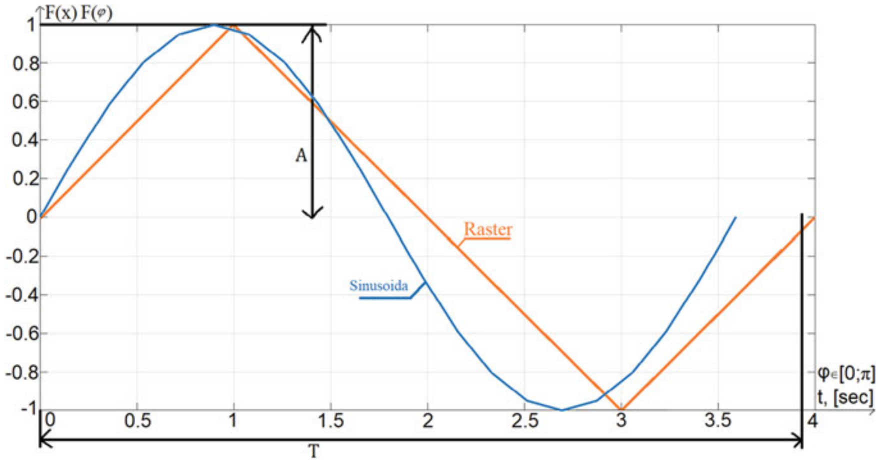


Fig. 3 Graph of successive triangular pulses and sinusoid

Fig. 4 Macrosection with penetrations during scanning the beam

In this paper, we propose to use a nominal sinusoidal scan rather than a raster scan and also reduce the sweep frequency, which in the end will give the best weld quality. The energy of the electron beam is described by the formula, which includes voltage, current beam and power along and across the seam; the formula was got from the work [13].

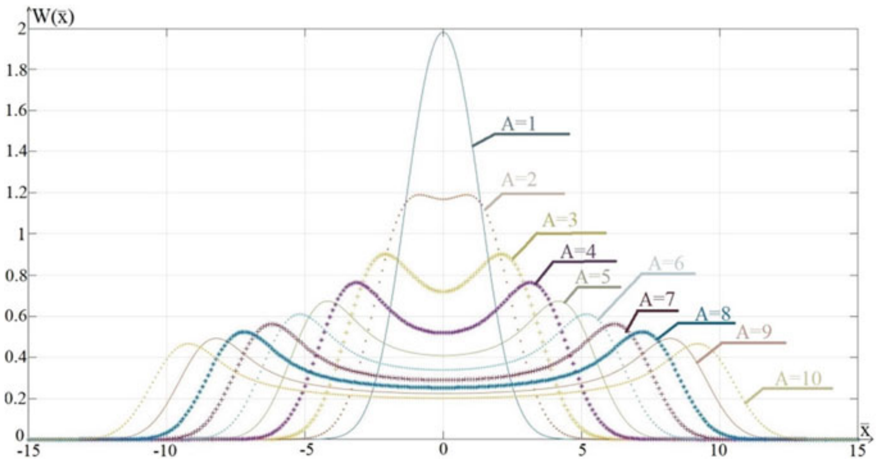


Fig. 5 Beam energy distribution on the part surface during sinusoidal scanning across the joint

To avoid taking into account the particular value of the standard deviation, we introduce dimensionless coordinates and amplitudes of the beam scanning, like $\bar{x}$ and $\bar{A}$, which we got from the formula in [13].

Based on the work [13], it was ascertained that the formula of energy distribution for a sinusoidal trajectory is as follows:

$$W(\bar{x}) = \int_{-\pi}^{\pi} \frac{1}{\sigma * \sqrt{2 * \pi}} * \exp\left(-\frac{(\bar{x} - (\bar{A}_x * \sin\varphi))^2}{2 * \sigma^2}\right) d\varphi, \quad (1)$$

The energy amplitude by X-axis ($\bar{A}_x$) is in the bounds from 1 to 10, $\bar{x}$ is in the bounds from -15 to 15, and the phase (φ) $-\pi$ to π.

The scanning path is shown in Fig. 5.

Figure 6 shows the curves for the event of a scanning electron beam path along the junction in the form of a raster under the influence of the normal energy distribution.

In work [16], a calculation method for determining the scanning amplitude was developed, which allows determining the shape of the scan and its dimensions, as well as the desired energy distribution on the surface.

The description of the functional through point and linear sources gave the same results, which made it possible to simplify the representation of the heat source in EBW. As a result, the software implementation of the simulation of the thermal process is greatly simplified.

Thus, at this stage, only one moving source is required for modeling, instead of 256, as it was in previous studies.

The authors obtained a graph for the scattering of temperatures normalized to the maximum, which has a characteristic minimum at point “b” in Fig. 7 [16]. The coordinates of this minimum coincide with the experimental dimensions of the scan path.

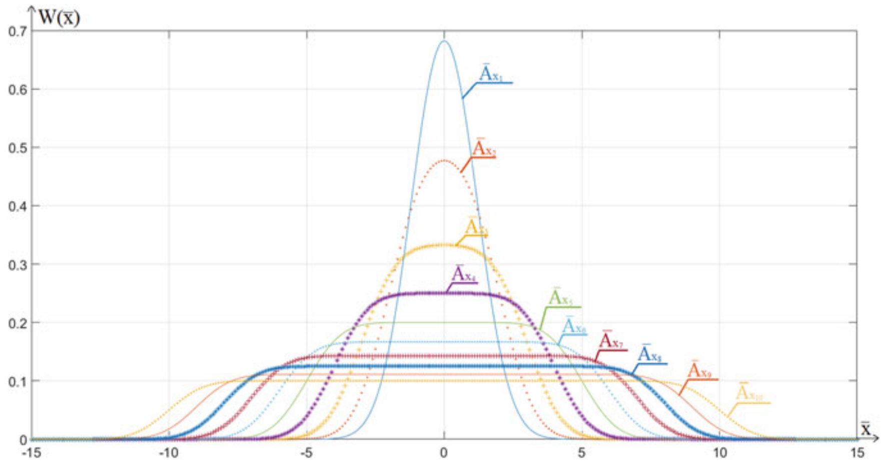


Fig. 6 Distribution of beam energy on the surface of the part during raster scanning across the joint

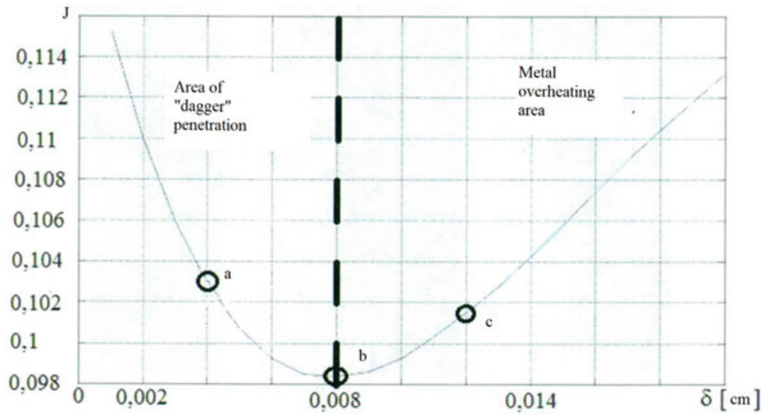


Fig. 7 The graph of the variation in scattering of the normalized temperature field depending on the interval between the electron beam positioning points in the form of a "Raster"

Based on the research results, a criterion was developed [17]. The optimality criterion is the minimum functional when varying the parameter for which a solution is being sought.

As a result of the simulation, a graph of the dependence of the function on the amplitude presented in Fig. 8 was obtained.

The graph is interesting in that it has a characteristic minimum. From the graph, it follows that the optimal value of the amplitude will be equal to four.

The proposed optimality criterion is attractive for several reasons:

- First: The nature of the criterion has a physical semantic orientation;

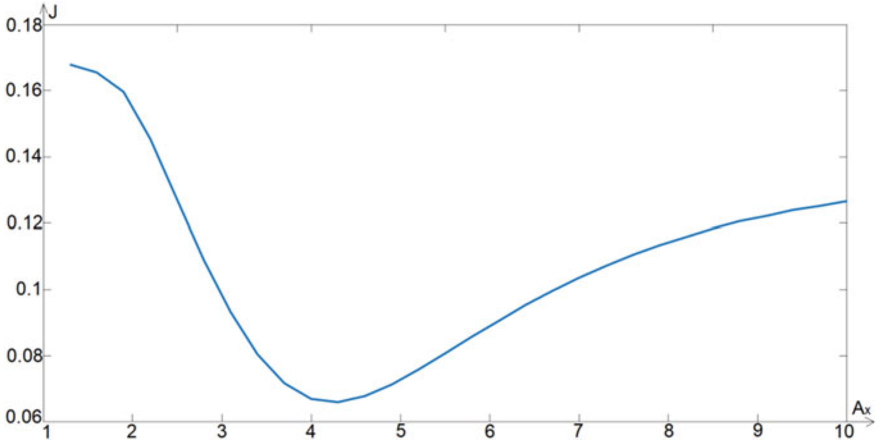


Fig. 8 Graph of variation of temperature field scattering as a function of amplitude

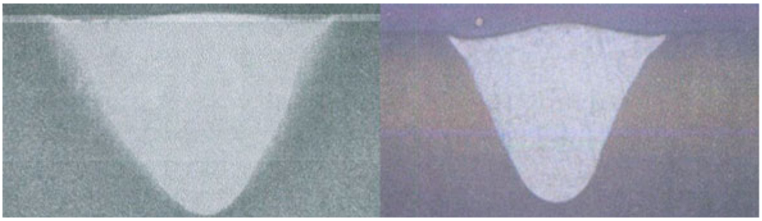


Fig. 9 Penetration macrosections

- Second: The criterion has extrema with characteristic values of variable parameters.

As a result of the conducted work, microsections were obtained considering the change in amplitude and with a low synchronization frequency, which made it possible to improve the quality of the weld. The results are presented in Fig. 9.

3 Conclusion

We developed a program to control the distribution of energy over the heated spot using different forms of scanning beam trajectory.

Determination of EBW parameters using numerical simulation of thermal processes will improve the quality and stability of the weld for structures made of new materials and various thicknesses.

References

1. Sobko SA, Kulikov VA (2013) The method of control of electron-beam welding. Russia, Moscow
2. Shcherbakov AB, Rodyakina RV, Novokreschenov VV, Lastovirya VN (2018) Technology of material processing. In: Equipment of electron-beam complexes, 3rd edn. Yurayt, Moscow, pp 114–130
3. Akopianc KS, Nesterenko VM (2002) Electron-beam welding of steels up to 600 m thick with longitudinal oscillations of the beam. Autom Weld 9:3–5
4. Ohmine M, Hiramoto S, Jamane J (1993) Fundamental study on the pulsed electron beam welding. In: International Institute of Welding (D IV-348-383), 13
5. Belenkiy V Ya (1986) Scanning of an electron beam along an x-shaped trajectory as a means of reducing defects in the root of a weld in EBW. Autom Weld 9:35–37
6. Barishev MS, Redchic AA, (1990) Optimization of welding modes by oscillating electron beam. In: International conference on welding of non-ferrous metals, vol 4, p 34
7. Zenker R (2002) Electron beam surface treatment and multipool welding – state of the art. In: Proceedings of the EBEAM. International conference on high power electron beam technologies, vol 10, pp 12-1–12-5
8. Shilov GA, Akopianc KS, Kasatkin OO (1983) Influence of the frequency and diameter of the circular scan of the electron beam on the penetration in EBW. Autom Weld 8:25–28
9. Akopianc KS, Emchenko-Ribko AV, Neporohin VYu (1989) Prevention of the formation of root defects in EBW with non-penetration of up to 60 mm depth. Autom Weld 4:30–34
10. Komiro Y, Punshon CS, Gooch TG (1986) Effects of process parameters on centerline solidification in EB weld. Metal Constr 2:104–111
11. Vasil'ev AM, Goncharov VA, Krivkov BG (1998) Electronic beam control device by welding shipbuilding industry. Series Welding 6:11–12
12. Mara CL, Funk ER, McMaster, RC (1974) Penetration mechanism of electron beam welding and spiking phenomenon. Weld J 53(6):246–251
13. Kurashkin SO, Laptenok VD, Murygin AV (2019) Analytical characteristics of the electron beam distribution density over the heated spot for optimizing the electron-beam welding process. In: IOP conference series: materials science and engineering, vol 681, p 012021
14. Seregin Yu N, Laptenok VD, Uspenskiy NV (2011) Experimental research on the optimization of the technology of electron-beam welding of aluminum alloys. Technologies and equipment of EBW-2011, pp 71–80
15. Seregin Yu N, Laptenok VD, Murygin AV (2019) Experimental research on electron-beam welding technology with a scanning electron beam. In: IOP conference series: materials science and engineering, vol 467
16. Seregin Yu N, Laptenok VD, Uspenskiy NV (2014) Software of electron beam distribution energy over the heated spot by EBW. Technologies and equipment of EBW-2014, pp 39–45
17. Seregin Yu N, Kurashkin SO (2017) Modeling of the EBW mode for predicting the parameters of the weld. In: MPEI collection of materials of 2nd international conference, pp 26–36
18. Mladenov G (2009) Electronic and ionictechnologies. Prof. Marin Drinov, Sofia
19. Shiller Z, Gaizig U, Panzer Z (1980) Electron beam technology. Energiya, Moscow
20. Yazovskikh VM, Trushnikov DN, Belenky V Ya (2004) The mechanism of secondary emission processes in electron-beam welding with electron beam modulation welding. Weld Int 18:21–27
21. Béraud N, Vignat F, Villeneuve F, Dendievel R (2014) New trajectories in electron beam melting manufacturing to reduce curling effect. Procedia CIRP 17:738–743

A Machine Learning Model for Predicting Fetal Hemoglobin Levels in Sickle Cell Disease Patients

Konstantinos Oikonomou, Kathleel Steinhöfel, and Stephan Menzel

Abstract Sickle cell disease is one of the commonest genetic diseases and is defined as a decrease in hemoglobin concentration in the blood. The main known factor that can alleviate the disease is the persistence of fetal hemoglobin (HbF), and thus the aim of our research is to build a model to predict the HbF% of patients based on the three regulating genes of the disease (BCL11A, Xmm1-HBG2 and HBS1L-MYB). A machine learning approach is employed in order to improve the accuracy of the model, with various algorithms of that type being explored. In the end, the K-nearest neighbors algorithm is chosen and an initial version of it is implemented and tested. Finally, the algorithm is optimized enabling our optimized model to predict the HbF% of a patient with 87.25% accuracy, a major improvement over the existing alternative that has a mean error of 336.33%. Furthermore, 93.45% of our predictions have a sheer error that is less than 0.5, and all these facts reinforce the strength of our model as a quick and accurate estimation tool for small and medium-sized clinical trials, where fast HbF% predictions can help adjust for genetic background variability that obscures test outcomes.

Keywords Sickle cell disease · Fetal hemoglobin · Machine learning prediction model

---

K. Oikonomou (✉) · K. Steinhöfel
Department of Informatics, King's College, Strand, London, UK
e-mail: k1896365@kcl.ac.uk.com

K. Steinhöfel
e-mail: kathleen.steinhofel@kcl.ac.uk.com

S. Menzel
Department of Molecular Haematology, King's College, Strand, London, UK
e-mail: stephan.menzel@kcl.ac.uk.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_10

1 Introduction

Sickle cell disease is one of the commonest genetic diseases with the levels of fetal hemoglobin (HbF) being one of the most important indications of its severity. However, high levels of HbF have been proven clinically beneficial for patients [1] and are associated with longer survival [2] and lower pain rates [3]. The persistence of HbF in patients beyond the first year of life and throughout adulthood is the main known factor that can alleviate sickle cell disease. Our main objective is the development of an accurate model to predict the levels of a patient's HbF% and thus the severity of sickle cell disease. A model of this kind could be especially powerful in smaller and medium-sized clinical trials and help geneticists explain human variation in health and disease through underlying genetic variability.

The development of a prediction model for various aspects of sickle cell disease has been a topic for several scientific research works. Work in [4] tries to predict the chances of a stroke for patients of the disease, using transcranial ultrasonography, while [5] focuses on predicting the adverse outcomes of the disease on children. Furthermore, [6, 7] seek to better define the phenotypes and genotype–phenotype relationships regarding the disease, work in [8] tries the same for β-thalassemia patients and [9] presents a pain prediction model for African-American adults affected by it. However, there is little research focused on the development of an accurate prediction model specifically for the HbF% of a patient, as the majority of previous work in this particular area focuses on the general connection of gene regulation with sickle cell. In [10] the authors confirm the effect of genes BCL11A and HBS1L-MYB on sickle cell anemia in African-Americans, while in [11, 12] the authors investigate the same for Tanzanian, British and Sri Lankan patients. A similar model as the one we are trying to develop was proposed by [13] and its 22% accuracy will be used as a comparison benchmark but it suggests a linear regression connection between the γ-chain genes and HbF. This is a major disadvantage in our opinion as it fails to take into account the effect of the γ-gene activation and the creation of young blood cells that also contain fetal hemoglobin (F-cells) and both will be incorporated in our proposed approach.

In this paper, the goal is to create a prediction model for sickle cell disease patients' HbF%, optimize it and improve on the accuracy of the already published model [13]. At first, the aforementioned model is implemented to be used as a comparison metric and an accuracy target for our model to beat. Then several algorithms are tested in an attempt to choose the most appropriate one as the base for our prediction model. Their results and characteristics are compared and the K-nearest neighbors algorithm is selected among them as the one indicating the highest potential accuracy. The K-nearest neighbors algorithm is then customized and optimized for our problem, and the final implementation of our algorithm is tested against the data set. Finally, our optimized model is compared with [13] in terms of accuracy between the predicted and actual values of HbF%.

This paper is organized as follows: Some background information regarding sickle cell anemia and fetal hemoglobin are discussed in Sect. 2. Section 3 contains a detailed description of our approach, both regarding the theoretical model and the development of the algorithm. On the other hand, the actual initial and final implementations are discussed in Sect. 4, along with an overview of the chosen algorithm. The accuracy results of our model are presented in Sect. 5, accompanied by a comparison with a previously implemented model and information regarding the model's optimization and cross-validation. Finally, Sect. 6 includes the conclusions of our research and proposals for future work on this topic. From this point on, the abbreviations HbF for fetal hemoglobin and F-cells for young blood cells containing HbF will be used for brevity's sake.

2 Background

Red blood cells (or erythrocytes) are the most common type of blood cell. Their main function is oxygen binding and transport to the tissues as well as the return of carbon dioxide from the peripheral tissues to the lung. In order to achieve this exchange, the erythrocytes contain a specific protein called hemoglobin. Hemoglobin is a metalloprotein and an erythrocyte that contains about 640 million hemoglobin molecules. Hemoglobin types found in adults are HbA ($\alpha 2\beta 2$), HbA2 ($\alpha 2\delta 2$) and HbF ($\alpha 2\gamma 2$). However, the primary hemoglobin during the fetal development period (2nd–3rd trimester of pregnancy) is HbF ($\alpha 2\gamma 2$).

2.1 Polymorphism and Hemoglobin Regulating Genes

The expression of hemoglobin genes is regulated by factors in the same gene locus (acting in cis) or those encoded in other regions (acting in trans, i.e. transcription factors). Either types of regulatory factors can be affected by DNA sequence polymorphisms, leading to changes in hemoglobin synthesis. Some of those polymorphisms can lead to the persistence of γ-globin and therefore HbF in adults. The presence of HbF is critical to the phenotype of sickle cell disease and because of this, the effort to induce HbF synthesis has begun as a therapeutic approach to these diseases as described by Makani et al. [11]. The three polymorphic loci that affect the expression of hemoglobin γ-chains are BCL11A, Xmm1-HBG2 and HBS1L-MYB. Their most representative variants are summarized in the following table.

The representative variants of Table 1 can have values 0, 1 or 2 and will be used as inputs for our prediction model.

Table 1 Representative variants

Variant	Gene
rs1427407	BCL11A
rs654816	BCL11A
rs66650371	HBS1L-MYB
rs7482144	Xmn1-HBG2

2.2 Sickle Cell Anemia

Anemia is defined as a decrease in hemoglobin concentration in the blood. Although normal values vary across laboratories, the indicative values for the diagnosis of anemia could be less than 13.5 g/dl in adult men and less than 11.5 g/dl in adult women. Sickle cell anemia/disease is a genetic hemoglobinopathy that is inherited with an autosomal recessive mode of inheritance. This disease is one of the commonest genetic diseases with over 300,000 annual births worldwide, of which about 70% occur in sub-Saharan Africa, where most of the affected children die before the age of 5. In the UK and other Western countries, the disease is present mostly through the African diaspora and shows significant clinical diversity.

3 Our Approach

A visualization of our theoretical model is given in Fig. 1.

Our approach takes into account both the young blood cells that contain fetal hemoglobin (F-cells) as well as the γ-gene activation that affect the HbF% of a patient. These two factors can independently affect the concentration of HbF on a patient but are in turn affected by the values of the HbF modifier genes, expressed through their representative variants. The arrows on Fig. 1 represent these relationships that our model will try to explore.

Furthermore, the interaction between the F-cells, the HbF% and the sickling rate forms a negative feedback loop. This happens because of the high concentration of HbF that leads to reduced hemolysis, which is the rupturing of red blood cells. This reduction causes less need for the creation of new blood cells that would contain fetal hemoglobin and would therefore be categorized as F-cells. However, the number of F-cells affects the percentage of HbF, and thus, a decrease in F-cells leads to a decline of HbF% which can subsequently cause increased hemolysis. The aforementioned interaction has been taken into account in our approach and is a key factor for the accuracy of our model. Finally, the straight line that connects the HbF modifier genes with the %HbF represents the linear regression proposed in [13]. It was implemented as a comparison tool for accuracy but is not part of our model.

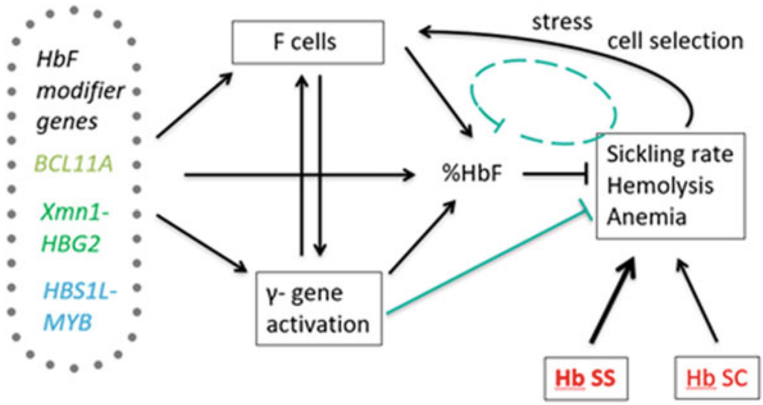


Fig. 1 Model visualization

3.1 Machine Learning Approach

A pure mathematical approach to our model was dismissed in the early development stages, as our theoretical model is very complicated and involves many unknown relationships. The efforts to try and seek the functions that represent these relationships would be extremely time-consuming and possibly not realistic, as our data suggest that most of these functions could be potentially discontinuous in multiple parts. Therefore, a machine learning approach was chosen instead with a supervised algorithm being the first choice.

3.2 Training/Test Set

The data set included 465 patients with complete data regarding the values of the representative variants, their HB and HbF levels. A fairly common strategy employed in supervised machine learning applications is to split the data set into two subcategories. More specifically, the training set consists of pairs of an input and an output vector. The model is then adjusted based on the results of the model's predictions against the actual output vectors that are used as targets. On the contrary, the test set's outputs are calculated by the trained model and can be used to measure its accuracy. In our case, 296 cases were used as the training set while the remaining 169 cases were the evaluation/test set. Various test runs occurred and in each, the size of the training and test sets was constant but the cases that formed the two sets were randomly selected in order to cross validate the model.

4 Implementation

4.1 Previous Model Accuracy

The model proposed in [13] was implemented and tested against the data set. This model is based on linear regression in order to predict the HbF% of a patient. The formula that was proposed to calculate the HbF% was

$$e^{1.89+0.14*G1+0.3*G2+0.13*G3+0.1*G4}$$

where G1, G2, G3 and G4 are the values of the representative gene variants summarized in Table 1. An initial observation is that due to the nature of the formula, the percentage levels that are closer to the edge of the range could potentially be misrepresented. The exponent is a first-degree polynomial function and this is bound to affect the accuracy of the model. The results confirm our initial reservations regarding its accuracy, as the model has a mean error of 336.33% over the 465 samples. This evidence further supports our initial assumption that linear regression is not the optimal method to support our model as it fails to take into account the effect of the γ-gene activation and the young F-cells.

4.2 K-Nearest Neighbors Algorithm Implementation

The K-nearest neighbors algorithm [14] is a supervised machine learning algorithm that can be used to solve both classification and regression problems. Our problem requires a regression approach, as our algorithm's output is the HbF% of a patient which is a real and positive number. This algorithm is based on the assumption that similar things exist in close proximity. It calculates the Euclidean distance

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$

when trying to predict the value of new input. In our case, the three dimensions are the γ-gene activation, the F-cells and the HbF%. The algorithm was implemented in Java but a high-level description of it in pseudocode exists in Fig. 2.

```
Load the data
For each example in the test set
    For each example in the training set
        Calculate the distance between the 2 examples
        Add the distance and the index of the training set example to an Array
    Sort the Array and pick the first K elements
    Find the mean of the corresponding training set values and add as predicted value
Compare the predicted values to the actual values
```

Fig. 2 K-nearest neighbors pseudocode

The value K represents the number of "Neighbors" that will be used as a comparison in our predictions. The algorithm was tested for various values of K, in order to find the optimal one, as explored in more detail in Sect. 5.2. In general, we can expect that as we decrease the value of K our predictions become more unstable but if we increase it disproportionally, we will begin to witness more errors in our predictions.

4.3 Model Data Processing

The final model was implemented in Java and employed an object-oriented programming approach. This allowed for a separate implementation of our prediction algorithm and accuracy test and led to easier manipulation of the data and cleaner maintenance and reusability for future versions of the model. The main class is called "Case" and contains instances of the other classes as well as useful data fields for the operations of the program. Figure 3 illustrates the steps of the model regarding Case, along with its class diagram. The cells highlighted in yellow represent the data fields or instances of classes that are different from the previous step.

At first, the data set is imported and the values for hbfp (HbF%), hb and g1, g2, g3, g4 (representative variants) are populated. The values of gamma and F-cells are initially empty, as they are calculated based on the values of the representative gene variables and the Hb levels of the patient during the next step. Each gene affects those values differently, through its representative variants. The application of the K-nearest neighbors algorithm follows, with a training set of 296 entries and a test set of the remaining 169 entries. Section 5 contains all the information regarding the choice of the number K, as well as data on test runs that were conducted with different-sized training/test sets. The model produces a prediction for each entry of the test set and its value is stored in the respective instance of Case in the calculated hbf data field. Finally, the accuracy of the predicted value is calculated and stored in the prediction acc data field. The mean accuracy of all predictions on the test set is calculated and the program terminates after printing that overall value for the entirety of the model.

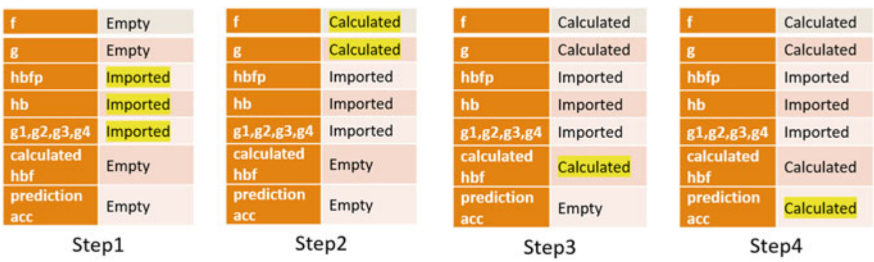


Fig. 3 Case diagram and execution steps

4.4 Alternative Algorithms Comparison

Various alternative algorithms were implemented as the base of our model and their accuracy was tested against the K-nearest neighbors algorithm. This was an especially important comparison as the base algorithm could potentially greatly affect the accuracy of the model. In order to have an initial metric of their accuracy, their standard deviation for our training set was compared and summarized in Table 2.

The comparison of the standard deviation presented in Table 2 is the first indicator that the nearest neighbors algorithm should be used as a base for our model since it has the lowest value among the ones tested. Although standard deviation does not guarantee that the chosen algorithm will be optimal, graphs such as those in Fig. 4 were also taken into account in the decision-making process, to ensure the best possible choice.

Table 2 Standard deviation of potential algorithms

Potential algorithm	Standard deviation
Decision tree	7.07
Gradient boosted trees	5.03
Linear regression	5.52
Nearest neighbors	4.57
Neural network	4.92
Random forest	4.96
Gaussian process	4.73

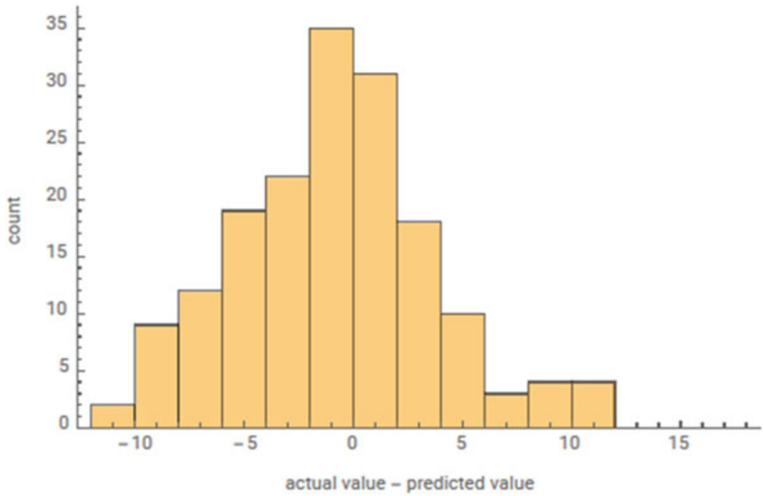


Fig. 4 Actual versus predicted output values of the count of cases

The perpendicular axis in Fig. 4 represents the difference between the actual and the predicted value of HbF% on a patient while the horizontal axis the count of cases with that difference. Even in this unoptimized state, the nearest neighbors algorithm shows promising results, with many predictions coming really close to the corresponding actual values. The aforementioned observations led us to the conclusion that the nearest neighbors algorithm is the optimal algorithm as the base for our model.

5 Results and Evaluation

5.1 Accuracy

The mean accuracy achieved with the use of 296 entries as a training set and 169 as test set is 87.25%. This is an excellent result that proves the validity of our theoretical model and our assumption that both the F-cells and the γ-gene activation greatly affect the percentage of HbF of a patient. Furthermore, the mean error of our model's predictions is just 12.75% and compares favorably to [13], which has a mean error of 336.33% and an accuracy of 22%. Moreover, only 11 cases have a difference greater than 0.5% between the predicted and real HbF%. That means 93.45% of our predictions have a sheer error that is less than 0.5. All these facts allow us to be extra confident that our model is safe to use, as there are no cases for which predictions are extremely wrong and potentially dangerous. There are two graphs presented below comparing the different implementations, in order to better visualize the accuracy of the respective model.

Figure 5 presents the linear regression model proposed in previous work versus our optimized model. The horizontal axis is the index number of the patient case while the perpendicular represents the HbF% values. Red dots mark the prediction values generated by the model while blue ones show the actual HbF% values. It is obvious that the linear regression approach lacks accuracy and in many cases the difference between the predicted and the actual value is quite pronounced. On the contrary, the very satisfactory accuracy of our model is apparent as the blue and red

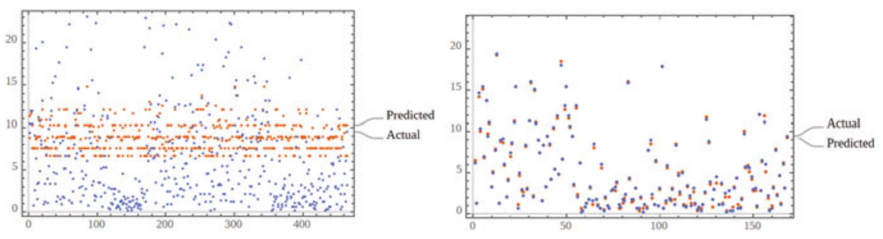


Fig. 5 Linear regression versus optimized nearest neighbors implementation

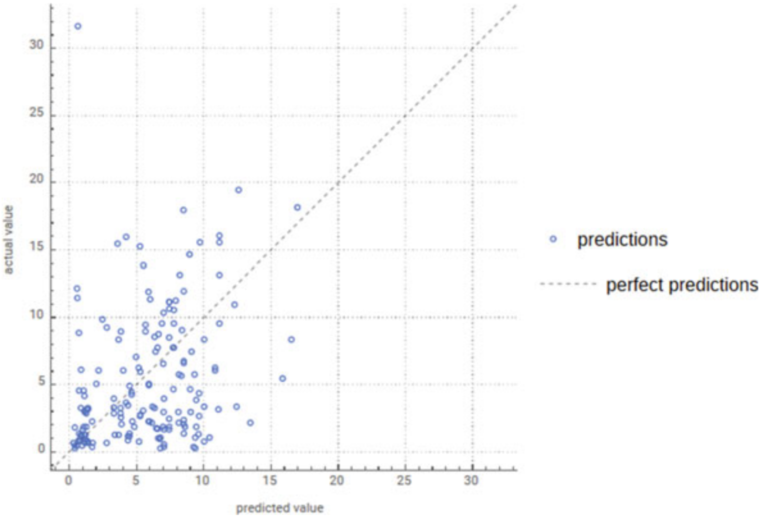


Fig. 6 Initial K-nearest neighbors implementation

dots belonging to the same pair are very close or even non-distinguishable from one another and the different pairs can be easily pinpointed. It is an obvious improvement over our unoptimized model shown in Fig. 6, as most of the cases where there was a sizeable difference between the predicted and the actual value have been eliminated and the overall mean error has been therefore greatly reduced.

The perpendicular axis in Fig. 6 contains the actual HbF% values while the horizontal represents the predicted ones. If a prediction is perfect it occupies a spot on the diagonal dotted line. Although far from perfect, even this initial implementation shows more promise than the linear regression in Fig. 5 and already has several cases where the predicted value is perfect or nearly perfect. This indicates that the chosen algorithm was the correct choice as the base for our model.

5.2 K-Value Evaluation

The nearest neighbors algorithm was tested for a range of values for the variable K in order to decide on the optimal one. This value represents the number of "Neighbors" that should be taken into account by the algorithm in determining the value of its prediction. The accuracy results for the different values of K are summarized in Table 3.

As mentioned before, there is an optimal value for K, which in our case is 4, while all other values that are higher or lower have decreased accuracy in their predictions. Despite the fact that 4 is an even number, there is no obstacle to choosing it, as our problem is a regression one and thus there are no tiebreakers that need resolving.

Table 3 K-value accuracy

K-value	Mean accuracy
1	85.35
2	86.29
3	86.39
4	87.25
5	86.43
6	85.87
7	84.84
8	84.64
9	84.36

5.3 Training Set Size and Cross-Validation

The algorithm was executed with varying numbers of entries for the training and the test set. It was observed that increasing the number of entries for the training set resulted in improved accuracy for the predictions of the algorithm. This implies that our model can further improve its accuracy by processing a larger data set than the one provided. Still, our data set of 465 patients proved more than enough to achieve 87.25% accuracy in our predictions. Moreover, all tests were conducted multiple times with random sampling for the training set, in order to cross validate the model and ensure the validity of its accuracy. All accuracy values presented in this paper are the mean accuracy of the respective test runs.

6 Conclusions

The objective of our research was to develop a prediction model in order to find the HbF% of patients suffering from sickle cell disease. Our aim was to explain a large proportion of HbF variability in sickle cell patients through common genetic variants and to build a polygenic score that can be calculated to predict HbF levels and, to a certain degree, disease severity.

The inputs of the model are the representative variants of four genes regulating sickle cell disease. After consecutive implementations and testing, our final model was based on the nearest neighbors algorithm and utilized data from 465 patients to make a regression prediction regarding their HbF%. With a test set of 169 cases, our model achieved 87.25% accuracy in its predictions, a significant improvement from the currently implemented model [13].

6.1 Model Applications

A major advantage of our model is that the high accuracy of its results allows it to be used as a quick estimation tool for small and medium-sized clinical trials. Another advantage is the small number of inputs that it requires to predict the HbF%. This makes our model easy to implement, as those inputs (representative variants, Hb levels) can be made readily available with common blood samples from the patients. Further beneficiaries of such a system will be researchers and clinicians conducting clinical or drug trials, in Africa, the UK, and elsewhere, where this score will help to adjust for genetic background variability that obscures test outcomes.

6.2 Future Work

The fact that the accuracy of our model increased as the size of the test set was incremented can motivate us to further expand our data set in the future, in order to provide an even more accurate prediction model. Furthermore, the possibility of more genes affecting the HbF% of a patient can be explored by enhancing our model in order to accommodate these extra inputs. Finally, if the size of the training/test sets increases notably, methods and techniques can be explored and implemented in order to improve the execution time of the model, if it increases disproportionately in that case.

References

1. Platt OS, Brambilla DJ, Rosse WF, Milner PF, Castro O, Steinberg MH, Klug PP (1994) Mortality in sickle cell disease–life expectancy and risk factors for early death. N Engl J Med 330(23):1639–1644
2. Platt OS, Thorington BD, Brambilla DJ, Milner PF, Rosse WF, Vichinsky E, Kinney TR (1991) Pain in sickle cell disease: rates and risk factors. N Engl J Med 325(1):11–16
3. Paikari A, Sheehan VA (2018) Fetal haemoglobin induction in sickle cell disease. Br J Haematol 180(2):189–200
4. Adams R, McKie V, Nichols F, Carl E, Zhang DL, McKie K, Hess D (1992) The use of transcranial ultrasonography to predict stroke in sickle cell disease. N Engl J Med 326(9):605–610
5. Miller ST, Sleeper LA, Pegelow CH, Enos LE, Wang WC, Weiner SJ, Kinney TR (2000) Prediction of adverse outcomes in children with sickle cell disease. N Engl J Med 342(2):83–89
6. Steinberg MH (2005) Predicting clinical severity in sickle cell anaemia. Br J Haematol 129(4):465–481
7. Rees DC, Williams TN, Gladwin MT (2010) Sickle-cell disease. The Lancet 376(9757):2018–2031
8. Razak SAA, Murad NAA, Masra F, Chong DLS, Abdullah N, Jalil N, Aziz NA (2018) Genetic modifiers of fetal haemoglobin (HbF) and phenotypic severity in β-thalassemia patients. Curr Mol Med 18(5):295–305

9. Gil KM, Carson JW, Porter LS, Scipio C, Bediako SM, Orringer E (2004) Daily mood and stress predict pain, health care use, and work activity in African American adults with sickle-cell disease. Health Psychol 23(3):267
10. Bae HT, Baldwin CT, Sebastiani P, Telen MJ, Ashley-Koch A, Garrett M, Bhatnagar P (2012) Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. Blood 120(9):1961–1962
11. Makani J, Menzel S, Nkya S, Cox SE, Drasar E, Soka D, Fegan G (2011) Genetics of fetal hemoglobin in Tanzanian and British patients with sickle cell anemia. Blood 117(4):1390–1392
12. Darshana T, Bandara D, Nawarathne U, de Silva U, Costa Y, Pushpakumara K, Wijayawardena M (2020) Sickle cell disease in Sri Lanka: clinical and molecular basis and the unanswered questions about disease severity
13. Gardner K, Fulford T, Silver N, Rooks H, Angelis N, Allman M, Rees DC (2018) g(HbF): a genetic model of fetal hemoglobin in sickle cell disease. Blood Adv 2(3):235–239
14. Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 4:580–585

Simulations of Instationary Schrodinger Equation with Coupled Time- and Space Splitting Methods

Jürgen Geiser and Mohammad Hajiketabi

Abstract In this paper, we present a coupled method of a time- and spatial-splitting approach. Such combined methods are used to reduce the computational amount of the time- and space approximation methods, while we could concentrate on each time or spatial part, which could also be done in parallel and reduce the computational time. We deal with the instationary Schrodinger equation, which are well known for modelling molecular-dynamical problems. Based on efficient combination of both time- and space splitting approaches, we could speed up the computational process and optimise each part in the time and space splitting to efficient solver methods.

Keywords Time-dependent schrodinger equation · Schwarz-Waveform relaxation method · Operator-splitting method · Gross-Pitaevskii equation · Time-space splitting method · Strang-Splitting method · Exact classical Schwarz-Waveform relaxation method

1 Introduction

We are motivated to model molecular problems related with Schrodinger equations. We apply highly oscillation energy potentials, which leads to stiff differential equations. Here, we propose different multiscale methods to overcome the delicate stiffness problems. We deal with the general nonlinear Schrödinger (or GPE), which is given as

J. Geiser (✉)
The Institute of Theoretical Electrical Engineering, Ruhr University of Bochum, Universitätsstrasse 150, 44801 Bochum, Germany
e-mail: juergen.geiser@ruhr-uni-bochum.de
URL: https://homepage.ruhr-uni-bochum.de/Juergen.Geiser/

M. Hajiketabi
Department of Applied Mathematics, Imam Khomeini International University, Qazvin 34149-16818, Iran

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_11

$$i\hbar \frac{\partial}{\partial t} \mathbf{u}(\mathbf{r}, t) = [\mathbf{u}, H], \tag{1}$$

$$\begin{aligned} i\hbar \frac{\partial}{\partial t} \mathbf{u}(\mathbf{r}, t) &= \Big(-\frac{\hbar^2}{2m} \nabla^2 + V_{ext}(\mathbf{r}) - \mu + \\ &+ \int \mathbf{u}^{\dagger}(\mathbf{r}', t)\; V(\mathbf{r} - \mathbf{r}')\; \mathbf{u}(\mathbf{r}', t)\; d\mathbf{r}' \Big) \mathbf{u}(\mathbf{r}, t). \end{aligned} \tag{2}$$

where $\hat{H} = -\frac{\hbar^2}{2m} \nabla^2 + V(x, t)$.

where m is the particle's reduced mass, i is imaginary unit, V(x,t) is its potential energy, ∇^2 is the Laplacian (a differential operator), and u is the wave function. For solving such a delicate Schrödinger equation, which is strongly coupled in time and space, see [10]. We consider a combination of time- and space splitting, which allows to reach higher accurate results in the time and also in the spatial dimensions, see [4] and reduce the computational amount with respect to the decomposition in smaller time and spatial domains.

The paper is outlined as following: The numerical methods for the time-space splitting is given in the Sect. 2. The numerical results are presented in the Sect. 3. In the final Sect. 4, we summarise our results.

2 Time-Space Splitting Method for Instationary Schrödinger Equations

Time-space splitting methods are important to reduce the computational amount in large scale computations of evolution equations, see [6, 7].

- For the Time-splitting methods, we have the idea to decouple full operators in the time evolutions to simpler operators, which are solved individual with fast adequate. Here, we can consider different splitting ideas:
  - Time-spectral splitting, see [4]
  - Operator-splitting, see [7]
  - Conservative splitting approaches, see [10]
- For the Space-splitting methods, we have the idea to decouple the full spacial domain in simpler spatial domains, which are solved individual, e.g., with parallel options, much more faster. Here, we can consider different splitting ideas:
  - Feti-methods, which are Lagrange multiplier based methods, see [12, 13].
  - Classical or optimal Schwarz-waveform-relaxation methods (SWR), which are iterative approaches, see [2, 6].

We apply the operator-splitting schemes as a time-splitting approach and the SWR methods as s spatial-splitting approach. Further, we improve the classical methods with the following algorithmical ideas:

- We apply so-called exact SWR method, which means, we embed the time-splitting approach as an exact method into the space splitting, see our novel Algorithm 1. Here, we have the benefit to skip one splitting approach and accelerate the schemes.
- We apply higher order time-discretization schemes to improve the ABA-splitting scheme, which means, we embed a CN (Crank-Nicolson)-splitting, which is conservative and stabilise the splitting approach for large scale computations, see [10], see our novel Algorithm 2.

In the following, we discuss two algorithms, see 1 and 2, which are applied to the harmonic oscillator potential for the Schrödinger equation.

Algorithm 1 *In this algorithm, we decompose space and apply an exact computation based on the exp-integrator for the spacial operator and the potential operator.*

1. *We initialise the vector* $u_{1,i}^0 = u_0(x_i),\ 1 \le i \le M/2 + \tilde{m}$ *and* $u_{2,i}^0 = u_0(x_i)$, $M/2 - \tilde{m} \le i \le M$.
2. *We have to apply the time-steps* $n = 1, \ldots, N$ *as an outer iteration* *We initialise the vector* $u_1^{(n+1,0)} = u_1^n$ *and* $u_2^{(n+1,0)} = u_2^n$.
3. *We apply the wave-form-iterations* $k = 1, \ldots, K$ *as an inner iteration.* **The application is an exact SWR method with exact integration of the term** *In this iterations, we have to apply the 2 steps for the 2 domains:*
4. *Step 1: Solve the* u_1*-part*

$$u_1^{(n+1,k)} = \exp(\boldsymbol{A}\Delta t)u_1^{(n)} + (\boldsymbol{A})^{-1}(\exp(\boldsymbol{A}\Delta t) - I)\frac{\boldsymbol{f}_1^{(n)} + \boldsymbol{f}_1^{(n+1,k-1)}}{2}, \quad (3)$$

where $u_1 = (u_{1,1}, \ldots, u_{1,M/2+\tilde{m}})^t$ *and the matrix is given as (explicit version)*

$$\boldsymbol{A} = \begin{bmatrix} -2a - bx_1^2 & a & 0 & 0 \ldots & 0 \\ a & -2a - bx_2^2 & a & 0 \ldots & 0 \\ 0 & a & -2a - bx_2^2 & a \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots \\ 0 & 0 & 0 & \ldots & -2a - bx_{M/2+\tilde{m}}^2 \end{bmatrix} \in \mathbb{R}^{M/2+\tilde{m}} \times \mathbb{R}^{M/2+\tilde{m}} \quad (4)$$

with $a = \frac{i}{2\Delta x^2}$ *and* $b = \frac{i}{2}$. *The right-hand side is given as* $\boldsymbol{f}_1^{(n+1,k-1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ a\ u_{2,M/2+\tilde{m}+1}^{(n+1,k-1)} \end{bmatrix} \in \mathbb{R}^{M/2+\tilde{m}}$, *where for the starting condition, we have* $\boldsymbol{f}_1^{(n+1,0)} =$

$$\boldsymbol{f}_1^{(n)} \; and \boldsymbol{f}_1^{(n)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ a \; u_{2,M/2+\tilde{m}+1}^{(n)} \end{bmatrix} \in \mathbb{R}^{M/2+\tilde{m}}$$

5. *Step 2: Solve the u_2-part*

$$u_2^{(n+1,k)} = \exp(\boldsymbol{B}\Delta t)u_2^{(n)} + (\boldsymbol{B})^{-1}(\exp(\boldsymbol{B}\Delta t) - I)\frac{\boldsymbol{f}_2^{(n)} + \boldsymbol{f}_2^{(n+1,k-1)}}{2}, \quad (5)$$

where $u_2 = (u_{2,M/2-\tilde{m}}, \ldots, u_{2,M})^t$ and the matrix is given as (explicit version)

$$\boldsymbol{B} = \begin{bmatrix} -2a - bx_{M/2-\tilde{m}}^2 & a & 0 & 0 \ldots & 0 \\ a & -2a - bx_{M/2-\tilde{m}+1}^2 & a & 0 \ldots & 0 \\ 0 & a & -2a - bx_{M/2-\tilde{m}+2}^2 & a \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \; \ddots & \vdots \\ 0 & 0 & 0 & \ldots & -2a - bx_M^2 \end{bmatrix} \in$$

$$\mathbb{R}^{M-(M/2-\tilde{m})+1} \times \mathbb{R}^{M-(M/2-\tilde{m})+1}, \quad (6)$$

with $a = \frac{i}{2\Delta x^2}$ and $b = \frac{i}{2}$. The right-hand side is given as $\boldsymbol{f}_2^{(n+1,k-1)} = \begin{bmatrix} a \; u_{1,M/2-\tilde{m}-1}^{(n+1,k-1)} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{M-(M/2+\tilde{m})+1}$ *where for the starting condition, we have*

$$\boldsymbol{f}_2^{(n+1,0)} = \boldsymbol{f}_2^{(n)} \; and \boldsymbol{f}_2^{(n)} = \begin{bmatrix} a \; u_{1,M/2-\tilde{m}-1}^{(n)} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{M-(M/2+\tilde{m})+1}$$

6. *If we have $k = K$ or $\max\{||u_1^{(n+1,k)} - u_1^{(n+1,k-1)}||, ||u_2^{(n+1,k)} - u_2^{(n+1,k-1)}||\} \leq \epsilon$ then we have to the result $u_1^{n+1} = u_1^{(n+1,k)}$ and $u_2^{n+1} = u_2^{(n+1,k)}$ go to Step 7, else we do $k = k + 1$ and go to Step 3.*
7. *If we have $n = N$, we are done otherwise we have to do $n = n + 1$ and goto Step 2.*

Algorithm 2 *In the second algorithm, in addition to space splitting, we decompose the spacial operator with the potential operator and apply the Strang-splitting* *method. The potential operator, we could compute fast with the spectral methods as an* exp*-function, while the spatial operator is computed with the CN method.*

1. *We initialise the vector* $u^0_{1,i} = u_0(x_i),\ 1 \le i \le M/2 + \tilde{m}$ *and* $u^0_{2,i} = u_0(x_i)$, $M/2 - \tilde{m} \le i \le M$.
2. *We have to apply the time-steps* $n = 1, \dots, N$ *as an outer iteration*
   We initialise the vector $u_1^{(n+1,0)} = u_1^n$ *and* $u_2^{(n+1,0)} = u_2^n$.
3. *We apply the wave-form-iterations* $k = 1, \dots, K$ *as an inner iteration.*
   The application is a CN method
   In this iterations, we have to apply the 2 steps for the 2 domains:
4. *Step 1: Solve the* u_1*-part*

$$\tilde{u}_1^{(n+1,k)} = \exp(\boldsymbol{A}_0) u_1^{(n,k)},\ 1/2 \textit{ time-step}, \tag{7}$$

$$\hat{u}_1^{(n+1,k)} = (\boldsymbol{A}_2)^{-1} \boldsymbol{A}_1 \tilde{u}_1^{(n+1,k)} + (\boldsymbol{A}_2)^{-1} \boldsymbol{f}_1,\ \textit{full time-step}, \tag{8}$$

$$\tilde{u}_1^{(n+1,k)} = \exp(\boldsymbol{A}_0) \hat{u}_1^{(n+1,k)},\ 1/2 \textit{ time-step}, \tag{9}$$

where $u_1 = (u_{1,1}, \dots, u_{1,M/2+\tilde{m}})^t$ *and the potential matrix is given as*

$$\boldsymbol{A}_0 = \begin{bmatrix} -bx_1^2 & 0 & 0 & 0 \dots & 0 \\ 0 & -bx_2^2 & 0 & 0 \dots & 0 \\ 0 & 0 & -bx_2^2 & 0 \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots \\ 0 & 0 & 0 & \dots & -bx^2_{M/2+\tilde{m}} \end{bmatrix} \in \mathbb{R}^{M/2+\tilde{m}} \times \mathbb{R}^{M/2+\tilde{m}} \tag{10}$$

and the space-dependent matrix is given as (explicit version)

$$\boldsymbol{A}_1 = \begin{bmatrix} 1-2a & a & 0 & 0 \dots & 0 \\ a & 1-2a & a & 0 \dots & 0 \\ 0 & a & 1-2a & a \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1-2a \end{bmatrix} \in \mathbb{R}^{M/2+\tilde{m}} \times \mathbb{R}^{M/2+\tilde{m}} \tag{11}$$

and the space-dependent matrix is given as (implicit version)

$$\boldsymbol{A}_2 = \begin{bmatrix} 1+2a & -a & 0 & 0 \dots & 0 \\ -a & 1+2a & -a & 0 \dots & 0 \\ 0 & -a & 1+2a & -a \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1+2a \end{bmatrix} \in \mathbb{R}^{M/2+\tilde{m}} \times \mathbb{R}^{M/2+\tilde{m}} \tag{12}$$

with $a = \frac{i\ \Delta t}{4\ \Delta x^2}$ *(full time-step based on the B operator) and* $b = \frac{1}{2} \frac{i\ \Delta t}{2}$ *(half time-step based on the A operator).*

The right-hand side is given as $\boldsymbol{f}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ a \ \frac{u^n_{2,M/2+\tilde{m}+1} + u^{(n+1,k-1)}_{2,M/2+\tilde{m}+1}}{2} \end{bmatrix} \in \mathbb{R}^{M/2+\tilde{m}}$

5. *Step 2: Solve the* u_2*-part*

$$\tilde{u}_2^{(n+1,k)} = \exp(\boldsymbol{B}_0) u_2^{(n,k)}, \ 1/2 \textit{ time-step}, \tag{13}$$

$$\hat{u}_2^{(n+1,k)} = (\boldsymbol{B}_2)^{-1} \boldsymbol{B}_1 \tilde{u}_2^{(n+1,k)} + (\boldsymbol{B}_2)^{-1} \boldsymbol{f}_2, \ \textit{full time-step}, \tag{14}$$

$$\tilde{u}_2^{(n+1,k)} = \exp(\boldsymbol{B}_0) \hat{u}_2^{(n+1,k)}, \ 1/2 \textit{ time-step}, \tag{15}$$

where $u_2 = (u_{2,M/2-\tilde{m}}, \ldots, u_{2,M})^t$ *and the potential dependent matrix is given as (explicit version)*

$$\boldsymbol{B}_0 = \begin{bmatrix} -bx^2_{M/2-\tilde{m}} & 0 & 0 & 0 & \ldots & 0 \\ 0 & -bx^2_{M/2-\tilde{m}+1} & 0 & 0 & \ldots & 0 \\ 0 & 0 & -bx^2_{M/2-\tilde{m}+2} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & \ldots & -bx^2_M \end{bmatrix} \in \mathbb{R}^{M-(M/2-\tilde{m})+1} \times \mathbb{R}^{M-(M/2-\tilde{m})+1} \tag{16}$$

and the space-dependent matrix is given as (explicit version)

$$\boldsymbol{B}_1 = \begin{bmatrix} 1-2a & a & 0 & 0 & \ldots & 0 \\ a & 1-2a & a & 0 & \ldots & 0 \\ 0 & a & 1-2a & a & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & \ldots & 1-2a \end{bmatrix} \in \mathbb{R}^{M-(M/2-\tilde{m})+1} \times \mathbb{R}^{M-(M/2-\tilde{m})+1} \tag{17}$$

and the space-dependent matrix is given as (explicit version)

$$\boldsymbol{B}_2 = \begin{bmatrix} 1+2a & -a & 0 & 0 & \ldots & 0 \\ -a & 1+2a & -a & 0 & \ldots & 0 \\ 0 & -a & 1+2a & -a & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & \ldots & 1+2a \end{bmatrix} \in \mathbb{R}^{M-(M/2-\tilde{m})+1} \times \mathbb{R}^{M-(M/2-\tilde{m})+1} \tag{18}$$

with $a = \frac{i \ \Delta t}{4 \ \Delta x^2}$ *and* $b = \frac{1}{2} \frac{i \ \Delta t}{2}$ *(half time-step based on the A operator).*

The right-hand side is given as $\boldsymbol{f}_2 = \begin{bmatrix} a \ \frac{u^n_{1,M/2-\tilde{m}-1} + u^{(n+1,k-1)}_{1,M/2-\tilde{m}-1}}{2} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{M-(M/2-\tilde{m})+1}$

6. *If we have* $k = K$ *or* $\max\{||u_1^{(n+1,k)} - u_1^{(n+1,k-1)}||, ||u_2^{(n+1,k)} - u_2^{(n+1,k-1)}||\} \le \epsilon$ *then we have to the result* $u_1^{n+1} = u_1^{(n+1,k)}$ *and* $u_2^{n+1} = u_2^{(n+1,k)}$ *go to Step 7, else we do* $k = k + 1$ *and go to Step 3.*
7. *If we have* $n = N$*, we are done otherwise we have to do* $n = n + 1$ *and go to Step 2.*

3 Numerical Results

In the following, we apply the different splitting methods in time and space for the instationary Schrodinger equation (1D Case).

We have the following Schrödinger equation, see also [5]:

$$i\hbar \frac{\partial}{\partial t} u(x,t) = -\frac{\hbar^2}{2\,m} \Delta u(x,t) + V(x,t) u(x,t), \ \ \Omega = [-10, 10], \ \ t \in [0, 10], \tag{19}$$

$$u_0(x) = u_{analyt}(x, 0), \Omega = [-10, 10], \tag{20}$$

$$u(x,t) = u_{analyt}(x,t), \ \ x \in \partial\Omega, \ \ t \in [0, 10], \tag{21}$$

where we have the harmonic oscillator potential $V(x,t) = \frac{k}{2} x^2$ and assume $\hbar = m = k = 1.0$.

The analytical solution is given with respect to the different wave-functions:

$$u_{n_x}(x,t) = \sqrt{\frac{1}{\sqrt{\pi}\ 2^{n_x}\ n_x!}} \exp(-i\ E_{n_x}\ t) \exp(-x^2/2)\ H_{n_x}(x), \tag{22}$$

where $E_{n_x} = \frac{(n_x \pi)^2}{2}$ and H_{n_x} are the Hermitian polynomials with $H_{n_x+1}(x) = 2x H_{n_x}(x) - 2n_x H_{n_x-1}(x)$ and $H_{-1}(x) = 0, \ H_0(x) = 1$. Further $n_x = 0, 1, \ldots,$ are the different wave-numbers, see [5], of the oscillator.

For the comparison of the exact and numerical errors, we deal with the following error measurement:

- L_1-error

$$||e_{\Delta x, \Delta t}(t)||_{L_1} = \Delta t\ \Delta x \sum_{n=1}^{N} \sum_{i=1}^{M} |u_{num}(x_i, t^n) - u_{exact}(x_i, t^n)|, \tag{23}$$

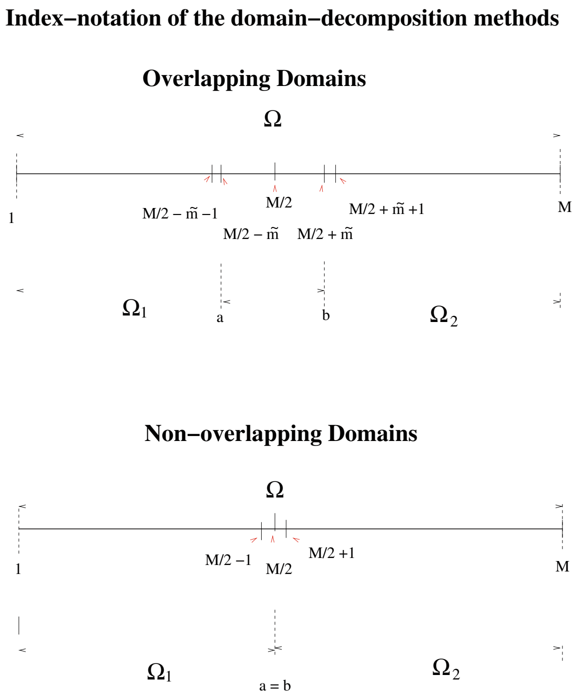


Fig. 1 Index-notation for the domain-decomposition method (for all SWR methods)

where we compute the error at end-time t, where we have the spatial grid size $\Delta x = 1/M$ and the time-step $\Delta t = 1/N$. In Fig. 1, we can see the indices based on the different domains.

We applied Algorithm 1 and obtained the results in Table 1 for two different sizes of overlaps and different time and space-step sizes. We can see, the increasing overlap reduces the repetition of the algorithm k.

The results of Algorithm 2 obtained in Table 2 for two different sizes of overlaps and different time and space-step sizes. From this Table, we see the convergence of the algorithm with respect to time and spatial steps.

In Table 3, we compare two algorithms with respect to their efficiencies and also we included the solution of the problem without any splitting. For just space splitting, we see the benefits of Algorithm 1, while we could distribute the computational amount. For both time- and space splitting, Algorithm 2, we have the best performance. We see that whatever the spatial intervals get bigger the performance of Algorithms 1 and 2 improves. This should also be true for the time scale. We also could apply splitting of more than 2 subdomains for space, then we could reduce the CPU time more.

Table 1 L_1-error for the Schrödinger equation in $t \in [0, 1]$ and $x \in [-10, 10]$ using the Algorithm 1 for two different sizes of overlapping 5 and 10

Time-step	Error	k	Error	k	Error	k	Error	k	Error	k	Error	k
1/5	1.5781e-002	4	1.5781e-002	2	3.9802e-003	4	3.9802e-003	3	9.9842e-004	6	9.9908e-004	4
1/10	1.4506e-002	4	1.4506e-002	2	3.6598e-003	4	3.6598e-003	3	9.1765e-004	6	9.1798e-004	4
1/20	1.3866e-002	2	1.3866e-002	2	3.4987e-003	4	3.4987e-003	3	8.7708e-004	6	8.7735e-004	4
1/40	1.3545e-002	2	1.3545e-002	2	3.4180e-003	4	3.4184e-003	3	8.5670e-004	6	8.5689e-004	4
1/80	1.3384e-002	3	1.3384e-002	2	3.3776e-003	4	3.3777e-003	3	8.4646e-004	6	8.4654e-004	4
Overlap	5		10		5		10		5		10	
Space-step	1/5				1/10				1/20			

Table 2 L_1-error for the Schrödinger equation in $t \in [0, 1]$ and $x \in [-10, 10]$ using the Algorithm 2 for two different sizes of overlapping 5 and 10

Time-step	Error	k	Error	k	Error	k	Error	k	Error	k	Error	k
1/5	2.8807e-002	4	2.8807e-002	3	1.8815e-002	4	1.8815e-002	3	1.6478e-002	4	1.6478e-002	4
1/10	1.7326e-002	2	1.7326e-002	2	6.9472e-003	4	6.9472e-003	3	4.5187e-003	4	4.5187e-003	3
1/20	1.4512e-002	2	1.4512e-002	2	4.2183e-003	4	4.2183e-003	3	1.6857e-003	4	1.6857e-003	3
1/40	1.3700e-002	2	1.3700e-002	2	3.5852e-003	4	3.5852e-003	3	1.0367e-003	4	1.0367e-003	3
1/80	1.3421e-002	2	1.3421e-002	2	3.4183e-003	4	3.4183e-003	3	8.8956e-004	4	8.8956e-004	3
Overlap	5		10		5		10		5		10	
Space-step	1/5				1/10				1/20			

4 Conclusion

We have presented a combination of time- and space splitting methods, which could reduce the complexity of the full problem and solve much more simpler differential equations. We present the application to a instationary Schrödinger equation, which is decomposed into the spatial and potential part. Further, we decompose the underlying space domain and solve them with SWR methods. To accelerate the parts, we embed exact solutions of the SWR methods and could speed up the algorithms. We could accelerate the speed of computation with respect to the decomposition of spatial domains, further, we also obtain a benefit in the time splitting. Such decomposition

Table 3 Comparison of Algorithms 1, 2 and the solution (without any splitting) on different space intervals in $t \in [0, 1]$. We fixed $\Delta t = 0.01$, $\Delta x = 0.05$ in all computations and reported L_1-errors and computational times

Space	Solution		Algorithm 1		Algorithm 2	
interval	L1-error	CPU (s)	L1-error	CPU (s)	L1-error	CPU (s)
[−10, 10]	8.4432e-004	1.0140e+000	8.4440e-004	1.2480e+000	8.7227e-004	8.2680e-001
[−30, 30]	8.4432e-004	1.1466e+001	8.4440e-004	1.4414e+001	8.7227e-004	5.7096e+000
[−50, 50]	8.4432e-004	3.9234e+001	8.4440e-004	3.9405e+001	8.7227e-004	2.1278e+001
[−70, 70]	8.4432e-004	1.2534e+002	8.4440e-004	8.1432e+001	8.7227e-004	5.8999e+001
[−90, 90]	8.4432e-004	3.0247e+002	8.4440e-004	1.5247e+002	8.7227e-004	8.7781e+001
[−110, 110]	8.4432e-004	8.5847e+002	8.4440e-004	2.3429e+002	8.7227e-004	1.6267e+002

methods in time and in space help to distribute the amount of computational work and preserve the numerical accuracy with sufficient higher order splitting schemes. In our future scope, we consider the combinations of the time-space splitting with respect to a general iterative splitting scheme, which could be analysed as fixpoint scheme.

References

1. Antoine X, Lorin E (2017) An analysis of Schwarz waveform relaxation domain decomposition methods for the imaginary-time linear Schrödinger and Gross-Pitaevskii equations. Numerische Mathematik 137:923–958
2. Antoine X, Lorin E (2019) On the rate of convergence of Schwarz waveform relaxation methods for the time-dependent Schrödinger equation. J Comput Appl Math 354:15–30
3. Antoine X, Lorin E (2019) Asymptotic convergence rates of Schwarz waveform relaxation algorithms for Schrödinger equations with an arbitrary number of subdomains. Multisc Sci Eng 1(1):34–46
4. Bao W, Jin S, Markowich PA (2003) Numerical study of time-splitting spectral discretizations of nonlinear Schrödinger equations in the semiclassical regimes. Siam J Sci Comput 25(1):27–64
5. Becerril R, Guzman FS, Rendon-Romero A, Valdez-Alvarado S (2008) Solving the time-dependent Schrödinger equation using finite difference methods. Revista Mexicana de Fisica e 54(2):120–132
6. Daoud D, Geiser J (2007) Overlapping schwarz wave form relaxation for the solution of coupled and decoupled system of convection diffusion reaction equation. Appl Math Comput Elsevier 190(1):946–964
7. Geiser J (2011) Iterative splitting methods for differential equations. Chapman & Hall/CRC Numerical Analysis and Scientific Computing Series, edited by Magoules and Lai

8. Geiser J (2018) Multiscale modelling and splitting approaches for fluids composed of coulomb-interacting particles. Math Comput Modell Dyn Syst Taylor & Francis 24(4):319–382
9. Geiser J (2016) Additive and iterative splitting methods for multiscale and multiphase coupled problems. J Coupled Syst Multisc Dyn 4(4):271–291
10. Geiser J, Nasari A (2019) Comparison of splitting methods for deterministic/stochastic Gross–Pitaevskii equation. Math Comput Appl 24:76
11. Jia H, Li K (2011) A third accurate operator splitting method. Math Comput Modell 53:387–396
12. Mathew TPA (2008) Lagrange multiplier based substructuring: FETI method. In: Domain decomposition methods for the numerical solution of partial differential equations. Lecture notes in computational science and engineering, vol 61. Springer, Berlin, Heidelberg
13. Toselli A, Widlund O (2005) Domain decomposition methods—algorithms and theory. Springer Series in Computational Mathematics, vol 34

Prediction of Movie Success Using Twitter Temporal Mining

Bushra Alhijawi and Arafat Awajan

Abstract This research work presents a movie success prediction mechanisms using Twitter mining. The proposed methods predict movie success in terms of rating and temporal popularity. We develop two models to achieve these goals. The first model, the rating prediction model (RPM), aims to predict the users' satisfaction with the product using an ensemble regression model. The second model is a temporal product popularity model (T-PPM) that aims to predict the product's temporal popularity using a random forest classifier. We collect a new dataset called TweetAMovie from IMDb and Twitter to evaluate the developed models. The comparative results against baseline methods demonstrate the superiority of the proposed models. On average, the RPM and T-PPM achieved 32.4% and 30.2% improvement in terms of accuracy. Also, the average precision and F-score results improved by T-PPM are 6.7% and 2.6%, respectively.

Keywords Review mining · Sentiment analysis · Twitter · Temporal · Rating prediction · Popularity prediction · Internet movie database

1 Introduction

Social media (e.g., Twitter, Facebook, YouTube) are powerful platforms where people can share their feedback about products and services. Businesses consider these reviews to enhance their services and improve the sales of their products [6, 9, 12].

B. Alhijawi (✉) · A. Awajan
Princess Sumaya University for Technology, Amman 11941, Jordan
e-mail: b.alhijawi@psut.edu.jo

A. Awajan
e-mail: awajan@psut.edu.jo

A. Awajan
Mutah University, Karak 61710, Jordan

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_12

Twitter, with over 319 million[1] monthly active users, covers a wide variety of people's thoughts and topics, making it a goldmine for companies to enhance their clout and reputation [5, 12]. Today, predicting the popularity and success of a product, service, or an event by analyzing tweets attracts the attention of researchers [1, 7, 10].

The movie industry is a big business carrying profits or losses up to millions of dollars. Whereas movies are a great source of entertainment for people and may affect their decisions, lifestyle, and political opinions [11]. Various online websites keep track of movies like Rotten Tomatoes and Internet Movie Database (IMDb). IMDb is the world's most popular and authoritative source for movie, TV, and celebrity content. IMDb tracks a movie's success in terms of revenue, ratings, and popularity. In addition, IMDb computes a generic and reliable rating reflecting the users' opinions on a movie using the users' ratings. Also, IMDb weekly ranks movies according to their popularity depending on statistical features.

This research work introduces a new challenge (i.e., temporal popularity) that can be exploited to provide business with recommendations of future popular movies. To the best of our knowledge, this is the first work in the movie success prediction field that predicts the popularity from the IMDb perspective. The core idea is to adopt the users' opinions about movies on Twitter to estimate the average satisfaction level and temporal popularity. The average satisfaction level refers to the users' average ratings for a particular movie, while the temporal popularity is a binary measure indicating that a particular movie is a trend within a specific period. We developed two predictive models for predicting the average rating and temporal popularity of a movie. Both models fuse features from two sources: IMDb (i.e., regular movie features) and Twitter (i.e., tweet-related features). The Rating Prediction Model (RPM) handles the rating prediction challenge as a regression problem using the cumulative tweet-related features. While the Temporal Popularity Prediction Model (T-PPM) utilizes temporal tweet-related features and handles the problem as a binary classification problem. We collect a new dataset called TweetAMovie to train and test the proposed models. The contributions of this research work can be summarized as follows:

– A new dataset (TweetAMovie) that consists of tweets and movie features.
– A rating prediction mechanism depending on public opinion.
– A temporal popularity prediction mechanism depending on public opinion.

This paper is organized as follows. Section 2 presents the background and literature review. Section 3 shows details related to the movie success prediction model development. Section 4 presents and discusses the experimental evaluation of the proposed models. Finally, Sect. 5 concludes the research proposal.

[1] https://www.imdb.com/interfaces/.

2 Related Works

Mining the textual data available on social networks to determine the public opinion toward an entity is crucial for people and businesses. People's decisions about purchasing a particular product may be affected by the opinion of others. Also, business analyzes the peoples' feedback about their products to make future decisions. One opinion from a single person is insufficient for action in various applications such as movie success [1, 7, 9, 11], tweets popularity [4, 6, 13], and hashtag popularity [6]. Thus, opinion mining applications require to analyze opinions of a large number of people. Machine learning algorithms such as Artificial Neural Network (ANN) [1–3], Linear Regression [1, 6], Decision Tree [1, 3, 11], Support Vector Machine (SVM) [1, 11], K-Nearest Neighbor (KNN) [11], Random forest [11, 13], Logistic Regression [3], and Naive Bayes [3] are the commonly used for developing the product success prediction models.

In the literature, movie success is measured from three preservatives: revenue [9, 11], rating [1, 3, 7], and view-based popularity [2]. In other domains such as event and social content (i.e., tweet and hashtag), the popularity is measured based on the tweets related to a particular product [3, 4, 6, 8, 13]. In this research work, we measure movie success in terms of rating and popularity. Popularity refers to the IMDb temporary popularity measure. To the best of our knowledge, T-PPM is the first work predicting the temporal popularity of the movie from the IMDb perspective. The Top-100 popular movies list on the IMDb website is weekly updated. Thus, our models require that we collect a new dataset, TweetAMovie, containing the temporal popularity feature. Also, the proposed models' input features differ from those used in recent methods. RPM depends on the release date, genre, popularity record, aggregated tweets count, retweets count, replies count and like count/movie, cumulative sentimental score/movie, and cumulative emotional score/movie features. T-PPM uses genre, release year, historical popularity record, temporal sentimental score/movie, and temporal emotional score/movie features as predictors. Table 1 summarizes the related works.

3 Movie Success Prediction Models

This section presents the developed movie success prediction models: Rating Prediction Model (RPM) and Temporal Popularity Prediction Model (T-PPM). We focus on predicting rating and popularity from the IMDb perspective by fusing features from two sources: IMDb and Twitter. The development process of RPM and T-PPM involves three phases: collecting and pre-processing data (Sect. 3.1), preparing features (Sect. 3.2), and developing model (Sect. 3.3) phases.

Table 1 Summarization of contributions

Contribution	Goal	Domain	Input data source	Input features	Method	Context data
Magdum and Megha [9]	–Movie sales –Movie success	Movie	–IMDb –Twitter	Bag of words and Hidden sentiment factor which detail the nature of sentiments on a particular aspect	–Sentiment probabilistic latent semantic analysis –Autoregressive sentiment –Aware model	Temporal
Latif and Afzal [7]	Movie success	Movie	–IMDb	Rating, MPAA rating, Genre, Awards, Opening weekend, Screens, Metascore, Votes count, Budget	–Classification (Logistic regression, Simple Logistic, Multilayer Perceptron, Decision Tree, Naive Bayes, PART)	None
Ahmed et al. [1]	–Movie sales –Movie rating	Movie	–IMDb –Twitter –YouTube	Genre, Budget, Screens count, Sequel, Aggregate Actor Followers, Views count, Likes count, dislikes count, Comments count, Sentiment Score	–Classification (Linear Regression, Decision Tree, ANN, SVM)	None
Katarla and Rao [6]	Content popularity	HashTag	–Stream of raw tweets	Tweets count, Retweets count, Maximum number of users' followers posting the hashtag, Timestamp ranking score, Impression count, Favorite count, Number of users posting the hashtag per hour, Number of long tweets including the hashtag per hour	–Classification (Linear Regression, SVM)	Temporal
Chen et al. [4]	Content popularity	Tweet, HashTag	–Twitter	Tweets, Users, and Time	Deep learning	Temporal
Mundra et al. [11]	Movie sales	Movie	–IMDb –Twitter	Genre, Average gross of director, Average gross of all the actors, PT-NT ratio and Rating	–Classification (Linear Discriminant Analysis, CART, KNN, SVM, Random Forest)	None

(continued)

Table 1 (continued)

Contribution	Goal	Domain	Input data source	Input features	Method	Context data
Chen et al. [3]	Product rating	Movie, Toys, Kindle	–Amazon data –Yelp data	Review	Deep learning	None
Lei et al. [8]	–Product rating	–	–Yelp data	Product features extracted from user reviews, Sentiment words that are used to describe the product features, Sentiment of a specific user on an item/product	Matrix factorization	None
Oliveira et al. [13]	Content popularity	Tweet	–Twitter	Followers count, Statuses count (tweets/retweets), Favorites count, Number of times the user was listed, Number of days of the account, If the user is verified, Hashtags count, URLs, Mentions, Length of the tweets, Words count, Is a tweet a reply, Tweet timestamp, Videos count, Photos count, Sentiment score	Random forest	None
Bai et al. [2]	Movie Popularity	Movie	–Douban –Youku	Historical view counts, Ratings, Ratings count, Release date, Language, Genre, Reputation and influence of actors and the film director	Deep learning	None

3.1 TweetAMovie Dataset Collection

TweetAMovie is a new movie rating and temporal popularity dataset generated from IMDb and Twitter. This dataset contains 6015 average ratings ranged between 1 and 9.4 on 6041 movies and 23 genres. Each movie belongs to at least one genre and at most three genres. We collected the most popular movie list for 6 weeks. In total, TweetAMovie includes 368726 English tweets composed of 4370 well-structured tweets and 364356 unstructured tweets. Our dataset comprises three files: MovieReg.csv, MovieTPop.csv, and Tweets.csv, which store movie regular features, temporal popularity of movies, and tweet-related features. We adopted an IMDb identifier as a movie id to facilitate additional metadata enrichment.

MovieReg.csv is collected from IMDb on 2/12/2019.[2] This file contains information about movies, including title, genre, release year, ratings, and the number of votes. MovieTPop.csv stores the Top-100 popular movie data. We collected the Top-100 popular movie list, which is weekly updated within the period 2/12/2019 and 13/1/2020. The data stored in Tweets.csv are collected from Twitter. TweetAMovie includes well-structured and unstructured tweets. The structured tweets were obtained by querying a series of regular expressions include "I rated" and hashtag "#IMDb" (e.g., "I rated Fast & Furious Presents: Hobbs & Shaw (2019) 8/10 #IMDb"). The unstructured tweets are extracted by querying the movie title. We faced a challenge when the movie title consists of words that may be used in several contexts in which unrelated tweets are retrieved. For example, for the movie "They Live" we obtained the tweet "**they said this? IN 2020???? what universe do they live in???**", which is unrelated to the movie. Therefore, we validated the tweet using IBM Watson Natural Language Understanding (IBM-W-NLU) service[3] by extracting its category to ensure that TweetAMovie dataset includes only movie-related tweets.

3.2 Feature Preparation

The feature preparation phase is responsible for preparing the input features of RPM and T-PPM. This phase includes four steps: sentiment analysis, emotion analysis, feature generation, and feature selection. We adopt IBM-W-NLU service to analyze the tweet sentiment and emotions (i.e., sadness, joy, fear, disgust, and anger) scores. However, IBM-W-NLU service was not able to correctly classify the structured tweets since it did not consider the rating data in the tweet (e.g., "8/10"). Thus, the tweet sentiment and emotions were computed based on the pre-defined rules. We set the sentimental and "Joy" emotion scores of tweets with [0–4] rating to $[-1-0.2]$ and [0–0.4], tweets with a rating of 5 to 0 and 0.5, and tweets with [6–10] rating to [0.2–1] and [0.6–1]. Other emotion scores are computed based on the "Joy" emotion

[2] https://www.imdb.com/interfaces/.

[3] https://www.ibm.com/cloud/watson-natural-language-understanding.

Table 2 Input features of each model

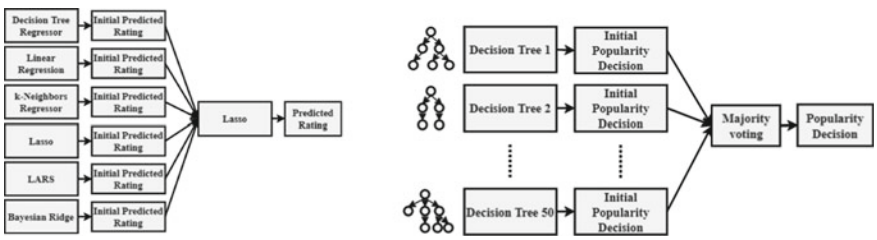
Rating prediction model	Popularity prediction model
• Rating of movie	
• Genre of movie	• Genre of movie
• Release year	• Release year
• Popularity record	• Historical popularity record
• Overall count of tweets	• Temporal count of tweets
• Overall count of retweets	• Temporal count of retweets
• Overall count of replies	• Temporal count of likes
• Overall count of likes	• Temporal sentimental score/movie
• Overall sentimental score/movie	• Temporal emotional score/movie
• Overall emotional score/movie	• Timestamp

score (1- score(Joy)). For example, the sentiment and emotion scores of "I rated Fast & Furious 6 (2013) 8/10 #IMDb" are [sentiment score:0.6, sadness:0.2, joy:0.8, fear:0.2, disgust:0.2, anger:0.2].

A different set of input features are prepared for each prediction model. The RPM uses cumulative tweet-related features, while T-PPM uses temporal tweet-related features. Also, the genre feature is transformed into binary features using one-hot encoding. Table 2 shows the prepared features for each model. We applied an oversampling method to the training data of T-PPM for addressing the problem of imbalanced data.

3.3 Proposed Prediction Models

Movie success will be measured based on rating and temporal popularity. The rating perspective reflects the users' satisfaction with a particular movie, while the popularity perspective is the movie's reputation within a certain period. RPM (Fig. 1a) predicts the users' satisfaction level with movies using an ensemble regression model which consists of six regression models, linear regression, decision tree regressor, K-neighbors regressor, lasso, Least-Angle Regression (LARS), and Bayesian ridge. Each of the regression models is trained on the same training data to produce initially estimated ratings. RPM uses these initial predicted results as input features to lasso regression that predicts the final decision. We selected and tuned these regression models experimentally. T-PPM (Fig. 1b) predicts the Top-100 popular movies within a certain period. This model uses temporal sentimental and emotional scores as predictor features. T-PPM is an ensemble classifier involving 50 decision trees. Each of the decision trees was trained on the same training data to produce initial popularity decision. Then, T-PPM produces the final decision using a majority voting mechanism.



(a) RPM Architecture. (b) T-PPM Architecture.

Fig. 1 Movie success prediction model

4 Experiments and Results

This section provides details of how the proposed movie success prediction models are tested. Various experiments were conducted for comparing the accuracy of the proposed methods against alternative baseline classification and prediction techniques using TweetAMovie data. The first experiment aims to compare the RPM with baseline regression techniques (i.e., Decision Tree Regressor (DTR), Linear Regression (LR), K-Neighbors Regressor (K-NR), Lasso, LARS, and Bayesian Ridge (BR)) in terms of accuracy. It is worth to notify that all models use the same set of input features. Three evaluation measures are used in this experiment: Mean Squared Error (MSE) and Mean Absolute Error (MAE), and the coefficient of determination (R2). The second experiment evaluates the performance of T-PPM and compares it to the performance of baseline classifiers (i.e., Logistic Regression (LogR), Random Forest (RF), Gaussian Naive Bayes (GNB), K-Neighbors Classifier(K-NC), Decision Tree (DT), and AdaBoost (AB)). In the second experiment, accuracy, precision, recall, and F-score are used as evaluation measures. Three experimental setups are considered: DEV1, DEV5, and DEV10. In DEV1, 70% and 30% of the dataset are randomly selected as training and testing data. Whereas the average of 100 runs/method is considered for comparison. The performance of the prediction methods is measured across fivefold cross-validation and tenfold cross-validation in DEV5 and DEV10, respectively.

4.1 Experiments Results

The results of the first experiment are shown in Table 3. On average, RPM outperforms other approaches by 11.9%, 7.2%, and 57.6% in terms of MSE, MAE, and R2. In terms of DEV1, the achieved accuracy by LR, Lasso, LARS, and BR is close to the achieved accuracy by RPM. However, the R2 results of RPM are much better than other methods indicating that RPM fits the data better than other alternative methods. RPM achieved fewer prediction errors under DEV2 and DEV10 setups. Also, the

Table 3 Results of the rating prediction accuracy

Experiment setup	Method	MSE	MAE	R2
DEV1	DTR	0.75952 ± 0.0315	0.66699 ± 0.0114	0.21623 ± 0.0159
	LR	0.58649 ± 0.0284	0.56742 ± 0.0103	0.39486 ± 0.0172
	K-NR	0.73165 ± 0.0307	0.64305 ± 0.0103	0.24501 ±0.0149
	Lasso	0.58647 ± 0.0284	0.56741 ± 0.0103	0.39488 ± 0.0172
	LARS	0.58649 ± 0.0284	0.56742 ± 0.0103	0.39486 ± 0.0172
	BR	0.58756 ± 0.0285	0.56806 ± 0.0103	0.39378 ± 0.0168
	RPM	**0.57077** ± 0.0271	**0.55986** ± 0.0106	**0.4111** ± 0.0167
DEV5	DTR	0.77947 ± 0.06315	0.67406 ±0.01959	0.12963 ± 0.07033
	LR	0.61048 ± 0.07615	0.57943 ± 0.0345	0.32051 ± 0.05955
	K-NR	0.77973 ± 0.068	0.66925 ± 0.01693	0.13138 ± 0.03748
	Lasso	0.61047 ± 0.07617	0.57943 ± 0.0345	0.32052 ± 0.05954
	LARS	0.61048 ± 0.07615	0.57943 ± 0.0345	0.32051 ± 0.05955
	BR	0.61202 ± 0.07903	0.58068 ± 0.03526	0.31918 ±0.0577
	RPM	**0.57516** ± 0.07658	**0.56122** ± 0.04062	**0.36046** ± 0.05339
DEV10	DTR	0.76943 ± 0.08882	0.67061 ± 0.03382	0.10983 ± 0.14063
	LR	0.60193 ± 0.113	0.57666 ± 0.05662	0.3166 ± 0.06367
	K-NR	0.75221 ± 0.11331	0.65327 ± 0.04321	0.13821 ± 0.08991
	Lasso	0.60191 ± 0.11299	0.57665 ± 0.05661	0.31662 ± 0.06366
	LARS	0.60193 ± 0.113	0.57666 ± 0.05662	0.3166 ± 0.06367
	BR	0.60218 ± 0.1109	0.57694 ± 0.05463	0.31556 ± 0.06255
	RPM	**0.56009** ± 0.10691	**0.5535** ± 0.05737	**0.36555** ± 0.04906

Table 4 Results of the second experiment

Experiment setup	Method	Accuracy	Precision	Recall	F-Score
DEV1	LogR	0.7054 ± 0.0306	0.0468 ± 0.0057	0.7419 ± 0.0904	0.088 ± 0.0104
	AB	0.9708 ± 0.0044	0.3766 ± 0.0486	0.7102 ± 0.0904	0.4864 ± 0.0528
	GNP	0.4988 ± 0.2126	0.0465 ± 0.0215	**0.7967** ± 0.1329	0.0813 ± 0.0353
	K-NC	0.957 ± 0.0047	0.1264 ± 0.038	0.2131 ± 0.0739	0.1576 ± 0.0489
	DT	0.9516 ± 0.0136	0.2685 ± 0.0612	0.7494 ± 0.0935	0.3879 ± 0.0656
	RF	**0.9858**± 0.002	**0.6851** ± 0.0835	0.4611 ± 0.0973	**0.5411** ± 0.0821
DEV5	LogR	0.6953 ± 0.0336	0.0454 ± 0.0075	0.7541 ± 0.08	0.0855 ± 0.0137
	AB	0.9655 ± 0.0197	0.39 ± 0.1929	0.6874 ± 0.1076	0.4682 ± 0.1462
	GNP	0.4678 ± 0.1842	0.0555 ± 0.0291	**0.8129** ± 0.1455	0.0885 ± 0.0362
	K-NC	0.9586 ± 0.008	0.124 ± 0.0449	0.1913 ± 0.0673	0.1484 ± 0.0495
	DT	0.947 ± 0.031	0.306 ± 0.1547	0.7228 ± 0.1286	0.3935 ± 0.1297
	RF	**0.9841** ± 0.0049	**0.6432** ± 0.2174	0.4287 ± 0.1489	**0.4782** ± 0.121
DEV10	LogR	0.6947 ± 0.0442	0.0454 ± 0.0119	0.7493 ± 0.1375	0.0855 ± 0.0219
	AB	0.9672 ± 0.026	0.4423 ± 0.2661	0.699 ± 0.2227	0.4948 ± 0.2219
	GNP	0.4903 ± 0.1882	0.06 ± 0.0448	**0.8031** ± 0.1268	0.0936 ± 0.0527
	K-NC	0.958 ± 0.0126	0.1376 ± 0.0993	0.2012 ± 0.1234	0.1569 ± 0.1
	DT	0.9431 ± 0.037	0.3017 ± 0.2135	0.7501 ± 0.1873	0.3902 ± 0.1877
	RF	**0.984** ± 0.0084	**0.6965** ± 0.3233	0.4415 ± 0.1953	**0.4975** ± 0.1988

results show that the accuracy and R2 of the alternative methods become worst under DEV2 and DEV10 setups.

Table 4 presents the gathered results of T-PPM and other classifiers. On average, T-PPM achieved 29.1%, 32.3%, and 3.3% higher accuracy for DEV1, DEV5, and DEV10. The results show that T-PPM achieved the highest precision, regardless of the considered experiment setup. Thus, T-PPM reduces the false positive, on average, by 6.9%, 6.2%, and 6.2% under DEV1, DEV5, and DEV10. In terms of recall, the LogR, AB, DT, and GNP method outperform T-PPM. However, their precision results are much lower than T-PPM, which indicates that those methods achieve high false positive. Therefore, the reader can observe that T-PPM has the best F-score results when compared to other methods.

5 Conclusion

This research presented two new movie success prediction models, specifically rating prediction and temporal popularity prediction models (RPM and T-PPM). Both models are trained and tested on a newly collected dataset called TweetAMovie. TweetAMovie includes regular movie features and tweet-related features. The regular movie's features are collected from IMDb, while tweet-related features are scraped from Twitter. The RPM considers the cumulative values of tweet-related features and the T-PPM depends on the temporal values of those features. The results showed that RPM and T-PPM achieved better performance compared to baseline methods in terms of accuracy. Also, the performance of T-PPM was evaluated in terms of recall, precision, and F-score. In the future, this work can be integrated with other systems (e.g., information retrieval and recommender systems) to improve their performance.

References

1. Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I (2015) Using crowd-source based features from social media and conventional features to predict the movies popularity. In: IEEE international conference on smart cDity/SocialCom/SustainCom (SmartCity), pp 273–278. IEEE
2. Bai W, Zhang Y, Huang W, Zhou Y, Wu D, Liu G, Xiao L (2020) DeepFusion: predicting movie popularity via cross-platform feature fusion. Multimed Tools Appl 79:1–18
3. Chen C, Zhang M, Liu Y, Ma S (2018) Neural attentional rating regression with review-level explanations. In World Wide Web Conference on World Wide Web—WWW '18, pp 1583–1592. ACM Press
4. Chen G, Kong Q, Xu N, Mao W (2019) NPP: a neural popularity prediction model for social media content. Neurocomputing 333:221–230
5. Doong SH (2016) Predicting twitter hashtags popularity level. In 49th Hawaii International Conference on System Sciences (HICSS), pp 1959–1968. IEEE
6. Katarla KV, Rao VV (2017) Analyzing and predicting hash tags popularity on twitter. Int J Comput Sci Inf Secur 15(11):9–16
7. Latif MH, Afzal H (2016) Prediction of movies popularity using machine learning techniques. Int J Comput Sci Netw Secur (IJCSNS) 16(8):127

8. Lei X, Qian X, Zhao G (2016) Rating prediction based on social sentiment from textual reviews. IEEE Trans Multimedia 18(9):1910–1921
9. Magdum SS, Megha JV (2017) Mining online reviews and tweets for predicting sales performance and success of movies. In: International conference on intelligent computing and control systems (ICICCS), pp 334–339. IEEE
10. Mazumder P, Chowdhury NA, Bhuiya MA-U-A, Akash SH, Rahman RM (2018) A fuzzy logic approach to predict the popularity of a presidential candidate. In: Modern approaches for intelligent information and database systems, pp 63–74. Springer, Berlin
11. Mundra S, Dhingra A, Kapur A, Joshi D (2018) Prediction of a movie's success using data mining techniques. In: Information and communication technology for intelligent systems, pp 219–227. Springer, Berlin
12. Noori M, Maktoubian J (2019) Business improvement approach based on sentiment twitter analysis: case study. EAI Endorsed Trans Cloud Syst 5(14):14–20
13. Oliveira N, Costa J, Silva C, Ribeiro B (2019) Retweet predictive model for predicting the popularity of tweets. In: 10th international conference on soft computing and pattern recognition (SoCPaR), pp 185–193. Springer, Berlin

The Model of Business Intelligence Development by Applying Cooperative Society Based Financial Technology

A. Al-Khowarizmi, Rahmad Syah, and Marischa Elveny

Abstract Technological advances seem to have no limit so that its power can penetrate many sectors, one of them is business. On the one hand, the development of financial technology has proven to be of benefit to consumers, business actors and the world economy, but on the other hand it has potential risks that if it does not adequately follow technological developments it could disrupt the financial system. Cooperatives are understood as legal entities that are established based on the principle of kinship and also adheres to the principles of social economy with the aim to improve their members in carrying out dominant transactions using cash, then modelled for the future will rely on cashless and FinTech virtual accounts. So that a new model is formed in the movement of business intelligence that is FinTech-based cooperative society that is able to increase the movement of cooperative society, reduce the operational costs of cooperative society and prosper the community with one mobile device.

Keywords Modelling · Business Intelligence · FinTech · Cooperative Society

1 Introduction

Currently the financial sector is competing in developing a new technology model called financial technology (FinTech) [1]. FinTech is the result of a product by combining financial services and technology which in turn changes the IT-based

A. Al-Khowarizmi
Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
e-mail: alkhowarizmi@umsu.ac.id

R. Syah
Informatics Department, Universitas Medan Area, Medan, Indonesia
e-mail: rahmadsyah@staff.uma.ac.id

M. Elveny (✉)
Data Science & Computational Intelligence Research Group, Universitas Sumatera Utara, Medan, Indonesia
e-mail: marischaelveny@usu.ac.id

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_13

business model [2]. FinTech uses information technology in the financial system that results in new services, products, technology or business models and has a good impact on monetary stability, smoothness, security and reliability of the payment system in the financial process, both debt and receivables, as well as financial system stability and efficiency [3, 4]. FinTech began to develop from conventional to Islamic [5]. FinTech developers are also competing to create applications that are different from the others [6, 7]. But the field of intelligence cannot be separated from FinTech, it is the development of business intelligence that can be applied in this technology [8] Even business intelligence has long been developed in the 1990s where the lack of hardware impeded the development of business intelligence so that with the emergence of mobile or smartphone-based devices the development of business intelligence developed rapidly and simultaneously implemented FinTech [9–11].

Phan on reference [12] conducted a study that FinTech influenced the development of banks, especially conventional banks that did not follow the development of FinTech, This research also explains the excellent benefits if banks implement FinTech based on the ratio of income from net interest to total assets, ratio of net income to total assets, ratio of net income to total equity, and returns on productive assets. In addition, Gai on Reference [13] conducted a survey regarding FinTech and explained that FinTech has become a popular term that describes the new technology being adopted by financial service institutions. This term covers a variety of techniques, from data security to delivery of financial services and has five technical aspects summarized namely security and privacy, data collection, hardware selection and development of infrastructure, information systems, and service models for customers so that every problem can be actively and provide solutions with existing FinTech.

In Germany FinTech was adopted by non-bank star up developers, the survey results showed that household levels of trust and comfort with new technology, financial literacy, and overall transparency impact its tendency to turn to FinTech [14]. But in Indonesia, FinTech developed starting with an online transportation application that was developed with payments through mobile devices (GO-PAY, OVO) and is now developing continuously with the involvement of the Financial Services Authority or Indonesia which is called OJK [15–17].

In addition, FinTech develops to the health service side, as explained by reference [18] FinTech must be able to facilitate health services both by registration life insurance services and making life insurance claims. FinTech also progressed towards education by paying tuition fees and education credits for students [19, 20]. FinTech changes the lifestyle of humans with a mobile device that can do whatever it wants [21, 22].

However, many researchers who developed FinTech and Business Intelligence Model from all fields are one of the concepts that are left behind, namely cooperatives. Cooperative is a family-based business or social principle [23]. Cooperatives have the principle of residual operating results, where profits are distributed based on the contribution of each member [24]. This is not the case for some FinTech applications.

From the principle of the cooperative created an interest in being modelled on FinTech. So that the process or business at FinTech makes a contribution to the system based on a formula that is modelled so that the principle of the remainder of the business is carried out at every financial bookkeeping cover. This can increase income both at FinTech managers, lenders or debt lenders and debt recipients as well as any payments incorporated in the FinTech.

This is the basis that requires the development of FinTech that is able to provide mutual benefits and make business and non-business communication sustainable. so that with the implementation of FinTech this can be increased to big data so that it can perform forecasting such as the application of the industrial revolution 4.0 in the business intelligence section [25, 26].

2 Material and Method

2.1 Cooperative Society World Data

In parts of the world, cooperatives are able to make a very significant contribution to the economy of every country. Cooperatives control strategic sectors and help determine economic policy. The following materials provided from the cooperative are some of the cooperative's success records obtained from various sources several years ago. The Association of Asian Confederation of Credit Unions provides data that in Asia there are 45.3 million active members of credit unions (savings and loans cooperatives). INAES also explained that in Argentina it has 11,357 cooperative groups with more than 9 million members or the equivalent of 22% of Argentina's population. Belgium also has 29,933 cooperative communities recorded in 2001. Canada, based on the survey results, explained that four out of every ten Canadians are members of at least one cooperative. 70% of Quebec's population are active members of cooperatives, while in Saskatchewan 56% of the population is registered as members. Data obtained by CONFECOOP explains that in Colombia, more than 4 million Colombians are members of cooperatives or the equivalent of 9.17% of the total population. The population of Costa Rica is more than 10% of the people are counted as members of the cooperative. In Finland, a cooperative called the S-Group has an active member of 1,468,572 people or the equivalent of 62% of the total population of Finland. Germany, there are 20 million active members of the cooperative, this represents a quarter of the total population. In Japan, one third of the population are active members of cooperatives. In India, more than 239 million people are members of cooperatives. Malaysia 5.9 million people or 24% of the total population are members of the Malaysian cooperative. In addition, New Zealand has 40% of the total adult population active as members of cooperatives. Singapore has 50% of the population or 1.6 million people who are members of cooperatives and in the United States, 4 out of 10 people are members of cooperatives or the equivalent of 25% of the population [27–30].

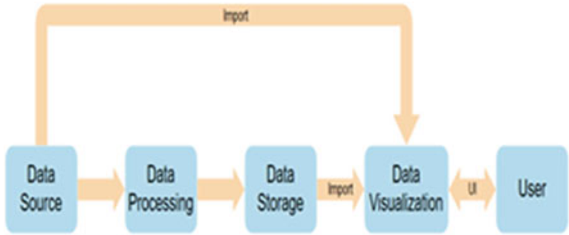


Fig. 1 Modelling data financial technology

2.2 *Relationships and BI Models to FinTech*

The data source is obtained in the form of a collection of information on the database server [31]. So that with the application of Business Intelligence in FinTech, it creates a new model which can support a data science concept [32]. With this paper, it is hoped that it can produce a standardized and IT-based model so that it can be used according to the needs of all industries and cooperative societies. The data model or system platform is most commonly used in structured conventional systems, but in data science it is mostly a database structure that relates to each other and provides information as illustrated in Fig. 1.

In Fig. 1 it is explained that this paper creates a model in which cooperative society is developed with the advancement of FinTech.

3 Result and Discussions

Cooperatives are understood as a legal entity that is established based on the principle of kinship and also adheres to the principles of social economy with the aim of bringing wealth to its members so that the benefits will be managed for the progress of cooperative performance and distributed to active members. At this time the cooperative was developed with the existence of an electronic cooperative (e-Cooperative) where cooperatives make loans online, while the borrower is not a member. This is the same as banks that are developing from internet banking, mobile banking (m-Banking) and e-Money to make payments.

In principle, in addition to the distribution of the Remnants of Business Results, cooperatives also have the principle of educational development and when the distribution of the Remnants of Operations Results of the distribution of the presentations in general cooperatives share profits to Members, Management, Employees, Trustees of educational contributions and reserve funds. However, by utilizing FinTech the cost distribution to management, employees and coaches is eliminated because all moves are based on the system. This certainly cuts operational costs. But the division is done based on member contributions to the FinTech application. To find a new model in FinTech the first time a simulation is performed on minimal members. The model offered in this system is summarized in Fig. 2.

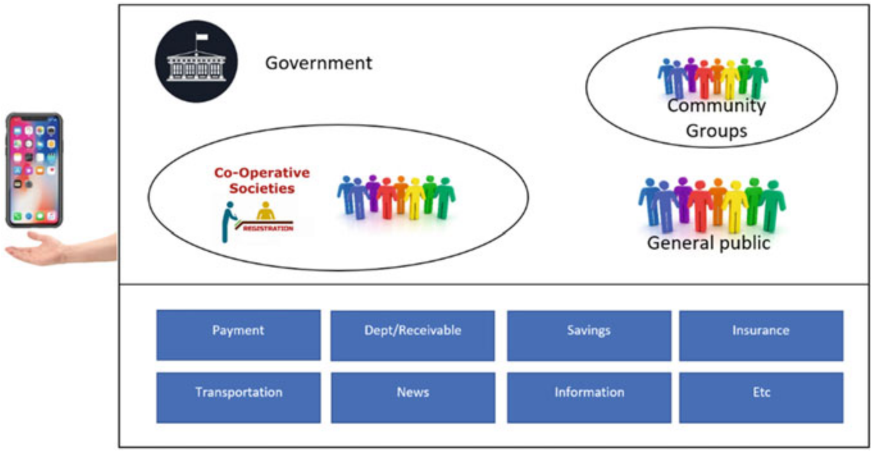


Fig. 2 Modelling by cooperative society-based FinTech

Along with the cooperative society in general the simulation process by implementing FinTech can be formulated based on Table 1.

From Table 1, it is simulated 5 users of FinTech-based Cooperative society who will be derived from the formula. Before distributing the Remaining Results of Operations, in accordance with the agreement of the Cooperative Society, the distribution of Remaining Results of Operations was also determined in advance, namely 70% for members, 10% for system management, 10% for educational development and 10% for reserve funds. So, from Table 1 you get the benefits of the formula (1).

$$LCS = DIR - VIR \tag{1}$$

where:

LCS is Profit of Cooperative Society.

DIR is Debt Interest Rates.

Table 1 Simulation data cooperative society-based FinTech

Member	Capital	Volunteer	Voluntary interest rates/year	Debt	Debt interest rates/year	Percentage/Contribution to the cooperative society
Member-1	1.000	10.000	1.200	–	–	–
Member-2	2.500	5.000	600	–	–	–
Member-3	3.000	–	–	15.000	1.800	83.3
Member-4	4.000	–	–	1.000	120	5.5
Member-5	500	–	–	2.000	240	11.11
SUM	11.000	15.000	1.800	18.000	2.160	100

VIR is Voluntary Interest Rates.

In detail, LCS in accounting is that the income account is reduced by the expense account and the tax account, so the net result of the Cooperative Society is obtained. After earning further profit divided according to the Cooperative Society agreement that is 70% for members, 10% for system management, 10% for educational development and 10% for reserve funds. So we need formula (2).

$$Lx = LCM * n\% \tag{2}$$

For Member:

$$LM = LCS * 70\% \tag{3}$$

For System:

$$LS = LCS * 10\% \tag{4}$$

For Education:

$$LE = LCS * 10\% \tag{5}$$

For Saving

$$LSa = LCS * 10\% \tag{6}$$

The next step is to distribute the Remaining Operations to members based on their contribution. Returning to Table 1, member-3 contributed 83.3% so member-3 received a remaining operating income of 209.92. The basis is based on the formula (7).

$$PM - i = LM * CCS \tag{7}$$

where:

PM − i is Member − i.

CCS is Contribution to the Cooperative Society.

Out of these formulas a new model, the Cooperative Society Based FinTech, is expected to increase business intelligence activities. Where with this model the Cooperative Society is more modern and suitable to be applied in various countries with the consideration that none of them violates state rules. In addition, Cooperative Society Based FinTech can cut operational costs and get a greater share of Remaining Results, this supports the welfare of the community. The difference between Cooperative Society Based FinTech with FinTech and General Cooperative Society is set out in Table 2.

Table 2 Excellence cooperative society based FinTech

No	Model	FinTech	Cooperative society general	Cooperative society based FinTech
1	Registration	√	√	√
2	Bank account	√	√	√
3	Savings/Capital		√	√
4	Loan	√	√	√
5	Activity a. Payment b. Insurance c. Merchant d. Education e. etc.	√ √ √ √ √		√ √
6	Remaining operating results		√	√

From Table 2 it can be seen that the Cooperative Society Based FinTech is able to create models to facilitate and prosper the community. In addition, this paper also explains that Cooperative Society must transform by presenting digital-based services such as those provided by FinTech. Cooperative society must adopt the development of information technology (IT), even though they are in remote areas. FinTech is aiming for markets in remote areas. Banking alone has learned a lot from the existence of FinTech. They have strengthened IT to be able to win the competition, so that the Cooperative Society should make a uniform effort. If it does not improve immediately, the Cooperative Society may be the first to be affected by digital disruption. Especially if there is still the mind of the Cooperative Society manager, who views FinTech as a threat. The cooperative society must instead view FinTech as a potential partner to work together. Many services can really be cooperated. Each has advantages that can complement each other, especially different loyal segments. So, Cooperative society must first improve through the adoption of information communication technology. Cooperative society is in a situation when the world is not only changing, but is developing rapidly through the establishment of the system for so many times contained in Science data. When the dominant transaction in cash, then in the future will rely on cashless virtual accounts and FinTech.

4 Conclusion

In summary, this paper develops mutual financial management progress, that is, FinTech is developed by implementing cooperative society or it can be called advancing cooperative society by utilizing FinTech. With the Cooperative Society Based FinTech able to reduce operational costs because all activities carried out by the system. Cooperative Society Based FinTech can also increase the movement of world money and serve humans with only one mobile device.

References

1. Anugerah DP, Indriani M (2018) Data protection in financial technology services (a study in indonesian legal perspective). Sriwij Law Rev 2(1):82. https://doi.org/10.28946/slrev.vol2.iss1.112.pp82-92
2. Alt R et al (2018) FinTech and the transformation of the financial industry. Electron Mark 28(3):235–243. https://doi.org/10.1007/s12525-018-0310-9
3. Le T-H et al (2019) Financial inclusion and its impact on financial efficiency and sustainability: empirical evidence from Asia. Borsa Istanbul Rev 19(4):310–322. https://doi.org/10.1016/j.bir.2019.07.002
4. Syah R, Nasution M, Nababan EB, Efendi S (2020) Optimization model under uncertainty for competitive merchant to customer behavior in financial technology using business metrics. 17(7). https://doi.org/10.1166/jctn.2020.9171
5. Rusydiana A (2019) How to develop sharia FinTech industry in Indonesia? interpretive structural model (ISM) approach. Al-Muzara'ah 6(2):117–128. https://doi.org/10.29244/jam.6.2.117-128
6. Luckandi D (2018) Analysis of payment transactions using FinTech at MSMEs in Indonesia. 4
7. Palmié M et al (2020) The evolution of the financial technology ecosystem: an introduction and agenda for future research on disruptive innovations in ecosystems. Technol Forecast Soc Change 151:119779. https://doi.org/10.1016/j.techfore.2019.119779
8. Fauzi F, Al-Khowarizmi AM (2020) The e-business community model is used to improve communication between businesses by utilizing. Jite 3(2):252–257
9. Fang LY et al (20187) Mobile business intelligence acceptance model for organisational decision making. Bull Electr Eng Informat 7(4):650–656. https://doi.org/10.11591/eei.v7i4.1356
10. Serrano W (2018) FinTech model: the random neural network with genetic algorithm. Proc Comput Sci 126:37–546. https://doi.org/10.1016/j.procs.2018.07.288
11. Duque HF et al (2018) Neural network model to re-rate and analyze consumer credit for FinTechs support vector machine. In: 13th Iberian conference on information systems and technologies (CISTI), pp 1–6. https://doi.org/10.23919/CISTI.2018.8399366
12. Syah R, Nasution MK, Elveny M, Arbie H (2020). Optimization model for customer behavior with MARS and KYC system. J Theoret App Inform Technol 98(13)
13. Gai K et al (2018) A survey on FinTech. J Netw Comput Appl 103:262–273. https://doi.org/10.1016/j.jnca.2017.10.011
14. Jünger M, Mietzner M (2020) Banking goes digital: the adoption of FinTech services by German households. Financ Res Lett 34:101260. https://doi.org/10.1016/j.frl.2019.08.008
15. Susilo AZ et al (2019) A comparative study of factors affecting user acceptance of GO-PAY and OVO as a feature of FinTech application. Proc Comput Sci 161:876–884. https://doi.org/10.1016/j.procs.2019.11.195
16. Iman N (2018) Is mobile payment still relevant in the FinTech era? Electron Commer Res Appl 30:72–82. https://doi.org/10.1016/j.elerap.2018.05.009
17. Hariyani I (2017) Legal protection and resolution of pm-FinTech service business disputes. J Legis Indones 14(3):345–358
18. Inkster B et al (2019) Improving insights into health care with data linkage to financial technology. Lancet Digit Heal 1(3):e110–e112. https://doi.org/10.1016/S2589-7500(19)30061-5

19. Davies E, Lea SEG (1995) Student attitudes to student debt. J Econ Psychol 16(4):663–679. https://doi.org/10.1016/0167-4870(96)80014-6
20. Q EG (2019) The effect of trust, perceived usefulness, and perceptions of ease of use on intention to behave in adopting financial technology. Universitas Lampung
21. Arif HFM (2019) Actualization of Zariah reasoning in anticipating crime through financial technology 18(2):848–867
22. Lee I, Shin YJ (2018) FinTech: ecosystem, business models, investment decisions, and challenges. Bus Horiz 61(1):35–46. https://doi.org/10.1016/j.bushor.2017.09.003
23. Suprayitno B (2012) Criticism of cooperatives (and their solutions) as a medium to encourage the growth of micro, small and medium enterprises. J Ekon dan Pendidik 4(2):14–35. https://doi.org/10.21831/jep.v4i2.608
24. Ganitri PT, Suwendra IW, Yulianthini NN (2013) The influence of own capital, loan capital, and business volume on the difference in operating results (SHU) in savings and loans cooperatives. Bisma J Manaj 4(2):105–113
25. Lubis AR et al (2019) Big data forecasting applied nearest neighbor method. In: ICSECC international conference on sustainable engineering and creative computing: new idea, new innovation, proceedings. https://doi.org/10.1109/ICSECC.2019.8907010
26. Al-Khowarizmi et al (2020) The effect of a secos in crude palm oil forecasting to improve business intelligence. Bull Electr Eng Informat 9(4):1604–1611. https://doi.org/10.11591/eei.v9i4.2388
27. Syah R, Nasution M, Nababan EB, Efendi S (2020) Knowledge Acceleration Estimator (KAE) model to customer behavior using business metrics. J Theoret App Inform Technol 98(08)
28. Castillo AE et al (2020) Management accounting: effect on key factors of competitive success in the cooperative sector. CIRIEC-Espana Rev Econ Publica Soc y Coop 9:119–126. https://doi.org/10.7203/CIRIEC-E.99.14087
29. Syah R, Elveny M, Nasution MKM (2020) Performance knowledge acceleration optimization with mars to customer behavior in merchant ecosystem. In: 3rd international conference on mechanical, electronics, computer, and industrial technology (MECnIT). Medan, Indonesia, pp 178–182. https://doi.org/10.1109/MECnIT48290.2020.9166604
30. Nicholls MG et al (2020) Resurrecting New Zealand's public healthcare system or a charity hospital in every town? Intern Med J 50(7):883–886. https://doi.org/10.1111/imj.14903
31. Syah R et al (2020) Optimization metrics model: the mobile wallet for merchant ecosystem. J Phys Conf Ser 1566:1. https://doi.org/10.1088/1742-6596/1566/1/012124
32. Lubis AR et al (2020) Optimization of distance formula in K-nearest neighbor method. Bull Electr Eng Inform 9(1). https://doi.org/10.11591/eei.v9i1.1464

A Computational Intelligent Cognition System Under Uncertainty

Ben Khayut, Lina Fabri, and Maya Avikhana

Abstract It has been explored the conception of implementation the system of the most important systemic psychological function of the Computational Brain (CB)—the System of Computational Situational Reasonable Cognition and Understanding of the reality Under Uncertainty with applying the Fuzzy Logic, Fuzzy Control, Computational Linguistics, Cognitive Psychology, Data Science, Computer Science at whole, oriented on introduction in the Artificial Super Intelligence Self-X system as one of the main components. Computational psychology is investigated and implemented on basis of the following CB's computational systemic mental situational functional processes of self-perception, self-inference, self-decision making, self-control, self-developing, intuition, self-awareness, self-consciousness, and self-understanding of reality. These processes are implemented on basis of the self-developing memory and modules, that use the self-computing computational models, computational mathematical modeling psychological situations under their time changes. The computed and identified psychological categories, properties, features, and essences of objects of reality are correlated with the corresponding subject area and are used by the mentioned processes for the intellectual analysis and modeling of the systemic situational reasonable cognition.

Keywords Computational · Super Intelligent · Cognition · System · Under Uncertainty

1 Introduction

1.1 Conception

The modern computational Machine Learning (ML) systems refer to statistical and expert *Narrow Artificial Intelligence* (NAI), and cannot *think* by itself and independently, be *conscious* and *cognize* themselves and the surrounding reality *without*

B. Khayut (✉) · L. Fabri · M. Avikhana
Intelligent Decisions Technologies Systems, 77711 Ashdod, Israel

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_14

using expert *human intelligence* in conditions of *uncertainty* and time *changes* of the *objects* and *situations* in the environment.

"Since *self-consciousness* can be defined as the 'ability to be aware of oneself,' we may assume that *self-cognition*, i.e., 'cognition of oneself,' has a meaningful connection with self-consciousness" [1].

Consequently, computational self-cognition [2] is realized on the basis of computational self-awareness [3], computational self-consciousness [4, 5], and computational self-understanding [2] of reality, which are one of the main functional properties of the *Computational Brain* (CB) (along with its auxiliary computed functional properties of self-developing [2], self-inference [6], self-decision-making [7], self-control [8], self-reasoning [9], self-thinking [10], intuition, wisdom, mind [11], and others) in conditions of uncertainty of situations and objects in *its* surrounding environment.

Understanding reality is associated with the interaction of people and *Computational Artificial Super Intelligence Self-X* (CASISX) systems [12].

Since the computational *understanding* of reality is mainly associated with the computing of the psychological properties of *awareness*, *consciousness*, and *cognition*, we will consider the models and methods of their self-developing on basis of opportunities of CASISX.

Computational cognition is *computational psychology* [13].

Consequently, systems capable of computing the psychological properties of *cognition* based on *systemic situational* self-development, self-organizing, self-awareness, self-consciousness, and self-cognition *at a whole* refers to the CASISX *system* (Fig. 1).

Cognition is "the mental action or process of *acquiring knowledge* and understanding through *thought*, *experience*, and the *senses*" [14].

"*Acquiring Knowledge* is a learning process that lasts a lifetime" [15].

Thought encompasses an "aim-oriented flow of ideas and associations that can lead to a reality-oriented conclusion" [16].

According to Wikipedia, the "*Experience* is the process through which conscious organisms perceive the world around them."

Likewise, the conscious and reasonable CASISX system can also perceive and cognize the world and itself by computing the aforementioned properties of awareness, consciousness, and cognition using smart modeling them by CB, and self-developing memory, self-changing models, and their systemically organized integrated functional processes, implemented in conditions of uncertainty and unknown in advance situations. These processes are supported by *auxiliary* systemic synchronous functional computing *subprocesses* of situational target *fuzzy control* of data [17], information [17], knowledge [6], models [8], processes [17], *fuzzy logic inference* [6], *making* of decisions [7], *reasoning* [7], *systemic think* [10], *planning* [7], and others.

The modern systems of computational Cognition, such as [1, 18, 19], and others, do not have the capabilities of CASISX, cannot self-aware, self-conscious, think independently, self-develop, and self-operate under uncertainty and in continuously time-changing situations.

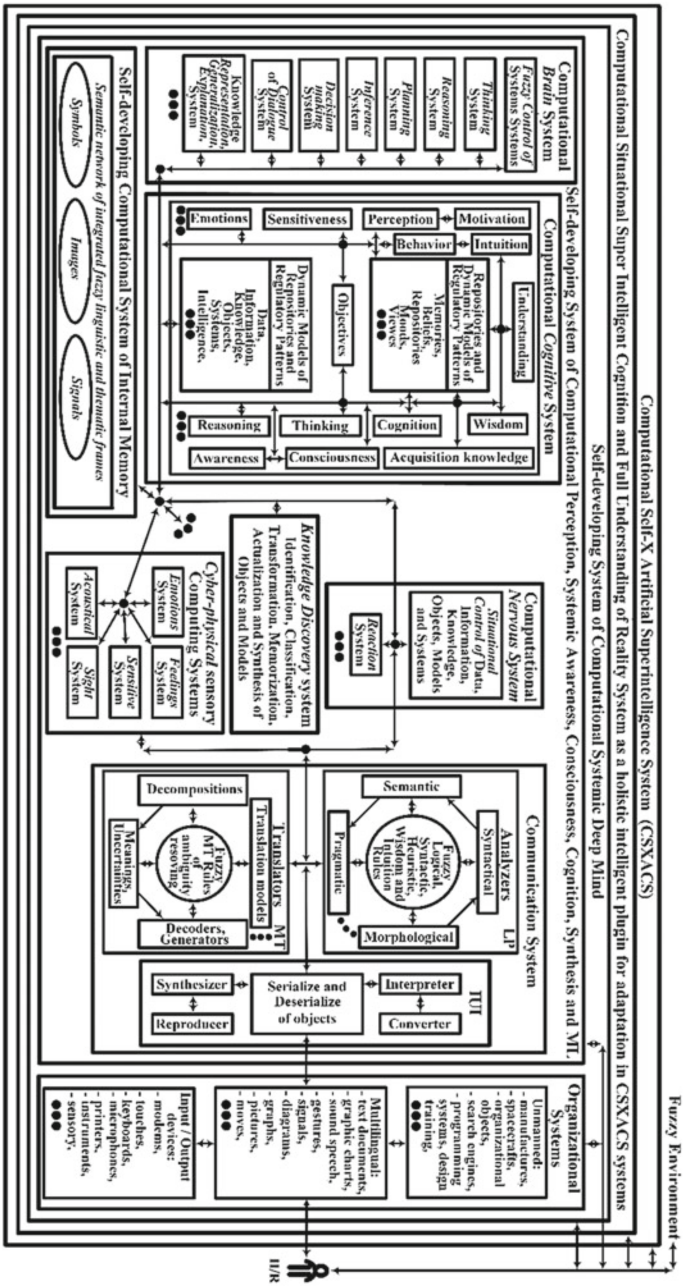


Fig. 1 The computer situational super intelligent cognition and full understanding system of the environment in conditions of its uncertainty

The offered paper examines the models, methodology, technology, system, and conception for implementation of one of the most important systemic psychological function of the CB—the *System of Computational Situational Super Intelligent Cognition and Understanding* (SCSSICUS) of the reality Under Uncertainty with applying the main aspects of Computational *Cognitive Science*, that includes the *Fuzzy Logic* [20, 21], *Situational Control* [22], *Linguistics*, *Cybernetics*, *Cognitive Psychology*, *Neural Science*, *Data Science*, *Computer Science*, *Systems Approach* at whole for the CASISX system.

1.2 Motivation, Novelty, and Contribution of the Study

The *motivation* in this paper lies in the need to make a breakthrough in the field of creating computational super-intelligent cognitive systems, that allow by themselves and independent functioning in conditions of uncertainty, taking into account the understanding of the current tasks and situations, for example, in the unmanned spacecraft, robotic systems, nuclear plants, and others, where is required high reliability, accuracy, adoption of extremely reasonable solutions, situational control, and others, what modern systems of NAI do not have.

The *novelty* and *contribution* of this study are in the development of unique (not statistical and non-probabilistic) *model*, *method*, *technology*, and the *concept* of creating *SCSSICUS*, functioning continuously, by *yourself* and *independently* (without *reprogramming* of them on basis of human intelligence) in *conditions* of Uncertainty and changes of the situations in the surrounding reality and of the objects in it, through of applying the System Approach, Fuzzy Logic, Situational Control, Linguistics, Cognitive Psychology, Neuroscience, Data Science, and Computer Science at whole, with the aim of introducing the *SCSSICUS* as a main computational and psychological component into the *CASISX* systems.

2 The Self-Developing Computational Systemic Cognitive Memory

2.1 The Model of Self-Developing Computational Systemic Cognitive Memory

The description of the Morphological, Semantic, Mental, and Subject Area Data, Information and Knowledge in the Cognitive Memory and its Self-developing Model are represented in [4], respectively, by Table 1 and Fig. 2. The cognitive self-developing model of cognitive memory is organized as a neural network of fuzzy sets of linguistic variables, and their values, that display the linguistic, subject area and psychological data, information, and knowledge in cognitive memory.

Table 1 The Linguistic, Thematic, and Mental structures of cognitive Data, Information and Knowledge in Systemic Cognitive Memory

	Domain model (P)					
	Language					
	Morphological model			Semantic model	Intelligent model	Psychological (mental) model
X (Corteges)	Lexis	Syntax	Grammar	Meanings Concepts	Knowledge	Mind Knowledge Knowledge Knowledge
R (Relationships)	Categories	Categories	Categories			
	Properties	Properties	Properties			
	Terms	Terms	Terms			
	Values	Values	Values			

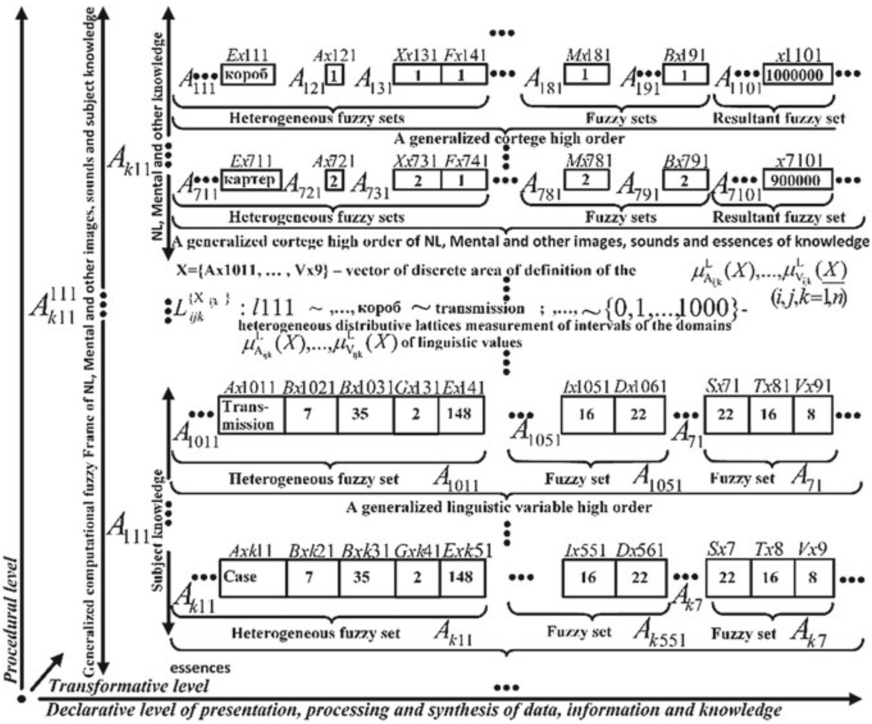


Fig. 2 The model of cognitive memory of computational intelligent cognition system

The computational cognitive system (Fig. 1) uses the model of cognitive memory (Fig. 2) for implementing the cognition process through CB's action orders via activating of the situational fuzzy control system, that is described in [4].

3 The Self-Developing Computational System Model of Situational Control of Mental Cognition

3.1 The Model of Situational Fuzzy Control of Mental Cognition

The model of situational fuzzy control of cognition (1) is described in [11] and includes, respectively, the computational cognition sub-model (2) and sub-model of situational control of the process of cognition (3) in conditions of uncertainty of the situations, objects, and states of the SCSSICUS in the environment.

$$\mu = < A_n^s, K_n^s, F_n^s, (S_i : Q_j \xrightarrow{x,u,w} Q_I : I_i), (\mu_R^Y(\mu_R^F(\mu_R^L(\mu_R^Z(\mu_R^D(\mu_R^M(\mu_R^T(\mu_R^H(\mu_R^C(\mu_R^O)))))))))) \tag{1}$$

$$\mathrm{M} = (\mathrm{A_n^s}, \mathrm{K_n^s}, \mathrm{F_n^s}, \mathrm{M_B^R}(\mathrm{X}), \mathrm{S}) \tag{2}$$

$$\mathrm{S} = (\mathrm{S_i} : \mathrm{Q_j} \Rightarrow \mathrm{Q_I} ; \mathrm{I_i}) \tag{3}$$

(4) is a goal function, oriented for implementation of cognition process, where

$$\mathrm{F_N^S} = \mathrm{u}(\mathrm{x}, \mathrm{w}) \tag{4}$$

x, w, and *u*, respectively, *definite, indefinite,* and *resultant* fuzzy controlled actions for performing the Computational Situational Systemic Cognition (*CSSC*) process, which brings the SCSSICUS from the *previous* state to a *new state* in the *arisen situation.*

4 The CSSC Modeling Under Uncertainty

4.1 The Model of CSSC Modeling Process

The CSSC *Model* displays the reasonable *psychological mechanism* of the CSSC *Modeling process.*

According to sub-models (2), (3), (4), and Fig. 1, the model (1) presents the process of the situational control of the CSSC, which is described in [11] by activation of CB's processes of inference, acquisition and actualization knowledge, planning and making decisions, reasoning, thinking, awareness, consciousness, and cognition.

Figure. 3 is described in [11] as the process of modeling of the CSSC (Fig. 3) using the mentioned models (1), (2), (3), goal function (4), cognitive measures of opportunities (5), cognitive mapping rules (6) at whole, via of activation of the mentioned above *CB's* processes (Fig. 1).

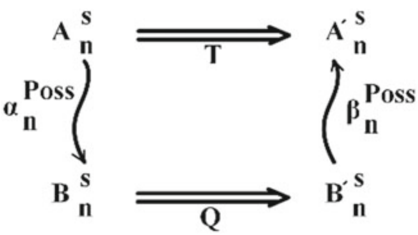


Fig. 3 Diagram of modeling of the CSSC

4.2 Modeling of the CSSC

The CSSC *Modeling* process, that is described in [11] by Fig. 3 is understood by us as a *cyclical* and *continuously* computational–mental *target action* of the CB, which uses the current situation, CSSC Model, cognitive memory system, and mind system (Fig. 1) at whole for the computational solution of cognitive psychological (mental) *tasks*, carried out by the *self-developing CB* through its: (a) cognitive psychological *memory*, (b) cognitive *model*, (c) computational psychological *properties* (thinking, awareness, consciousness, and others), and (d) making reasonable *solutions* in acquire *new* and use *existing* knowledge in order to intelligently *understand* the *new state* of *oneself* and the surrounding *reality* in conditions of *uncertainty* of these *states.*

The *rules* (5) are described in [11] and they are used for mapping the resultant values of fuzzy logic membership functions $\boldsymbol{\mu}_b^q$ using their $\boldsymbol{\mu}_a^q$ input values and *measures of opportunities* $\boldsymbol{\varphi}_l^q$ by *fuzzy matching* of fuzzy sets in the fuzzy *relation* R_k^X (Fig. 4a, b).

$$\mu_b^q(X) = \mu_a^q(X) \circ \varphi_l^q \tag{5}$$

Under the *fuzzy matching* of fuzzy sets, we mean the action, performed by procedures with the frame $\Phi_{L_{ijk}}^q : U_k^m \leftrightarrow V_k^l$ of the cognitive memory, which generate the *matching* of base *fuzzy sets* U_k^m and V_k^l in SCSSICUS. This matching of base fuzzy sets is implemented by using of compositional mapping *rules* $\mu_{B_{L_{ijk}}}^q(X) =$

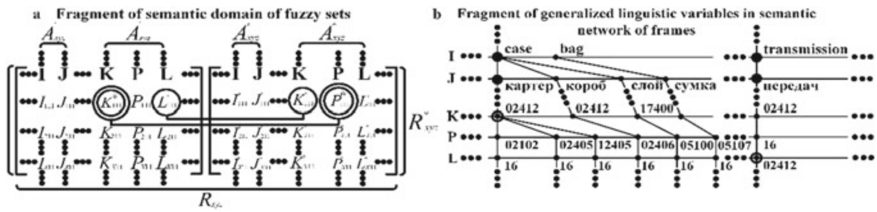


Fig. 4 **a** Human; **b** Machine levels of modeling of the CSSC

$\mu^q_{A_{L_{ijk}}}(X) \circ \Phi^q_{L_{ijk}}$ where: $\mu^q_{B_{L_{ijk}}}(X), \mu^q_{A_{L_{ijk}}}(X)$ are, respectively, the resultant and initial *membership functions* in the considered generalized fuzzy *relation* $R^{X_{ijk}}_{ijk}$, $X = \{x_{ijk}\}$ is a *vector* of the discrete domain of definition of resultant and initial membership functions, ∘- is a *sign* of the computational mapping, $\Phi^q_{L_{ijk}}$ is a fuzzy *matching* in the procedure of CSSC, $A_{L_{ijk}}$, $B_{L_{ijk}}$—are the heterogeneous multidimensional *fuzzy sets*, L_{ijk}—are the heterogeneous distributive *lattices* of measurement intervals of the domains of definitions of considered membership functions, *q*—determine the mentioned above *levels*, respectively, on *sublevels* of *modeling* of *RX*-codes, *USK*—Universal Semantic Code, and *SF*—Semantic Frames, *l, k, m = (1, n).*

4.3 *The Model of Computational Mapping Rules for the CSSC Modeling*

The computational cognitive mapping rules *model* (6) are presented, described in [11], and used for the CSSC Modeling.

$$\underbrace{\underset{L_X^R}{\tilde{M}_B^R(X)}}_{\text{New state of A}} = \bigvee \underbrace{\Big(\bigwedge \underset{L_X^R}{\tilde{M}_A^R(X)} \circ \text{Poss } a / a' \Big)}_{\text{Current state of A}} \tag{6}$$

In [11] is described the example (Fig. 4a, b) of modeling of CSSC for processing of texts of English–Russian languages in the task of removing their *lexical ambiguity* in the Tractor Engineering Subject Area.

The (7) represents the computational cognitive mapping rule in the *model* (6), which

$$\text{IF } ((K^{\bullet}_{xyz} = P^{\bullet}_{xyz}) \text{ AND } L^{\circ}_{xyz} = K^{\circ}_{xyz})) \text{ THEN } ((I_{xyz} = I^{\circ}_{xyz}) \text{ AND } J_{xyz} = J^{\circ}_{xyz})) \tag{7}$$

means: "***if*** two words are included in sentence in both languages and the code of *right compatibility* of word, located to the *left* of a group of two words in *another* language ***is equal*** to *conceptual* code of the word, located to the *right* of a group of two words in the *output* language ***and*** the code of *left compatibility* of word, located to the *right* of a group of two words in the *output* language is ***equal*** to *conceptual code* of word, located to the *left* of a group of two words in *input* language, ***then*** these *pairs* of words of *input* and *output* languages *correspond* and are *identical* each to other at the level of conceptual *disambiguation* in mapping process" [11].

5 Conclusion

The modern computing *ML* systems refer to statistical and expert *NAI,* and cannot think systemically, be conscious and cognize itself, and the surrounding reality by yourself (without using in them expert human intelligence) in the conditions of uncertainty and unexpectedly time-changing situations and states of SCSSICUS in the surrounding environment.

The *choice* of the *CASISX system* in this paper was made on the basis of objective studies of the state and testing of the capabilities of the modern NAI, based on the use of the ML methodology, where the *models* cannot be *retrained* on *their own* in conditions of uncertainty and changes in time of situations and objects in the surrounding reality. Since the functionality of the *CASISX* system is implemented based on self-awareness, self-consciousness, self-cognition, self-understanding, and other intelligent psychological features and properties of the CB—the capabilities of the system become like human intelligence.

For solving the mentioned problems of the NAI, this paper has been offered the model, method, technology, system, and conception of implementation of one of the most important systemic *psychological function* of the *CB*—the *System of Computational Situational Super Intelligent Cognition and Understanding of the reality Under Uncertainty* with applying the Fuzzy Logic, Situational Control, Linguistics, Cognitive Psychology, Data Science, Computer Science at whole, oriented on introduction this *Cognition* system in the *CASISX system* as a plug-in, as, for example, in Super intelligent unmanned robotic, spacecraft, and other computational reasonable systems.

References

1. Takuma O, Junichi T (2016) Development of self-cognition through imitation behavior. In: 7th annual international conference on biologically inspired cognitive architectures, vol 88. Elsevier, NY, USA pp 46–51
2. Khayut B, Fabri L, Avikhana M (2018) A self-developing computational system of full awareness and understanding of reality. In: ISAE-MAICS conference. Spokane, USA, pp 37–42
3. Khayut B, Fabri L, Avikhana M (2017) Modeling of computational perception of reality, situational awareness, cognition and machine learning under uncertainty. In: Intelligent systems conference. London, UK, pp. 331–339
4. Khayut B, Fabri L, Avikhana M (2020) The reasonable and conscious understanding system of reality under uncertainty. J Circuits Syst Signal Process 14:296–308
5. Khayut B, Fabri L, Avikhana M (2020) Toward general AI: consciousness computational modeling under uncertainty. In: 2nd International conference on mathematics and computers in science and engineering (MACISE 2020) Madrid, Spain, pp 90–97
6. Khayut B (1989) Modeling of fuzzy logic inference in decision-making system. In: Modeling systems, institute of mathematics of the moldavian academy of science, vol 110, pp 134–143
7. Khayut B, Fabri L, Avikhana M (2013) Modeling, planning, decision-making, and control in fuzzy environment. In: Advance Trends in Soft Computing, vol 312, Springer, USA, pp 137–143

8. Khayut B, Fabri L, Avikhana M (2014) Intelligent multi-agent fuzzy control system under uncertainty. J Comp Sci Inform Tech 4(18):369–380
9. Khayut B, Fabri L, Avikhana M (2014) Knowledge representation, reasoning and system thinking under uncertainty. In: 16th International conference on computer modeling and simulation, Cambridge, UK, pp 119–128
10. Khayut B, Fabri L, Avikhana M (2014) Modeling of intelligent systems thinking in complex adaptive systems. In: International Conference on complex adaptive systems, USA, pp 93–100
11. Khayut B, Fabri L, Avikhana M (2016) Modeling of computational systemic deep mind under uncertainty. In: 8th International conference on complex adaptive systems, USA, pp 253–258
12. Bostrom N (2014) Superintelligence: Paths, Dangers Strategies. Oxford University Press, UK
13. Wikipedia, Computational cognition and computational modeling
14. Site: Dictionary.com (2020) Cognition. Oxford University Press, Lexicon
15. Site: Bartleby.com (2020) Acquiring Knowledge Essay. Bartleby Research
16. Maric J (2005) Klinicka psihijatrija. Nasa kanjira
17. Khayut B, Pechersky U (1987) Situational data control. Deposited manuscript, Moscow, Viniti, p 29
18. Jiawei Z (2019) Cognitive functions of the brain: perception, attention and memory. IFM Lab Tutorial Series, vol 6
19. XioLan F, LianHong C, Ye L, Jia J, WenFeng C, Zhang Y, GuoZhen Z, YongJin L, ChangXu W (2014) A computational cognition model of perception, memory, and judgment. In: Science China information sciences, China, vol 57, pp 1–15
20. Zade L (1956) Fuzzy Sets. In: Information and control, vol 8, USA, pp 338–359
21. Zade L.: The Concept of a Linguistic Variable and its Application to Approximate Reasoning, Information Sciences, vol 14, pp. 141–164, USA (1995).
22. Pospelov D (1986) Situational control: theory and applications. Moscow, Science, p 288

State-of-the-Art of the Impact of HIV and Its Treatment on the Voice of PLHIV

Aliou Badji, Youssou Dieng, Ibrahima Diop, Papa Alioune Cisse, and Boubacar Diouf

Abstract Patients living with HIV (PLHIV) can be subjected to various conditions and complications due to, on the one hand, HIV infection and, on the other hand, the side effects of the drugs used for treatment. In this work, from a review of the literature, we studied the influence of this immunosuppressive disease and its therapy on the voice of HIV-positive individuals. This work was undertaken to understand the impact of HIV and drugs on voice parameters in order to assess the feasibility of identifying patients using voice biometrics and to propose solutions. From this study, it appeared to us that HIV itself as well as the associated opportunistic diseases and the drugs used for its treatment can cause vocal disorders in people who have contracted the virus. Solutions for the identification of PLHIV through their voice were also discussed in this work.

Keywords HIV · PLHIV · ASR · ARV · Voice

1 Introduction

The identification of PLHIV is perceived by those involved in the fight against HIV as one of the main issues to be resolved in order to improve monitoring. This is the reason

A. Badji (✉) · Y. Dieng · I. Diop · P. A. Cisse
LI3, Assane Seck University of Ziguinchor, Ziguinchor, Senegal
e-mail: A.BADJI3992@zig.univ.sn

Y. Dieng
e-mail: ydieng@univ-zig.sn

I. Diop
e-mail: ibrahima.diop@univ-zig.sn

P. A. Cisse
e-mail: papa-alioune.cisse@univ-zig.sn

B. Diouf
Enda-Santé Ziguinchor, Ziguinchor, Senegal
e-mail: boubacar.diouf@enda-sante.org

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_15

why Senegal has taken the option of strengthening its information system by deciding in PSN-2018 [1] to set up a unique identification code for individuals benefiting from monitoring services. In the same perspective and in order to contribute to the fight against the loss of sight of certain patients in health structures, a study on the identification of these by biometrics is being carried out. The type of biometrics in question is voice biometrics, which is based on automatic speaker recognition (ASR) technology. The latter is defined by Atal in [2], as "any decision-making process which uses some characteristics of the speech signal to determine whether a particular person is the author of a given utterance". With this method, the recognition (identification) of HIV-positive patients will be done by means of their voice. Thus the confusion between patients on ARV treatment and those who are not yet will be removed.

This article has three sections in addition to the introduction and conclusion. In the first section, we briefly present the different sources of voice variability in ASR. In the second section, we will talk about the impact of the immunosuppressive virus and medications on the voice of HIV-positive individuals. The third section is devoted to discussion; we will talk about the possible solutions to be put in place for the platform for the identification of PLHIV using their voice.

2 The Variability of the Voice in RAL

ASR is a technology that consists of extracting the vocal characteristics of the speech signal specific to each individual and creating a "speaker model" which will be then used to identify it. The expression "speaker model" is, by analogy to the fingerprint, what one might call voiceprint. However, because of the physiological characteristics that the voice possesses and which are linked to the anatomy of the phonatory system and on which are grafted behavioral characteristics such as speech, accents, emotions, humor, etc., some authors speak of voice signature rather than voiceprint. This property of the voice to change depending on certain factors has a strong impact on the reliability, efficiency and performance of ASR systems.

2.1 Technology-Related Variability

The sources of voice variations related to technology are a set of factors external to the voice signal and which have negative effects on its quality and integrity. More precisely, these elements are used for the acquisition and/or transmission of the signal as well as the environment in which it was obtained. Among these elements we can cite acquisition and/or transmission devices: transmission channel, microphone integrated into one (cell, mobile or landline telephone), microphone; the environment: acoustic room, office, market, airport, beach, etc.

These elements generally generate components or phenomena (background noise, aliasing, reverberation, rapid temporal variation, etc.) which mix with the signal and strongly affect its fullness.

2.2 Conversation-Related Variability

This type of voice variability results from a voice interaction between two or more people or between a person and a machine and the context. The same person speaks differently depending on his interlocutors; it can be much more comfortable talking with a familiar person in public or with a stranger; the context: she can speak differently, she is reading, talking on the phone, attend a meeting, to argue, to interact with a machine; she can raise or lower her voice depending on whether the place is quiet or noisy; the spoken language or dialect: a person's accent is better understood with their mother tongue than with another language, etc.

2.3 Speaker-Related Variability

By pronouncing a given word, the same person does not produce the same sound in an identical way also by two different people. For illustration, in the figures below, we have two audio recordings of the same sound under the same conditions from the same person (Fig. 1) and two other recordings of the same sound under the same conditions from two different people (Fig. 2). Looking at and analyzing these figures, we see that there are always differences in the amplitude and duration of the signals generated. The human voice, therefore, varies from one person to another but also for the same person. The variations of the voice for the same speaker are called

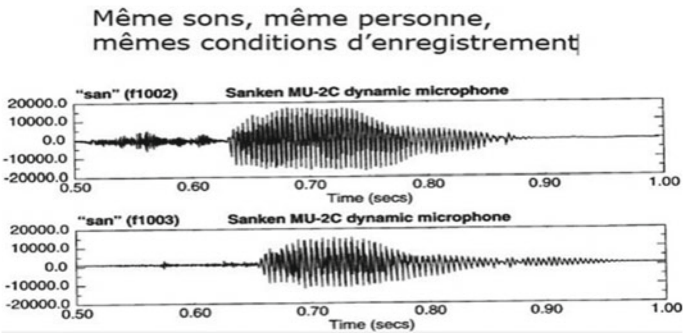


Fig. 1 Intra-speaker variability (*Source* http://outilsrecherche.over-blog.com/pages/Notes_311_Decodage_du_Signal_de_la_Parole-3082466.html)

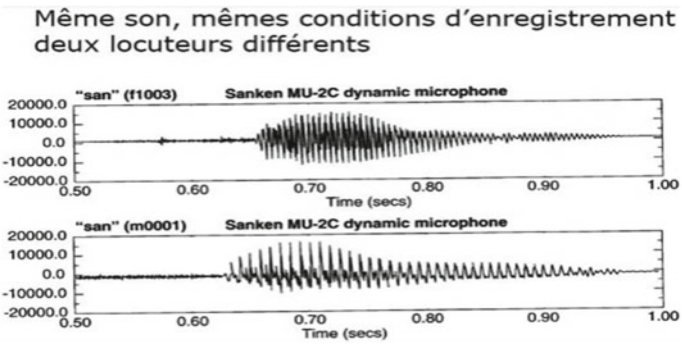


Fig. 2 Interlocutor variability (*Source* http://outilsrecherche.over-blog.com/pages/Notes_311_Decodage_du_Signal_de_la_Parole-3082466.html)

variability intra-speaker, and in the case of different speakers, we speak of variability interlocutor.

3 Does HIV and its Treatment Have an Impact on the Voice?

The difficulty in coming into contact with PLHIV, in collecting their audio recordings or in accessing data concerning them; the lack of adequate tools for analyzing the voice and diagnosing its disorders in a more concise manner; and the embarrassment to carry out the study and the experiments directly on the subjects concerned led us to follow an approach to carry out this work. For this, we have followed an approach consisting of studying the influence of HIV and its treatment on the voice of seropositive people in a general and non-specific way by making a state-of-the-art of the previous relative works reported in the literature. Thus, in this section, we wish, from a review of the medical literature and the information provided by the review prescriber on vocal attacks by certain drugs, to understand the impact of HIV and its therapy on the voice of PLHIV.

3.1 Difference Voice, Speech as Well as Their Disorders

First of all, we want to remember that voice and speech, often considered by some people to be the same thing, are in fact different but have the same channel. In our daily life, we talk, sing, shout, whisper, cough, etc. Each of these cases translates an expression of the voice into a different form and depends on our need to communicate, our state, our emotions, the context, etc. The voice can be defined as the set of sounds produced by the vocal cords in vibratory movement under the influence of air coming from the lungs. As for speech, it is the transformation (articulation, modulation) of these sounds at the level of the vocal tract to give intelligible sounds. The vocal tract is made, among other things, of the oral, nasal and pharyngeal cavities which act as resonators and modulators. More comprehensively, "*voice is the set of sounds produced orally and speech the set of voice sounds used in a language*" (Authôt.com). From these definitions, we realize that there is a distinction not only in the definition but also between the organs involved in the production of these two elements.

3.2 Neurological Disorders and Voice Disorders

Neurological disorders are diseases of the central or peripheral nervous system. They affect the brain, spinal cord, cranial nerves, peripheral nerves, nerve roots, vegetative nervous system, neuromuscular junction and muscles (WHO). In the summary of their article [4], JL St Guily and B. Roubeau mention the following: "*Motor process which underlies speech, phonation, like any motor function, results from the coordination of the activity of muscles located at different levels of the body….*". This, therefore, means that the coordination of the organs involved in the production of speech is under the control of the central and peripheral nervous systems. Therefore, an affection of these will have repercussions on the phonation as well as on the resulting speech.

Studies on dysarthria and, more particularly, those including dysphonia exist and are numerous and diverse. They show, for the most part, that neurological damage does not only cause communication problems but also can lead to disturbances in the production of the voice. However, what does this have to do with HIV?

3.3 HIV and Voice Disorders in PLHIV

The attack of the central and peripheral nervous systems by HIV is not a trivial matter in the scientific community. The WHO itself already emphasizes that bacterial infections such as tuberculosis, viral infections such as HIV and parasitic infections such as malaria can affect the nervous system. However, we know that neurological disorders are due to damage to the nervous systems, so from what precedes, we can immediately assume that HIV can have an influence on the voice since it affects the central nervous systems and peripheral. However, this is neither sufficient nor convincing and even less conclusive. Numerous studies have shown the influence of HIV infection on the central, peripheral and mixed nervous systems. A significant part of them highlighted the relationship between immunosuppressive disease and speech and voice disorders. M. Mathew and J. Bhate state in the summary of their preliminary study [5] on voice disorders in seropositive individuals that by infecting cells of the brain, HIV damages the central and peripheral nervous systems and leads to the consequence of communication disorders from the early stages of infection. Also in their work, their study of eight individuals living with HIV revealed, after subjective analysis by a speech therapist and an objective analysis of acoustic characteristics, that vocal parameters such as pitch, intensity, quality of voice, the maximum duration of phonation, etc. disturbed almost all participants even in the absence of direct pathology to the phonation mechanism. Since this was a preliminary study, they say it is too early to conclude that there is a correlation between vocal deviations and immune conditions.

In another study [6] by the same authors on the same participants, tests were performed to assess aphasia, dysarthria, articulation, fluency of speech and voice disturbances. The results obtained illustrate, with some variation in patients, the presence of problems with oral motor function (restricted tongue movements as well as tremor and dyskinetic movements of the lips), voice disorders (low intensity, quality of speech) disturbed voice, reduction of the maximum duration of phonation, etc.) and dysarthria (affection of the parameters of the reflex, respiration, larynx, tongue, etc.); fluidity and articulation being intact. Another study [7] with slightly more participants (15 seropositive patients) by the same authors also shows a significant disturbance of speech and laryngeal parameters followed by respiratory and reflex functions.

From the work of these authors, we clearly feel that the link between HIV and voice disorders is very real. On their three works, we always note a disturbance of vocal and laryngeal parameters which, however, intervene in the phonation mechanism. What is even more interesting are the results of their first preliminary study which reveal an affection of the vocal characteristics in the absence even of direct pathology to the process of phonation. This suggests a vocal dysfunction of neurological origin due to HIV.

This information further strengthens the connection between HIV and voice disorders. Laryngeal opportunistic infections, themselves, can cause problems with the larynx and this will consequently lead to dysfunctions in the phonatory process. Everything that has just been summarized from the work of these authors shows that HIV can have an influence on the voice of seropositive patients by causing lesions in the central and peripheral nervous systems which cause neurological disorders which in turn affect certain levels (respiration, larynx, etc.) involved in the voice production mechanism. The simple fact that the results of their work still reveal alterations in vocal acoustic parameters as well as in the larynx is one of the proofs that testifies to the existing relationship between HIV and vocal disorders.

HIV can also, through associated opportunistic infections, cause voice disorders. This is because patients living with HIV may present with various problems such as oral, ear and pharyngeal problems [9]. These conditions include oral and pharyngeal candidiasis, aphthous ulcers, Kaposi's sarcoma, hairy leukoplakia and herpes simplex and can, along with neoplasms, cause dysarthria, dysphonia, hoarseness, etc. [9]. There are also respiratory infections such as pneumocystis carinii pneumonia, tuberculosis, and various bacterial, viral and fungal infections which are among the diseases most associated with HIV [9]. For illustration, many studies report cases of hoarseness of the voice or dysphonia in individuals carrying the virus and presenting opportunistic infections such as Kaposi's sarcoma [10, 11], cytomegalovirus infection of the laryngeal nerve or larynx [12, 13], disseminated histoplasmosis [14], infection with *Penicillium marneffei* [15, 16], thyroid infection with *Pneumocystis jirovecii* [17], laryngeal cryptococcosis [18], chronic laryngitis herpes simplex [19], invasive aspergillosis [20], laryngeal leishmaniasis [21], etc.

In summary, after having unveiled the link between HIV and voice disorders in infected individuals, we infer that HIV and associated opportunistic diseases may have an impact on the voice of PLHIV.

3.4 Medicines and Voice Disorders in PLHIV

Certain drugs induce damage to the vocal cords by various mechanisms such as the deposition of substances on the mucous membrane, dryness of the mucous membranes and decrease in the lubrication of the vocal cords, formation of edemas, necrosis, etc., paralysis of the vocal cords, and so on. These vocal cord attacks appear within a period varying from a few hours to a few months and their recovery occurs after stopping the drug, and this also within a period varying between a few hours to several months (review prescriber). This, therefore, means that, throughout the period covering the taking of the drug, the voice of the people concerned will be affected.

Medicines used in the treatment and monitoring of patients living with HIV can be grouped into two classes: antiretrovirals and medicines for the treatment of opportunistic infections or other related diseases. Antiretrovirals are drugs that reduce the HIV viral load or even make it undetectable, strengthen the immune system of the infected person and prevent the AIDS phase from developing. When the HIV infection reaches the AIDS phase or the patient has other conditions, antiretroviral therapy alone is no longer sufficient. A combination of the two classes of drugs is therefore needed to treat both HIV and related diseases. We do not have enough information and articles on the relationship between antiretrovirals and voice disorders.

Based on the information provided by the review, we also infer that the drugs used for the treatment of certain diseases present in individuals who have contracted HIV may also have influences on the voice of PLHIV.

4 Discussion

One of the weak points of automatic speaker recognition technology is the variability of the voice. This strongly impacts the reliability, efficiency and performance of ASR systems. Based on the foregoing (section above), identifying PLHIV through their voice may at first appear less relevant or even ineffective. However, this is not the case because the change in an individual's voice is inherent in this technology. Therefore, this variation should always be taken into account when implementing ASR. Thus, to enhance the robustness, credibility and security of an ASR system, it is often recommended to use it with another identification/authentication factor or to update the speaker models.

In this section, we discuss the different possible solutions for the system based on the recommended methods for enhancing the security, reliability and accuracy of authentication/identification systems based on voice biometrics.

4.1 The Updated Models

This solution consists of periodically updating the models of patients already enrolled in the database. At each update, different options are possible: either delete all the old models and create new ones with the new and old training data, or readjust all the models with the new training data, or re-estimate all models with new training data only. The first option is the best because it combines current enrollment data with old training data. This will give a voice of the same patient at different times and circumstances and consequently will allow learning of the change of the voice; that is to say, to encourage the system to "learn" the variation of the voice of a person. This will make it easier to recognize a patient even if the patient's voice is affected. This is perfectly suited as an outcome for the identification system for PLHIV given the context in which it will be used. Indeed, during the whole period covering the treatment, the patient's voice may remain affected because even if the patient does not come across drugs that can damage the vocal cords, HIV itself as well as the associated opportunistic diseases can cause the impact. With this option of updating patient models, the system will be much more robust as it will learn how patients' voices vary.

However, this method can come up against an obstacle: the problem that can arise from the refusal of patients to cooperate with each update. In fact, with each modification of the models, the patients have to pronounce a sequence of words to first be identified, and then doing the same thing to be enrolled again may annoy some; as a result, they may refuse. So for this solution to work well, good patient cooperation is needed. For that, the time between two updates must be well chosen. As the duration of an appointment is specific to each person and therefore varies from one patient to another, this update can be done at each appointment or after every two, three or four appointments for patients whose time interval between these is short. Also to collect a maximum of audio samples, the recordings during the recognition phase must be added to the training data.

4.2 Double Identification

This solution consists of double identification of the patients. It is particularly recommended for transaction or payment systems using voice biometrics to enhance reliability and security. Even if safety is not essential for the PLHIV identification platform, we will offer it as a solution because it will help strengthen the distinction between patients on antiretroviral treatment and those who have not yet been treated. For this method, the identification process is done by first identifying the patient using their voice. Then if he is recognized by the system, he will be asked to provide his telephone number or that of his national identity card previously registered during the enrollment phase. If it is recognized again then it is taken care of; otherwise, he is not on antiretroviral therapy; therefore, the system would have mistakenly recognized him. Likewise, if the patient is rejected by the system during the voice recognition phase and accepted during the second identification, he is then considered to have been taken care of and that the system would have wrongly refused; otherwise he is not on antiretroviral therapy. With this solution, if a patient is wrongly accepted or rejected during the speech recognition phase, the second factor will allow it to be detected and corrected.

The problem with this method is that voice identification will lose its meaning regardless of a patient's voice condition because patient recognition relies only on the second factor. However, it remains relevant insofar as it will strengthen the robustness and reliability of the identification of PLHIV through their voice.

5 Conclusion

HIV infection can affect the voice of HIV-positive patients by causing damage to their central and peripheral nervous systems. Similarly, the associated opportunistic diseases and their treatment can also affect the voice of patients. In view of this, possible solutions were discussed in the previous section for identifying PLHIV using their voice in order to deal with recognition problems that may arise from vocal disorders. These solutions consist of either updating the patient models or opting for a double identification. For the first solution, the best option for updating is to re-estimate all models by combining the new and old training data to allow the system to learn how the voice varies.

Since HIV infection can cause voice and communication disturbances in HIV-positive patients, we wonder whether it is not possible to use voice and speech parameters to diagnose it. In our next work, we would like to provide an answer to this question.

References

1. RAPPORT-Annuel-CNLS-2018–1.pdf. Consulté le: oct. 20, 2020. [En ligne]. Disponible sur: https://www.cnls-senegal.org/wp-content/uploads/2019/07/RAPPORT-Annuel-CNLS-2018-1.pdf
2. Atal BS (1976) Automatic recognition of speakers from their voices. Proc IEEE 64(4):460–475, avr. 1976. https://doi.org/10.1109/PROC.1976.10155
3. Hansen JHL, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag. https://doi.org/10.1109/MSP.2015.2462851
4. St Guily JL, Roubeau B (1994) Voies nerveuses et physiologie de la phonation. Encycl Méd-Chir Oto-Rhino-Laryngol
5. Mathew M, Bhat J (2007) Voice disorders in HIV-infected individuals: a preliminary study. Int J STD AIDS 18(11):732–735. https://doi.org/10.1258/095646207782212306
6. Mathew MM, Bhat JS (2008) Profile of communication disorders in HIV-infected individuals: a preliminary study. J Int Assoc Physicians AIDS Care 7(5):223–227. https://doi.org/10.1177/1545109708320682
7. Mathew MM, Bhat SJ Profile of speech impairments in HIV- infected individuals: a preliminary study, p 5
8. Verdolini K, Rosen C, Branski RC (2006) Classification manual for voice disorders-I, special interest division 3, voice and voice disorders. Am Speech-Lang Hear Div
9. McCabe P, Sheard C, CodeC (2009) What do speech pathologists know about HIV? Int J Speech Lang Pathol 1:9–18. https://doi.org/10.3109/14417049909167149
10. Watson JR, Granoff D, Sataloff RT (2004) Dysphonia due to Kaposi's sarcoma as the presenting symptom of human immunodeficiency virus. J Voice Off J Voice Found 18(3):398–402. https://doi.org/10.1016/j.jvoice.2004.04.007
11. Kadar SQMA, Nah AAS, Yunus MRM, Vivekanandan A, Ismail F (2015) Kaposi sarcoma of the tonsils and pulmonary tuberculosis secondary to HIV infection. Int Med J 22(6)
12. Small PM, McPhaul LW, Sooy CD, Wofsy CB, Jacobson MA (1989) Cytomegalovirus infection of the laryngeal nerve presenting as hoarseness in patients with acquired immunodeficiency syndrome. Am J Med 86(1):108–110
13. Levine PA, Marelli RA, Biddinger PW, Gluckman JL (1998) Cytomegalovirus infection of the larynx in the acquired immunodeficiency syndrome. Otolaryngol Neck Surg 106(3):296–301
14. Sanmani L, Randall CJ, Palfrey J, Rowen D (2011) Hoarseness of voice in an AIDS patient: a rare presenting feature of disseminated histoplasmosis. Int J STD AIDS 22(2):115–116
15. Devi SB, Chanu KJ, Singh KB (2017) Hoarseness due to penicillium marneffei infection-an unusual presentation
16. Nong D, Li J, Nong H, Xu Z, Liu X (2007) Pharyngeal and laryngeal infection of penicillium marneffei in patients with acquired immunodeficiency syndrome. Chin J Mycol 4:3
17. Zavascki AP, Maia AL, Goldani LZ (2007) Pneumocystis jiroveci thyroiditis: report of 15 cases in the literature. Mycoses 50(6):443–446
18. Zeglaoui I, Belcadhi M, Mani R, Sriha B, Bouzouita K (2009) Laryngeal cryptococcosis revealing AIDS: a case report. Rev Laryngol Otol Rhinol 130(4–5):307–311
19. Yeh V, Goldstein NS, Hopp ML, Meyer RD (1994) Herpes simplex chronic laryngitis and vocal cord lesions in a patient with acquired immunodeficiency syndrome. Ann Otol Rhinol Laryngol 103(9):726–728
20. Florent M et al (2001) Dysphonia and dysphagia as primary manifestations of invasive aspergillosis. Eur J Clin Microbiol Infect Dis 20(6):441
21. Canovas DL, Carbonell J, Torres J, Altés J, Buades J (1994) Laryngeal leishmaniasis as initial opportunistic disease in HIV infection. J Laryngol Otol 108(12):1089–1092

Specifics of Digitalization Strategies for Risk Management Systems as a Transformation Factor of Russian Companies' Management Systems in the Context of the COVID-19 Pandemic

Mikhail Vladimirovich Khachaturyan and Evgeniia Valeryevna Klicheva

Abstract In the modern environment, the digitalization of management mechanisms is becoming an effective tool for transforming the companies' organizational systems. A particularly fast pace of introducing digital technology is translated into creating and developing risk management systems in the context of the coronavirus pandemic, despite the related crisis that had a serious impact on both the Russian and global economies. It is important to note that, for most company owners and managers, the crisis became a signal of the need for management systems to become more transparent and efficient. Another important prerequisite for transformation is the trend to change the processes of interaction at both the owners-to-managers and company-to-customers and counterparties level. In the context of the pandemic, it became clear that the digital transformation of risk management systems would improve the operational quality of organizational structures and the quality of interaction with customers, thus increasing their satisfaction and loyalty.

Keywords Specifics · Strategies · Digitalization · Management · Risks · Transformation · Russian companies

1 Introduction

In the development of digital management systems, company owners and managers implement a sequence of steps related to the transformation of infrastructure, service provision, business process reengineering, data management, security, customer

M. V. Khachaturyan (✉)
Bauman Moscow State Technical University, Moscow, Russia
e-mail: khachaturyan@bmstu.ru

E. V. Klicheva
Plekhanov Russian University of Economics, Moscow, Russia
e-mail: evklicheva@mail.ru

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_16

interaction, and HR development. Each transformation step poses complex challenges for owners and managers and is associated with new types of risks; therefore, the ability of the company's risk management system to cope with such challenges efficiently determines the efficiency of its organizational structure.

The main element of increasing the efficiency of interaction between the company's owners and managers to identify, analyze, and assess new types of risks arising in the context of the economic crisis caused by the corona-virus pandemic have become information and communication technology or digital technology. It is its introduction both in management systems in general and risk management systems in the quarantine conditions that has allowed many Russian companies to develop their ecosystems, by bringing together employees, partners, and consumers into an integrated space. On the other hand, the logic of the company owners' response to the economic crisis development confirmed the authors' assumption that the process of implementing electronic and digital risk management systems in most organizational structures of Russian companies are fragmented in nature, which in the contemporary context, reduces the effectiveness of satisfying the customers' needs. This situation requires an integrated approach to introducing digital technology for risk management in the process of developing and implementing strategies by Russian companies. According to the authors, one of the areas for solving this problem may be expanded implementation of digital systems for risk management and analysis in the service sector as well as expanding the list of services provided by both the public and private entities in electronic format.

It is obvious that, before the pandemic, the process of introducing digital and information technology into organizational structures and risk management systems was associated with many difficulties and risks. According to the authors, the key problem, that has retained its relevance in the context of the pandemic, is the technological and economic complexity of digital technology. Another important problem is the need to provide access to data obtained while implementing digital technology for managing the risk for a significant number of stakeholders involved. Therefore, the need to implement effective data visualization and management tools is of importance. It is also important to note that, in the con pandemic conditions, it became obvious that the key feature of risk management systems and tools based on digital technology for strategic decision-making should be their easy implementation and application efficiency. The key task of these systems should be expanding the owner's role in the processes of monitoring and evaluating the company's management system effectiveness and the external environment of both the company and the owner's field of activities [1–3].

The logic of the economic crisis caused by the pandemic confirmed the authors' assumption that despite a significant expansion of the risk range that can be monitored by digital technology, neither Russian nor foreign companies have developed generally accepted methods and frameworks for conceptual building strategies for managing such risks.

2 Role and Significance of Digital Technology in Strategic Management and Risk Management Systems of Russian Companies in the Context of Pandemic

In this regard, the solution analyzed in this paper for implementing the structure of a risk management system based on digital technology allows for a more clear assessment of the threats and risks the activities of Russian companies are prone to. It is worth noting that the use of the structure proposed by the authors also allows expanding the company owners' and managers' ability to monitor the external environment factors acting for the company and the scope of the owner's activities that affect the company's operations. Obviously, this aspect largely facilitates the planning and visualization of management decisions and contributes to identifying and eliminating the smallest "bottlenecks" in the company management systems. The above feature is a favorable distinction of the suggested structure as compared to the existing counterparts characterized by significant complexity and details that reduce the usability and value of such systems for owners and managers. Another important advantage of the suggested structure of the risk management system is that it makes it easier to monitor implementing key aspects of the company's strategy.

In the context of the economic crisis caused by the pandemic, it became clear that digital technology has a long life cycle at both implementation and operation stages, which makes it possible to expand the scope of tasks solved by it. Therewith, the structure created based on digital technology shall feature transformation flexibility, which facilitates and accelerates the company's response to changes in the external environment and the owner's field of activities. On the other hand, when developing a risk management system, it shall be borne in mind that every project implemented by a company is subject to many technical and technological iterations. However, the risk management system itself shall be resilient to such changes. In the context of the pandemic, it became clear that the risk management system should be a link between owners, managers, and executors, thereby reducing the gap between the estimated and actual results.

In its definition of digital technology in company management, the Organization for Economic Cooperation and Development (OECD) emphasizes that digital technology is a tool for using information and communication technology to provide the owner, the company management system, and other stakeholders with more convenient access to data on the company's standing and the risks that affect the activities of both the company itself and its owner and stakeholders as well as their ability to directly participate in the processes of managing such risks [4].

The international literature on management provides another definition of digital technology in company management as a social and technical system consisting of people, technology, organizational structures, and processes [5].

According to another definition that is common in global management science, digital technology in company management is a continuous process of optimizing services and management—by transforming internal and external relations using technology, the Internet, and new information communication means [6].

The above definitions emphasize the role of digital technology in company management in general and in risk management in particular to transform the relationship between the owner, managers, and stakeholders, which allows creating the company's ecosystem with the ability to self-develop through continuous optimization.

In this regard, it is natural to conclude on the complex nature of the interaction and integration of owners and managers in the processes of introducing digital technology into company risk management systems, and the risk management system structure suggested by the authors is intended to simplify such systems. This paper considers three groups of factors on which the author's structure of the risk management system is focused: people, processes, and technology. In the context of the pandemic, it became clear that integrating the authors' risk management system structure into the development and implementation of the company's strategy would serve as a link between these three components in ensuring the effectiveness of the management system.

3 Research Method

The study presented in this paper is a qualitative analysis of the practice of introducing digital technology into the Russian companies' management systems and an analysis of domestic and foreign management literature. In the analysis, we studied the digitalization strategies of the management systems of Russian corporations. This allowed for the understanding of which strategies were more successful in terms of the company owners' and management systems' response to the risks associated with the pandemic and the economic crisis that has followed.

An important task in preparing the paper was obtaining enough data regarding the content and implementation mechanism of management strategies in general and the Russian companies' risk management in the pandemic. On the one hand, a very large sample selection could complicate researching the pandemic. On the other hand, an insufficient amount or variability of data could result in one-sided conclusions and recommendations. Therefore, as the main research method, we chose that of collecting and analyzing qualitative data based on open-source analysis and comparing this data with the theoretical principles as outlined in the Russian and foreign literature on management. It was the comparison of the theory and practice of Russian companies' operations that made it possible to form well-grounded conclusions and recommendations.

4 Analysis of the Content of Processes of Introducing Digital Technology in Russian Companies' Strategic Management and Risk Management Systems in the Context of the Pandemic

The analysis of the Russian and foreign management literature allowed the authors to put forward the assumption that company owners, managers, and other stakeholders involved in introducing digital technology into strategic management systems in general and risk management systems in particular need simpler and more effective tools for visualizing both the introduction and implementation processes as such, as well as assessing the functional effectiveness of such systems. This prompted the authors to seek analogies for a comparative analysis of the practice of solving problems like those discussed in this study.

During the research, various analogies were examined, and ultimately, the authors concluded that the engine operation process might be the most suitable of the entire range. In its functioning, any engine driving a certain object shall collide with opposing forces acting on this object; these forces are called the resistance that the object driven by such an engine shall overcome with the appropriate thrust to achieve its goal.

Based on the above analogy, the authors could build up a concept that reflects the processes of introducing digital technology in strategic management systems in general and risk management systems and overcoming the associated difficulties. This is a model of digital transformation of strategic management systems in general and risk management systems.

As the basis for implementing this model, the authors propose to use mechanisms of control and audit of business processes, people, and technology which form the basis of any company's operations, by supplementing them with means of objective owner monitoring and communication between the company's owners and managers.

Based on the results of assessing the functional effectiveness of Russian companies' strategic management systems in the context of quarantine measures and the economic crisis caused by the pandemic, the authors concluded that the use of such digital visualization of strategic management and risk management would reduce the severity of the crisis phenomena for the owners of many companies. In addition, introducing such digitalization patterns for risk management systems facilitates the company's management functioning. In most cases, it is easier for managers to deal with graphical presentations than with huge tables, long texts, etc.

Therewith, introducing digital technology, which has proven its effectiveness in the context of the pandemic, into strategic management and risk management systems is fraught with many difficulties and much resistance.

The digitalization of strategic management business processes obviously conflicts with traditional management mechanisms, breaking down the boundaries between services and departments within one company and often between different companies, and forming a new corporate ecosystem. It is obvious that the current differences in legal, regulatory, and administrative mechanisms that exist both in individual departments of the same company and in different companies grouped into a single ecosystem can impede information and service flow through new network management channels.

Another important factor in the digitalization process is effective coordination between services and departments, as the responsibility for the management system operation is often fragmented and distributed between several management levels and the owner.

The costs of the development, implementation, and maintenance of management system digitalization (e.g., costs of software, hardware, and staff training) can be a significant factor impeding these processes. Another limiting factor is the issues of measuring the ratio of the costs and the benefits from management system digitalization. Despite the fact that some benefits can be seen in clear measurable parameters (such as the number of staff and reduction of overhead costs), many of them cannot be identified with confidence in the same way, since they are qualitative in nature and, therefore intangible, and their future translation is not predictable (e.g., improving the quality of service, new services, responding to customer needs, or saving the costs associated with using non-digital tools).

It is important to note that the presence or absence of trust in implementing innovative approaches has always been an important factor in developing management systems. Therefore, the issue of trust in digital technology is one of the most important elements of its implementation and performance assessment as part of strategic management and risk management systems. These complications are due to the challenges of ensuring the security of data of individuals and legal entities, the company's customers, and problems associated with possible theft of personal data. To help overcome trust issues, we need to develop mechanisms providing broad confidence, to protect the company's customers from unauthorized electronic disclosure of their personal information, including the transfer of such data both between company's departments and companies that are part of a single ecosystem or between a company and government authorities.

It is obvious that the scheme proposed in this paper for implementing digital technology in Russian companies' strategic management and risk management systems is an ecosystem that connects people, processes, and technology. These three components work in close coordination when implementing any project within the company. To achieve the overall goals and certain performance criteria, a common practice is to establish an independent supervisory board that monitors and reviews the progress of each project—on a regular basis. These comments are brought to the project managers' attention with the purpose of taking specific control measures to make corrections and adjustments. It is establishing such a supervisory body that brings together the company's owners and managers that will solve the problem of integrating owners into the management system and expanding the range of tools available to them to identify, analyze, assess, and manage risks associated with both the company's operations and the scope of their activities that are not directly related to the company.

5 Conclusion

The structure of implementing digital technology in Russian companies' strategic management and risk management systems, as presented in this paper, is a tool that describes projects for the digitalization of management systems in a technologically neutral and abstract manner, based on a simple and understandable analogy. Such a presentation allows company owners and persons who make strategic decisions to see and understand the problems faced by the management system digitalization as well as to ensure a qualitative choice of forms and methods of overcoming such problems.

Besides, this can reduce significantly the gap between the company's policy-makers and executors, as the general presentation of projects provides greater transparency and clarity of decisions made.

References

1. Lee O, Baby D (2013) Managing dynamic risk in global IT projects: agile risk management using the principles of service-oriented architecture. Int J Inf Technol Decis Mak 12(6):1121–1150
2. Bradley J, Loucks J, McCaulay J, Noronha A, Wade M (2015) Digital vortex—how digital disruption is redefining industries. Global Centre for Digital Business Transformation, http://www.cisco.com/c/dam/en/us/solutions/collateral/industry-solutions/digitalvortex-report.pdf
3. Hussain Q, Chang E, Hussain F, Dillon T (2007) Ascertaining risk in financial terms in digital business ecosystem environments. In: Proceedings of the 2007 inaugural IEEE international conference on digital ecosystems and technologies
4. UN E-Government Survey–2008 (2008) From E-Government to Connected Governance, UNDESA
5. Anttiroiko A-V (2008) Electronic government: concepts, methodologies tools, and applications. Univ Tampere, Finland
6. Gartner Group (2000) Key Issues in E-Government Strategy and Management. Research Notes, Key Issues, 23 May 2000

Application of Adjusted Differential Evolution in Optimal Sensor Placement for Interior Coverage

Adis Alihodzic, Damir Hasanspahic, Eva Tuba, and Milan Tuba

Abstract It is well known that determining visual sensors in 2D space can be often modeled as an Art Gallery problem. Tasks such as surveillance dictate the coverage of the interior of a non-convex polygon with the optimal number of sensors. The optimal sensor placement is a difficult combinatorial optimization problem, and it can be formulated as seeking the smallest number of sensors obliged to cover every point in a heterogeneous setting. In this article, we propose a suboptimal deterministic algorithm, as well as an adapted differential evolution algorithm for tackling sensor placement. Both versions of novel algorithms have been implemented and tested over hundreds of random polygons. According to the outcomes presented in the experimental analysis, it can be noticed that the approach based on differential evolution beats the deterministic technique as well as other stochastic optimization algorithms for practically all instances.

Keywords Computational geometry · Swarm intelligence · Visibility · Sensor placement · Computer vision · Differential evolution algorithm

A. Alihodzic (✉) · D. Hasanspahic
University of Sarajevo, Sarajevo, Bosnia and Herzegovina
e-mail: adis.alihodzic@pmf.unsa.ba

D. Hasanspahic
e-mail: damir.hasanspahic@gf.unsa.ba

E. Tuba
Singidunum University, Belgrade, Serbia
e-mail: etuba@acm.org

M. Tuba
State University of Novi Pazar, Novi Pazar, Serbia
e-mail: mtuba@acm.org

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_17

1 Introduction

Sensor deployment is one of the most challenging research fields in computer vision [14]. It can be described as automatically estimating sensor locations in order to observe complex scenes. Commonly, sensor deployment is being considered in three dimensions, but in some applications, it can also be viewed in two dimensions. For instance, in the case of buildings, which are being modeled as objects obtained by extrusion. Since the sensors are omnidirectional or rotating, for tasks such as surveillance, the interior coverage (IC) by the optimal number of sensors can be modeled by the widespread Art Gallery Problem (AGP), which requires seeing the interior of a polygonal environment with the smallest number of sensors, where sensors are usually represented as the points in a 2D plane [2]. The art gallery problem dates back to the 1970s, and it was one of the earliest and most significant problems in computational geometry, sensor placement, architecture, placement of radio antennas, urban planning, ultrasonography, mobile robotics, and other branches of science and industry [5]. In computational geometry, it presents a visibility problem of placing at least one security guard to cover every area of a museum or gallery [2]. Since the optimal sensor placement (OSP) represents the process of finding the minimal number of sensors that are sufficient to cover every point in the environment, we can say the both AGP and OSP are very similar to each other. Art gallery problem in the original form presents an intractable NP-hard problem which was based on determining smallest number of security guards sufficient to see every point in an n-sided two-dimensional polygon P with or without holes [3, 4, 6–8, 10, 13]. In this paper, we suggest two classes of algorithms such as a suboptimal deterministic sensor location method (SLM) as well as an adjusted discrete version of differential evolution (ADDE) for tackling the sensor placement problem. Before we apply the mentioned algorithms to tackle IC of a polygon, we should exploit preprocessing techniques for a polygon decomposition into convex components, which will be covered by a minimal number of sensors. Since, optimal sensor placement is an intractable problem, in this article, we propose a discrete differential evolution nature-inspired swarm intelligence technique to search for the suboptimal solutions to reach fast convergence and reduce computational time [15]. In order to show the power of proposed techniques, they have been tested on 127 various randomly generated polygons. The results produced in the experimental analysis were compared with ones reached by two well-known metaheuristics such as simulated annealing (SA) and particle swarm intelligence (PSO), which we also have been implemented for comparison purposes. From the analysis of results, it can be concluded that the ADDE method is the best technique, and it yields the best solutions considering both accuracy and convergence speed.

The rest of the paper is organized as follows. Mathematical problem formulation is described in Sect. 2. In Sect. 3, we propose a suboptimal deterministic method. The details of our adjusted discrete differential evolution method (ADDE) are presented in Sect. 4. Experimental and comparative results of applying our algorithms are presented in Sect. 5. Finally, conclusions and suggestion for future work are discussed in the last section of the paper, Sect. 6.

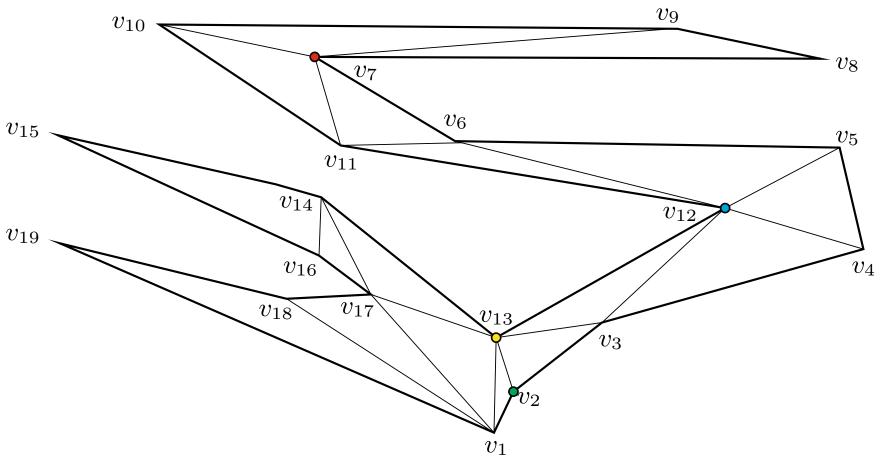


Fig. 1 The interior of the polygon P has been covered by vertices v_2, v_7, v_{12}, and v_{13}

2 Mathematical Problem Formulation

In this section, we introduce a mathematical model for seeking an optimal number of sensors required to perform interior covering (IC) of a non-convex 2D polygon. Foremost, we will briefly introduce notation to facilitate exposure. For any two distinct points v_1 and v_2 in the plane, we denote by $\overline{v_1 v_2}$ the segment whose two endpoints are v_1 and v_2. A planar polygon P presents a closed plane figure whose boundary is composed of segments $\overline{v_i v_{i+1}}$ $(i = 0, 1, \ldots, n-1)$, where $v_n = v_0$. Also, a polygon P is simple if it is not self-crossing and has no holes. A vertex v of a polygon P is reflex if the internal angle at v is greater than 180°. A planar polygon P is concave (non-convex) if there are two points u and w inside of P such that the segment $\overline{uw}$ is not entirely contained in the P. A planar polygon P is convex it is not concave. Also, a concave polygon must have at least four sides, and it always has at least one reflex interior angle, that is, an angle with a measure that is between 180° and 360° exclusive. Any point u in P is said to be visible from any other point w in P if the segment $\overline{uw}$ does not intersect the exterior of P as well it is entirely contained in P. For any point $u \in P$, the set of all points in P which are visible from a vertex u is called the visibility region of u, and we denote that set by $F(P, u)$. If the point u is a vertex of the polygon P, i.e., exists some index $k \in \{1, 2, \ldots, n\}$ such that $u = v_k$, then we call the subset $F(P, u)$ of P fan F_k, where the vertex v_k denotes the fan vertex of the set F_k. On the other hand, let u is not a vertex of the polygon P. Then, the set $F(P, u)$ is called a region under surveillance from the point u. In order to perform the coverage of interior by using sensors, we should make polygon decomposition into a set of nonoverlapping convex parts C_j such that their union is the entire region of polygon P. There are several ways how to accomplish dividing a simple non-convex polygon into nonoverlapping convex components [9]. In this paper, the partitioning of a polygon has been executed by exploiting triangulation. To efficiently perform triangulation, we have implemented a very efficient algorithm whose time complexity is proportional to the $O(n \log n)$ [2]. Since the triangulation of a simple non-convex polygon P composed of n vertices produces $n-2$ triangles or components, we will first create those components $C_1, C_2, \ldots, C_{n-2}$ such that the fan F_j ($j \in \{1, 2, \ldots, n\}$) contains the indexes of components which are visible from the vertex v_j. Now, optimal sensor placement problem can be expressed as seeking the smallest number of sensors which can see all components. In the following, we consider the creating of fans for a simple non-convex polygon P composed of 19 vertices as depicted in Fig. 1. It is not hard to notice that, for example, for the vertex v_7, a fan F_7 has indexes $\{2, 3, 4, 6\}$, because from the vertex v_7, the fan F_7 sees these components (triangles) C_2 ($\triangle v_6 v_7 v_{11}$), C_3 ($\triangle v_7 v_{10} v_{11}$), C_4 ($\triangle v_7 v_9 v_{10}$), and C_6 ($\triangle v_7 v_8 v_9$). The contents of the other fans can be determined quite analogously, for example, the fan F_6 has these elements: $\{1, 4, 5, 6\}$. As we can witness from Fig. 1, exactly four sensors as v_2, v_7, v_{12}, and v_{13} were enough to being achieved optimal sensor placement, because all triangles are covered by them. Also, the polygon P can be covered by sensors v_{10}, v_{11}, v_{13}, and v_{19}. In general, for some types of polygons there is more than one optimal solution.

3 A Suboptimal Sensor Location Method for IC

In this part, we will sketch in detail a suboptimal Sensor Location Method (SLM) for interior covering (IC). By taking into account the above-introduced definitions, the main idea of our deterministic method has been summarized by Algorithm 1. At the beginning of the algorithm, we determine the fan F_{i_1}, which contains the largest number of component indexes covered by the vertex v_{i_1}. In that case, we save the number i_1 as the index of the first sensor which covers a specific interior of a polygon. After that, we update the remaining sets F_j by removing from them all the elements which appear in the set F_{i_1}, i.e., we make difference $F_j \leftarrow F_j \setminus F_{i_1}$ for all fans. It implies that the set F_{i_1} becomes empty, so it is no longer considered. For non-empty updated fans F_j, we repeat the same procedure as at the beginning of the algorithm, i.e., we select the fan F_{i_2} which has the most elements, and then take that index i_2 be the index of the second sensor. It is clear now that the sensor with index i_1 covers more components than the sensor with the index i_2. By repeating the mentioned procedure, all fans F_i will be empty after a certain number of iterations, which is an indicator for the end of the algorithm. Now, the generated numbers $i_1, i_2, \ldots, i_k$ present the indexes of vertices from which the complete inside of the polygon P can be seen, where k denotes the number of sensors required to cover IC. From the pseudo-code presented in Algorithm 1, we can see that the method stops when all fans become empty. In other words, since the union of fans F_i ($i = 0, 1, \ldots, n-2$) denotes the indexes of components from which the original polygon P is composed, it is easy to conclude that the algorithm ends as soon as all components are covered.

Algorithm 1 A suboptimal sensor location method (SLM) for IC

Set $n_s \leftarrow 0$, $S \leftarrow \emptyset$, where n_s is a number of sensors and S is their list.
Determine a triangulation of n-sided non-convex simple polygon P. Let us denote the obtained triangles as components $C_1, C_2, \ldots, C_{n-2}$.
For each vertex v_i $(i = 0, 1, \ldots, n-1)$ determine the fan F_i by adding indexes of components C_j into F_i which are completely visible from the vertex v_i.
Initialize the list of all fan's indexes with $F \leftarrow \{F_0, F_1, \ldots, F_{n-1}\}$.
while $n_s \neq n$ **do**
    From the list F, find the fan that has most elements and denotes its index by i.
    Put to the list S the sensor $v_i \in P$ which was referred to the biggest founded fan F_i from the previous step.
    From all fans F_j remove the elements which were appeared to the set F_i, i.e., $F_j \leftarrow F_j \setminus F_i$.
    Set $n_s \leftarrow n_s + |F_i|$ and remove the fan F_i from the list F.
end while

4 An Adjusted Discrete Differential Evolution for IC

In this part, we will outline in detail an adjusted version of discrete differential evolution (ADDE) algorithm applied for seeking the optimal number of sensors to tackle IC problem. Due to the shortage of room for writing, at this place, we will not describe the basic structure of the DE algorithm. More details related to the DE algorithm for combinatorial problems can be found in the paper [11]. The main problem for all swarm intelligence techniques is how to design an agent for a particular type of problem. Specifically in this article, an agent is defined as a vector $\mathbf{v} = (v_{i_1}, v_{i_2}, \ldots, v_{i_d})$ of dimension d, where the binary coordinates v_{i_k} are related to indices of fans, while d is a number of fans. To further speed up the search process of the algorithm, instead of looking at n fans for a polygon composed of n vertices, we will consider only those fans that are obtained after elimination. Elimination is performed as follows. If an arbitrary vertex i can see all the components contained in the fan F_j, then it is said that the fan F_i covers the fan F_j, so the fan F_j can be removed from the set of all fans, which directly impacts on the reducing size of fans. It implies that after elimination, a dimension d is strictly less than n. For example, for the polygon P shown in Fig. 1, the number of fans after elimination is reduced from 19 to 5, so the agent can be represented by the decision vector $\mathbf{v} = (v_{i_1}, v_{i_2}, \ldots, v_{i_5})$ of dimension 5. In particular, after elimination remains these fans F_2, F_6, F_7, F_{12}, and F_{13}, so the decision vector $\mathbf{v}$ becomes $\mathbf{v} = (v_2, v_6, v_7, v_{12}, v_{13})$. Then an algorithm determines which components v_{i_j} will select $(v_{i_j} = 1)$ to cover the interior of the polygon. For the polygon P in Fig. 1, the algorithm selects these components 2, 7, 12, and 13, so the decision vector $\mathbf{v} = (1, 0, 1, 1, 1)$ represents an optimal solution. This further means that the fans F_2, F_7, F_{12}, and F_{13} perform IC by using the components contained in them. Another problem that arises here is how to treat the agents that are unable to cover the interior of a polygon. For example, if the algorithm generates the solution $\mathbf{v} = (1, 0, 0, 1, 1)$, then the components F_2, F_{12}, and F_{13} are not able to cover interior of the polygon. We will call this solution an infeasible solution. To enable the algorithm to operate simultaneously both with feasible and infeasible solutions, we apply Deb's rules as in the paper [1]. Thus, at the beginning of the algorithm, none type of solution is preferred, while at the end of the algorithm, only feasible solutions with a minimum number of non-zero elements (ones) are taken into consideration. Since an optimal sensor placement is a combinatorial problem, and the original DE algorithm was invented to deal with continuous problems, within the DE algorithm, after the agents are initialized, each component v_{i_j} of an agent $\mathbf{v} = (v_{i_1}, v_{i_1}, \ldots, v_{i_d})$ is updated according to the following function:

$$v_{i_j} = \begin{cases} 1, & \textbf{if}\, r < \frac{1}{1+e^{-v_{i_j}}} \\ 0, & \textbf{otherwise} \end{cases} \tag{1}$$

where r presents a random number uniformly distributed from (0,1). Once the agent was created, it is further being passed through a fitness function f, which returns the number of sensors required to treat the interior of a polygon. Those solutions which are fittest (they generate the smallest number of sensors) under Deb's rules are being moved to the succeeding generations of the algorithm.

5 Experimental Results

In the experimental study, we have compared our proposed ADDE algorithm with our deterministic version as well as with two other standard stochastic methods such as SA and PSO. The goal of the application our methods was to minimize the number of sensors required to perform interior coverage (IC). Our algorithms have been applied to 127 randomly generated polygons, which are produced with our random polygon generator similarly as in paper [12]. Each instance is called RI-k-i, where k denotes the size of the ith instance. The coordinates of points (x, y) are chosen from [0, 500]. The proposed approaches have been implemented in C# programming language. All tests were performed on Intel Core i7-3770K @3.5GHz with 32 GB of RAM running under the Windows 10×64 OS. To compare the ADDE approach with the PSO, and the SA, the function evaluation was computed as $N \times G$, where N is the population size, while G is the maximum number of generations. Each algorithm was executed parallel in 30 independent series using multi-threads. For the algorithms PSO ($N = 100$, $G = 200$) and ADDE ($N = 10$, $G = 2000$) was allocated 20,000 functional evaluations. The learning parameters of the ADDE, such as C_r and F, are set to 0.4, 0.4, respectively. Also, the parameters of PSO such as α, β, and the initial velocities of particles v_{min}, v_{max} were set to 2.0, 2.0, and -6, 6, respectively. In the case of SA, the temperatures T_0 and T_{stop} were set to 1.0 and 10^{-10}, while temperature-reducing parameter α was set to 0.99.

The number of sensors and the corresponding computational times found by our algorithms are presented in Tables 1 and 2. In Table 1, to check the quality of obtained solutions as well as computational times, the algorithms were selected and tested through five groups of randomly generated polygons, where each group was

Table 1 The simulation results provided by proposed algorithms for 25 randomly distributed instances of the simple non-convex polygons over 30 independent runs

Rand. instan.	Algorithms													
	SA				PSO				ADDE				SLM	
	Min.	Avg.	SD	Time	Min.	Avg.	SD	Time	Min.	Avg.	SD	Time	Ns	Time
RI- 20-1	4	4	0	0.01	4	4	0	0.12	4	4	0	0.15	5	0.01
RI- 20-2	3	3	0	0.01	3	3	0	0.10	3	3	0	0.12	4	0.01
RI- 20-3	2	2	0	0.01	2	2	0	0.08	2	2	0	0.10	3	0.01
RI- 20-4	4	4	0	0.01	4	4	0	0.17	4	4	0	0.13	5	0.01
RI- 20-5	3	3	0	0.01	3	3	0	0.09	3	3	0	0.13	4	0.01
RI- 60-1	11	12.57	0.72	0.13	9	9.27	0.44	0.54	9	9	0	0.49	11	0.15
RI- 60-2	9	11.17	0.64	0.08	8	8.67	0.47	0.52	8	8	0	0.51	9	0.13
RI- 60-3	10	11.03	0.75	0.13	9	9.03	0.18	0.54	9	9	0	0.48	11	0.13
RI- 60-4	11	12.53	0.92	0.11	10	10.13	0.34	0.54	10	10	0	1.18	12	0.09
RI- 60-5	9	10.63	0.71	0.08	8	8.03	0.18	0.46	8	8	0	0.47	10	0.15
RI- 100-1	17	19.50	1.02	0.25	14	14.80	0.48	1.04	14	14	0	0.87	16	0.36
RI- 100-2	20	22.40	1.28	0.19	16	17.00	0.63	1.06	16	16	0	0.86	18	0.30
RI- 100-3	16	19.87	1.36	0.17	14	14.77	0.67	0.96	14	14	0	0.88	18	0.31
RI- 100-4	19	20.80	1.19	0.19	13	13.90	0.65	1.00	13	13	0	0.86	14	0.39
RI- 100-5	18	21.00	1.24	0.19	15	15.70	0.82	1.46	15	15	0	0.88	16	0.37
RI- 300-1	–	–	–	–	49	54.10	2.29	6.82	46	46.23	0.42	4.44	52	4.28
RI- 300-2	–	–	–	–	50	54.53	2.33	6.78	44	45.33	0.60	4.28	50	4.10
RI- 300-3	–	–	–	–	51	57.27	2.66	6.96	45	45	0	4.02	49	4.43
RI- 300-4	–	–	–	–	47	50.90	2.07	6.34	43	43.93	0.36	4.51	50	4.22
RI- 300-5	–	–	–	–	49	53.90	2.34	6.86	44	44	0	4.13	48	3.87
RI- 500-1	–	–	–	–	115	126.57	5.00	25.68	77	77.80	0.75	12.42	87	13.99
RI- 500-2	–	–	–	–	93	99.57	3.96	19.78	70	71.80	0.87	11.04	85	12.80
RI- 500-3	–	–	–	–	93	99.97	3.09	19.97	70	71.77	0.72	11.66	78	14.30
RI- 500-4	–	–	–	–	103	113.43	5.33	25.29	74	75.17	0.69	12.09	85	15.11
RI- 500-5	–	–	–	–	102	110.13	4.95	27.85	68	69.50	0.62	11.86	77	14.01

Table 2 The mean number of sensors and mean time processing provided by proposed algorithms for 127 randomly generated instances through 30 independent series

No.	SA			PSO			**ADDE**				SLM	
vert.	Best	Avg.	Time	Best	Avg.	Time	Best	Avg.	Std.	Time	Best	Time
20	**3.43**	**3.43**	**0.01**	**3.43**	**3.43**	0.11	**3.43**	**3.43**	**0**	**0.15**	3.77	0.01
60	9.89	11.32	0.08	**9.07**	9.24	0.52	**9.07**	**9.13**	**0.04**	**0.54**	10.04	0.12
100	18.96	21.62	0.19	**14.88**	15.67	1.10	**14.88**	**14.90**	**0.04**	**1.09**	16.28	0.32
300	–	–	–	50.10	54.68	6.89	**44.00**	**44.54**	**0.38**	**4.31**	48.90	4.28
500	–	–	–	99.78	108.26	23.14	**72.74**	**75.45**	**0.80**	**11.56**	81.17	13.56

composed of five randomly generated polygons containing 20, 60, 100, 300, and 500 vertices, respectively. Also, in the same table for the metaheuristics, SA, PSO, and ADDE, the best results, mean results, standard deviation, and computational time were denoted by Min., Avg., SD, and Time, respectively. On the other hand, since the SLM is a deterministic method, it is driven once, and its results have been recorded by the parameters NS (number of sensors) and Time. From the results shown in Table 1, we see that all methods work equally well, where the SA and SLM are the fastest. For polygons composed of 60 vertices, the SA (if we consider the parameter Min.) and SLM produce similar results. In contrast, the best results were generated by PSO and ADDE, where ADDE is a more stable method (SD = 0). The ADDE is also stable for polygons composed of 100 vertices, for which SA is the most unstable and gives the worst results, while PSO gives better results than SLM, and slightly worse than ADDE on average. By increasing the number of vertices, e.g., for polygons composed of 300 and 500 vertices, SA fails to find a feasible solution because it has been stuck in some local optima. On the other hand, for the instances, RI-300, PSO, and SLM work very similarly and yield slightly worse outcomes as opposed to the ADDE method. For the instances RI-500, the ADDE method remains superior to others as before, while the PSO returns worse results than the SLM method, even in the best case. To cover all polygons from all groups shown in Table 1, in the best case, our ADDE approach requires 713 sensors, while both SLM and PSO need 817 and 884 sensors, respectively. Based on this, it follows that ADDE consumes 104 sensors less than the SLM method, and 171 sensors less than the PSO method. Also, the difference in the number of sensors is evident in the average case, where ADDE takes 94.47 sensors less than SLM method. Particularly superiority comes to the fore with an increase in the size of polygon. In reality, reducing the number of sensors has many implications, and some of them are less money, energy savings, and hardware components. To show the real robustness of the proposed methods, we tested them for a dataset composed of 127 randomly generated polygons, and the obtained results were saved in Table 2. The results show that the ADDE produces in each case better outcomes compared to other algorithms. Namely, the mean number of sensors linearly increases concerning the total number of vertices (n), which implies that the growth rate of sensors is noticeably slower at the ADDE method compared to other approaches. Also, from the results shown in Table 2, we can see that the average number of sensors consuming by the ADDE is less or equal to $\lceil * \rceil \frac{n}{6}$. Based on the simulations, it can be concluded that the ADDE is a robust method which provides quality solutions when addressing the sensor placement problem.

6 Conclusion

In this paper, we studied the optimal sensor placement problem for IC and proposed two versions of algorithms for its solving. Quality of our methods was tested throughout 127 randomly generated instances. Based on the simulation results, it can be concluded that our ADDE method is convenient for this task, and it produces excellent overall performance. Also, the ADDE approach proved to be robust, in the sense that it was able to tackle different instances from a broad range of randomly generated. Since the preprocessing of the ADDE is computationally expensive for large-scale instances, in future work, we will investigate the efficient techniques from computational geometry to tackle these drawbacks. Also, we will apply more metaheuristics, as well as other types of polygons, such as orthogonal polygons, von Koch polygons, and other types.

References

1. Alihodzic A (2016) Fireworks algorithm with new feasibility-rules in solving uav path planning. In: 2016 3rd International Conference on Soft Computing Machine Intelligence (ISCMI), pp 53–57. https://doi.org/10.1109/ISCMI.2016.33
2. de Berg M, Cheong O, van Kreveld M, Overmars M (2008) Computational geometry: algorithms and applications, 3rd edn. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77974-2
3. Bjorling-Sachs I, Souvaine DL (1995) An efficient algorithm for guard placement in polygons with holes. Discret Comput Geom 13:77–109. https://doi.org/10.1007/BF02574029
4. Chvátal V (1975) A combinatorial theorem in plane geometry. J Comb Theory, Ser B 18(1):39–41. https://doi.org/10.1016/0095-8956(75)90061-1
5. Elnagar A, Lulu L (2005) An art gallery-based approach to autonomous robot motion planning in global environments. In: 2005 IEEE/RSJ international conference on intelligent robots and systems, pp 2079–2084. https://doi.org/10.1109/IROS.2005.1545170
6. Hoffmann F, Kaufmann M, Kriegel K (1991) The art gallery theorem for polygons with holes. In: Proceedings 32nd annual symposium of foundations of computer science, pp 39–48. https://doi.org/10.1109/SFCS.1991.185346
7. Katz MJ, Roisman GS (2008) On guarding the vertices of rectilinear domains. Computat Geometry 39(3):219–228. https://doi.org/10.1016/j.comgeo.2007.02.002
8. Lee D, Lin A (1986) Computational complexity of art gallery problems. IEEE Trans Inf Theory 32(2):276–282. https://doi.org/10.1109/TIT.1986.1057165
9. O'Rourke J (1998) Computational geometry in C. Cambridge University Press, Cambridge
10. O'Rourke J, Supowit K (1983) Some np-hard polygon decomposition problems. IEEE Trans Inf Theory 29(2):181–190. https://doi.org/10.1109/TIT.1983.1056648
11. Pan QK, Tasgetiren MF, Liang YC (2008) A discrete differential evolution algorithm for the permutation flowshop scheduling problem. Comput Ind Eng 55(4):795–816. https://doi.org/10.1016/j.cie.2008.03.003
12. Sadhu S, Hazarika S, Jain KK, Basu S, De T (2012) Grp_ch heuristic for generating random simple polygon. In: Combinatorial algorithms, pp 293–302. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35926-2
13. Schuchardt D, Hecker H (1995) Two np-hard art-gallery problems for ortho-polygons. Math Logic Quart 41(2):261–267. https://doi.org/10.1002/malq.19950410212
14. Scott WR, Roth G (2003) Jean-François: view planning for automated three-dimensional object reconstruction and inspection. ACM Comput Surv 35(1):64–96. https://doi.org/10.1145/641865.641868
15. Storn R, Price K (1997) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11(4):341–359. https://doi.org/10.1023/A:1008202821328

Minimizing Costs of Transportation Problems Using the Genetic Algorithm

Marwan Abdul Hameed Ashour, Alyaa Abdulameer Ahmed, and Iman Amer Hameed Al-dahhan

Abstract The research aims to minimize the total costs and reach an optimal solution in transporting the gasoline product from the main warehouses in Baghdad Governorate to filling stations: by relying on the dataset obtained from the Oil Products Distribution Company of the Iraqi Ministry of Oil, by using the traditional method (linear programming) and the modern method (genetic algorithm), and then compare them to help decision-makers make the right decision. The preference was achieved for the modern method, which was able to make a slight improvement in the final results. This is because the mathematical model of the problem is a linear model, where the cost in the technique of linear programming (1,424,165) dinars and in the technique of the genetic algorithm (1,424,157) dinars.

Keywords Transportation cost · Linear programming · Evolutionary algorithms

1 First Section

1.1 Introduction

The transportation problem is a special type of linear programming problem that describes the general model in the distribution of any type of commodity from any group of supply centers, which are called the sources to reception centers which are called destinations. In that way, it minimizes the total cost of distribution, given that the entire request is received from the source for distribution. Any transportation problem is solved after obtaining the basic parameters (supplies, requests, unit costs

M. A. H. Ashour (✉) · A. A. Ahmed
Statistics Department, College of Administration & Economic, University of Baghdad, Baghdad, Iraq
e-mail: dr_marwan@uobaghdad.edu.iq

I. A. H. Al-dahhan
Continuing Education Center/ University of Baghdad, Baghdad, Iraq

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_18

per unit), to build a special model as it is arranged in one schedule to solve the problems that occur during the process of distributing units from sources to destinations, and the most important of these problems may be the lack of distribution directly from the source to the destination as the commodity or goods may pass through stages along the way and through certain transportation points. which leads to different shipping costs from one point to another and It should not exceed the upper limits of the specified final cost to solve problems that occur during the process of distributing units from sources to destinations and the most important of these problems [1]. We must shed light on who handled the transportation problem for the first time by (Hitchcoc) in 1941. In a historical study to distribute a product from several sources to many regions which is the first contribution to solving the transportation problem.; (Koopmans) in 1947. Another important study is (the best use to study the transport problem). Then many scientists then took care of such problems; (Haly) in 1962 who studied the problem of solid transport, which is known as multi-commodity transport, (srdjevic) in 1997 who studied standard linear programming for the transport problem [2, 14]. The transportation problem is a very important aspect because it is used in a wide range of real-life problems in our daily life. And it seeks to achieve the total demand for individuals companies or the concerned authorities when the supplies are known. This is done through several methods including (Vogel technique, Least Cost technique, Northwest corner technique) Taking into account the balance between rows and columns in all ways in order to obtain the maximum possible profit. This is done by defining the planned goals such as delivering products or goods to different consumers and other important goals that help to reach the best solution to the problems that the transport process thus help decision-makers to reach the best possible methods [3]. Also, a complete study was conducted on the transportation of damaged materials or deteriorated. With the goal to minimize the cost of transportation and reduce fractures that occur during transportation. As time and speed play a decisive role in it, increasing the speed minimizes the total time for transportation. But this creates a great risk of increasing fractures which leads to more losses. Therefore, this type of problem is called multi-target transport, where the primary goals are set to reach the minimum fraction in the least possible time. The study also sought to reduce the total distance for transporting eggs from different cities while maintaining the specified goals and not focusing on one goal which often minimizes transportation costs, by following two stages The first is to conduct a survey of the requirements of the beneficiary and the second use the Tora program to find the best solution in reducing the distance and time [4]. In a brief study on the use of the Vogel method, which is considered one of the most important methods known to solve the problem of transportation and the most efficient to reach optimal solutions and to minimize the cost of shipping goods from one location to another until the needs of each region are met, It was done by relying on the penalty cost which means the difference between the two lowest costs in each row and column. And then identifying the row or column that contains the largest penalty cost. It was found that the Vogel method is highly efficient in solving all transport problems, but it stands out more clearly in the small problems of the many iterations generated during the solution to reach the optimal solution [5]. Three variables have also been used to minimize the cost of transportation using four methods and then compare them to find the best methods. It is possible to meet the demand at specific points and choose the optimal solution from among the available alternatives. It was concluded that the Modi method is the best in solving problems that have several variables, and the best solution was reached within a shorter period of time than the other methods use [6]. In a study of the problem of the transport network with concave cost, the genetic algorithm was used to solve this type of problem in order to reach an optimal solution, by generating a preliminary solution for the population and then evaluating the fitness function of each chromosome from the population and choosing the good chromosomes. To create a new child that works at least like the parents work to reach an acceptable solution and the results confirmed the correctness of the algorithm. Because it enables us to get a better solution and perhaps get a global solution as we have run it several times as the algorithm has proven the speed of access to a solution in large networks [7]. Evolutionary algorithms have also been used in the problem of multi-purpose transmission as a linear improvement problem and by obtaining new generations of solutions by re-installing and transforming that occurs in the algorithm until the optimal solution is reached [8].

1.2 Research Method

Data were collected from the Oil Products Distribution Company as it is one of the formations of the Ministry of Oil and responsible for the process of distributing products and gas in the country from gasoline storage depots to filling stations to achieve the goals and development plans of the Ministry of Oil.

1.3 Optimization

Optimization is a mathematical technique related to finding the maximum or minimum set of variables, where a group of improvement techniques competes to reach the best solutions that decision-makers need to solve a problem to maximize profit or minimize losses. Based on the principle of improvement to make the appropriate decision, the principle of improvement appeared in the 1940s. As many researchers have come up with different solutions to linear optimization problems mathematically, to improve one or more goals under a set of specific circumstances, whether the restrictions are in the form of equality or inequality. Depending on the fitness function problem of a specific developmental algorithm, which represents an area of solution to the problem and is a global improvement tool for many problems, the optimization technique can be applied well to the transportation problem in order to minimize the cost of shipping goods and to reach an optimal solution or a better alternative [9].

1.4 Transportation Problem

The transportation problem is a special type of improvement problem characterized by its ability to minimize the cost of shipping a certain commodity from production centers to marketing centers to the consumer and to determine the quantities to be transferred from the source of supply to the source of the purpose in order to minimize the cost of shipping and find the best way to meet a specific order and satisfy the consumer's need. The transportation problem also focuses on achieving the demand for individuals after determining the necessary supplies, taking into account the total time in the transportation process, and trying to overcome crises and obstacles that hinder the arrival of goods to the consumer, such as traffic or difficult climatic conditions and other crises for the purpose of reaching more efficient solutions, other methods of improvement are considered, the most important of which is the genetic algorithm. Below is the general model of the transport problem [3].

Minimize $Z = \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij} Z_{ij}$.

Subject To

$$\sum_{i=1}^{n} Z_{ij} \leq si, \quad \text{i} = 1, 2, 3, ..., \text{n} \tag{1}$$

$$\sum_{j=1}^{m} Z_{ij} \geq, \quad \text{j} = 1, 2, 3, \ldots, \text{m} \tag{2}$$

C_{ij}: It represents the transportation cost per unit
Si: supply
dj: demand.

1.5 Genetic Algorithm

The genetic algorithm is the most widely used algorithm in computational methods and solving multi-objective optimization problems to generate optimal solutions to complex problems by means of certain factors inspired by nature such as mutation and selection that work to improve the solution in a real-world window [10]. To reach the optimal solution, which means the most appropriate or best solution to solve a specific problem, as many decision-makers need to know the most appropriate solutions to achieve the goals of the company or institution, which is often a profit maximization or loss reduction, and that is by trying repeatedly until the global optimal solution is reached, which is considered the best. Kinds of solutions absolutely [9]. The idea of a genetic algorithm is based on the Darwinian principle which means survival of the fittest among living organisms and this means that they have an opportunity to stay better than others and have the ability to adapt to evolving conditions and their offspring may inherit the same traits. This leads to the production of future generations more fit with the possibility of genetic mutations randomly in one of the organs when reproduction and this results in opportunities for the continuation of suitable individuals in the long term. The mutation introduces a new genetic material to the population as it replaces one or more decision variables with a new, random solution and is created by GA is called a chromosome, which consists of a series of genes that represent a possible solution to the improvement problem. The value of fitness determines the ability of individuals to survive [11]. It has also been applied in many fields such as design, real estate, environment, and construction, in an attempt to reduce project completion time and costs [12]. The following figure shows how the genetic algorithm works.

1.6 The Relationship Between the Genetic Algorithm and the Transfer

The transport problem is one of the linear improvement problems that seek to find an optimal solution and one of the most important modern methods used to improve the solution of transport problems are the evolutionary algorithms, especially the genetic algorithm to reach the best possible results and achieve the goal of the institution or company as soon as possible, and that is what we will see in our research in solving the problem of transferring a product Gasoline from the main depots to the filling stations, using the genetic algorithm, at the lowest possible cost (Fig. 1).

Through the information obtained to transfer the gasoline product from the warehouses to the stations on both sides of Karkh and Rusafa In order to secure petroleum products for citizens, a schedule of costs assigned to transport the product to each station is organized and the supply and demand are shown as follows (Table 1):

1.7 Solution

According to the general model of the transportation problem, we are building a special model for the problem of transporting the gasoline product from the main warehouses in Baghdad Governorate to the six fuel filling stations and determining the value of the goal function and restrictions and then solving the model in two ways and comparing them by relying on a table (1–1) as follows:

$$\begin{aligned}
\text{Min Z} &= 836 \times 11 + 1097 \times 12 + 1202 \times 13 + 1254 \times 14 + 1045 \times 15 \\
&+1620\times 16 + 560\times 21 + 2717 \times 22 + 2926 \times 23 + 3396 \times 24 + 2665 \times 25 \\
&+3762 \times 26 + 508 \times 31 + 2874 \times 32 + 1855 \times 33 \\
&+1515 \times 34 + 1672 \times 35 + 1881 \times 36
\end{aligned}$$

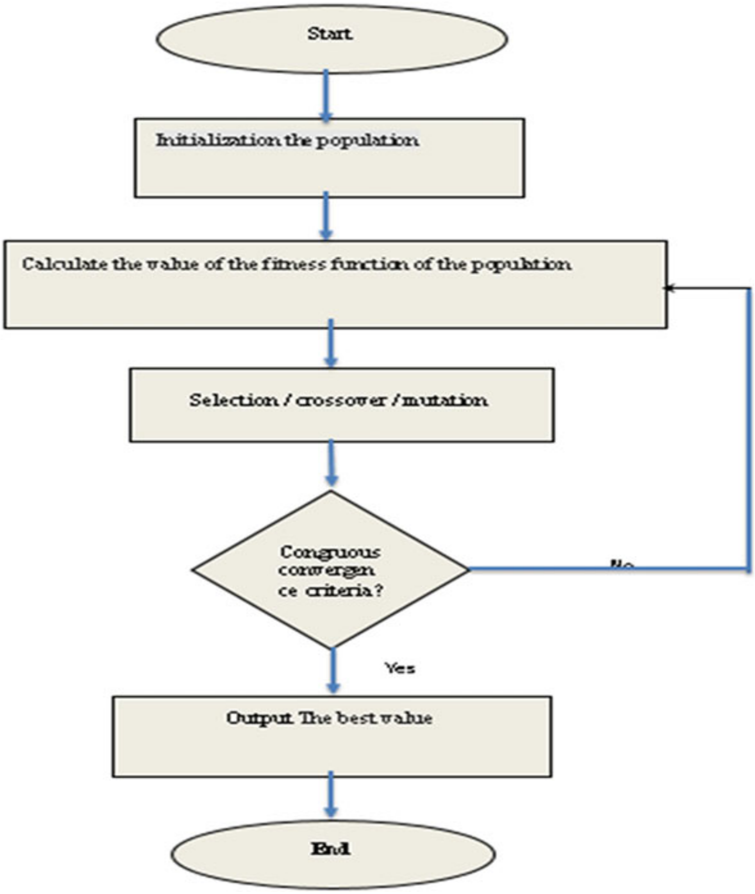


Fig. 1 Steps to implement the genetic algorithm [13]

Tabel 1 Transport costs and assigned quantities for six stations from three warehouses

Stations Ware houses	S1 almathnaa	S2 AlMansour	S3 AlMustansiriya	S4 AlIdrisi	S5 AlKilani	S6 albnuk	supply
D1 aldawruh	836	1097	1202	1254	1045	1620	440
D2 alkarkh	560	2717	2926	3396	2665	3762	504
D3 alrusafa	508	2874	1855	1515	1672	1881	560
demand	150	255	160	150	265	115	

Subject To

$$836 \times 11 + 1097 \times 12 + 1202 \times 13 + 1254 \times 14 + 1045 \times 15 + 1620 \times 16 \le 440$$
$$560 \times 21 + 2717 \times 22 + 2926 \times 23 + 3396 \times 24 + 2665 \times 25 + 3762 \times 26504 \le 504$$
$$508 \times 31 + 2874 \times 32 + 1855 \times 33 + 1515 \times 34 + 1672 \times 35 + 1881 \times 36 \le 560$$
$$836 \times 11 + 560 \times 21 + 508 \times 31 \ge 150$$
$$1097 \times 12 + 2717 \times 22 + 2874 \times 32 \ge 225$$
$$1202 \times 13 + 2926 \times 23 + 1855 \times 33 \ge 160$$
$$1254 \times 14 + 3396 \times 24 + 1515 \times 34 \ge 150$$
$$1045 \times 15 + 2665 \times 25 + 1672 \times 35 \ge 265$$
$$1620 \times 16 + 3762 \times 26 + 1881 \times 36 \ge 115$$

1.8 Results

The mathematical model was solved by two methods (linear programming and the genetic algorithm) to minimize the final cost of transporting the gasoline product and the following was reached.

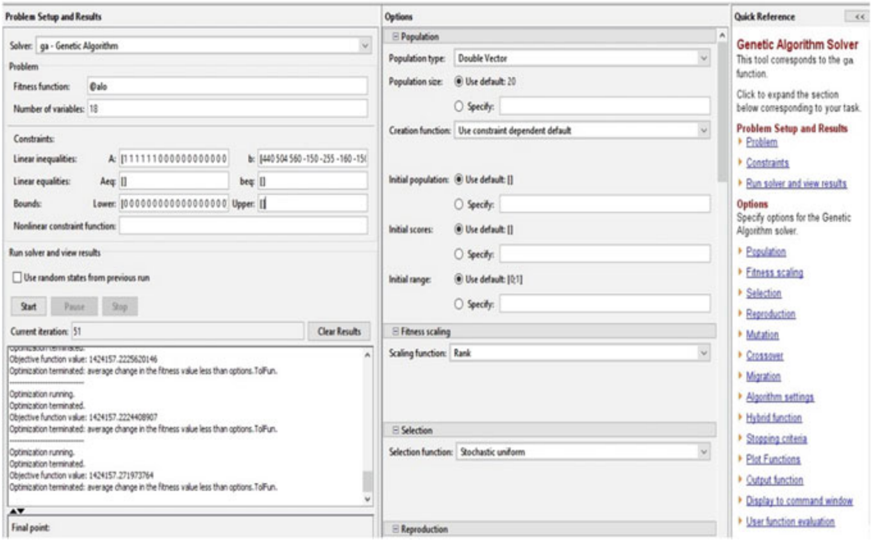
First: Linear programming technique.

After applying the mathematical model in (Win.Q.S.B) program, the following results were reached (Table 2).

Table 2 The optimal solution results

Decision variable	Optimal solution	constraint	Shadow price	constraint	Shadow price
X11	0	X31	55	C1	0
X12	255	X32	0	C2	409
X13	160	X33	0	C3	0
X14	0	X34	0	C4	0
X15	25	X45	240	C5	0
X16	95	X36	115	C6	0
X21	0	X24	0	C7	0
X22	0	X25	0		
X23	0	X26	0		
X24	0	X31	55		
X25	0	X32	0		
X26	0	X33	0		

Table 3 Shows the results of the solution using the genetic algorithm method in the MATLAB program



The final cost for transporting the gasoline product from warehouses to stations is (1,424,165) dinars.

Second: The genetic algorithm.

The mathematical model was applied in the program (MATLAB) using the genetic algorithm method to try to improve the solution resulting from the linear programming method and to make sure whether there was a better solution or not and the following results were reached (Table 3).

1.9 Discussion

The transportation problem was solved in the technique of linear programming and the genetic algorithm, and then the comparison between them as the results showed that a slight improvement was made to the cost of transport in the technique of the genetic algorithm, where the final cost in the linear programming method was (1,424,165) dinars, whereas the genetic algorithm method (1,424,157) dinars which confirms the advantage The genetic algorithm method is to minimize the total cost of transportation and to reach the best solution and achieve the goal for the beneficiary, and through this, the following is concluded:

1. 1.The results proved that the final cost of using the genetic algorithm method is slightly less than the linear programming method

2. The results demonstrated the possibility of improving the solution and reaching the optimal solution by repeating the attempts in the technique of the genetic algorithm
3. The research discussed two methods of improvement to solve the transportation problem and minimize the cost of transporting the gasoline product to the stations, and then compare them.

References

1. Hillier FS (2015) Introduction to operations research, Tata McGraw-Hill Education
2. Sharma G, Abbas SH, Gupta VK (2015) Solving multi-objective transportation problem to reduce transportation cost and time. J Advc Math 11(1)
3. Gupta R, Gulati N (2019) Survey of transportation problems. In: International conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE, pp 417–422
4. Mohammed NAR, Lahji AA, Syed JK (2013) Transportation of eggs: a case study using multi-objective transportation problem (MOTP). J Inf Oper Manag 4(1):327
5. Korukoğlu S, Balli S (2011) A improved vogel's approximatio method for the transportation problem. Math Comput App 16(2):370–381
6. Joshi RV (2013) Optimization techniques for transportation problems of three variables. IOSR J Math 9(1):46–50
7. Paşa T (2018) The genetic algorithm for solving the non-linear transportation problem. Rev Air Force Acad 2:37–44
8. Bharathi K, Vijayalakshmi C (2016) Optimization of multi-objective transportation problem using evolutionary algorithms. Global J Pure App Math 12(2):1387–1396
9. Talukder S (2011) Mathematicle modelling and applications of particle swarm optimization
10. Rivera G et al (2020) Genetic algorithm for scheduling optimization considering heterogeneous containers: a real-world case study. Axioms 9(1):27
11. Bzorg-Haddad O et al (2017) Meta-heuristic and evolutionary algorithms for engineering optimization, Wiley, Hoboken, NJ
12. Okagbue HI, Peter NJ, Akinola AO, Iroham CO, Opoko AP (2020) Smart review of the application of genetic algorithm inconstruction and housing. Int J Advc Trends Comp Sci Eng
13. Junbin Z et al (2014) Genetic algorithm particle swarm optimization based hardware evolution strategy. WSEAS Trans circuits Syst 2:274–283
14. Ashour MAH (2019) Optimum cost of transporting problems with hexagonal fuzzy numbers. J Southwest Jiaotong Univ 54(6). https://doi.org/10.35741/issn.0258-2724.54.6.10
15. Ashour MAH, Al-Dahhan IAH, Al-Qabily SMA (2019) Solving game theory problems using linear programming and genetic algorithms. In: International conference on human interaction and emerging technologies. Springer, Cham, pp 247–252. https://doi.org/10.1007/978-3-030-25629-6_39
16. Hassun DS, Jaber A, Ibrahim N (2013) Operation Research, aljazirat library
17. Hamada, A (2018) Solve the problem of direct transportation in order to find the optimal solution some oil depots in Baghdad are case studies: Al-Muthanna J Adm Econ Sci. https://doi.org/10.18018/MJAES/2018-8/36
18. Mondal, RN et al (2015) An innovative method for unraveling transportation problems with mixed constraints. Am J Math Stat 5(4):190–195 https://doi.org/10.5923/j.ajms.20150504.04

Analysis of the Most Promising Parametric Linear Methods for Studying Intellectual Competence

Yana Ivanovna Sipovskaya

Abstract The article examines parametric linear models of the productivity of the intellectual activity of older adolescents. The study involved 100 older students (15 years old). The methodological bases of the research are: "Conceptual synthesis", "Moods", "Methods for diagnosing the level of reflexivity", "Comparison of similar drawings", and "Interpretation". The results of the study allow us to conclude that among all the methods of statistical analysis used in the study, data normalization using z-scores, discriminant (stepwise) analysis, factorial (exploratory and confirmatory) analysis with varimax rotation, and regression (multiple) analysis, the following method showed the best results: factorial analysis. The factor structure of intellectual competence can be described by indicators of conceptual (conceptual), intentional (mentality and beliefs), and metacognitive abilities (reflexivity and reflectivity). Thus, factor analysis is the most flexible model for studying fuzzy sets and nonlinear connections of psychological constructs on the example of indicators of intellectual competence. However, the reliability of the models obtained (variance explained) is insufficient for methods that are more reliable. The facts obtained direct us to find more suitable analysis tools.

Keywords Data analysis · Parametric · Linear methods · Intellectual competence · Older adolescents

1 Introduction

The most promising approach to considering intellectual competence is to understand it as a special form of organizing individual mental experience [1, 2]. From our point of view, conceptual and intentional abilities are of particular interest.

Y. I. Sipovskaya (✉)
Institute of Psychology of the Russian Academy of Sciences, Yaroslavskaya st, 13, 129366 Moscow, Russia
e-mail: syai@mail.ru

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_19

This understanding of intellectual competence is consistent with the results of various psychological studies. In particular, it is noted that the achievement of competence presupposes the formation of systems of practical and mental skills that a person masters in the course of a long, "deliberate practice" [3]. The fact is emphasized that experts acquire intellectual competence only after spending a lot of time in favor of studying the relevant subject area, accumulating decision-making experience, etc. The hallmarks of "deliberate practice" are a high level of motivation for learning (cognitive need); constant feedback on the assessment of the correctness of their actions or their erroneousness (criticality, reflexivity); depth and thoroughness of material processing; and initiative and independence. In recent years, the key importance of motivation and value orientations in the development of intellectual competence has been demonstrated [4].

Thus, the components of intellectual competence are:

(1) Conceptual abilities, which provide a special type of organization of subject knowledge about the subject area and ways of transforming this knowledge in solving the assigned tasks [5, 6] etc.;
(2) Metacognitive abilities responsible for intellectual self-regulation [1];
(3) Intentional abilities, manifested in individually specific cognitive preferences and underlying the accumulation of implicit knowledge, on which, in turn, explicit, explicated knowledge is formed [7].

Thus, the analysis of research in the field of the psychology of competence made it possible to formulate a hypothesis that the key components of competence are conceptual, metacognitive, and intentional abilities. Accordingly, an urgent scientific task is to reveal the structure of intellectual competence, to show what is the relationship of its constituent components, and how important is the measure of their participation in effective intellectual activity. It is for the purpose of a more complete definition of the construct "intellectual competence" that this study was conducted, during which conceptual (conceptual), metacognitive, and intentional abilities as components of intellectual competence were studied in more detail (Fig. 1):

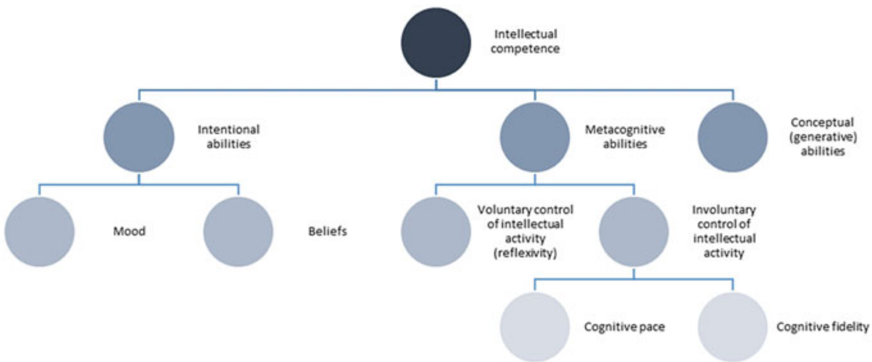


Fig. 1 The structure of intellectual competence

This approach certainly does not exclude the role of other components. However, this model is due to the idea that for the formation of intellectual competence, a necessary condition is the presence of subject knowledge, which would be generalized to the level of concepts (mental formations within an individual conceptual experience). At the same time, we assumed that, according to Vygotsky [8], the ability to reflect conceptually, to operate with generalized concepts is the highest stage of intellectual development.

Concepts are integral cognitive structures that are characterized by different ways of coding information, the hierarchical nature of the organization of semantic features, and the complexity of the cognitive composition, including the inclusion of sensory-emotional impressions [5]. The participation of sensory-emotional impressions in the composition of intellectual competence is a prerequisite for tacit knowledge, which, in our opinion, is an indicator of intentional abilities.

Thus, there is reason to assume the presence of a certain complex of related mental formations that characterize intellectual competence, including intellectual competence, which is formed in the process of school education among students, i.e. school competence. In the presented study, the "conceptual, metacognitive and intentional structure" of intellectual competence is analyzed, which implies the consideration of the distinguished abilities as components of the construct of intellectual competence.

1.1 Research Questions

Theoretical hypotheses of this study:

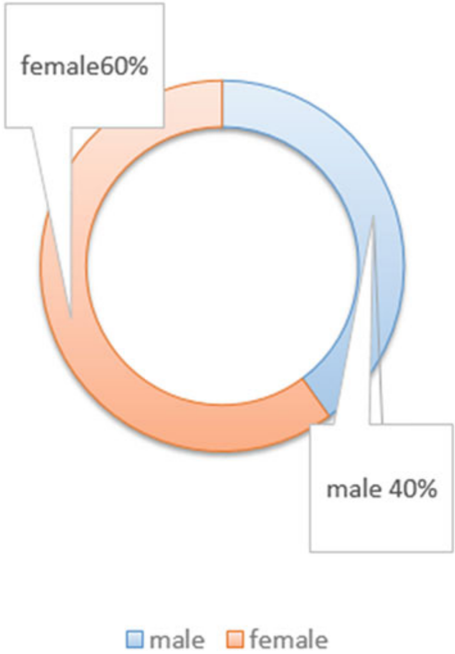
- Indicators of intellectual competence are associated with conceptual abilities; metacognitive abilities as regulatory mechanisms of intellectual activity; intentional abilities;
- The structure of intellectual competence consists of conceptual, metacognitive, and intentional components.

1.2 Purpose of the Study

The subject of this study is the indicators and composition of the construct of intellectual competence, the object of the study is students of the ninth grade of secondary schools, whose intellectual competence is formed in the process of schooling.

Purpose of the study: disclosing the specifics of the construct of intellectual competence in older adolescence using a statistical analysis of research indicators.

Fig. 2 Study participants. Sex differentiation



2 Study Participants

Pupils of ninth grade took part in the study, totaling 100 older teenagers (Figs. 2 and 3).

3 Research methods

3.1 Methodology for Diagnosing Conceptual Abilities: "Conceptual Synthesis" [9]

Indicator: the level of formation of conceptual abilities (the success of constructing really absent connections between concepts).

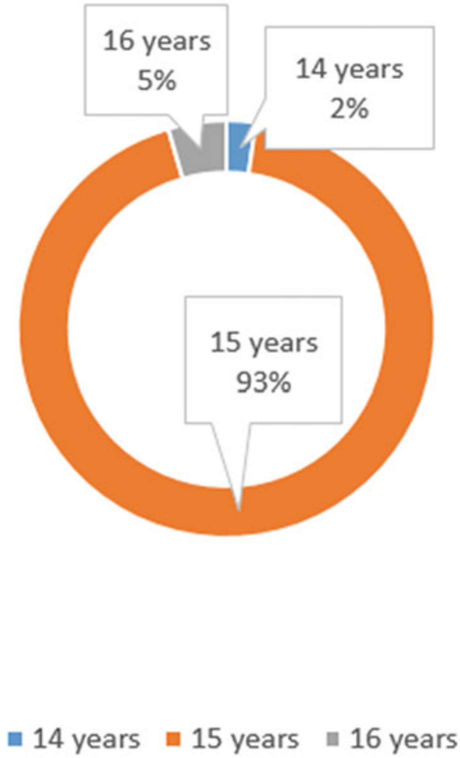


Fig. 3 Study participants. Age differentiation

3.2 Methods for Identifying Metacognitive Abilities:

3.2.1 *"Methods for Diagnosing the Level of Reflexivity"* [10]

Indicator: the level of formation of reflexivity as an aspect of metacognitive (voluntary) abilities.

3.2.2 *"Comparison of Similar Drawings" by J. Kagan* [11]

Indicators of impulsivity/reflexivity as an aspect of metacognitive (involuntary) abilities: (1) latent time of the first response (sum); (2) total number of errors/

3.3 The Author's Method of Diagnostics of Intentional Abilities: "Mood"

Indicators: (1) mentality; (2) beliefs.

3.4 Author's Methodology for Assessing Intellectual Competence: "Interpretation"

Along with an overall assessment of the complexity of the text, a more thorough analysis of the texts was undertaken. The units of the analysis were sentences as text units.

Indicators of the methodology: "Interpretation": the score received by each participant in the study. In particular, a measure of the complexity of the generated text (mental narrative) was assessed; the number of sentences of different types.

4 Results

Despite the data normalization procedure, the distribution does not differ significantly from the normal only for the variables describing metacognitive abilities. Although referring to a significant number of study participants (sample size) and the procedure of artificial data normalization, one can risk applying parametric methods of statistical processing (discriminant, factorial, and regression types of analysis).

4.1 Discriminant Analysis of Conceptual, Metacognitive, and Intentional Components of Intellectual Competence

To determine which indicators—conceptual, metacognitive, or intentional abilities—have predictive power in relation to indicators of intellectual (school) competence, discriminant (stepwise) analysis was carried out.

Three canonical functions were obtained (Table 1) as a result of carrying out

Table 1 Discriminant structure of intellectual competence, including conceptual, metacognitive, and intentional abilities

Descriptors	Functions	λ-Wilks	F-inclusion	p-level
Metacognitive abilities: cognitive accuracy	1	0.684	1.954	0.026
Metacognitive abilities: reflexivity	2	0.490	1.789	0.010

discriminant analysis for five classes.

As follows from Table 1, two types of abilities predict manifestations of intellectual (educational) competence in older adolescence: cognitive accuracy and voluntary metacognitive abilities, and the level of reflexivity. Indicators of conceptual and intentional abilities do not predict the manifestation of learning competence in older adolescence.

The results obtained testify in favor of the complexity of the structure of the construct of intellectual competence—its heterogeneity. The multilevel principle of organizing intellectual competence in older adolescence is emphasized by the size of λ-Wilks, which vary from function to function.

Thus, the manifestation of the productivity of intellectual activity determines reflexivity. In our previous studies [2], the importance of reflexivity for the productivity of intellectual activity was reflected. At the same time, speaking of metacognitive abilities, one should first point out their regulatory function. Our study revealed the descriptive significance of voluntary intellectual control—reflexivity. These data correlate with the studies of A.V. Karpov [10], which postulate the importance not so much of knowledge itself—cognitive abilities—as the ability to manage one's knowledge and organize it to achieve specific goals—metacognition.

At the same time, we were able to identify the descriptive role of involuntary metacognitive abilities in relation to intellectual competence. In the studies of E.Yu. Savin [12], involuntary intellectual control was not correlated with the manifestations of intellectual competence measured in the older group—students and their professors. On the other hand, studies by M.A. Kholodnaya and O.G. Berestneva [13] found that successful senior students are characterized by reflexivity, measured by J. Kagan's method. In this regard, various options are possible to explain the identification of the descriptive role of involuntary metacognitive abilities in relation to intellectual competence. The likelihood of insufficient sensitivity of the applied methodology or an incorrectly accepted unity of the cognitive style "reflexivity/impulsivity" and involuntary intellectual control, meanwhile, disappears.

However, such an explanation is possible, according to which in older adolescence, in connection with the formation of a developed conceptual apparatus, the verbal and arbitrary becomes the leading regulatory channel instead of the earlier figurative and involuntary type of control of intellectual activity. In the situation of the presented empirical research, it may be appropriate to speak of an insufficient or only an initial stage of the transition from involuntary intellectual control to voluntary. So, E.V. Volkova [14] points to the factors of the immaturity of mental constructs, when two options for interpreting the obtained fact are possible: at the maturity of the constructs (it becomes complete differentiated and integrated and isolated in one's experience) and with the immaturity of constructs (the situation of their formation and fragmentation, non-integration and weak differentiation in one's experience).

In this case, it seems more appropriate to talk about the second version of the explanation of the obtained facts, in view of the peculiarities of the age of the study participants (older adolescence as a period of ontological crisis). In older adolescence, there is a transition from one stage of intellectual development to another through a kind of "shapeless intellectual education" [14], when old connections between constructs are destroyed and fundamentally new formations are created that are adaptive in relation to the changed living conditions of the subject of ***activity,*** forming new forms and styles of behavior.

Another way of arguing such results: insufficient automation of metacognitive abilities in older adolescence. This "immaturity" of metacognitive abilities leads to the fact that the study participants "regress" to onto and phylogenetically earlier forms of interaction with the environment [15]. In this situation, the primary place is occupied by involuntary mental processes, while arbitrary regulatory instruments remain in secondary roles.

Thus, in accordance with the results obtained, there is reason to conclude that by the age of adolescence, metacognitive thinking is sufficiently formed to ensure a qualitative increase in the intellectual resources of a teenager. By this age stage, it becomes possible to turn the mental space most completely, although in this area some heterogeneity is found.

4.2 Factor Structure of Intellectual Competence (Conceptual, Metacognitive, and Intentional Components)

Factor analysis (principal component analysis with rotation) was applied in view of the fact that we assumed the existence of a relationship between the variables. This was substantiated theoretically and practically on older samples. In addition, the selection of orthogonal factors does not correspond to the topic of our study.

The procedure of factor analysis itself was carried out in two stages: factor analysis with the participation of a general indicator of intellectual competence and conceptual, metacognitive, and intentional abilities and factor analysis of particular manifestations of intellectual competence (indicators of the ability to generate sentences of various types) and conceptual, metacognitive, and intentional abilities.

During the statistical analysis of data by the method of factor analysis, it was demonstrated that for the first stage of factor analysis, a two-factor model is distinguished (33% of the variance). The results of factor analysis are presented in Table 2.

Table 2 The factor structure of intellectual competence (general indicator), including conceptual, metacognitive abilities

Indicators	Factor–1	Factor–2
Conceptual abilities	0,253	0,679
Intellectual competence	0,765	0,177
Reflexivity	0,468	
Cognitive pace		−0,304
Cognitive fidelity		−0,436

Notes Highlighted loads > | 0.3 |

As follows from Table 2, reflexivity is included in one factor with the general indicator of intellectual competence (general indicator). The second factor is conceptual abilities and both variables of involuntary metacognitive abilities—temporal and precise characteristics. Moreover, attention is drawn to the great importance of the accuracy characteristics of involuntary metacognitive abilities in comparison with temporal ones.

In the statistical analysis of data by the method of factor analysis, it was demonstrated that for the second stage of factor analysis, a five-factor model (53% of the variance) is distinguished, shown in Table 3.

Turning to the conversation about factorization of particular indicators of intellectual competence and conceptual, metacognitive, and intentional abilities, it should be pointed out that, in contrast to the factorization of these same abilities and the general indicator of intellectual competence, in this case the participation of voluntary metacognitive abilities (reflexivity) was not revealed.

First of all, it is worth commenting on the negative load of the indicators of involuntary metacognitive abilities. It is explained by the negative characteristic that was chosen for them—errors as well as the amount of time spent on the task. Semantically inverting these characteristics, it turns out that conceptual abilities are characterized by accurate and rapid task completion as characteristics of involuntary metacognitive abilities. Moreover, accuracy is a more significant characteristic in

Table 3 The factor structure of intellectual competence (particular indicators), including conceptual, metacognitive, and intentional abilities

Indicators	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Conceptual abilities	0.166	0.569	0.548	0.178	0.101
Reflexivity	0.292		0.108		
Cognitive pace		0.144	−0.571		0.13
Cognitive fidelity	−0.151	−0.151	−0.327		
Ability to generate sentences of emotional-evaluative personality type	0.252	0.742		−0.152	0.379
Ability to generate sentences of emotional-evaluative content type	0.718	0.215	−0.187	−0.187	
Ability to generate sentences of interpretive type	0.831	0.339		0.183	0.169
Ability to generate sentences of factual type				0.792	
Ability to generate sentences of arguing type	0.522		0.453	0.187	0.259
Ability to generate interrogative sentences	0.307	0.720			−0.127
The ability to generate sentences of a systematizing type					0.639

Notes Highlighted loads > | 0.3 |

comparison with speed, which brings us back to the conversation about cognitive styles, where reflectivity characterizes intellectual productivity to a greater extent than impulsivity.

Thus, the results allow us to formulate an attitude toward hypotheses: the factor structure of intellectual (school) competence (measured in terms of narrative-interpretation—"Interpretation") includes indicators of conceptual abilities (in terms of conceptual abilities measured by the methodology "Conceptual synthesis"), indicators of metacognitive abilities (in terms of the cognitive style "impulsivity/reflexivity", as measured by the "Comparison of similar pictures" method for involuntary metacognitive abilities and "Methods for diagnosing the degree of development of reflexivity" for arbitrary metacognitive abilities) but does not include indicators of intentional abilities (in terms of mentality and beliefs, measured by the method of "Mind").

4.3 Regression (Multiple) Analysis of the Components of Intellectual Competence (Conceptual, Metacognitive, and Intentional Components)

In order to study the possible influence of conceptual, metacognitive, and intentional abilities on the manifestation of intellectual competence, a regression analysis of the data was carried out. For the analysis, indicators of categorical abilities, voluntary and involuntary metacognitive abilities, and intentional abilities (moods), an indicator of intellectual competence (in points), private indicators of intellectual competence (sentences of various types) were taken.

Two equations were obtained, where:

$\widehat{Y}$—dependent variable represents intellectual competence (overall score);

$\widehat{Y_1}$—dependent variable represents intellectual competence (sentences of different types), D (Da-Dg).

X_1, X_2, …, Xn are independent variables, namely:

X_1 is an indicator of conceptual abilities, A;

X_2 is an indicator of intentional abilities—mentality, Ca;

X_3 is an indicator of involuntary metacognitive abilities—cognitive pace, Bb:

$$\hat{Y} = -0,410 + 0,030X_1 + 0,010X_2 + 0,001X_3, \tag{3.1}$$

$$\hat{Y} = 2,176 + 0,33X_2 \tag{3.2}$$

As you can see from Fig. 4, describing even normalized data by means of regression Eqs. (3.1) and (3.2) does not seem credible.

We conduct further analysis in order to check the adequacy of the resulting model (indicators of intellectual competence and intentional abilities), i.e. compliance with

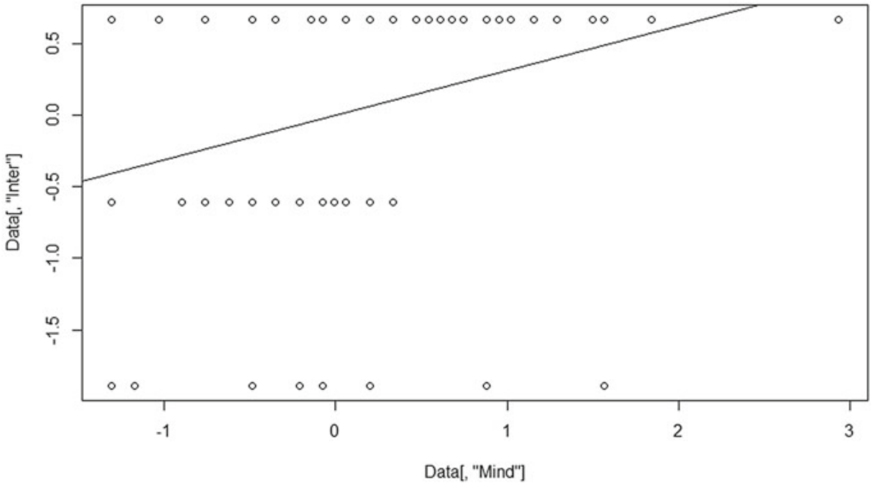


Fig. 4 Graphical representation of the second regression equation with the variables “intellectual competence” as the dependent and “mentality” as the independent variable

the condition of homoscedasticity of errors (constancy of the conditional variance of a vector or a sequence of random variables, and in this case, homogeneous variability of observation values, expressed in the stability of the variance of the random error of the regression model—the variances are the same at all moments of measurement) with zero mathematical expectation, the residuals (Figs. 5–8).

The first shalom of this analysis was the consideration of the indicators of residues in comparison with the obtained values (Fig. 5).

Based on the results shown in Fig. 5, there is reason to conclude that the observed residuals with zero mean are not random and a systematic relationship between the values of the random term in any two observations exists.

Also, as a test of the constructed linear model, a graph was plotted of the dependence of the quantiles of the residuals against the quantiles, which could be expected provided that the residuals are normally distributed, and in the conditions of regression modeling, a normal distribution of errors is assumed (Fig. 6).

In accordance with the assumptions of the model, the points on the graph should lie as close as possible to the regression line, some insignificant deviations can be expected only at the ends, but in this case, the deviations are quite large and there is no normality of the residuals. The model is incorrect.

Fig. 7 presents the analysis of the dependence of the residuals on the shoulder, as one of the options for detecting outliers in the linear regression model, where this indicator, i.e. shoulder, describes how far one covariate is from other to changes in real data.

Based on the analysis of the residuals from the shoulder, where the spread of normalized residuals should not change depending on the shoulder, in the obtained facts there is a decrease in variability, which indicates heteroscedasticity.

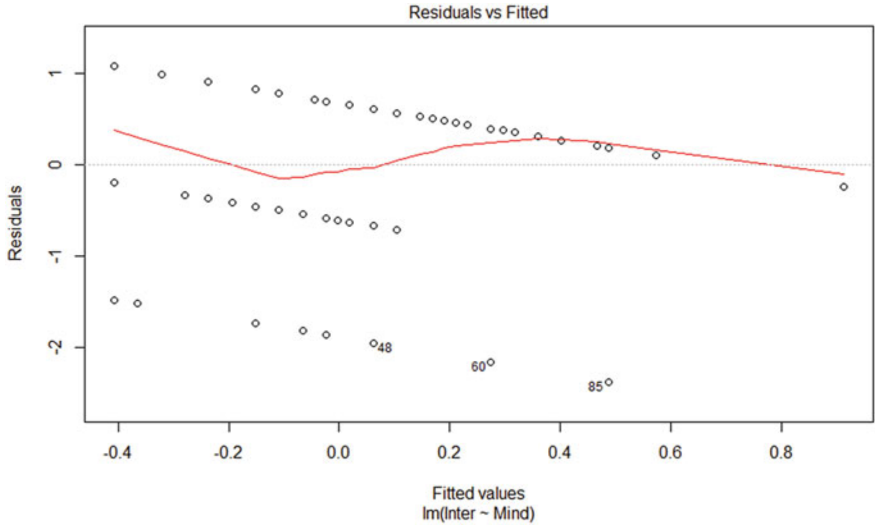


Fig. 5 Graphical presentation of the analysis of the indicators of residuals (indicators of mood) in comparison with the obtained values

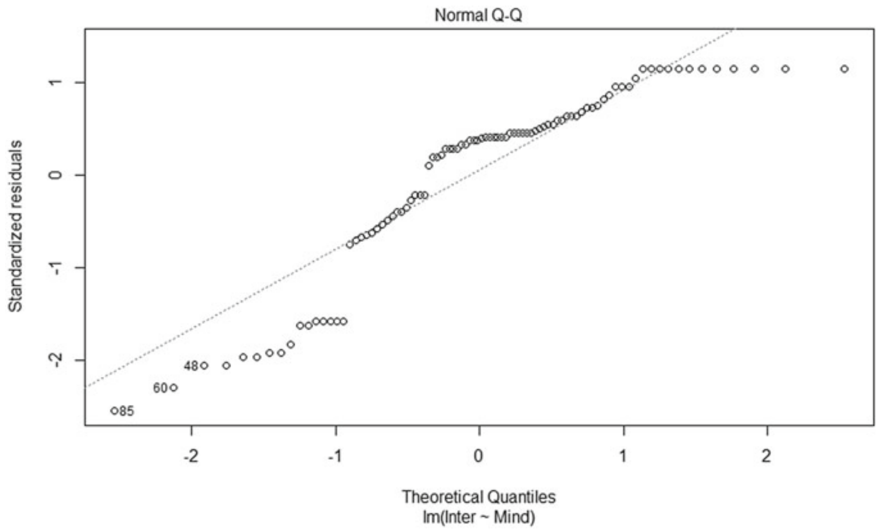


Fig. 6 Graphical representation of residual quantiles versus quantiles

In addition, the study data does not go beyond Cook's line (red dashed line), which indicates that this variable does not significantly affect the results of the regression.

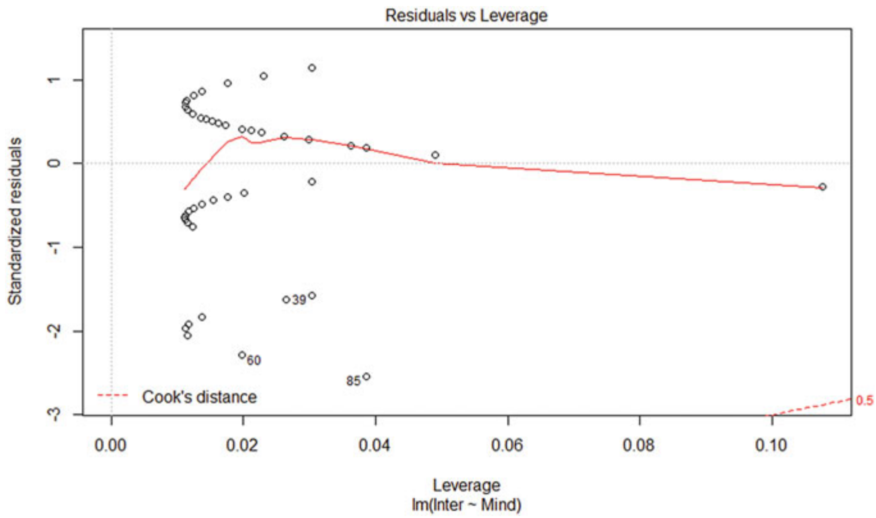


Fig. 7 Graphical representation of leverage balance

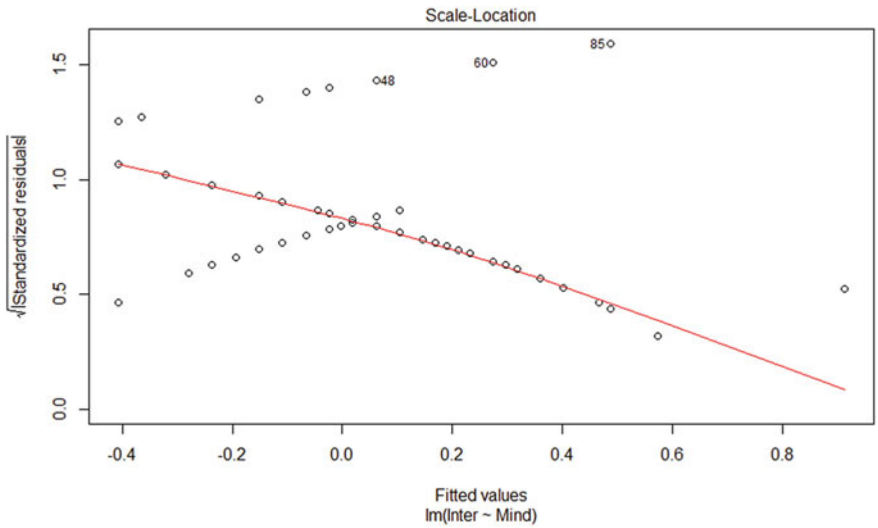


Fig. 8 Graphical representation of the scale location

We also analyzed the scale location plot, which considers the square root of the absolute value of the standardized residuals (Fig. 8). In the resulting graph, the X-axis shows the values obtained in the model, and the Y-axis is the square root of the standardized residuals.

Thus, the description of data by means of regression equations does not appear to be reliable.

5 Findings

Thus, the data allow us to draw a number of general conclusions.

1. Among all the methods of statistical analysis used in the study, data normalization using z-scores, discriminant (stepwise) analysis, factorial (exploratory and confirmatory) analysis with varimax rotation, and regression (multiple) analysis, the following method showed the best results: factorial analysis.
2. The use of regression models of indicators of intellectual competence, the problem of the adequacy, and reliability of the selected equations was discovered, which might indicate the nonlinear nature of the mental and the weak suitability of such methods for analyzing measurement results in the humanities.
3. Conducted using statistical methods, the analysis made it possible to obtain the following significant psychological results regarding the relationship between the indicator of intellectual competence, measured in terms of the characteristics of the mental narrative, with indicators of conceptual abilities, voluntary and involuntary metacognitive abilities in terms of the level of formation of reflexivity and reflectivity, respectively, intentional abilities as the ability to rely on intuitive attitudes in looking for an answer in a situation of uncertainty. Thus, intellectual competence is a factor in the integration of intellectual abilities of different levels of organization of mental activity at the stage of senior adolescence.
4. The most predictive in relation to the assessment of the level of intellectual competence in older adolescence are indicators of metacognitive abilities (reflexivity as one of the types of voluntary regulation of intellectual activity and reflexivity is an indicator of involuntary intellectual control).
5. The factor structure of intellectual competence can be described by indicators of conceptual (conceptual), intentional (mentality and beliefs), and metacognitive abilities (reflexivity and reflectivity).
6. Regression equations of indicators of intellectual competence include conceptual (conceptual) abilities, intentional abilities (mentality), and metacognitive (reflectivity cognitive tempo) abilities.

Acknowledgments State assignment for research work of the IP RAS No. 0159-2019-0008, Institute of Psychology of the Russian Academy of Sciences).

References

1. Kholodnaya MA (2019) Psychology of intelligence, paradoxes of research: a textbook for bachelors and graduate students, 3rd edn. Yurayt Publishing House, Moscow
2. Sipovskaya YaI (2016) Conceptual, metacognitive and intentional abilities in the structure of intellectual competence: diss Cand psychol. sciences, Moscow
3. Ericsson KA, Krampe RT, Tesch-Romer C (1993) The role of deliberate practice in the acquisition of expert performance. Psychol Rev 363–406
4. Raven J (2002) Competence in modern society: identification, development and implementation (translated from English). Kogito-Center, Moscow
5. Kholodnaya MA (2012) The psychology of conceptual thinking: from conceptual structures to conceptual abilities. Institute of Psychology RAS, Moscow
6. Sternberg R (Ed.) (2002) Practical intelligence. SPb.: Peter
7. Sultanova LB (2004) The problem of implicit knowledge in science. USPTU Publishing House, Ufa
8. LS Vygotsky (Ed.) (2005) Educational psychology. AST, Moscow
9. Kholodnaya MA (2002) Psychology of intelligence: paradoxes of research. SPb.: Peter
10. Karpov AV (2003) Reflexivity as a mental property and the method of its diagnosis. Psychol J 24(5):45–57
11. Kagan J, Rosman BL, Day DAJ, Phillips W (1964) Information processing in the child: significance of analytic and reflective attitudes. Psychol Monogr 78(1)(Whole No 578):1–37
12. Savin EYu (2002) Conceptual and metacognitive experience as the basis of intellectual competence. Diss Cand Psychol Scie, Moscow
13. Kholodnaya MA, Berestneva OG, Kostrikina IS (2005) Cognitive and metacognitive prerequisites for intellectual competence in the field of scientific and technical activity. Psychol J 26(1):51–59
14. Volkova EV (2011) Psychology of special abilities: differential-integration approach. Publishing House “Institute of Psychology RAS”, Moscow
15. Aleksandrov YuI, Svarnik OE, Znamenskaya II, Kolbeneva MG, Arutyunova KR, Krylov AK, Bulava AI (2017) Regression as a stage of development. Publishing house of the Institute of Psychology of the Russian Academy of Sciences, Moscow

Enabling Cloud Computing to Facilitate Health Analytics Application from Local Hospitals in Thailand

Praowpan Tansitpong

Abstract This study investigates patterns in electronic medical records (EMRs) in Thailand in terms of prescription and treatment cost from patient data with identification diagnosis for cancer, hypertension, and diabetes. This study developed a comparison model of implementing multiple cloud computing platforms in tracking and monitoring medical records for hospitals with limited database and analysis capacity. This study also suggested an application of health data analytics in identifying prescriptions that violated prescription guidelines for patients with chronic diseases.

Keywords Electronic medical records · Healthcare operations · Prescription pattern

1 Introduction

With the growing demand for health services due to the availability of access to healthcare at large after the Thailand Health Reform in 2004, regional hospitals in Thailand are able to reach more local patients through multiple government-subsidized benefit plans which enable large data collection of electronic medical records (EMRs). While most of the hospitals are implementing traditional database management, these hospitals are facing challenges in handling big data and choosing appropriate technology to improve health service performance. EMRs facilitate various usage in the medication process, including patients, doctors, nurses, etc. This medical technology today has been developed substantially as the availability of patient data information increases dramatically. Former methods of recording information in a paper-based document may not be efficient for searching patient's information. The EMR system can be accessed and recorded by multiple users at the same time. The users can access the network from any place. The system provides an organized document than a

P. Tansitpong (✉)
NIDA Business School, National Institute of Development Administration (NIDA), 118 Seri Thai Rd, Bangkok 10240, Thailand
e-mail: praowpan.tan@nida.ac.th

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_20

paper document and gives a high performance of storing, managing, and retrieving information.

This study focuses on the full spectrum of government-funded programs by gathering data at the hospital level. The purposes of this study are to investigate variations in prescriptions related to benefit plans, categorization by frequency, and variety of prescriptions under government guidelines. This study aims to identify significant similarities and differences in prescribing medications and the cost of treatments in multiple benefit groups and summarizes decisions for prescriptions. Thailand has introduced Universal Coverage Scheme (UCS) in 2001 which provided access to healthcare to approximate 47 million or 75% of the population, complementing the existing Civil Servant Medical Benefit Scheme (CSMBS), which includes approximately 4 million government and state enterprise employees, and Social Security Scheme (SSS) which is mandatory health insurance for 10 million employees in private sectors. Under the establishment of the Universal Coverage Scheme, 99.5% of the Thai population has health protection coverage [1]. According to Brady et al. [22], benefits are included in the three benefit schemes. The highest spending is found in the Civil Servant Medical Benefit Scheme (CSMBS) list of guidelines, which includes all treatment and intervention, all prescriptions within and beyond the government suggested list, and the lower spending is the Universal Coverage Scheme (UCS) [2]. Social Security Scheme (SSS) spending is found to be between the other two programs [3, 4]. The goal of this study is to utilize data mining techniques to identify prescription patterns in health-administered programs and to clarify whether health insurance and benefit plans play role in changing the decision of drug choices for treating chronic diseases patients in Thailand.

2 Literature Review

Electronic medical records (EMRs) facilitate modern medical workers in understanding the health process. The records also provided opportunities for medical society to observe the medical process and improve efficiency and utilization [5–9]. While most of the studies in the past were focusing on identifying phenotypes, seeking algorithms for comorbidity, and exploring personalized prescriptions for individuals, less attention was paid to understand the patterns in prescription behavior and equity of care. By using data analytics in health applications, previous findings suggested applications such as identifying phenotypes, seeking algorithms for comorbidity, and personalized prescriptions [10–13]. While some operational decision patterns cannot be observed or classified during normal practice, the EMRs usage reveals some structured decisions made by physicians using data analytics [14–17].

As chronic diseases have become the major national concern, monitoring the standard of treatment requires a cooperative database shared among national public health administer, healthcare providers, insurance companies, and pharmacies. Cancer treatment is the most difficult process and yields the highest cost of treatments [18]. As oncology is seeking the best delivery of care, the current challenges are to find the most efficient and appropriate cancer care. Less than 1% of the population is commercially insured, but they account for approximately 10–12% of healthcare expenditures. The cost of cancer care is rising at an unsustainable rate of 15–20% annually [19]. Up to 85% of cancer care delivery was provided in community-based medical oncology practices. A study of reimbursement medical expenditure [20] suggested medical oncology practices assume an increasing degree of responsibility for navigating patients through the complex, the availability of electronic medical records (EMRs), and the actionable information that can be mined from clinical databases to understand the underlying pattern in past cancer treatment.

In individual healthcare services, the patient medical data rate is growing up every single day. The data such as the patient registration, laboratory test, prescription, and payment collected in the hospital's database can be potentially very large if the number of visiting patients per day is in the order of several hundreds. The larger the number of patients, the larger is the number of resources required for the healthcare IT service. Most of the hospitals in Thailand use an EMR system, which is a standalone operation that causes a high maintenance cost. Big data ecosystem refers to the architecture components with their operational models in big data systems in terms of storage, processing, and analytic. For storage, the traditional framework or software use in storing the data is not optimized for storing big data. For example, a rational database management system (RDBMS) requires that all the clearly attributes are clearly defined, and also the type of data [21]. Managing health information can be a great challenge as a criterion of the collected data can be changed to match the business needs; hence, a suitable storage type to reduce the workload of the server and increase system efficiency is required [22]. An example of a structure of the medical data process is shown in Fig. 1. For processing, the data batch and streaming are two types of processes in big data. These types of data take a long time in processing because it has to access all files and compute large data. In contrast, the streaming process computes data independently in real time. For analytic purposes, the amount of data in big data need to be studied and processed before mining the data for patterns [23].

Thailand has introduced Universal Coverage Scheme (UCS) in 2001 which provided access to healthcare to approximate 48 million or 75% of the population, complementing the existing Civil Servant Medical Benefit Scheme (CSMBS), which includes approximately 5 million government and state enterprise employees, and Social Security Scheme (SSS) which is mandatory health insurance for 11 million employees in private sectors. Under the establishment of UCS, CSMBS, and SSS, 99.5% of the Thai population has health protection coverage. Healthcare costs can be attached to extensive medical expenses for patients [24–26]. Some argue that healthcare has become expensive in the United States because of the use of more expensive technology [27, 28]. Decisions made by physicians in the treatment process are critical; however, one physician may have different decisions than others. Participation with patients does share influences in the medical treatment decision-making process [29–32]. Due to rising prescription drug costs, the literature provided supports that tier of pharmacy benefit plays a significant role in prescription [33–35]. Such arguments shaped a focal question of this study to seek underlying patterns of prescription

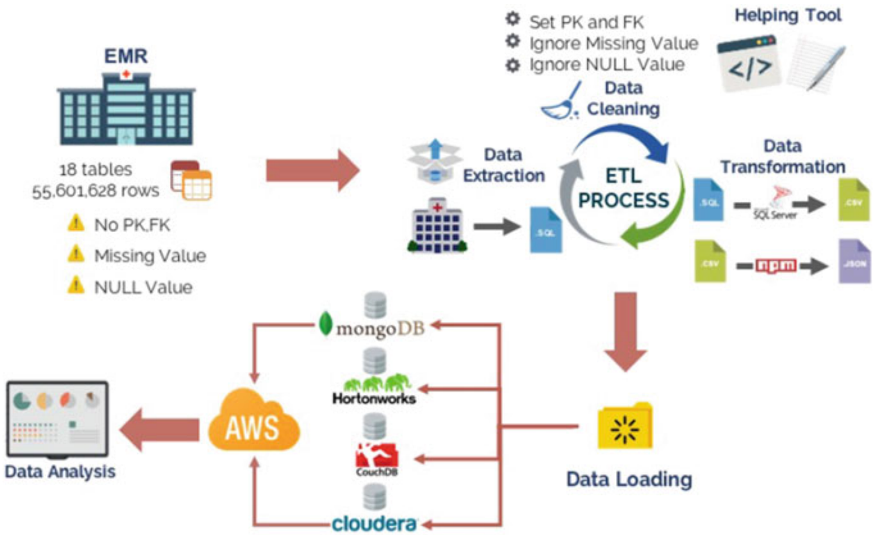


Fig. 1 Structure of data process

drugs made by physicians. Decision patterns can be formulated from clusters found in the usage of EMRs. This study aims to identify significant similarities and differences in prescribing medications and the cost of treatments in multiple benefit groups and summarize decisions for prescriptions.

3 Methodology

3.1 Data Collection

This study combines big data EMRs and deploys the data with cloud computing techniques in two phases. In the first phase, the EMR data were deployed in comparative programs of SQL database, Cloudera and Hortonworks Sandbox, and independent NoSQL databases, MongoDB and CouchDB. This process of testing data stability contributed to the programs for the most outstanding performance option, such as the speed of execution time, the stability of the service, and the program's ease of use will be chosen for the later implementing on computing process of this study was that disease-related cases could be a determining factor of inequality treatments on prescription for the patient a given insurance or treatment plan. Figure 1 summarizes the software that is adopted in this study. This system can be divided into three main parts: storing data, MapReducing, and deploying data. The process began with receiving hospital data management of patient data. In this process, the information is collected in .csv (comma-separated value) format. The next process is followed by implementing MongoDB as a core software to organize each patient data appropriately. Robomongo software, a shell-centric across platform MongoDB management tool, is used to display all fields via Admin GUI. MongoDB is classified as a NoSQL database that is able to handle real-time, scalable, and tremendous data size by contributing a shorter time to process or retrieve the data. The next process is to transfer patient data from the hospital into MongoDB and convert the .csv file to .JSON file, in which the values are sliced and stored as the attribute's values column.

The data collected from the regional part of Thailand is shown in Fig. 2. The electronic records were transferred from a local store database to an assigned internet protocol (IP) address. The database consisted of 18 isolated tables in structured query language (SQL) format. In the data cleaning process, tables from patient positive identification, the diagnosis codes, treatments and prescriptions, test results, and other related health information contain patient records under three main

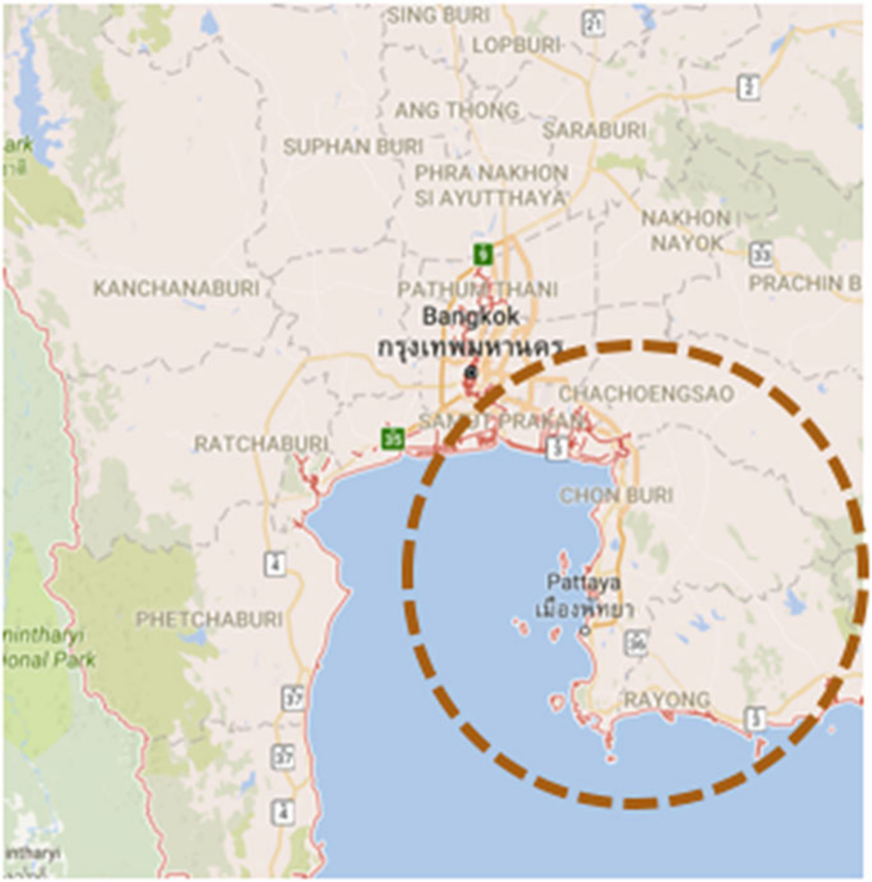


Fig. 2 Location of data collection

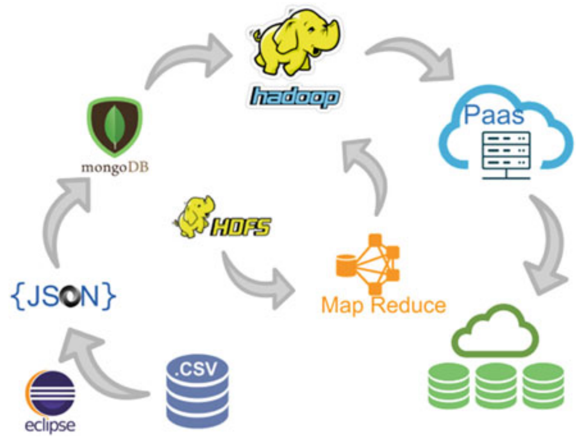


Fig. 3 Example of software used in the process of data extraction

government-subsidized coverage programs (Security Scheme (SSS), Civil Servant Medical Benefit Scheme (CSMBS), Universal Coverage Scheme (UCS)) and two others (private insurance holder and out-of-pocket payers). Electronic information service was created to cross-list chronic disease patients with multiple arrays of International Statistical Classification of Diseases and Related Health Problems (ICD-10) codes for the key selected chronic diseases (hypertension, cancer, and diabetes).

The overview of the data migrating process is summarized in Fig. 3. The script formats were transformed from SQL to JavaScript Object Notation (JSON). The data cleaning process includes dropping missing values and null values. Some translations from string scripts of prescription dosage and units are required. The cleaning process also includes the separation of number and text units. Translation requires open-source software (e.g., MongoDB, Hortonworks, CouchDB, and Cloudera) to convert files in comma-separated values (.csv) format to JSON. The relationship diagram among 18 tables is shown in Fig. 3. The cloud service of Amazon Web Service (AWS) is chosen for deploying data and managing this data ecosystem.

The following process is to deploy a java program that determined the values by focusing on comparing the number of commas with the columns contained in one table and checking values in specific columns with the real data by using the Eclipse development environment. To have a desirable performance in data processing algorithm, this study utilizes MongoDB and Hadoop technology to generate MapReduce function by running on different nodes of clusters, or "Hadoop Distributed File System". This process is designated to execute the data in parallel between two software. In this approach, the primary keys are tallied with the relevant values and reduced into one column. Due to a conflict during the Mongo–Hadoop connector installation that led to the failure in applying MapReduce, this study also compares other programs such as Mortar and Hortonworks. In the last process, the system is designed to be a platform-as-a-service host on cloud service to access and store large data organization. This process of cloud storage has been shown to reduce the resources used in hospitals, especially on the budget spending on the data storage due to the scalability and flexibility.

3.2 Data Query for Cancer, Hypertension, and Diabetic Patients

The electronic records were transferred from a local store database that includes 55,601,628 rows from 18 tables in structured query language (SQL) format. The data were primarily stored separately for departmental use. The data conversion process included identification of primary key (PK) and foreign key (FK) in all tables to access information in diagnosis and health-benefit details. The overview of the data migrating process is summarized. The script formats were transformed from SQL to JavaScript Object Notation (JSON). The structure of the data is adjusted by setting the PK and FK from the original data in order to relate each table. To set the PK and FK, the most unique attribute was set to be the PK of each table and if the PK of the same table contains another table, that attribute was set to be FK for another table instead of PK. One example of known PKs are "HN" or patient identification number, "c diag" or diagnosis number, and "s pharm" or medicine number, etc. An example of the filing structure diagnosis process is shown in Fig. 4. The data cleaning process includes dropping missing values and null values. Some translations from string scripts of prescription dosage and units are required. Data query for cancer patients is shown in Fig. 5.

3.3 Data Cleaning Process

The cleaning process also includes the separation of number and text units. Translation requires open-source software (e.g., MongoDB, Hortonworks, CouchDB, and Cloudera) to convert files in comma-separated values (.csv) format to JSON. The relationship diagram among 18 tables is shown in Fig. 6. Cloud service endorsed by Amazon Web Service (AWS) is chosen for deploying data and managing this data ecosystem. By using the extracted dataset of patients, the top three medicines are dispensed each year from 2014 to 2016. Focusing on treatment prescriptions, the National List of Essential Medicines (NLME) provides a list of drugs that are necessary for the prevention and control of all major diseases; the list was first introduced in 1972 and the current update was made in 2016. The initial aim for NLME was to prevent the irrational use of medicines and to control the overall cost of prescribed drugs. However, the CSMBS program allows three physicians to *co-endorse the use of the drug outside the list if required for treatments*. Violations could be seen in three ways: positive violations (did not comply with the guideline), neutrals or follow the guideline, and negative violations (override the guideline and prescribed others).

Table Name: diagnosis				Table#1		
Table Type: Master						
Description: -						
Field Name	**Type**	**Length**	**Description**	**Key**	**Reference**	**Null**
c_diag	nvachar	5		PK		
diag_seq	int	4		PK		
c_dept	nvarchar	2				
c_sub_dept	nvarchar	2				
diag_name	nvarchar	250				
diag_name_thai	nvarchar	250				
group_icd	int	4				
valid_sex	nvarchar	1				
valid_fr_age	int	4				
valid_to_age	int	4				
c_icd_grp1	nvarchar	3				
c_icd_grp2	nvarchar	3				
c_icd_grp3	nvarchar	3				
status	nvarchar	1				
upd_date	int	4				
upd_time	int	4				
upd_by	int	4				
comorbidity_flag	nvarchar	1				✓
external_flag	nvarchar	1				✓
select_chk	nvarchar	1				✓
c_diag_como	nvarchar	5				✓

Fig. 4 Example of filing structure diagnosis process

```
select s_pharm ,c_pharm,trade_name,count(distinct hn)as amount from
(select distinct d.c_diag , d.diag_seq , d.diag_name,d.diag_name_thai,m.hn ,f.c_pttype,
f.pttype_name,p.s_pharm, p.c_pharm,p.trade_name,pl.sale_price,p.presentation,ph.order_date,ph.s_doctor
from phipdl pl
join phipd ph on pl.hn = ph.hn and pl.an = ph.an and pl.nb_flag = ph.nb_flag
     and pl.ph_order_date = ph.ph_order_date
     and pl.ph_order_time = ph.ph_order_time
join msg_finaldiag m on m.hn = pl.hn and m.an = pl.an
join diagnosis d on d.c_diag = m.c_diag and d.diag_seq = m.diag_seq
join phmaster p on p.s_pharm = pl.s_pharm
join fright f on f.c_pttype = ph.c_pttype) a
where c_pttype = 1500 and diag_name_thai like'%มะเร็ง%' and order_date  between 25540101 and 25541231
group by s_pharm,c_pharm,trade_name
order by amount DESC
```

Execute | Save as... | Explain | Format | or create a | New query

Fig. 5 Query of cancer treatment

Presentation
100 MG
15 MG
5 MG/ML
50 MG/ML
30 MG/ML
137.155 MG
0.05 MG/ML
10 MG/ML
100 MG/2 ML
200 MG
400 MG
200 MG
0.736111111
350 MG
300 MG
300 MG

→

Presentation	unit
100	MG
15	MG
5	MG/ML
50	MG/ML
30	MG/ML
137.155	MG
0.05	MG/ML
10	MG/ML
100	MG/2ML
200	MG
400	MG
200	MG
0.73611111	
350	MG
300	MG
300	MG

Fig. 6 Example of cleaning units

3.4 Relational Diagram and Violation Mapping

A relational diagram shown in Fig. 7 provides a linkage of patient identification to diagnosis and prescriptions. For observations in terms of prescriptions, in-patient visits were extracted for all diagnosis codes. Prescription records were tabulated

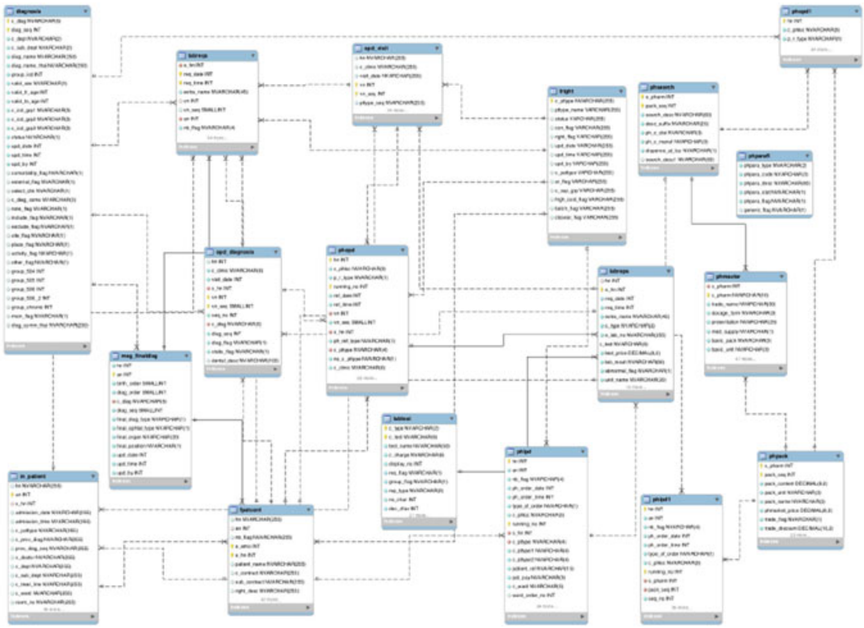


Fig. 7. A complete relational diagram

into specific codes for diabetes (E.08x–E11x), essential primary hypertension, and a number of patients in both groups. The table of guideline violations is extracted from the prescription table and administered program. The violation mapping is done by taking the flag information from three administered programs with the label "Y" shown in Fig. 8. The conceptual model visualized relationships of the impact of the number of patients, cost of treatment, and total dispenses on the number of violations. Regression analysis is used to test the impact of the number of cases, cost of treatment, and amount of prescription items on the number of violations that physician has violated from the guideline.

c_pharm	trade_name	natio	thirty_	social_	gover	unit_price	Diabetes_30	Cancer_30	Hyper_30	Diabetes_ssn	Cancer_ssn	Hyper_ssn	Diabetes_go	Cancer_gov	Hyper_gov
1DAPS100		A	Y	Y	Y	3.2	23	0	11	5	0	7	6	2	1
1DICLO500		A	Y	Y	Y	1	23	0	11	5	0	7	6	2	1
1DILTI120	CARDIL	N		Y	Y	139	23	0	11	5	0	7	6	2	1
1EENO160	SUPRALIP	N		Y	Y	90	23	0	7	5	2	7	6	0	2
1CANDE8	BLOPRESS	N		Y	Y	42	0	0	19	0	0	25	2	2	2
1ENTA200	COMTAN	G			Y	26	23	0	7	5	2	5	6	0	0
1ESOME20	NEXIUM	N			Y	183	23	0	7	5	2	5	6	0	0
1ETORI60	ARCOXIA	N		Y	Y	353.	23	0	7	5	2	5	6	0	0
1FERRS200		A	Y	Y	Y	5.5	23	0	7	5	2	5	6	0	0
1FEXO60	TELFAST	D			Y	34	23	0	7	5	2	5	6	0	0
1FULLER		A	Y	Y	Y	128.4	23	0	7	5	2	5	6	0	0
1GABA100	NEURONTIN	D	Y	Y	Y	26	23	0	7	5	2	5	6	0	0
1GABA600	NEURONTIN	D	Y	Y	Y	26	23	0	7	5	2	5	6	0	0

Fig. 8 Identification for violations

4 Results

The electronic records were transferred from a local store database that has 55,601,628 rows from 18 tables in structured search language (SQL) format. The dataset was primarily stored separately for departmental use. The conversion process included the identification of primary key (PK) and foreign key (FK) to access information in diagnosis and health-benefit details.

A descriptive summary statistics suggested per-visit cost of treatments covered by government-subsidized programs. The violation is coded as 0 if the prescription is listed in the NLME list, −1 if the physician did not prescribe the prescription in step with the guideline (no prescriptions), and +1 if the physician has overridden the rule by prescribing other medications apart from the NLME list. A descriptive summary shows the most of two overridden frequencies from two administered programs and also the maximum of three violations from the guidelines. The statistic summary for chronic disease treatments is displayed in Table 1, which summarizes the proportion of the highest five prescriptions. The descriptive summary also suggested the uniqueness of Thailand's healthcare benefits coverage characteristic in covering a spectrum of administered health services (USC, SSS, and CSMBS).

For observations in terms of prescriptions, in-patient visits were extracted for all diagnosis codes. The regression model is utilized to identify variations in groups with the reimbursement allowance program. The model estimates violation prediction from the cost of treatment and total dispenses in multiple health benefit plans. The association between the programs and disease characteristics was determined by the tendency of prescription violations. The regression results (Table 2) reported significant coefficients with $p < 0.05$ and ordinary errors were summarized during this section. The regression results suggested a significant positive coefficient of the number of diabetic cases and also the violation of cancer prescriptions consistent with the guideline (0.278***) in the universal health coverage (UCS) program. The results also suggested a significant positive impact of the number of patients on the violation (or overriding prescriptions) on hypertension (0.300**) and diabetes treatments (0.206***). These results of positive coefficients may be observed overridden guideline which aligns with the availability of physicians to co-endorse the use of the drugs outside the list if required for treatments. The estimates obtained from the

Table 1 Summary of the number of cases, cost of treatment, total dispenses, and violations

Variable	Obs	Mean	Std. Dev	Min	Max
Diabetes	918	15.96083	241.8615	0	7334
Cancer	918	13.7432	208.2378	1	6315
Hypertension	918	17.52283	477.8835	0	14,473
Cost of treatment (THB)	918	150.7206	440.4111	0.7	5007.6
Total dispenses	918	15.7658	13.0741	1	55
Total violations	918	−0.3551198	1.129669	−3	2

Table 2 Regression results

	Number of violations
No. of cases	
Diabetes	0.206***
	(0.0534)
Cancer	0.278***
	(0.0746)
Hypertension	0.300**
	(0.0927)
Cost of treatment ($)	0.000237
	(0.000180)
Total dispenses	−0.103**
	(0.0364)
_cons	−1.733***
	(0.145)
N	918
R^2	0.447
adj. R^2	0.444
F	145.4

violation prediction model indicated the full cost of treatment and total dispenses positively impact the number of violations. However, the coefficient estimates a negative relationship between cancer patients and the tendency of violations in Thailand's health benefit program. This result suggests that the overriding in prescription decisions and physicians are not influenced by written guidelines but health program has influenced the system by increasing frequency of violations.

5 Conclusion

This study has introduced methods for the integration of big data technologies and cloud computing in practical use and reasonable support in migrating data from a local hospital's database to the cloud's database. The suggestion focused based on each software performance and funding optimization on an actual cloud service deployment. This study investigates prescription diversity in patients diagnosed with cancer symptoms by mapping prescription patterns from electronic medical records to understand the impact of prescription diversities on patients. This study is able to confirm that although guidelines for prescriptions are provided, the decisions were influenced by health benefit programs in treatments of chronic diseases.

Standard care procedures and prescription patterns are significantly unique among government-subsidized benefit schemes in Thailand (Universal Coverage Scheme

(UCS), Social Security Scheme (SSS), Service Medical Benefit Scheme (CSMBS)), and private insurance. The analyses also reported non-uniform prescription patterns of three major non-communicable diseases (diabetes, cancer, and cardiovascular) and three local diseases located in the in-patient and outpatient database. As the decision in prescribing treatment for cancer is known, large variation in drug choice, eligibility of reimbursement, and the results from the regression model suggested that physicians are more likely to ignore guidelines for Universal Coverage Scheme, however, they are more likely to comply with the guidelines when prescribing prescriptions to the government benefit holders. The regression results gave examples that physicians are more likely to override guidelines on hypertension patients. This study suggests options for using cloud technology in managing health analytics of local EMR. The results from this study portrayed a usage of EMR in deploying cloud services in comparing cloud computing software that showed benefits in speed, availability of execution time, stability, ease of querying, reduce the cost of initial database investment, and obtain scalability and usability of database management compared to traditional systems. Moreover, this study is able to confirm an application of the health informatics process in terms of understanding prescription patterns in the Thailand medical service.

Records for diagnostic identification (ICD-10) and prescriptions were collected from the eastern region of Thailand along with prescription guidelines provided by the public health administration. The results confirmed that deductible levels do not have an impact on the cost of treatment but unique characteristics of prescription frequencies indicated differentiation in the benefit schemes. Most of the prescriptions were being selected, some prescription items violated the administered guidelines. The results from regression analysis also suggested that violations are significantly different when it comes to treatment programs. The results from this study are able to suggest an application for health informatics that although guidelines for prescriptions are provided to physicians, they prefer individual knowledge in prescribing treatments on chronic diseases, and violations could be found and compared among chronic disease treatments. The implication of this study should be able to shed light on future research to explore further the interface between an economic model of social health benefit and coverage in conjunction with healthcare service operations.

References

1. Suraratdecha C, Saithanu S, Tangcharoensathien V (2005) Is universal coverage a solution for disparities in health care?: findings from three low-income provinces of Thailand. Health Policy 73(3):272–284
2. Teerawattananon Y, Tantivess S, Yothasamut J, Kingkaew P, Chaisiri K (2009) Historical development of health technology assessment in Thailand. Int J Technol Assess Health Care 25(S1):241–252
3. Tansitpong P, Chaovalitwongse W (2018) An investigation of differentiated prescription decision on profitability: a case study from Thailand. Int J Healthc Manag 11(1):44–51

4. Ross SE, Moore LA, Earnest MA, Wittevrongel L, Lin CT (2004) Providing a web-based online medical record with electronic communication capabilities to patients with congestive heart failure: randomized trial. J med Internet Res 6(2)
5. Ralston JD, Rutter CM, Carrell D, Hecht J, Rubanowice D, Simon GE (2009) Patient use of secure electronic messaging within a shared medical record: a cross-sectional study. J Gen Intern Med 24(3):349–355
6. Wald JS, Businger A, Gandhi TK, Grant RW, Poon EG, Schnipper JL, Middleton B (2010) Implementing practice-linked pre-visit electronic journals in primary care: patient and physician use and satisfaction. J Am Med Inform Assoc 17(5):502–506
7. Simon SR, Kaushal R, Cleary PD, Jenter CA, Volk LA, Orav EJ, Burdick E, Poon EG, Bates DW (2007) Physicians and electronic health records: a statewide survey. Arch Intern Med 167(5):507–512
8. King J, Patel V, Jamoom EW, Furukawa, MF (2014) Clinical benefits of electronic health record use: national findings. Health Servc Res 49(1pt2):392–404
9. Kohane IS (2011) Using electronic health records to drive discovery in disease genomics. Nat Rev Genet 12(6):417
10. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Jensen LJ (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 7(8):e1002141
11. Wilke RA, Xu H, Denny JC, Roden DM, Krauss RM, McCarty CA, Savova G (2011) The emerging role of electronic medical records in pharmacogenomics. Clin Pharmacol Ther 89(3):379–386
12. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM (2014) A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 21(2):221–230
13. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17(3):37–37
14. Piatetsky-Shapiro G, Brachman RJ, Khabaza T, Kloesgen W, Simoudis E (1996) An overview of issues in developing industrial data mining and knowledge discovery applications. In: *KDD*, vol 96, pp 89–95
15. Milovic B (2012) Prediction and decision making in health care using data mining. Kuwait Chap Arabian J Bus Manag Rev 33(848):1–11
16. Ţăranu I (2016) Data mining in healthcare: decision making and precision. Database Syst J 6(4):33–40
17. Hess LM, Michael D, Mytelka DS, Beyrer J, Liepa AM, Nicol S (2016) Chemotherapy treatment patterns, costs, and outcomes of patients with gastric cancer in the United States: a retrospective analysis of electronic medical record (EMR) and administrative claims data. Gastr Cancer 19(2):607–615
18. Sprandio JD (2010) Oncology patient-centered medical home and accountable cancer care. Commun Oncol 7(12):565–572
19. Lin HC, Wang Z, Boyd C, Simoni-Wastila L, Buu A (2018) Associations between statewide prescription drug monitoring program (PDMP) requirement and physician patterns of prescribing opioid analgesics for patients with non-cancer chronic pain. Addict Behav 76:348–354
20. Sánchez-de-Madariaga R, Muñoz A, Lozano-Rubí R, Serrano-Balazote P, Castro AL, Moreno O, Pascual M (2017) Examining database persistence of ISO/EN 13606 standardized electronic health record extracts: relational vs. NoSQL approaches. BMC Med Inform Decis Mak 17(1):1–14
21. Chen R, Enberg G, Klein GO (2007) Julius–a template based supplementary electronic health record system. BMC Med Inform Decis Mak 7(1):10
22. Brady TM, Neu AM, Miller ER III, Appel LJ, Siberry GK, Solomon BS (2015) Real-time electronic medical record alerts increase high blood pressure recognition in children. Clin Pediatr 54(7):667–675

23. Ruiz P, Venegas-Samuels K, Alarcon RD (1995) The economics of pain: mental health care costs among minorities. Psychiatr Clin North Am 18(3):659–670
24. Balsa AI, Cao Z, McGuire TG (2007) Does managed health care reduce health care disparities between minorities and Whites? J Health Econ 26(1):101–121
25. Szczepura A (2005) Access to health care for ethnic minority populations. Postgrad Med J 81(953):141–147
26. Hanchate A, Kronman AC, Young-Xu Y, Ash AS, Emanuel E (2009) Racial and ethnic differences in end-of-life costs: why do minorities cost more than whites? Arch Intern Med 169(5):493–501
27. Roos NP, Shapiro E, Tate R (1989) Does a small minority of elderly account for a majority of health care expenditures?: a sixteen-year perspective. Milbank Q 347–369
28. Braddock CH, Fihn SD, Levinson W, Jonsen AR, Pearlman RA (1997) How doctors and patients discuss routine clinical decisions: informed decision making in the outpatient setting. J Gen Intern Med 12(6):339–345
29. Flynn KE, Smith MA, Vanness D (2006) A typology of preferences for participation in healthcare decision making. Soc Sci Med 63(5):1158–1169
30. Fraenkel L, McGraw S (2007) What are the essential elements to enable patient participation in medical decision making? J Gen Intern Med 22(5):614–619
31. Amjad H, Carmichael D, Austin AM, Chang CH, Bynum JP (2016) Continuity of care and health care utilization in older adults with dementia in fee-for-service Medicare. JAMA Intern Med 176(9):1371–1378
32. Motheral B, Fairman KA (2001) Effect of a three-tier prescription copay on pharmaceutical and other medical utilization. Med Care 39(12):1293–1304
33. Ganther-Urmie JM, Nair KV, Valuck R, McCollum M, Lewis SJ, Turpin RS (2004) Consumer attitudes and factors related to prescription switching decisions in multitier copayment drug benefit plans. Am J Manag Care 10(3):201–208
34. Priya K, Joy N, Thottumkal AV, Warrier AR, Krishna SG, Joseph N (2018) Impact of electronic prescription audit process to reduce outpatient medication errors. Indian J Pharm Sci 79(6):1017–1022
35. Winter JD, Kerns JW, Winter KM, Sabo RT (2019) Increased reporting of exclusionary diagnoses inflate apparent reductions in long-stay antipsychotic prescribing. Clin Gerontol 42(3):297–301

Neural Network-Based TID Controller for Wheeled Mobile Robot Trajectory Tracking

Najva Hassan and Abdul Saleem

Abstract This brief proposes a novel control scheme for tracking a differential drive wheeled mobile robot. A neural network-based tilt integral derivative controller is introduced to track the wheeled mobile robot. Conventional control algorithms fail to satisfy the desired performance in terms of accuracy, total control effort, and maximum control input. By properly training a neural network, optimal values of the parameters of the tilt integral derivative (TID) controller are obtained. The proposed controller is used to track circular, lemniscate, straight line, and B spline trajectories. Performance of the proposed controller is compared with conventional TID and PID controllers using different performance analysis indices. Simulation studies show that the integral squared error and total control effort are small for the proposed controller as compared to the TID and PID controllers.

Keywords Trajectory tracking · Mobile robot · PID controller · TID controller · NN-based TID controller · Kinematics · Tracking accuracy

1 Introduction

Robots are replacing humans in various fields in order to achieve accuracy, efficiency, and speed as part of modernization. The terrestrial robots are classified as wheeled, stationary, and legged robots. Wheeled mobile robots (WMR) are less complex, cheaper, and have less balance issues as compared with legged robots. Motion control algorithms are required to track the trajectory planned by the guidance systems. Highly robust control algorithms ensure a safe navigation in restricted environments.

In literature, a number of control algorithms have been proposed. In [1], an output feedback controller is proposed using adaptive sliding mode controller and is robust against un-modeled dynamics. But the tracking speed is affected by the measurement noise. In [2], the desired trajectory is exponentially converged with minimum tracking

N. Hassan (✉) · A. Saleem
Department of Electrical and Electronics Engineering, Government Engineering College (Affiliated to APJ Abdul Kalam Technological University), Thrissur, Kerala, India

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_21

error but model uncertainties, effect of saturation, and actuator failures during motion are not considered. A controller in which the gain of the proportional controller is adaptively tuned is developed in [3]. The tracking accuracy is poor in the presence of noise and also stability is not ensured. An adaptive PID controller [4] is used to increase the flexibility of the control system. Only the tracking accuracy is focused rather than considering the stability of the vehicle body control system. Fuzzy logic is used to design a controller for the safe Robotino navigation [5]. A fuzzy-based PID controller [6] is proposed to track the path of a WMR. Tracking accuracy is improved while compared to the PID controller but the time taken to track the desired trajectory is more. Fractional order calculus and sliding mode controller are combined in a F-AIHMC [7]. High accuracy and high precision motion of the robot are ensured here. A combination of PI controller and neural network controller is used in [8] to control the omni-directional WMRs and the tracking error is reduced as compared to conventional PID. A backstepping and fuzzy sliding mode controller (BFSMC) is proposed in [9]. The desired trajectory is tracked even in the presence of external disturbances and model uncertainties.

In this work, an NN-based TID controller is developed for trajectory tracking of WMRs and the performance is evaluated. PID controllers are widely used in industrial process control owing to its simplicity and easy implementation. TID controllers are popular due to better capability for disturbance rejection, simpler tuning method, and lesser control effort as compared to PID controllers. This paper compares the performance of the proposed controller with these two popular controllers. WMR kinematics is explained in Sect. 2. The proposed controller is described in Sect. 3 followed by the results in Sect. 4 and finally ends up with conclusion in Sect. 5.

2 Kinematic Model of a Robot

A differential drive mobile robot (DDMR) consisting of two driving wheels and one castor wheel is considered. Let v_l and v_r be the left and right wheel velocities, respectively. Then the robot velocity is the average of the two wheel velocities and the angular velocity of the DDMR is

$$\omega = \frac{v_r - v_l}{L} \tag{1}$$

where L is the distance between two wheels. From Fig. 1 and from the definition of angular velocity, the kinematic equations of a DDMR can be obtained as

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} v\cos\theta \\ v\sin\theta \\ \omega \end{bmatrix} \tag{2}$$

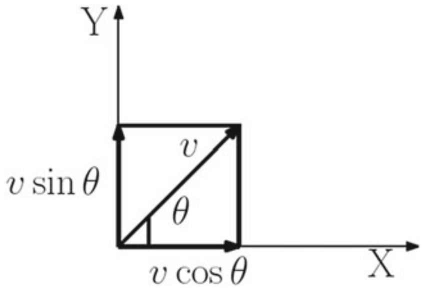


Fig. 1 Differential drive kinematics

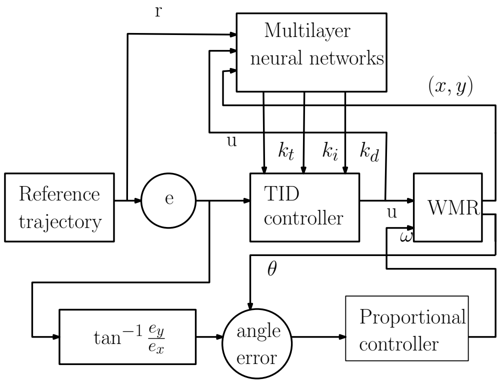


Fig. 2 Proposed controller for trajectory tracking

Consider a DDMR with angular velocity ω and the linear velocity v, respectively. The point $(x_r(t), y_r(t))$ on the reference trajectory is the desired point. The robot which is currently at $(x(t), y(t))$ should track this reference point. The error in position and orientation is defined as

$$\begin{bmatrix} x_e(t) \\ y_e(t) \\ \theta_e(t) \end{bmatrix} = \begin{bmatrix} x_r(t) - x(t) \\ y_r(t) - y(t) \\ \theta_r(t) - \theta(t) \end{bmatrix} \quad (3)$$

3 Proposed NN-Based TID Controller

The proposed methodology uses a neural network (NN)-based TID controller for the effective trajectory tracking of a WMR. The optimal parameters for the TID are generated using a neural network. The proposed structure of the neural network-based non-linear TID is shown in Fig. 2. The objective of trajectory tracking problem is to make the errors in distance and the deviation angle zero. The error in distance and deviation in azimuth angle are obtained from (4) and (5), respectively

$$d_e = \sqrt{(x_r - x)^2 + (y_r - y)^2} \quad (4)$$

The errors in position and orientation are fed back to two TID controllers (one for x and other for y) and to a proportional controller, respectively.

$$\phi_e = \tan^{-1}\frac{y - y_r}{x - x_r} \quad (5)$$

The errors in x- and y-positions are provided as the input signals to the proposed NN controllers. The deviation in azimuth is adjusted to zero by a proportional controller. The tracking errors are described by $e = \begin{bmatrix} x_e & y_e & \theta_e \end{bmatrix}$. A TID controller whose parameters are optimized using the neural networks is used to track the desired trajectory with minimum control input and tracking error. The TID is similar to that of a PID [10], but the proportional gain is substituted with a tilted mode consisting of a transfer function $1/s^{1/n}$.

The neural networks trained using backpropagation algorithm are used to find the optimal K_i, K_d and K_t of a TID controller. The inputs of the neural networks are the reference trajectory, control input, and current position of the robot. There are 15 neurons in the hidden layer and the output layer consists of three neurons which corresponds to K_t, K_i and K_d. Inputs to the TID controller are the errors in x- and y-positions given by $x_e(t)$ and $y_e(t)$, respectively. The control inputs are the linear velocity v and angular velocity ω. The accuracy of the proposed NN-based TID controller is evaluated using different performance analysis indices such as integral squared error and maximum of absolute tracking error. The integral squared error is calculated by

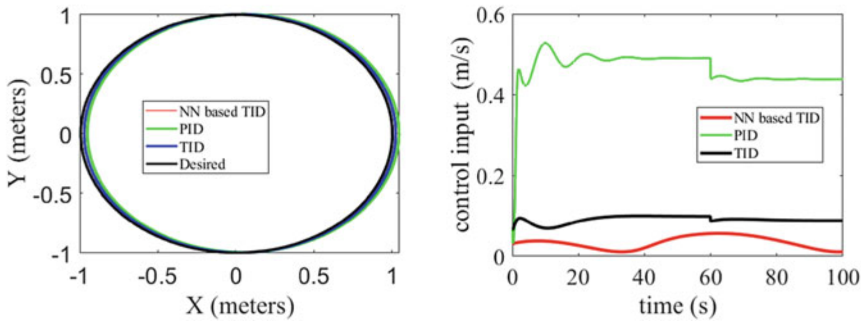
$$ISE = \int_0^t e(t)^2 dt \quad (6)$$

The integral of absolute error denotes the accumulated error and is computed as

$$IAE = \int_0^t |e(t)| dt \quad (7)$$

4 Results and Discussions

The effectiveness and the feasibility of the NN-based TID controller for trajectory tracking of a WMR are evaluated by MATLAB/SIMULINK simulations. This section describes the key findings. The model of the differential WMR described in Sect. 2 and the NN-based controller presented in Sect. 3 are developed using MATLAB 2020a. Neural networks are trained using backpropagation algorithm. Simulations are carried out by considering various trajectories such as circular trajectory,



(a) Tracking of circular trajectory

(b) Control history

Fig. 3 Performance evaluation of controllers (circular trajectory)

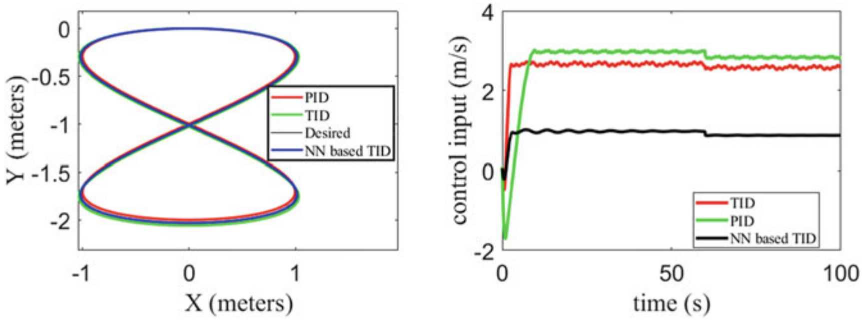
Table 1 Comparison of controllers (circular trajectory)

Controller	Maximum control	Total control effort	ISE	IAE
PID	0.5	21.69	4.25	2.21
TID	0.1	6.23	3.38	1.62
NN-based TID	0.05	3.01	2.89	1.042

lemniscate trajectory, B spline trajectory, and straight line trajectory as the reference trajectories.

4.1 Case 1: Circular Trajectory

A circle of radius 1m is considered as the reference trajectory. The trajectories are tracked using PID-, TID-, and NN-based TID controllers are shown in Fig. 3a. It is evident from this figure that the NN-based TID controller tracks the desired trajectory exactly whereas the tracked trajectories by other controllers slightly deviate from the desired trajectory. Figure 3b shows the control history of each controllers. From Table 1, it is very clear that the maximum control, total control effort, ISE, and IAE are very small as compared to PID and TID controllers. Thus, for a circular trajectory, the proposed NN-based TID requires lesser control and achieves negligible small displacement error and so is superior to PID and TID controllers.



(a) Tracking of lemniscate trajectory

(b) Control histroy

Fig. 4 Performance evaluation of controllers (lemniscate trajectory)

Table 2 Comparison of controllers (lemniscate trajectory)

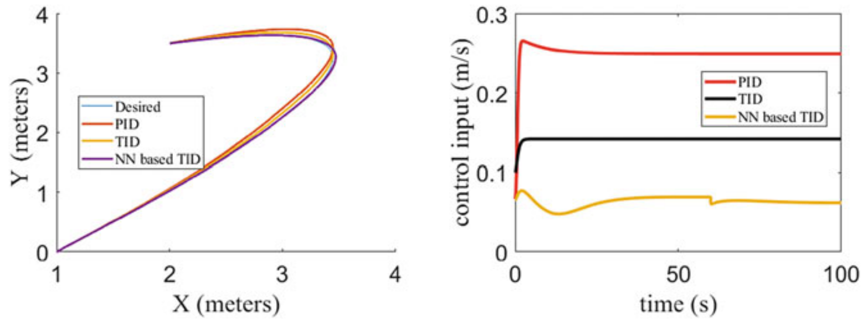
Controller	Maximum control	Total control effort	ISE	IAE
PID	3.1	85.5	6.98	2.01
TID	2.8	64.62	4.21	1.74
NN-based TID	1.12	28.23	2.21	1.21

4.2 Case 2: Lemniscate Trajectory

The trajectories tracked by PID-, TID-, and NN-based TID controllers are shown in Fig. 4a. The displacement errors made by PID and TID controllers are high as compared with the proposed NN-based TID controller. As shown in Table 2, controls required for PID and TID are 3.1 and 2.8, respectively, as compared to a smaller value of 1.12 for the proposed method. In Fig. 4b, the control histories for each of these controllers are shown. It is obvious from Table 2 that the total control effort, ISE, and IAE also are very small for the proposed method.

4.3 Case 3: B Spline Trajectory

A DDMR cannot take sharp turn. The performance of DDMR with a sharp turn can be analyzed with a B spline trajectory. The B spline trajectory is tracked by using the three control algorithms. As can be observed from Fig. 5a, the displacement error with the NN-based TID controller is less as compared to other controllers. The initial part of the trajectory is a straight line and all the three control algorithms are able to track the desired trajectory. However, due to the sharp turn the PID and TID



(a) Tracking of B spline trajectory (b) Control histroy

Fig. 5 Performance evaluation of controllers (B spline trajectory)

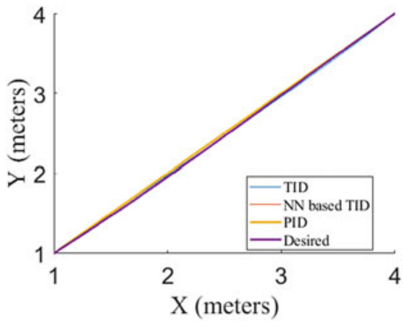
Table 3 Comparison of controllers (B spline trajectory)

Controller	Maximum control	Total control effort	ISE	IAE
PID	0.25	15.14	1.92	0.96
TID	0.13	7.63	1.53	0.42
NN-based TID	0.075	2.38	1.11	0.24

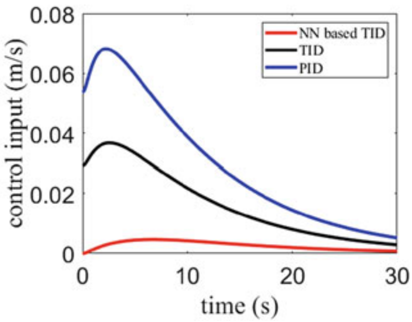
controllers deviate from the desired trajectory as it reaches the goal position. The NN-based TID controller reports a highest tracking accuracy. From Table 3, it can be concluded that the proposed method possesses superior performance in terms of control input and displacement error.

4.4 Case 4: Straight Line Trajectory

This section considers a straight line trajectory. The trajectories plotted by the NN-based TID controller, the PID controller, and the TID controller are shown in Fig. 6a. The tracking error is least for the NN-based TID controller. The error in tracking a straight line is small compared to the error occurred while tracking circular trajectory, B spline trajectory, and lemniscate trajectory. The control history shown in Fig. 6b and the performance comparison as per Table 4 clearly illustrate that the proposed NN-based TID comes with a better performance. The tracking error is reduced by nearly 50% by proposed controller when compared to the PID and TID controllers. The maximum control input is only 1.12 in the proposed controller which is very small as compared to TID (2.8) and PID (3.1). Analyzing the performance on simulation, NN-based TID controller is more advantageous in terms of accuracy, control input, and total control effort.



(a) Tracking of straight line trajectory



(b) Control history

Fig. 6 Performance evaluation of controllers (straight line trajectory)

Table 4 Comparison of controllers (straight line trajectory)

Controller	Maximum control	Total control effort	ISE	IAE
PID	0.06	5.42	1.23	0.32
TID	0.03	3.03	1.13	0.26
NN-based TID	0.005	1.08	1.01	0.14

5 Conclusion

In this paper, a neural network-based TID controller for the trajectory tracking of a differential drive mobile robot is presented. A neural network is used to tune the TID parameters. Simulations are done for circular, lemniscate, straight line, and B spline trajectories. The performance of the proposed controller is evaluated with the conventional TID and PID controllers. The simulation results illustrated that the proposed method exhibited superior performances with respect to tracking error, total control effort, and maximum control input.

References

1. Asif M, Memon AY, Khan MJ (2016) Output feedback control for trajectory tracking of wheeled mobile robot. Intell Autom Soft Comput 22(1):75–87
2. Zohar I, Ailon A, Rabinovici R (2015) Trajectory tracking and related application robot. Auton Syst 59(6):343–353
3. Al Khatib EI, Wasim MF (2015) A comparison of adaptive controllers for wheeled mobile robots. In: ISMA
4. Zhao P, Chen J, Song Y, Tao X, Xu T, Mei T (2016) Design of a control system for an autonomous vehicle based on adaptive-PID. Int J Adv Robot Syst 9(2):44

5. Al-Dahhan MRH, Ali MM (2016) Path tracking control of a mobile robot using fuzzy logic. In: Proceedings of the 13th international multi-conference on systems, signals and devices (SSD), pp 82–88, Mar 2016
6. Tiep DK, Lee K, Im D-Y, Kwak B, Ryoo Y-J (2018) Design of fuzzy-PID controller for path tracking of mobile robot with differential drive. Int J Fuzzy Logic Intell Syst 18(3):220–228
7. Ma L, Sun H, Song J (2020) Fractional-order adaptive integral hierarchical sliding mode control method for high-speed linear motion of spherical robot. IEEE Access PP(99):1–1
8. Domingos RMM, da Cruz G (2018) Neuroevolved controller for an omnidirectional mobile robot. In: IEEE Latin American conference (LACCI)
9. Wu X, Jin P, Zou T, Qi Z, Xiao H, Lou P (2019) Backstepping trajectory tracking based on fuzzy sliding mode control for WMR. J Intell Robot Syst 96(1):109–121
10. Saleh AL, Hussain MA, Klim SM (2018) Optimal trajectory tracking control of WMR using FOPID controller. J Univ Babylon Eng Sci 26(4)

Do You Trust Me? Value and Governance in Data Sharing Business Models

Ruben D'Hauwers and Nils Walravens

Abstract Collaborations to share data can enable private companies to create new business models and improve products and services. For public authorities, it can lead the way toward a data-driven policy. Yet, sharing data poses different challenges, such as risk to comply with legalities, commercial risk, quality of data, privacy and ethical risk, data ownership etc. Managing data within the borders of a firm has been explored thoroughly in the literature of data governance. Yet, data sharing poses new challenges, as it extends to collaboration within an ecosystem. The business model literature is not investigating yet the concept of data sharing. Thus, this paper investigates the different data-sharing business model factors. The research looks into the limitations of current research, which it aims to overcome by combining the insights from the data governance, data sharing literature and business model literature. The factors influencing data-sharing collaborations are (1) value creation and capturing, (2) data governance, (3) ecosystem trust and (4) data trust.

Keywords Data sharing · Business model · Data governance · Inter-organizational · Value · Control · Trust · Ecosystem

1 Introduction

Sharing data poses different opportunities for companies to create new business models and services [1–5] as data collaborations can facilitate the discovery of new insights, faster decision-making and increase innovation [4, 6]. For governments, it can lead the way toward evidence-based policymaking, which is the process of using (big) data in the policymaking process and improving services [4, 5, 7, 8]. Many challenges are preventing data sharing in ecosystems [6]. First, identifying

R. D'Hauwers (✉) · N. Walravens
IMEC-SMIT-VUB, Pleinlaan 9, Brussel, Belgium
e-mail: ruben.dhauwers@imec.be

N. Walravens
e-mail: nils.walravens@vub.be

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_22

and finding an agreement with a data-sharing partner is a challenge, as well as ensuring qualitative data by cleaning and processing the data [4]. In the pricing data result into complex negotiations, valuing data needs to be overcome [4]. Many companies are reluctant to share data due to a lack of trust [6, 9–12]. Sharing data causes commercial risk, as companies refrain from sharing sensitive information which might reduce a competitive advantage [7, 8, 13, 14]. Complying with the legal requirements discourages companies from sharing data due to unclear legal frameworks [4], and privacy laws such as the general data protection regulation (GDPR) [15, 16]. All these result in high transaction costs [7]. Previous research on data sharing has mainly focused on dyadic buyer–supplier relationships [6, 17, 18] which does not consider the complexity and dynamics of sharing data in ecosystems [6, 19]. A data-sharing business model framework for the inter-organizational setting was developed in this work based on data governance, data sharing and business model literature.

2 Research Questions

The covered literature reviews indicate the need for further research in data sharing and data governance for data ecosystems [20–22]. Abraham et al. [22] formulated questions concerning inter-organizational data sharing: How do organizations retain control over their data and design governance in inter-organizational relationships while deconstructing data silos? How are value and trust created in data collaborations? The business model literature can provide a new perspective on these questions [23] resulting in the research question '*What factors determine data sharing business models?*'.

3 Methodology

A systematic literature review [24] identifies and evaluates available research and interprets it by applying it to a particular research question to identify any gaps in current research and to provide a new framework. We aimed together knowledge in existing literature with keyword searches focused on 'data governance', 'business model', 'data sharing', 'data exchange', 'inter-organizational relationships' and 'data marketplace' with the most citations and mentioned in literature reviews. A limitation to the study was that the authors did not aim to cover the literature exhaustively. Additionally, papers were added based on citations in other papers, based on the snowball sampling method [25]. A difference was made between factors determining the 'intra-organizational scope' on a project- or firm-level [26] and the 'inter-organizational scope', which encompasses different firms or an ecosystem of firms [26]. The creation of the conceptual framework was based on merging and relating key factors in the literate [27]. A division between value and control as proposed by Ballon [23] was used.

4 Literature Review

4.1 Data-Sharing Literature

The literature on data sharing looks into factors that increase data sharing [6, 11, 28–31], and into data market characteristics that influence data sharing [7, 12, 14, 32–36]. Trust between parties can support the creation of data collaborations [6, 9, 11, 12, 14, 29, 30, 35, 36]. Trust relates to the provenance of the data good, decreasing asymmetric information [33] as well as power asymmetry [37] and the market type [6, 7, 9, 32, 36]. As companies do not want to lose bargaining power, they might avoid sharing information [38, 39]. Data governance parameters such as standards, metadata, legalities ensure the quality of data [7, 11, 31, 33]. Additionally, the ownership of data [29–32] concerns who can use and publish the data. Control over data [11, 14, 33, 36] prevents spreading sensitive commercial data [8, 13] and privacy violations [15, 16]. Also, business transactions have an impact on data sharing, such as revenue models [9, 30, 31, 36]; cost models and transaction costs [7, 30]; and revenue sharing [6, 36].

4.2 Data Governance Literature

The data governance covers intra-organizational parameters within one organization on a project- or firm-level [22–24, 26] and lacks the inter-organizational scope. The papers were selected based on the number of citations complemented with papers that already included inter-organizational factors [22, 36, 40, 41]. The goal of data governance is to ensure that an organization has a good data policy ensuring the control of data quality in a company-wide setting: policies and processes/governance mechanisms [21, 22, 42–44], organization including roles and responsibilities and leadership [21, 22, 42–45], data standards [21, 22, 42–45], metrics and monitoring [21, 22, 43], legalities and conformance [22, 40, 43] and personal information risk [40]. Parameters concerning the inter-organizational scope are data ownership [40], type of data sharing [41], provenance [40] and conformance regarding personal information [40].

4.3 Business Models Literature

As data-sharing business models consider intangible goods which are more difficult to protect [33], the current business model literature does not cover business models for data sharing to date. First, public value examines from the perspective of the end-user and the wide public [46]. Creating value [23, 46–63] in the case of data sharing is determined by the use cases of data that are relevant for the stakeholders [40]. As an example, if a company in the logistics sector knows data will be used to optimize the arrival times of containers, the value of data can be based on this use case. Capturing the value of data requires revenue models [23, 46–50, 52–56, 61–63]. The value of data is not always recognized between companies, which makes the pricing of data challenging [12, 64, 65]. Finally, cost models and cost-sharing models determine the costs in the business model, and how this can be shared between the different partners [23, 46–57, 59–63]. The cost of data relates to different aspects, such as storage of data, energy, maintenance and software [65].

5 Factors Determining Data Sharing

Based on the outcomes of the literature review, a novel framework (Table 1) incorporating the different factors of data sharing is proposed and discussed below.

Table 1 Data sharing business models framework

Value		Data governance	
Value creation	Ecosystem value	Policy	Policies and processes
	Application domain		Organization
	Value proposition	Quality	Standards and monitoring
Financial	Revenue (sharing) model	Conformance	Legalities
	Cost (sharing) model		Data sharing risk
Ecosystem trust		**Data trust**	
Value network	Data sharing model	Collaboration	Data sharing agreement
	Competition versus cooperation		Interoperability
Power structure	Data ownership	Provenance	Quality control
	Customer ownership		Control of data
	Data capacity		

5.1 *Intra-organizational*

Value: Value creation is the reason why companies will want to share data. The value of data is challenging to determine [4, 11, 12, 31] but can be established by the application domain [40], which shows how the data will be used and creates value to the ecosystem and users. Ecosystem value [46] is the value of the data to the overall ecosystem as if no ecosystem value is created, and data sharing on an inter-organizational level will be impossible. The intra-organizational value is the value proposition to individual organizations [52, 53, 61]. The financial variables are constituted by the revenue model, revenue-sharing model and cost-sharing model. Revenue models determine how much a business can price the product and whether it can capture the value of the product [23]. The revenue and cost-sharing models refer to agreements on sharing revenue and costs among the actors [23].

Data governance To ensure the quality of data, data governance is at the core of data sharing. It includes making data available for reuse, curation, preservation and quality assurance [66]. To ensure the quality of data, standards [21, 22, 42, 43] and policies [21, 22, 43–45] around the creation, development, control, management and audit of data are needed. Also, conformance and compliance with legal requirements are critical [40]. This is especially true in the case of personal information [33, 67], where the sharing of aggregated data could help alleviate privacy concerns [68].

5.2 *Inter-organizational*

On the inter-organizational level, both parameters are based on trust [6, 11, 14, 29, 30, 35, 36]. Before data sharing between two or more parties can occur, they must reach a point of sufficient mutual trust to establish governance policies and agreements [9, 12, 29].

Trust in the ecosystem: Data sharing depends on the collaborative or competitive nature of the value network [11]. Different models of data sharing occur, such as in hierarchical, bazaar, market or network models [41] or based on the market access and architecture [12]. Firms may refrain from sharing confidential information to avoid losing bargaining power [39, 69] or a powerful player may use their power to force other companies to share information [38]. Thus, the power structure is crucial to engendering trust in the ecosystem and depends on three factors. First, the data ownership rights describe who owns and uses the data [26, 36, 40]. Second, customer ownership refers to the establishment of direct relations with end customers [23]. Third, data capacity is the way in which intelligence is distributed across different players in the ecosystem [23], based on algorithms and skills.

Trust in the data: Data trust refers to the trust one can put in sharing and utilizing the data of its counterparts [9, 12, 29]. Before data sharing between two or more parties can occur, a certain level of trust must be established to agree on data-sharing agreements [29]. In order to be able to exchange information and services, interoperability and provenance are required. Provenance means to trace the history of the data life cycle in a transparent way [33, 40] while interoperability ensures machine readability [4]. Standardizing metadata [65] or utilizing technologies like blockchain can support both. When combining data sources, anonymized personal information may be identified [70], potentially causing the risk of re-identification leading to privacy issues.

6 Conclusion

Based on a literature review questioning 'what are the factors determining successful data sharing business models', four factors part of a novel data-sharing business model framework were identified: value creation, data governance, trust in the ecosystem and trust in the data. Abraham et al. [22] formulated how data governance and data sharing can be performed in an inter-organizational setting. These questions concern the control over data, governance designs, collaborations, value created and trust creation in inter-organizational data collaborations. Further research will need to investigate in depth how data sharing business models can answer these questions. Limitations of this research are that the body literature was not analyzed exhaustively, and has not been applied in a real-life example. From a practical perspective, the framework can be used to (1) analyze existing data-sharing ecosystems and business models and (2) utilize the design of new data-sharing ecosystems and related business models. The practical use and validation of this framework are subject to further research.

References

1. Hansen JV, Hill NC (1989) Control and audit of electronic data interchange. MIS Q 13(4):403–414
2. European Commission (2020) https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en. Accessed 18 Sept 2020
3. European Commission (2020) https://ec.europa.eu/digital-single-market/en/guidance-private-sector-data-sharing. Accessed 18 Sept 2020
4. BVDA (2020) https://www.bdva.eu/sites/default/files/BDVA%20DataSharingSpace%20PositionPaper_April2019_V1.pdf. Accessed 21 Sept 2020
5. GSMA (2020) https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/06/Mobile-Data-for-Social-Good-Report_29June.pdf. Accessed 18 Sept 2020
6. Kembro J, Naslund D, Olhager J (2017) Information sharing across multiple supply chain tiers: a Delphi study on antecedents. Int J Prod Econ 193(C):77–86
7. Martens B (2020) The economics of business-to-government data sharing. European Commission, Seville, JRC119947
8. Klein T, Verhulst T (2017) Access to new data sources for statistics: business models and incentives for the corporate sector. OECD statistics working papers, June 2017, OECD Publishing

9. Richter H, Slowinski P (2019) The data sharing economy: on the emergence of new intermediaries. Int Rev Intellect Prop Compet Law 50(1):4–29
10. European Commission (2020) https://op.europa.eu/en/publication-detail/-/publication/8b8776ff-4834-11e8-be1d-01aa75ed71a1. Accessed 20 Sept 2020
11. Dahlberg T, Nokkala T (2019) Willingness to share supply chain data in an ecosystem governed platform—an interview study. In: Bled eConference
12. Spiekermann M (2019) Data marketplaces: trends and monetization of data goods. Intereconomics 54:208–216
13. Jarman H, Luna-Reyes LF, Zhang J (2016) Private data and public value: governance, green consumption, and sustainable supply chains. Springer International Publishing, New York, pp 1–23
14. Agahari W (2020) Platformization of data sharing: multi-party computation (MPC) as control mechanism and its effect on firms' participation in data sharing via data marketplaces. In: Bled eConference, pp 691–704
15. Khurana M, Mishra P, Singh A (2011) Barriers to information sharing in supply chain of manufacturing industries. Int J Manuf Syst 1(1):9–29
16. Sayogo DS, Zhang J, Pardo TA, Tayi GK, Hrdinova J, Andersen DF, Luna-Reyes L (2014) Going beyond open data: challenges and motivations for smart disclosure in ethical consumption. J Theor Appl Electron Commer Res 9(2):1–16
17. Autry CW, Williams BD, Golicic S (2014) Relational and process multiplexity in vertical supply chain triads: an exploration in the US restaurant industry. J Bus Logist 35(1):52–70
18. Caridi M, Moretto A, Perego A, Tumino A (2014) The benefits of supply chain visibility: a value assessment model. Int J Prod Econ 151:1–19
19. Kembro J, Selviaridis K (2015) Exploring information sharing in the extended supply chain: an interdependence perspective. Supply Chain Manag Int J 20(4):455–470
20. Bruhn J (2014) Identifying useful approaches to the governance of indigenous data. Int Indig Policy J 5(2):1–32
21. Panian Z (2010) Some practical experiences in data governance. World Acad Sci Eng Technol 62:939–946
22. Abraham R, Brocke J, Schneider J (2019) Data governance: a conceptual framework, structured review, and research agenda. Int J Inf Manag 49:424–438
23. Ballon P (2007) Business modelling revisited: the configuration of control and value. Info 9(5):6–19
24. Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Engineering 2:1051
25. Morgan D (2008) Snowball sampling. In: The SAGE encyclopedia of qualitative research methods. SAGE Publications Inc., Thousand Oaks, pp 816–817
26. Tiwana A, Konsynski B (2010) Complementarities between organizational IT architecture and governance structure. Inf Syst Res 21:288–304
27. Miles MB, Huberman AM (1994) Qualitative data analysis: an expanded sourcebook, 2nd edn. SAGE Publications, Thousand Oaks
28. Elgarah W, Falaleeva N, Saunders C, Ilie V, Shim J, Courtney J (2005) Data exchange in interorganizational relationships: review through multiple conceptual lenses. Data Base Adv Inf Syst 36:8–29
29. Allen C, Des Jardins TR, Heider A, Lyman KA, McWilliams L, Rein AL, Schachter AA, Singh R, Sorondo B, Topper J, Turske SA (2014) Data governance and data sharing agreements for community-wide health information exchange: lessons from the beacon communities. EGEMS (Wash DC) 2(1):1057
30. Dinalog (2020) https://www.dinalog.nl/wp-content/uploads/2020/08/Dinalog_Whitepaper-Data-Infrastructure_DEF.pdf. Accessed 18 Sept 2020
31. Eckartz SM, Hofman WJ, Van Veenstra AF (2014) A decision model for data sharing. In: international conference on electronic government, EGOV 2014, vol 8653. Springer, Heidelberg

32. Stahl F, Schomm F, Vossen G et al (2016) A classification framework for data marketplaces. Vietnam J Comput Sci 3:137–143
33. Koutroumpis P, Leiponen A, Thomas L (2017) The (unfulfilled) potential of data marketplaces. ETLA working papers no 53
34. Pavlou PA (2002) Institution-based trust in interorganizational exchange relationships: the role of online B2B marketplaces on trust formation. J Strat Inf Syst 11(3–4):215–243
35. Pavlou PA, Gefen D (2004) Building effective online marketplaces with institution-based trust. Inf Syst Res 15(1):37–59
36. Schreieck M, Wiesche M, Krcmar H (2016) Design and governance of platform ecosystems—key concepts and issues for future research
37. Mason-Jones R, Towill DR (1999) Using the information decoupling point to improve supply chain performance. Int J Logist Manag 10(2):13–26
38. Ballou RH, Gilbert SM, Mukherjee A (2000) New managerial challenges from supply chain opportunities. Ind Mark Manag 29(1):7–18
39. Christopher M, Jüttner U (2000) Supply chain relationships: making the transition to closer integration. Int J Logist Manag 3(1):5–23
40. Lee S et al (2018) Designing data governance in platform ecosystems. HICSS
41. van den Broek T, van Veenstra AF (2015) Modes of governance in inter-organizational data collaborations, ECIS 2015 completed research papers, pp 0–12
42. Kim H, Cho (2018) Data governance framework for big data implementation with NPS case analysis in Korea. J Bus Retail Manag Res 12(3)
43. Cheong L, Chang V (2007) The need for data governance: a case study. In: ACIS 2007 proceedings—18th Australasian conference on information systems
44. Weber K (2007) A model for data governance—organizing accountabilities for data quality management. Gesellschaft für Informatik Bonn, pp 347–362
45. Cohn BL (2015) Data governance: a quality imperative in the era of big data, open data, and beyond. J Law Policy Inf Soc 10(3):811–826
46. Walravens N, Ballon P (2013) Platform business models for smart cities: from control and value to governance and public value. IEEE Commun Mag 51:72–79
47. Mahadevan B (2000) Business models for internet-based E-commerce: an anatomy. Calif Manag Rev 42:55–69
48. Alt R, Zimmermann P (2001) Preface: introduction to special section—business models. Electron Mark 11
49. Applegate LM (2001) Emerging e-business models: lessons from the field. Harvard Business School, working paper no 9-801-172
50. Rappa M (2020) http://ecommerce.ncsu.edu/business_models.html. Accessed 21 Sept 2020
51. Chesbrough H, Rosenbloom RS (2002) The role of the business model in capturing value from innovation: evidence from Xerox corporation's technology spin-off companies. Ind Corp Chang 11(3):529–555
52. Osterwalder A (2004) The business model ontology: a proposition in a design science approach. PhD thesis, HEC Lausanne
53. Osterwalder A, Pigneur Y (2010) Business model generation: a handbook for visionaries, game changers, and challengers (self-published)
54. Morris M, Schindehutte M, Allen J (2005) The entrepreneur's business model: toward a unified perspective. J Bus Res 58(6):726–735
55. Johnson M (2010) Seizing the white space: business model innovation for growth and renewal. Harvard Business Press, Boston
56. Johnson MW, Christensen CM, Kagermann H (2008) Reinventing your business model. Harv Bus Rev 86(12):50–59
57. Grefen P, Lüftenegger E, Comuzzi M, Weisleder C (2013) The service dominant business model a service focused conceptualization. Beta working paper series
58. Weking J, Kowalkiewicz M, Böhm M, Krcmar H (2018) Archetypes for industry 4.0 business model innovations. In: Proceedings of the 24th Americas conference on information systems (AMCIS), pp 1–10

59. Solaimani S, Bouwman H (2012) A framework for the alignment of business model and business processes: a generic model for trans-sector innovation. Bus Process Manag J 18:655–679
60. Wirtz BW, Daiser P (2017) Business model innovation: an integrative conceptual framework. J Bus Model 5(1):14–34
61. Al- MM, Avison D (2010) Developing a unified framework of the business model concept. Eur J Inf Syst 19(3):359–376
62. Bonaccorsi A, Giannangeli S, Rossi C (2006) Entry strategies under competing standards: hybrid business models in the open source software industry. Manag Sci 52(7):1085–1098
63. Wiener M, Saunders C, Marabelli M (2019) Big-data business models: a critical literature review and multi-perspective research framework. J Inf Technol 35(1):66–91
64. Spiekermann M, Wenzel S, Otto B (2018) A conceptual model of benchmarking data and its implications for data mapping in the data economy. In: Multikonferenz wirtschaftsinformatik, pp 314–325
65. Khatri V, Brown C (2010) Designing data governance. Commun ACM 53:148–152
66. Mayernik M (2015) Research data and metadata curation as institutional issues. J Assoc Inf Sci Technol 67
67. Martin K (2015) Ethical issues in the big data industry. MIS Q Exec 14(2)
68. Elliot M, O'hara K, Raab C, O'Keefe C, Mackey E, Dibben C, Gowans H, Purdam K, McCullagh K (2018) Functional anonymisation: personal data and the data environment. Comput Law Secur Rev 34(2):204–221
69. Seidmann A, Sundararajan A (1998) Sharing logistics information across organizations: technology, competition and contracting. In: Kemerer CF (ed) Information technology and industrial competitiveness. Kluwer Academic Publishers, Boston, MA, pp 107–136
70. Porter C (2008) De-identified data and third party data mining: the risk of re-identification of personal information. Shidler J L Com Tech 5(1)
71. Osborne Clarke LLP (2020) https://publications.europa.eu/en/publication-detail/-/publication/d0bec895-b603-11e6-9e3c-01aa75ed71a1/language-en. Accessed 20 Sept 2020
72. Robert KY (2014) Case study research design and methods, 5th edn. Sage, Thousand Oaks, CA. Can J Progr Eval
73. Tallon P (2013) Corporate governance of big data: perspectives on value, risk, and cost. Computer 46:32–38
74. Timmers P (1998) Business models for electronic markets. J Electron Mark 8:3–8
75. Weill P, Vitale MR (2001) Place to space: migrating to eBusiness models, 1st edn. Harvard Business School Press, Boston
76. International Data Spaces (2020) https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-Reference-Architecture-Model-3.0.pdf. Accessed 24 Sept 2020
77. Allee V (2000) Reconfiguring the value network. J Bus Strat 21:36–39
78. Prahalad CK, Hamel G (1990) The core competence of the organization. Harv Bus Rev 6(3):79–92

Knowledge Transfer in Intercultural Technical Communication in View of Translation Synergetic Paradigm

Lyudmila Kushnina, Irina Perlova, and Kristina Permiakova

Abstract The article discusses the problem of technical translation in the view of the translation space concept and the situation-reflective model of translation decision-making. Translation space is regarded as a synergetic model of translation, a space–time continuum in which heterogeneous, explicit and implicit meanings are transposed, synergized and harmonized from one linguo-culture to another. The cornerstone of the situation-reflective model is the interaction of the algorithm as a process of finding an optimal solution and heuristics as a directed search for a translation solution. The authors formulate the purpose of the study as the rationale for macro- and micro-translation strategies that provide an effective transfer of technical knowledge. The results of the research have revealed that the translator's actions are based on cognitive verbal-cogitative activity, which involves forming a conceptual image of the text as a compressed expression of meaning in the translator's mind. The focus of the research is the work of a translator—linguist who is not an expert in the technical field of communication. The achievement of a high-quality, harmonious translation of a technical text by such a translator requires both the mastery of a synergetic translation model and the mastery of a situation-reflexive model of translation decision-making, which together lead to the realization of a synergetic strategy for the transfer of technical knowledge through the interaction of languages and cultures. The macro-strategy is to identify the technicity category in the source text and to recreate it in the translation text. The success of macro-strategy means achieving synergetic effects, accompanied by the increment of new meanings acceptable in the host technical culture. Micro-strategies are focused on solving a specific translation

L. Kushnina · K. Permiakova
Perm National Research Polytechnic University, Komsomolsky Prospekt, 29, 614990 Perm, Russia
e-mail: lkushnina@yandex.ru

K. Permiakova
e-mail: kpuctu@rambler.ru

I. Perlova (✉)
Department of Foreign Languages, Linguistics and Translation, Perm National Research Polytechnic University, Komsomolsky Prospekt, 29, 614990 Perm, Russia
e-mail: perl_telecom@mail.ru

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_23

problem, overcoming certain technical difficulties encountered by the translator in each field of translation space.

Keywords Transfer of technical knowledge · Technicity category · Technical translation · Translation space · Synergy · Harmony · Translation strategy

1 Introduction

The methodological basis for this paper is the concept of translation space, a synergetic model of translation developed by one of the article authors. If the hypothesis about translation space existence was first formulated by Kushnina in 2003 [1], its concept field is still being formed up to now on the material of various discourses and functional styles.

Within the framework of this article, we are interested in the problems of interlinguistic and intercultural technical communication, which can be traced from the perspective of the modern research paradigm—knowledge transfer (KT), which allows a new look at the translator's linguistic personality and the results of their professional activity.

The idea of KT in the humanities has been intensively developed by Russian (Alekseeva, Mishlanova, Demyankov, etc.) [2, 3] and foreign scholars [4].

According to Alekseeva and Mishlanova, if knowledge is the main value of modern society, KT becomes a kind of innovation chain that includes, along with mechanical movement and development, '…transformation and interaction of knowledge and cognition' [2, p. 25]. According to the researchers, KT is related to the concept of communication, mainly its role in the mental activity of an individual as a way of solving a cognitive problem'.

The novelty of the study is that the idea of knowledge transfer as a common humanitarian problem is still not sufficiently understood in the translation of a special technical text. In this connection, we consider it important to highlight the importance of the work outcomes for the further study of translation patterns in the field of technical communication.

Starting their job, a translator faces many innovative tasks, especially if the translator—linguist deals with a scientific or technical text.

As our pedagogical experience shows, translators—linguists have real challenges in translating special technical texts, and this is the reason why some research containing models, methods and translation algorithms is required to help budding translators overcome these difficulties. One of the ways suggested by the authors of the article is for translators to master a synergistic model of translation, as well as micro- and macro-strategies that ensure the successful transfer of technical knowledge from one linguo-culture to another.

The following issues will be considered sequentially:

1. The need to distinguish between scientific and technical types of translation.

2. Summary of the translation space concept as a synergetic model of translation and as a methodology of the research being carried out.
3. Analysis of technicity category as a key parameter in the translation of a technical text.
4. Empirical analysis of the translated technical text from the standpoint of achieving harmony—a qualitative translation in the light of the translation space concept.

2 Problem Statement and Research Goals

Turning to the issue of technical translation, we consider it to be a special type of translation, different from other types. For a long time, scientists have been using the term 'scientific-technical translation', merging two disparate areas of translation into one, probably appealing to the notion of scientific-technical progress.

Meanwhile, Lotman made a clear distinction between the two areas: 'Technology and science are indeed closely linked. And they are not connected because they are synonymous, but because they are antitheses' [5, p. 67].

More recently, there has been growing support among specialists for the idea of distinguishing between the translation of scientific and technical texts [2, 5]. Accordingly, the transfer of KT in translation has also its peculiarities.

The aim of the study is to validate translation strategies that will lead to a high-quality (harmonious) translation of the technical text, providing an effective transfer of specialized knowledge in the technical field of communication. Understanding KT, following Alekseeva and Mishlanova, not as usual communication represented by mutual movement, but as the component of the translator's thinking activity in the process of solving a cognitive problem, we analyse the strategy of general and mental activity of the translator-linguist, who is not an expert in the field, unlike the author and the recipients. As Chudinov emphasizes in the preface to the recently published work of the mentioned authors 'Transfer of knowledge in the humanities', they 'managed to get away from the concept of transfer as simple linear transfer and come to its understanding within the framework of complex cyclic interpretations' [2, 5]. At the same time, the translator has a high level of competence in translation technology of traditional text types—narrative, description and argument [6, p. 26, 7]. Further research in this direction will make it possible to recognize the cognitive mechanisms of the translator-linguist's verbal and cogitative activity, which determine their technical KT strategy.

3 Methodology and Research Methods

The starting point of our research is that translation itself is regarded as a process of languages and cultures interaction taking place in a certain space–time continuum—a chronotope where one can distinguish spatial fields and spatial meanings on the one hand, and temporal meanings, on the other. Spatial fields include, firstly, the fields of subjects of translation (author, translator, recipient), secondly, text fields (energy and phatic); the core is the content of the text. Each field has its unique meaning. The temporal meanings are most clearly manifested in the phatic field as a field of culture, which is due to the cultural markedness of the representatives, forming the concept 'time'. This means that, along with translation space, there is translation time as a linguo-cultural category, and their totality produces a translation chronotop. From the translation time point of view, there are temporal meanings distinguished, which manifest themselves differently in the original and target texts. Temporal meaning is the actual, cultural value of time for different subjects of translation communication. Temporal meaning is the actual, cultural value of time for different subjects of translation communication. It is not a question of objective measurement of time, but of subjective perception by the individual, which is reflected in the language. For example, the expression 'два часа ночи' (two o'clock in the morning) in Russian will have the form 'deux heures du matin' in French (literally: 'два часа утра' (two o'clock in the morning). At the same time in French, there are words 'jour, mati, nuit, soir' (day, morning, night, evening).

This system of interacting, mutually intersecting temporal-spatial meanings functions in the translation space, which we regard as a synergetic model of translation.

According to the concept we are developing, translation takes place in the process of transposing meanings from one linguo-culture to another, and the result of translation is such a text that becomes a fact of the host culture, i.e. the recipient perceives the translation as natural, made in the mother tongue.

The key idea of our concept is based on the synergetic ideology first articulated by Haken for natural sciences [8] and later developed by other scientists [9, 10]. In our research, we have relied on the results of synergetic studies in the field of linguistics [11–13]. The study of the achievements in synergetics, as well as the formation of a synergetic worldview, has enabled us to put forward the idea that meaning transposition in translation represents fundamentally a synergetic process.

The synergetics of translation studies self-development and self-organization of heterogeneous explicit-implicit meanings in the translation space of texts and/or discourses of interacting languages and cultures, and the translation itself appears as a system, the cognitive mechanism of which is meanings synergy. The result of the synergetic effect is the increment of new meanings that are acceptable in the host culture. In this case, the translation is considered harmonious, which implies recognition of semantic harmony between the source and translation texts.

We have introduced the category of harmony into the system of translation quality assessment, where on the opposite pole there is disharmony between source and translated texts. The traditional criteria of adequacy and equivalence occupy an intermediate position.

As the analysis of empirical material, as well as our pedagogical and translation practices, has shown, the achievement of translation adequacy and/or equivalence is not sufficient to make the text of translation a fact of another culture or even enrich it.

There needs to be another parameter to assess the quality of the translation. Thus, we have chosen the category of harmony as the axiomatic dominant of translation, the ideal the translator should aspire to. But there may be errors and inconsistencies in the translation process which we have referred to as disharmony. As a result, a system has been established to evaluate the quality of translation, including disharmony, adequacy, equivalence and harmony.

The self-organization of translation space is manifested in the fact that the relationships between semantic fields are not rigidly deterministic, but probabilistic. This means that, in each case, a unique configuration of translation space emerges, due to a combination of linguo-mental, linguo-cultural, intra-intertextual and intra-interdiscursive factors. This also means that there may be an infinite number of harmonious translations. And if it is true that ‘manuscripts do not burn’, then the fact that translations are obsolete is undeniable. Translation as a synergetic process captures the interaction and interpenetration of national consciousness, languages and cultures into its orbit.

We see translation space not only as an abstract translation model based on a system of arguments that reveal the patterns of synergetic processes but mainly as a reliable empirical model, components of which guide the translator’s cognitive activity.

The results of research on KT in translation process are particularly significant in this paper. Demyankov defines the KT as follows: ‘In translation studies, the term KT is used in the meaning “transfer” of some sign as an element of some sign structure and as a potential of form and function being an element of another sign structure’ [14, p. 6].

We can conclude that translation maneuvers can be presented in the form of translation strategies inherent in a certain type of text. We are interested in the translation strategies of the technical text used by the translator-linguist. The thesis study of Ushakova carried out under our supervision [15] is dedicated to this problem.

The author of the thesis was able to identify a textual category of technicity, as a means of reflecting technical knowledge in the professional technical/philological thinking of communicators. The technical knowledge acquired by participants in a communicative situation is considered on the one hand, from the point of view of the subjects: author–expert, translator–non-expert, recipient–expert, non-expert, semi-expert; on the other hand, from the point of view of the text typology: technical text as such (technical documents) and technical text of a mixed type (scientific-technical and advertising-technical texts).

Ushakova has developed a situation-reflective model of translation decision making, based on the interaction of algorithm and heuristics.

Taking into account that the technical text is, in principle, difficult for a translator-linguist, who quite often does not have a command of vocational subject content, its categorical-conceptual apparatus and, above all, terminological vocabulary of technical communication [16], we assumed that Ushakova’s model of technical translation, presented through the prism of the KT concept in the view of translation synergy, will outline not only the strategy of technical translation but also some of the linguo-cognitive mechanisms of the translator’s verbal and cogitative activities.

4 Results and Discussions

In the course of empirical analysis, we relied on the synergetic model of translation—translation space, as well as the key concept of the situation-reflective model of translation decision-making, represented in the work of Ushakova—the textual category of technicity. We suggested that this would allow us to talk about the implementation of a synergetic strategy of transferring technical knowledge from one linguistics culture to another.

Our analysis will be based, on the one hand, on the identification of a text translation macro-strategy according to which the text as a whole can be considered to be harmonious, and on a micro-strategy translation assuming that separate parts of the text can be harmonious, while others are equivalent or adequate. There have not been noticed any cases of discordant translation in the test material. We have assumed and repeatedly stated before that from a general theoretical point of view, the entire text from the beginning to the end cannot be translated harmoniously.

Let us look at a technical text from the IT field. A copy of the source text is provided in French and its translation in Russian (with English version).

Depuis le lancement de l'iPhone par la célèbre firme de Cupertini, il ne se passe pas une année sans qu'on nous annonce que cette fois-ci pas de doute, c'est bien l'année du mobile. Smartphones de plus en plus performants, explosion du nombre d'applications dans les stores, avènement des tablettes; le marché de la mobilité a été l'un des plus dynamiques de ces 10 dernières années. Et 2015 a marqué un tournant pour les investissements publicitares tant attendus.

После того как одна небезызвестная компания из Купертино выпустила айфон, каждый год мы слышим, что именно он станет годом мобильных технологий. Благодаря расширению возможностей смартфонов, огромному количеству приложений и росту популярности планшетов, рынок мобильных технологий стал одним из самых быстроразвивающихся рынков за последнее десятилетие. И в 2015 году наступил переходный момент для долгожданных рекламных инвестиций.

After a famous company in Cupertino released an iPhone, every year we hear that it's going to be the year of mobile technology. Thanks to the expansion of smartphones, a huge number of applications, and the growing popularity of tablets, the mobile technology market has become one of the fastest-growing markets in the last decade. And in 2015 came the moment of transition for the long-awaited advertising investments.

Let us analyse this translation from a macro-strategic point of view. The translation of the technical text will be harmonious when the translator can recreate the technicity category, which can be predicted while formulating a translation task, being compressed knowledge of the linguistic and extralinguistic components of the source text. In other words, the transfer of technical knowledge during translation is carried out primarily at the macro-level, which is evaluated on the basis of the identification of a technicity textual category.

This process takes place in the translation space, where the translator's field and the individual-imaginative meaning formed therein play a special role. In our previous works, we stated that the unit of meaning in translation is the image-gestalt appearing in the translator's mind [17]. Researchers' reference to the nature of the concept showed that it is conceptualization and categorization that give rise in the language consciousness to concepts that are verbalized in the text as their lexical representatives. Given this we can say that the conceptual image of the original is formed in the translator's mind, and the image is also verbalized in the text. It is the conceptual image as a compressed expression of the text meaning that will enable the translator to formulate the translation task of a particular text. In this case, it can look like this: *an ever-growing market for iPhones, tablets, different applications mean an on-coming period of mobile technology.*

When analysing the text of the translation, we note that this translation task has been reflected in the text, and consequently the translator has managed to recreate the technicity category. This means that the text of the translation is harmonious on the macro-level as a result of macro-strategy use. Micro-strategies for technical translation are determined by examining the complexities of structural and semantic elements of the text and identifying possible ways to overcome them.

Let us analyse micro-strategies using another technical text from the same IT area. The original fragment is like this:

Sopra Banking Amplitude est une solution qui s'appuie sur un traitement automatisé, et qui offre une disponibilité multi-canal 24/7. Au travers d'une démarche d'urbanisation renforcée Sopra Banking Amplitude est le Coeur de une votre système d'information bâti sur des concepts durables. Votre solution évolue au rythme de votre dévelopement stratégique.

The translated part is below:

Sopra Banking Amplitude является решением, основанным на автоматической обработке данных и предлагает круглосуточный многоканальный режим работы (24/7). Используя сервис-ориентированную архитектуру, Sopra Banking Amplitude становится сердцем вашей информационной системы, построенной с учетом долговременных перспектив. Ваше решение развивается в соответствии с вашей стратегией.

Sopra Banking Amplitude is a solution based on automated data processing and offers round-the clock (24/7) multi-channel availability. Through an enhanced urbanization approach Sopra Banking Amplitude is the heart of your information system built on long-term concepts. Your solution evolves according to your strategic development.

We will analyse the micro-strategies of the translation from the position of technical knowledge transfer in translation space. To do this, we need to identify the difficulties encountered by translators and the opportunities they have used to achieve harmony.

From a theoretical point of view, in the translation space of a technical text, these difficulties can be found in all its fields: in a semantic field where the main difficulty is the translation of terms and special abbreviations, as well as the names of organizations and firms; in the author-expert field, dealing with the terms in question with a certain intended meaning and sense; in the fatal field, where complexity depends on differences in the interpretation of the same terms in different cultures, and the task of the translator is to harmonize them with the host culture norms; in the field of the expert-recipient, as well as the recipient-non-expert, who is predicted by the translator, but who can put in the translated terms the values similar to those of the author as well as different to them; in the field of a translator—non-expert in the technical sphere of communication, in whose consciousness a conceptual image is generated; in the energy field which reflects non-verbal elements of a text (graphical elements, diagrams, drawings, tables, screenshots of screens, etc.). As we can see, each field of translation space can contain its complexities and meanings which the translator obtains, and only the synergy of meanings, accompanied by the increment of new, significant meanings for a host culture, will lead to its harmonious translation.

Our attention has been drawn to the following translation solution, which is undoubtedly a successful example to overcome the difficulties of understanding and translating special terminology. In the original version, we read *Au travers d'une une démarche d'urbanisation renforcée…* The translator presents this idea as follows: *Используя сервис-ориентированную архитектуру… (Using a service-oriented architecture…).*

First of all, it is worth explaining that the meanings of the original terms have been redefined by the translator, which led to the creation of a synergetic effect and further harmonious translation since the terms perfectly fit into the technical term system of the Russian language. In this case, it is a terminological complexity, which we mentally observe both in the meaningful field of translation space and in the field of the author. Comparing these fragments of the original and its translation, we note that the translator did not use adequate vocabulary matches, did not rely on translation equivalents, but managed to form a conceptual image of the source text in the mind, and then to verbalize this image in translation.

In the same fragment, the translator harmoniously translated the combination '*une disponibilité multi-canal 24/7*' as follows: '*круглосуточный многоканальный режим работы (24/7)' (round-the clock 24-h multichannel operating mode).* Due to the absence of the word 'day' in French, the notation 24/7 is used. The translator not only translated this combination with the word 'round-the-clock' ('круглосуточный'), but also left the notation in the Russian translation, which is, in our opinion, superfluous.

We have thus observed the implementation of macro-strategies and micro-strategies for the translation of the technical text, which provides a reliable transfer of knowledge in the technical sphere of communication. As is shown above, these strategies are based on the cognitive speaking-thinking activity of the translator, which is aimed at forming in their mind a conceptual image of the text as a unit of meaning analysis, and then on synergy and harmonization of meanings in the translation space.

In previous examples, we analysed technical texts translated by translators-linguists. Now there is an abstract of the technical text in French and its translation by a graduate student who specializes in the same subject area. The title of the original text is ‘Développement d’un nouveau capteur de gaz basé sur la détection à large bande micro-onde’ (‘Development of a new gas sensor based on microwave broadband detection’). The original text of the abstract is

Résumé

Nous présentons une méthode originale de détection de gaz. Notre capteur se compose d’un materiau sensible introduit dans une structure coaxiale. Un champ électromagnetique (micro-onde) émis à travers le capteur par un alalyseur de réseau vectoriel sollicite le materiau sensible soumis à gaz. La variation de la réponse du capteur observé est due aux variations des propriétés diélectriques du materiau sensible induite par l’absorbtion du gaz. L’étude fait apparaître que le capteur à base d’oxyde SrTiO3 est quantitatif et selectif pour trois gaz différents (eau, étanol, toluène).

The translated text is

Мы представляем оригинальный метод обнаружения газа. Наш датчик состоит из чувствительного материала, заключенного в коаксиальную структуру. Электромагнитное поле (микроволновое), излучаемое через датчик посредством анализатора векторных цепей, воздействует на чувствительный материал, подверженный воздействию газа. Различия в наблюдаемой реакции датчика обусловлены изменениями диэлектрических свойств чувствительного материала под воздействием абсорбции газа. Исследование показывает, что датчик на основе оксида SrTiO3 является количественным и селективным для трех различных газов (вода, этанол, толуен).

We present an original method of gas detection. Our sensor consists of a sensitive material encased in a coaxial structure. The electromagnetic field (microwave) emitted through the sensor through a vector circuit analyser acts upon the sensitive material exposed to the gas. The difference in the observed sensor response is due to changes in the dielectric properties of the sensitive material when exposed to gas absorption. The study shows that the oxide-based sensor SrTio3 is quantitative and selective for three different gases (water, ethanol, toluene).

We will analyse the ways to express technicity category in translation. At the grammatical level, the text does not present difficulties, as it consists mainly of simple sentences. Special technical terms are presented at the lexical level, which shows that the text of the translation is oriented towards the targeted recipients-specialists in applied physics. The text uses both general technical terms: ‘détection de gaz’ (обнаружение газа, gas detection), ‘capteur’ (датчик, sensor), ‘materiau’ (материал, material), ‘l’absorbtion’ (абсорбция, absorption), ‘proprieties’ (свойства, properties, etc.) and special terms: ‘structure coaxiale’ (коаксиальную структуру, coaxial structure), ‘étanol’ (этанол, ethanol), ‘toluene’ (толуен, toluene, etc.). The text contains an attitudinal vocabulary ‘*une méthode originale quantitatif et selectif’ (оригинальный количественный и селективный метод, quantitative and selective method)* to reveal the characteristics of the technical product. Concerning synergetic effect arising in the translation process, we state that only a technical specialist could translate the commonly used word ‘réponse’ (literally 'ответ' (answer) as 'реакция' (response), when it is a question of 'sensor response'.

Evaluating the technical translation performed by the specialist in general, we note that it is harmonious because the technicity category is also reflected here and is understood by the Russian-speaking recipient. We believe that harmony in technical text translation is manifested when it comes to the expression of a text-based technicity category and the achievement of synergetic effect. It is conceivable, however, that some parts of the text have been translated adequately and equivalently.

So according to the synergetic concept of translation that we are developing, the modelling of translation actions is not linear, and we relate it to the traditional criteria for achieving qualitative translation—adequate or equivalent translation, but «vector-like», i.e. the result of a differently directed movement of translator's thoughts: to the understanding and transposition of meanings both in text fields and in the fields of translation communication subjects, i.e. the combination of texto-centric and anthropocentric approaches, which we relate to the translation harmony.

The key idea of our work is that translation space is created by the translator. Innovation in any field, including translation, is carried out by a creative linguistic persona. In this regard, we believe that the study of KT in translation, based on the subject of translation communication, is also promising.

5 Conclusion

In this article, we have attempted to link the current problem of technical KT with the translation of a text in the view of cognitive synergetic paradigm.

Our research is based on the synergetic concept of translation—the translation space, which is a heterogeneous space of meanings that the translator studies to harmoniously transpose these meanings into another language and another culture. We have focused on technical communication and technical culture, which is embodied in different ways in the culture of people, in the professional culture of specialists who are experts in their field but who turn to translators-linguists (not experts) in certain situations. The transfer of technical knowledge during translation is carried out successfully, provided the translator masters synergetic methodology and understands the text components forming the technicity category. This process can be analysed by identifying macro- and micro-strategies for translation that lead to harmonious translation and reliable transfer of technical knowledge, which eventually results in the technical progress of interacting linguo-cultures.

References

1. Kushnina LV (2003) Gipoteza o formirovanii i dinamicheskom razvitii perevodcheskogo prostranstva [Hypothesis on the formation and dynamic development of translation space]. Bull Tyumen State Univ (Tyumen) 4:144–149 (in Russian)
2. Alekseeva LM, Mishlanova SL (2019) Transfer znaniia v gumanitarnykh naukakh [Transfer of knowledge in the humanities]. Publishing House of Perm National Research Polytechnic University, Perm (in Russian)
3. Demyankov VZ (2017) Transfer znanii i kognitivnaia manipuliatsiia [Knowledge transfer and cognitive manipulation]. Cogn Linguist Issues 4:5–9 (in Russian)
4. Olmos- J, Castro- E, D'Estea P (2014) Knowledge transfer activities in social sciences and humanities: explaining the interactions of research groups with non-academic agents. Res Policy (Elsevier) 43(4):696–706
5. Lotman YM (2010) Unpredictable mechanisms of culture. Publishing House of TLU Press, Tallinn (in Russian)
6. Aumüller M (2020) Text type. https://www.lhn.uni-hamburg.de/node/121.html. Accessed 20 Sept 2020
7. Ogneva EA (2020) Transfer kulturno-markirovannogo znaniia [Transfer of culture-specific knowledge. Scientific result. Questions of theoretical and applied linguistics] 6(2):38–51 (in Russian)
8. Haken G, Haken-Krell M (2002) Secrets of perception (trans: from German by Logunov AR). Moscow
9. Castro- E, Fernández I, Pérez- M, Criado-Boado F (2008) The knowledge transfer from the humanities: possibilities and characteristics. ARBOR Cienc Pensam y Cult 184:619–636
10. Landry R, Amara N, Ouimet M (2007) Determinants of knowledge transfer: evidence from Canadian university researchers in natural sciences and engineering. J Technol Transf 32:561–592
11. Myshkina NL (2018) Vnutrenniaia zhizn teksta: mekhanizmy, formy, kharakteristiki [The inner life of the text: mechanisms, forms, characteristics]. Publishing House of Perm University, Perm (in Russian)
12. Moskalchuk GG (2003) Struktura teksta kak sinergeticheskii protsess [Text structure as a synergistic process]. URSS, Moscow (in Russian)
13. Drozhashchikh NV (2012) Vvedenie v dinamicheskuiu sinergetiku iazyka [Introduction to dynamic language synergetics]. Publishing House of Tyumen State University, Tyumen (in Russian)
14. Demyankov VZ (2016) O kognitsii, kulture i tsivilizatsii v transfere znanii [On cognition, culture and civilization in knowledge transfer]. Cogn Linguist Issues 4:5–9 (in Russian)
15. Ushakova AO (2017) Tekstovaia kategoriia tekhnichnosti i ee funktsionirovanie v originale i perevode [Textual category of technicity and its functioning in original and translated]. Dissertation abstract. Izhevsk (in Russian)
16. Serova T, Perlova I, Pipchenko E, Chervenko Y (2020) Intellect and intellectual speech-thinking activity of technical university students in integrative professionally-oriented technology of foreign language learning. In: Anikina Z (ed) Integrating engineering education and humanities for global intercultural perspectives. IEEHGIP 2020. Lecture notes in networks and systems, vol 131. Springer, Cham. https://doi.org/10.1007/978-3-030-47415-7_42. Accessed 20 Sept 2020
17. Kushnina LV (2004) Iazyki i kultury v perevodcheskom prostranstve [Languages and cultures in translation space]. Publishing House of Perm State Technical University, Perm (in Russian)

User Experience Framework for Teachers Using Mobile Technologies in Resource-Constrained Environments

N. R. Makama, C. J. van Staden, and A. Botha

Abstract The use of mobile technologies at schools in teaching and learning is emerging, enabling teaching and learning to happen anywhere anytime. CSIR implemented a project in resource-constrained environments schools specifically in Gauteng, Limpopo and Northwest Province in South Africa. Teachers were trained to use mobile technologies in teaching and learning. The study focussed on identifying the components and factors that may have an influence on the user experience (UX) of teachers using mobile technologies in a resource-constrained environment. The study used an explorative qualitative research strategy to explore the users' opinions, beliefs and perception through a questionnaire, where a thematic analysis approach was applied to analyse the qualitative data. The paper proposes a user experience conceptual framework for teachers using mobile technologies.

Keywords User experience · Mobile technology · Context · Resource-constrained environments · User experience components · Teachers · ICT in education

1 Introduction

The use of information and communication technology (ICT) has emerged globally. The occurrence of ICT is believed to have influenced the development of society including schooling [1]. Internet and mobile technologies are considered ICT tools, making it possible for schools to access e-education using mobile technologies in teaching and learning anywhere anytime [2, 3]. Access to technology in schools is not only limited to urban schools, it also includes the schools located in rural areas. According to Chinapah and Odero [2] education has the potential to improve development in rural areas, hence the importance of education transformation in rural schools. "Rural schools are based in rural environments and are considered to be

N. R. Makama (✉) · C. J. van Staden · A. Botha
School of Computing, Science Campus, University of South Africa, 28 Pioneer Ave, Florida Park, Roodepoort 1709, South Africa
e-mail: Vstadcj1@unisa.ac.za

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_24

resource-constrained environments, because they often lack proper road infrastructure, have low tele-connectivity, market growth is very low and have a low income" [4, 5]. Further limitations include limited access to electricity, low levels of literacy, a lack of infrastructure, lack of technology and technical constraints, all of which require people to be innovative [3].

There's a challenge in the adoption of the use of technologies in schools, compounded by the lack of infrastructure, shortage of qualified teachers capable of using the technology, technophobic teachers and lack of experienced teachers in using the technologies for teaching and learning [6]. As mobile technology becomes increasingly used in schools, and teachers are expected to conduct teaching and learning in the classrooms using mobile technologies, it is important to ensure that teachers are comfortable with using mobile technologies and that the user experience (UX) of teachers using mobile technologies is enhanced. Negative user experiences have been reported and have led to limited participation and engagement, resulting in the ineffective utilisation of the expected pedagogical gains and organisational benefits as well as the abandonment of using technologies [7]. As a result, it is suggested that the UX of the teachers be acknowledged in order to provide an optimal learning environment [8]. It is against this backdrop that the following research question has been identified: *What are the components and factors of user experience that are relevant to teachers using mobile technologies in resource-constrained environments?* By exploring the user experience of teachers using mobile technologies in resource-constrained environments, in Gauteng, North West and Limpopo the ICT4E project was implemented in rural schools. The paper will present a literature review, research methodology followed and the findings in an attempt to answer the research question.

2 Literature Review

In 2016, the South Africa's Council for Scientific and Industrial Research (CSIR) signed an agreement with the Department of Rural Development and Land Reform (DRDLR) for the implementation of the Information Communication and Technology for Education (ICT4E) project in rural schools, with the aim of incorporating the use of mobile technology in teaching and learning [9]. The project was implemented in 24 rural schools in 7 of the 9 provinces in South Africa. The provinces included Limpopo, Gauteng and North West [10]. The IDEA lab at the University of the Free State was appointed to facilitate the training of teachers through a short learning programme known as the teacher professional development (TPD) course [11]. The TPD course included skills, knowledge and strategy improvement in ICT, where teachers were trained to use the mobile technologies (tablets) in teaching and learning and were awarded certificates for completing the course [12].

Training teachers to use mobile technologies in the classroom is important, and therefore, teachers are expected to apply their technical skills effectively when using the technology at schools [9]. The user experience of the teachers who participated in the ICT4E needs to be evaluated with an intention to identify factors that may have an influence on the use of mobile technologies in a resource-constrained environment. The following topics are covered in this section: the conceptualisation of UX and the components of UX, exploring the UX components, user, system, and context components.

2.1 User Experience

User experience provides factors that are used to determine the UX of the user such as emotions, attitude, perceptions and the user's expectations [13, 14]. As the teachers apply their skills in using the technology in the classrooms one should also take into consideration the "beliefs, attitude and anxiety levels of teachers when integrating mobile technology into their classrooms" [9].

According to Kaasinen et al. [15] UX is a distinctive concept, as it can be explored based on what the researcher wants to focus on. As distinctive as is the concept of UX, there's no agreement on the definition of UX; it is viewed from different perspectives because each product or system may have different goals for UX [16–18]. User experience is defined as "the degree of positive or negative emotions that can be experienced by a specific user in a specific context during and after product use and that motivates for further usage" [19]. User experience is described by the International Organisation for Standardisation's current ISO standard 9241-11 on ergonomics of human–system interaction as "a person's perception and responses resulting from the use and/or anticipated use of a product, system or service" [20].

The concept of UX is known to have several different definitions by various researchers [21–25]. These definitions share common denominators which were identified as components, namely user, system, environment or context.

Kuniasky [23] simplifies the concept of UX, as being used to evaluate the effectiveness and efficiency of a system through the user's feedback by way of emotions, behaviour, attitude, expectations and perception of the system. This study adopted the definitions proposed by Santo [19] and Scapin [26], which state that characteristics such as perception, emotions, attitude, behaviour and expectations are incorporated when users interact with the system, which results in a UX being formed. In general, UX is about evaluating the system through the user's feedback when interacting or after interacting with the system. In this study, the system that is of interest is mobile technology.

2.2 Exploring the UX Components

Every system is designed for a different purpose, hence the UX in all systems may also differ; as a result, the components of each system may differ depending on the purpose of the system [16]. A "**component**" is defined as "any element of context [used] to combine the previously used terms of factors, components, dimensions, aspects, state, and environment to under same umbrella" [27]. For the purposes of this study, the word "**components**" will be used as defined above and factors will be used to refer to the subcomponents. UX factors "can be used to describe the situation in which a person felt a particular UX" [28]. For the purposes of this study, the characteristics of the factors will be interrelated to the factors that affect components of the UX. Various literature works have identified different UX components. Table 1 shows the explicitly identified components and Table 2 shows the implicitly identified components.

Table 1 Explicitly identified UX components

References	Identified components	Factors linked to the components
[29]	Framework identified the influential components of UX, namely **user**, **product**, **environment** and **designer**	The identified components illustrate the user–product–environment interaction, which is equivalent to the teacher–technologies–school structure. Perception, expectations, hedonic were also identified in the framework as relevant to UX evaluation
[30]	Established the three design components framework, namely **compositional**, **aesthetic** and the **aspect of meaning**	Related factors: usability, behavioural characteristics, satisfaction, aesthetic, perception and emotions relevant to the evaluation of the UX of the teachers
[31]	The framework identified the **system** component	The factors that were identified as influencing the UX in relation to the system component include useful, usable, desirable, valuable, findable, accessible, credible
[32]	The framework identified three components of UX, namely **emotional user reaction**, **instrumental quality perception** and **non-instrumental perception** which occurred when there is an interaction	Factors of the identified components such as efficiency, flexibility and usability are suitable for evaluating the technologies, behaviour, emotions and perceptions, and were used to evaluate the teachers' UX of technologies
[33]	The framework demonstrates the interaction with the product in a particular context. The components **user**, **product** and **context** were identified	Factors include experience, expectations, skills, usability, aesthetic characteristics and contexts where the interaction occurs. These factors are suitable for the evaluation of the UX of teachers at schools using the technologies
[34]	The factors identified in the framework are **user, system** and **context**	Factors include attitudes, expectation, knowledge and motivation. The product includes people, infrastructure and services involved during interaction. The social context includes physical context and task context

Table 2 Implicitly identified UX components

Reference	Identified UX components	Factors linked to the UX components
[22]	User, system, context	The user's internal state includes expectations, motivation, mood and predispositions. The characteristics of the system: functionality, usability, complexity and motive. The environment where interaction occurs: organisation, voluntariness of use
[34]	User, social factors, cultural factors, context of use, product	The state of the user includes mental and physical state, which influences the user's perception. The product, infrastructure involved during interaction. The environment where interaction occurs, social and temporal factors. The influence other people have concerning the user's expectations or willingness to participate in social situations
[34]	User, context, system	The user's mental state includes attitudes, expectation, knowledge and motivation. The product, people, infrastructure, services involved during interaction. The social context, physical context and task context
[28]	User, system, context, social, aesthetic	The user's mental and physical state includes motivation to use the system, mood and expectations. The properties of the designed system influence the user's perception. UX depends on where interaction occurs. Different contexts: physical, social, task, information and technical context
[35]	Aesthetic pleasure, semantic interpretation, emotional response	Cognitive response: product is perceived through senses such as pleasure, subjectivity and objectivity. Affective response: moods, feelings in response to the interaction with the product. Involves the attributes of the product such as the shape, colour and qualities of the product (performance, function)
[26]	Aesthetics, emotions, observing, questionnaires, heuristics	Involves psychological and social measurement of the user, also physiological measurement (heart rate, skin response, eye tracking). Emotions such as feelings, behaviour

This study acknowledged the three components: user, system and context, as the dominant components in the literature reviewed. It is understood that the evaluation of UX is incomplete without these components [36]. Therefore, this study adopted the three components:

- User component

In the UX context, a "user" is defined as a person who interacts with the system, where factors such as the user's needs, expectation, motivation, perception and mental state are evoked during and after the interaction [28, 37]. Another definition of user "refer[s] to the entity (individual or social group) interacting with the product" [38]. Users are at the centre of the UX, as UX is evaluated through the user's response after interacting with the system, and without the user, there is no UX [39]. For the purposes of this study, the researcher referred to the user as the teacher.

- System component

A system is described as "the application or device under examination which the people interact with" [38]. "A system is sometimes considered as a product or as the services it provides" [20]. The system is amongst the foremost components and UX is not complete without it. For UX to occur, there must be an interaction between a user and the system [39]. This study used the term "system" to refer to mobile technologies. Mobile technologies are mobile devices that enable a user to move around and include devices such as tablets, PDAs, smartphones and portable computers [40].

- Context component

Context refers to the physical environment where the user interacts with the system [22, 28]. According to [2], context can be either an object, person or place that forms part of the interaction between the user and the system. In a situation where the user uses mobile technologies as a tool to access information, context can be understood as the location of use, which includes the physical place and social conditions where the interaction occurs [2]. For this study, the context being the school (resource-constrained environment).

The objective of this literature study was to explore the UX components with the aim of identifying the factors that may affect the UX of the teachers using mobile technologies in resource-constrained environments. Three components—user, system and context—were identified, including the subcomponents (factors) of each component. The literature study guided the design of the conceptual framework. The conceptual framework makes it easier for the researcher to define the concept of the problem under investigation by showing the variables' relationship using a graphical or narrative technique [41]. Table 3 illustrates the summary of the conceptual framework, constructed from various literary works.

Table 3 Conceptual framework summary

Components	Subcomponents—factors	References
User: teachers	Needs, perceptions, attitude, expectation and experiences	[25, 28, 31, 32, 35, 38, 42–46]
System: mobile technologies	Usability, hedonic and aesthetic attributes, functionality, control and ownership, flexibility, credibility, valuable, desirable, technological skills, critical thinking, problem-solving skills, creativity and qualified teachers	[6, 12, 26, 31, 35, 42, 47–53]
Context: resource-constrained environment	Physical context, social context, task context, technical and information context, policy implementation and training	[5, 6, 27, 33, 37, 39, 48, 54, 54–56]

3 Research Methodology

An explorative qualitative research strategy was implemented in this research, as explorative research focusses on exploring the phenomenon with the aim of identifying the problem and gaining insight. The aim of using an explorative qualitative strategy is to identify the variables that may be the source of the problem. Explorative research complements qualitative research, as qualitative research enables the researcher to explore the study and answers the explorative questions [57]. According to Kalu and Bwalya, qualitative research is used in a human behavioural field in exploring the experiences of the human, beliefs, values and culture through the in-depth exploration of the literature, therefore qualitative research is useful in exploring a complex phenomenon [58].

Explorative research is explored qualitatively using qualitative data collection methods [57, 59]. Qualitative research enables participants to express their perspectives, beliefs, opinions and experiences [60]. This research strategy is relevant to the study as the perspective, belief, opinion and experience of the teachers are important. A questionnaire was distributed to 45 teachers, of which 30 teachers responded and only 27 completed questionnaires were used because three of the questionnaires were spoilt. Data were collected through open-ended and closed-ended questions to get the participants' opinions about the identified factors that may have an influence on the UX of the teachers using mobile technologies. The factors resulting from the literate study are presented in a conceptual framework.

4 Data Analysis and Findings

Factors that are considered to have an influence on the UX of teachers were identified in the literature study; therefore examining the teachers' perception, beliefs and opinions about the factors may contribute to answering the research question. The questions that were used in a questionnaire were piloted and tested by four experts in the field of UX and mobile technologies testing if the questions were understandable, transparent and equitable for the participants. The questionnaire consisted of four sections: A, B, C and D. Section A was used to collect the demographic information of the participants, while Sections B, C and D were in relation to the components user, system and context with its factors, respectively. The purpose of Sections B, C and D was to find out how the factors of the components influence the participants when using mobile technologies in teaching and learning activities in resource-constrained environments. A five-point Likert scale was used in Sections B, C and D (1 = Strongly agree, 2 = Agree, 3 = Not sure, 4 = Disagree and 5 = Strongly disagree). Teachers were also expected to comment next to each question, to gain insight and for qualitative analysis. A total of 27 responses were used to analyse the data.

The participants were aged between 21 and over 50, where the majority of participants (44%) rated themselves as highly skilled or average. Seventeen (17) of the participants were female and ten (10) males.

This study applied the thematic analysis, using the Atlas.ti tool to analyse the raw data and manage the collected data, using codes and themes [60]. Data were transcribed verbatim to ensure that the participant's responses were not misinterpreted. A total of 27 participants' scripts were uploaded to the Atlas.ti tool, where quotations, codes and themes were linked to the documents and produced the network diagram that is used to present the analysed data visually.

The participants' feedback revealed whether they agreed with the literature study or not, where all factors were tested in a questionnaire. As it has already been mentioned that the study used a five-point Likert scale, in this scenario the results interpreted indicate either agree or disagree (where strongly agree and agree results are combined, same with strongly disagree and disagree results).

4.1 Results: User (Teacher)

Ninety-two percent (92%) agreed that technologies need to be user-friendly (ease of use). Teacher-2 indicated that "ease of use to teachers is very important to improve learners' skills". Ninety-two percent (92%) agreed that they perceive technologies as useful in teaching and learning. Teacher-23 agrees that "[p]ositive perception will build confidence in the teacher". Approximately ninety-six percent (96%) agreed (10) or agreed (16) that the attitude of the teacher towards the use of the technologies is very important. Teacher-7 said that "[o]nce the teacher is positive also the learners will have a positive attitude". Ninety-six percent (96%) strongly agreed that the experience of teachers when using the technologies at school is very important. Seventy percent (70%) disagreed that gender influences the experience of teachers when using mobile technologies at school. Sixty-three percent (63%) disagreed that age influenced the use of technologies at school. Teacher-8 indicated "No gender equity in technology" and Teacher-20 "Age is not an aspect to influence usage of tablets". Sixty-three percent expect the use of technologies to support their teaching needs. Teacher-23 expected the technologies to "reduce the workload".

4.2 Results: System (Mobile Technologies)

Eighty-three percent (83%) agreed that they find the use of technologies efficient. Teachers' feedback: "The easier the tablet, the more efficiently they can be used". Eighty-one percent (81%) agreed that the visual attributes of the technologies do have an influence on the user experience of the teachers. Seventy percent (70%) agreed that functionalities are supposed to be easily accessible. Teacher-7 indicated "I'm very satisfied with the functions of the tablets." Teacher-14 indicated that the appearance influences their use of the technology: "the appearance and features influence the use of technology". Eighty-nine percent (89%) agreed that having control and ownership motivates the teachers to navigate the technologies freely. Teacher-2 felt that their "information is confidential". Ninety-six percent (96%) agreed that the flexibility of the technologies gives them the ability to work anywhere. Teacher-7 reported that the technology "works everywhere doesn't keep the teacher at hostage". Hundred percent (100%) agreed that the technologies should be reliable. Teacher-23 said that "results produced by the tablets must be valid". Eighty-nine percent (89%) agreed that the use of technologies adds value to teaching and learning. Teacher-9: "tablets add value to my lesson, they are valuable". Eighty-nine percent (89%) expect that the technologies should be pleasing to interact with. Teacher-14: "[i]f its attractive to use, they'll be enjoyable to use them". Ninety-six percent (96%) agreed that teachers should know how to operate the technologies for teaching and learning purposes. Teacher-4: "4th Industrial revolution technology is on the way, more training for all the teachers is required". Ninety-two percent (92%) agreed that teachers need critical thinking skills to use the technologies. Teacher-24 agreed that "[c]ritical thinking is required". Fifty-five percent (55%) disagreed that they should have problem-solving skills to work on technical tasks in teaching. Teacher-9: "[t]eachers are not technicians". One hundred percent (100%) agreed (8) that creative skills are required to apply knowledge when using technologies. Teacher-15: "[r]equired to think creative in order to apply knowledge". One hundred percent (100%) agreed that teachers' professional development is important to accelerate the use of technologies at schools. Teacher-1: "[d]evelopment is important to accelerate the use of tablets".

4.3 Results: Context (Resource-Constrained Environment-School)

Ninety-two percent (92%) agreed that the environment where teaching and learning are happening is important. Teacher-19: "[p]oor environment result in poor experience in using the tablets at schools". Seventy-four percent (74%) agreed that instructions from the school-governing body and/or principals on how the technologies should be used may influence their experience in the use of the technologies. Teacher-19: "[t]he opinion and instructions from all stake holders' influence teachers experience". Approximately fifty-nine (59%) disagreed that multitasking has an influence and may negatively affect concentration. Teacher-23: "[t]eacher should be able to deal with multiple tasks". Ninety-six percent (96%) agreed that the availability of ICT services at schools using technologies is important. Teacher-27: "ICT must provide with service". Eighty-nine percent (89%) agreed that improper implementation of ICT frameworks may influence the experience of the use of technologies in school. Teacher-12: "[s]ome policies work against tablets". Eighty-nine percent (89%) agreed that the skills of teachers in ICT are important. Teacher-4: Training is required "in order to use the gadgets more effectively".

This evidence supports the theoretical basis of the study [26]; the factors identified in the literature study are summarised in Table 3. However, the feedback implies that characteristics of experience factor such as age and gender do not influence the user experience of the teachers. The problem-solving skills factor is necessary to have but does not have an influence on resolving the technical issues experienced with the technologies. The task context characteristic focus is important to have, but characteristics of task context factor such as multitasking do not affect the use of technologies in teaching and learning.

Factors such as critical thinking and creativity shared the same characteristic: Innovative. Since critical thinking has other characteristics that influence the use of technology, it inherited the innovative characteristic and creativity was excluded. Factors such as desirable and hedonic share the same characteristics, including pleasing and attractive. Since hedonic has other characteristics that influence the use of technology, it inherited the pleasing and attractive characteristics and the desirable factor was excluded in the revised conceptual framework. The proposed conceptual framework is presented in Fig. 1.

The outcome of this paper is the conceptual framework presenting the components and factors that were identified in the literature study, which may have an influence on the UX of teachers using mobile technologies at schools for teaching and learning. The conceptual framework can be used by researchers to identify the concept of the problems in the study where teachers are using mobile technologies in teaching and learning in a resource-constrained environment.

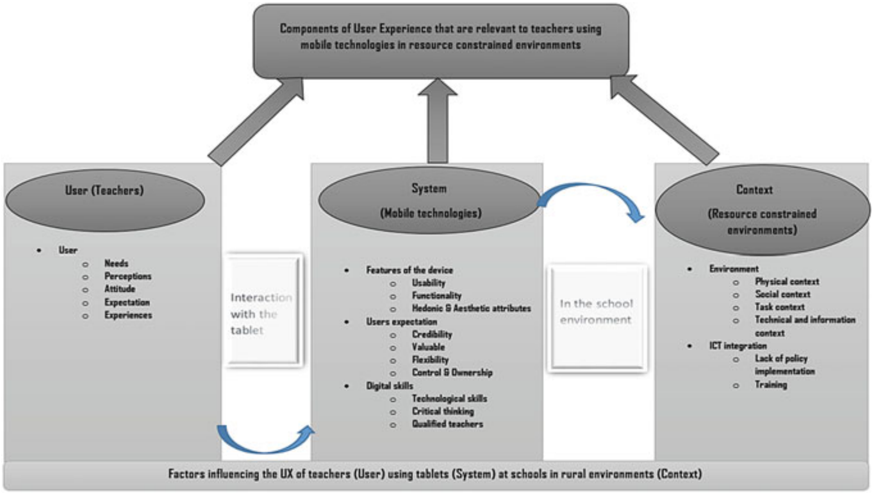


Fig. 1 Proposed conceptual framework

5 Conclusion

It is without a doubt that the use of mobile technology at school is improving and transforming the learning process in education and it is embraced by both teachers and learners. However, the use of technology cannot be fully implemented and adopted in schools if the user experience of either teachers or learners is negative. The research to examine the UX of the teachers is necessary and the conceptual framework is significant in the field of study. It will assist the researchers who are interested in understanding the factors that may influence the use of mobile technologies in schools in a resource-constrained environment.

Acknowledgements This paper is based on the thesis for the Master of Science in Computing degree at the University of South Africa (UNISA).

References

1. Bornman E (2015) Information society and digital divide in South Africa: results of longitudinal surveys. Inf Commun Soc. https://doi.org/10.1080/1369118X.2015.1065285
2. Chinapah V, Odero JO (2016) Towards inclusive, quality ICT-based learning for rural transformation. J Educ Res 2(5). https://doi.org/10.3126/jer.v5i0.15733
3. Mbebe TM (2017) Analysing the developmental role of ICT in the case of Bakgoma community library. Masters dissertation, Management and Economic Sciences. Stellenbosch University
4. Herselman M, Botha A, Maremi K, Dlamini S, Marais M (2020) Mobile technology affecting teaching and learning in rural schools. In: Conference: 16th international conference on mobile learning 2020, pp 23–30

5. Mamba MSN, Isabirye N (2015) A framework to guide development through ICTs in rural areas in South Africa. Inf Technol Dev 21(1):135–150. https://doi.org/10.1080/02681102.2013.874321
6. Mabila J (2017) A sustainability framework for mobile technology integration in resource-constrained schools: a case study in South Africa. Doctoral dissertation, Information Systems. UNISA
7. WDSglobal (2010) Mass adoption of smartphones eroding carrier profitability. http://www.wdsglobal.com/news/archive/2010/20100420/smartphones-eroding-carrier-profitability.pdf
8. Davis R, Wong D (2007) Conceptualizing and measuring the optimal experience of the eLearning environment. Decis Sci J Innov Educ 5(1):97–126
9. Herselman M, Botha A, Dlamini S, Marais M, Mahwai N (2019) Findings from a mobile tablet project implementation in rural South Africa. In: 15th international conference mobile learning, pp 49–56. https://doi.org/10.33965/ml2019_201903L007. https://www.researchgate.net/publication/333271018_FINDINGS_FROM_A_MOBILE_TABLET_PROJECT_IMPLEMENTATION_IN_RURAL_SOUTH_AFRICA
10. Herselman M, Botha A, Mayindi D, Reid E (2018) Influences of the ecological systems theory influencing technological use in rural schools in South Africa: a case study. In: 2018 International conference on advances in big data, computing and data communication systems (icABCD), pp 1–8
11. Botha A, Herselman M, Musgrave S, Jaeschke G (2017) Dimensional evaluation of a rural mobile learning teacher professional development curriculum. In: Conference: Unisa ISTE, Krugerpark, South Africa. https://www.researchgate.net/publication/322640943_Dimensional_evaluation_of_a_rural_mobile_learning_teacher_professional_development_curriculum
12. Mabila J, Herselman M, Van Biljon J (2016) Digital content and sustained use in integrating tablet technology into teaching in resource constrained environments in South Africa: educators' views. University of South Africa, Council for Scientific and Industrial Research, South Africa. https://researchspace.csir.co.za/dspace/bitstream/handle/10204/9251/Mabila_17748_2016.pdf?sequence=1&isAllowed=y
13. Alhussayen A, Alrashed W, Mansor EI (2015) Evaluating the user experience of playful interactive learning interface with children. In: 6th international conference on applied human factors and ergonomics (AHFE 2015) and the affiliated conferences, AHFE 2015. Procedia Manufacturing, pp 2318–2324
14. Kuusinen K, Väätäjä H, Mikkonen T, Väänänen K (2016) Towards understanding how agile teams predict user experience. In: Integrating user-centred design in agile development, pp 163–189. https://www.researchgate.net/publication/308839307_Towards_Understanding_How_Agile_Teams_Predict_User_Experience
15. Kaasinen E, Roto V, Hakulinen J, Heimonen T, Jokinen JPP, Karvonen H, …Turunen M (2015) Defining user experience goals to guide the design of industrial systems. Behav Inf Technol 34(10):976–991. https://doi.org/10.1080/0144929X.2015.1035335
16. Roto V, Rautava M (2008) User experience elements and brand promise. http://www.academia.edu/16090594/User_experience_elements_and_brand_promise
17. Van Schaik P, Aranyi G (2014) Research white paper model-based user experience engineering. Teesside University
18. Nakamura WT, De Oliveira EHT, Conte T (2019) Negative emotions, positive experience: what are we doing wrong when evaluating the UX. Glasgow, Scotland, UK, pp 1–6. https://doi.org/10.1145/3290607.3313000
19. Santosa PI (2016) Measuring user experience during a web-based survey: a case of back-to-back online surveys. Int J Adv Sci Eng Inf Technol 2(6):339–344
20. ISO 9241-11:2018 (2018) Ergonomics of human-system interaction—part 11: usability: definitions and concepts. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en
21. Hassenzahl M (2008) User experience (UX): towards an experiential perspective on product quality. In: IHM, Semantic Scholar, pp 11–15. https://doi.org/10.1145/1512714.1512717

22. Hassenzahl M, Tractinsky N (2006) User experience a research agenda. Behav Inf Technol 25(2):91–97. https://doi.org/10.1080/01449290500330331
23. Kuniavsky M (2010) Smart things: ubiquitous computing user experience design. British Library Cataloguing-in Publication Data, London
24. Law EL, Roto V, Hassenzahl M, Vermeeren AP, Kort J (2009) Understanding, scoping and defining user experience: a survey approach. In: CHI, Semantics Scholar. https://doi.org/10.1145/1518701.1518813
25. Tullis T, Albert B (2013) Measuring the user experience, collecting, analysing and presenting usability metrics, 2nd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
26. Scapin DL, Senach B, Trousse B, Pallot M (2012) User experience: buzzword or new paradigm? In: ACHI 2012—5th international conference on advances in computer-human interactions, pp 336–341
27. Jumisko-Pyykkö S, Vainio T (2010) Framing the context of use for mobile HCI. Int J Mob Hum Comput Interact 2:1–28. https://doi.org/10.4018/jmhci.2010100101
28. Roto V, Law E, Vermeeren A, Hoonhout J (2011) User experience white paper, bringing clarity to the concept of user experience. http://www.allaboutux.org/files/UX-WhitePaper.pdf
29. Abbasi MQ, Lew P, Rafique I, Li Z (2012) User experience evolution lifecycle framework. Int J Ind Manuf Eng 6(1). International Scholarly and Scientific Research & Innovation. https://waset.org/publications/8683/user-experience-evolution-lifecycle-framework
30. Kort J, Vermeeren APOS, Fokker JE (2007) Conceptualizing and measuring user experience, pp 57–64
31. Morville P (2004) User experience design. Semantic Studios. http://semanticstudios.com/publications/semantics/000029.php
32. Mahlke S (2008) Visual aesthetics and the user experience. In: The study of visual aesthetics in human-computer interaction, Research Gate
33. Arhippainen L (2009) Studying user experience: issues and problems of mobile services case ADAMOS: user experience (im)possible to catch. Doctoral dissertation, University of Oulu, Finland
34. Roto V (2007) User experience from product creation perspective. Nokia Research Center, Research Gate. https://www.researchgate.net/profile/Gavin_Macarthur/publication/257048076_Making_User_Experience_a_Business_Strategy/links/02e7e52441da6bbdcc000000.pdf#page=37
35. Gentner A, Bouchard C, Favart C (2013) Investigating user experience as a composition of components and influencing factors. In: 5th international congress of international association of societies of design research, Tokyo
36. Keskinen T (2015) Evaluating the user experience of interactive systems in challenging circumstances. Dissertation in School of Information Sciences, University of Tampere
37. De Kock E (2017) User experience of academic lecturing staff in the use of a learning management system tool: a case study at an open distance learning institution in South Africa. Masters dissertation, UNISA. http://uir.unisa.ac.za/bitstream/handle/10500/24965/dissertation_de%20kock_e.pdf?sequence=1&isAllowed=y
38. Mashapa J (2009) An investigation into user interface factors impacting on user experience: pastel accounting study. Masters dissertation, Nelson Mandela Metropolitan University
39. Mashapa J (2013) A model managing user experience. Doctorate dissertation, Nelson Mandela Metropolitan University
40. Huang G, Tsai C (2011) Research trends in mobile and ubiquitous learning: a review of publications in selected journals from 2001 to 2010. Br J Educ Technol 42(4):65–70. https://doi.org/10.1111/j.1467-8535.2011.01183.x
41. Adom D, Hussein E, Joe A (2018) Theoretical and conceptual framework: mandatory ingredients of a quality research. Int J Sci Res 7:438–441
42. Chan F, Johansson S (2016) Evaluation of user experience on a mobile application. KTH Royal Institute of Technology, School of Information and Communication Technology, Stockholm, Sweden

43. Maguire M (2013) Using human factors standards to support user experience and agile design. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39188-0_20
44. Langenhoven C (2016) User experience of learners in technology-facilitated learning in a resource deprived context. Masters dissertation, University of Pretoria. https://repository.up.ac.za/bitstream/handle/2263/52939/Langenhoven_User_2016.pdf?sequence=1
45. Chiu KF, Churchill D (2015) Adoption of mobile devices in teaching: changes in teacher beliefs, attitudes and anxiety. Interact Learn Environ. http://hdl.handle.net/10722/222576
46. Tan J (2009) FOUUX—a framework for usability & user experience. Master's thesis, Blekinge Institute of Technology, Sweden. https://www.diva-portal.org/smash/get/diva2:829950/FULLTEXT01.pdf
47. Bidin S, Ziden AA (2013) Adoption and application of mobile learning in the education industry. In: 6th international conference on university learning and teaching. Procedia Social Behav Sci 90:720–729
48. Blackboard (2008) Teaching in the 21st century, a review of the issues and changing models in the teaching profession. https://www.blackboard.com/resources/k12/k12_teaching21st_final.pdf
49. Petrie H, Bevan N (2009) The evaluation of accessibility, usability and user experience. In: Stepanidis C (ed). The universal handbook. CRC Press, Boca Raton
50. Brown TH, Mbati LS (2015) Mobile learning: moving past the myths and embracing the opportunities. Int Rev Res Open Distrib Learn 16(2):115–135
51. Hlagala AR (2015) Mobile educational technologies currently used as a means to enhance teaching and learning in privileged high school. Masters dissertation, Information Science, UNISA
52. Chisholm L (2011) The challenge of South African schooling: dimensions, targets and initiatives. http://transformationaudit.org/blog/wp-content/uploads/2012/02/Review-The-challenge-of-South-African-schooling.pdf
53. Becta (2010) 21st century teacher: are you ready to meet the challenge. Leading next generation learning. https://www.erte.dge.mec.pt/sites/default/files/Recursos/Estudos/21st_century_teacher.pdf
54. Botha A, Herselman M (2017) Teachers become cocreators through participation in a teacher professional development (TPD) course in a resource constraint environment in South Africa. Electron J Inf Syst Dev Ctries. https://researchspace.csir.co.za/dspace/bitstream/handle/10204/9970/Botha_20061_2017.pdf?sequence=1&isAllowed=y
55. Donohue D, Bornman J (2014) The challenges of realising inclusive education in South Africa. S Afr J Educ 34(2)
56. Nkula K, Krauss K (2014) The integration of ICTs in marginalized schools in South Africa: considerations for understanding the perceptions of in-service teachers and the role of training. In: Conference: 8th IDIA conference at: Port Elizabeth, South Africa
57. Swanson AC (2015) Exploratory research. Research gate
58. Kalu FA, Bwalya JC (2017) What makes qualitative research good research? An exploratory analysis of critical elements. Int J Social Sci Res 5(2):43–56
59. Malhotra NK (2010) Marketing research: an applied orientation, 6th edn. Pearson Publication, London
60. Young B, Hren D (2006) Introduction to qualitative research methods. Methods in Research on Research (MiRoR). http://miror-ejd.eu/wp-content/uploads/sites/34/2017/03/Introduction-to-qualitative-research-methods.compressed.pdf

Two-Phase Approach for Solving the Rich Vehicle Routing Problem Based on Firefly Algorithm Clustering

Emir Žunić, Sead Delalić, Dženana Đonko, and Haris Šupić

Abstract The Vehicle Routing Problem (VRP) is an important optimization problem, the solution of which brings great savings to the company. Finding the optimal solution is significantly hampered by the introduction of realistic constraints such as time windows, capacity, customer-vehicle restrictions, and more. The paper presents a two-phase approach to solving the problem of vehicle routing with the fulfillment of several realistic conditions. The approach consists of customer clustering based on the firefly algorithm and process to solve rich VRP based on the created clusters. The algorithm was implemented in the real world and tested in some of the largest distribution companies in Bosnia and Herzegovina. The algorithm showed quality results in relation to the previously used methods, and in relation to the manual division of customers by the distribution manager.

Keywords Vehicle routing problem · Rich vehicle routing problem · Cluster-based vehicle routing problem · Firefly algorithm · Clustering · Firefly clustering · 2-opt

E. Žunić (✉) · S. Delalić
Info Studio d.o.o. Sarajevo, Sarajevo, Bosnia and Herzegovina
e-mail: emir.zunic@infostudio.ba

S. Delalić
e-mail: delalic.sead@pmf.unsa.ba

E. Žunić · D. Đonko · H. Šupić
Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
e-mail: ddonko@pmf.unsa.ba

H. Šupić
e-mail: hsupic@etf.unsa.ba

S. Delalić
Faculty of Science, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_25

1 Introduction

Transport management is an important factor in the business of many companies. By optimizing transportation, significant savings are achieved and the efficiency of the company is increased. This problem has been the focus of a large amount of research since the first mention in the 1950s.

Distribution companies have special importance in transport optimization. They receive hundreds of orders on a daily basis, and it is necessary to deliver orders in a timely manner to preserve the company's reputation. Distribution companies make deliveries from central warehouses called depots. Each distribution center has a predetermined fleet of vehicles, and a set of customers who need to be delivered during the day. It is necessary to determine the vehicle that will deliver the order for each customer, and to determine the exact order of delivery within each vehicle. In doing so, the goal is to minimize the total cost of transportation. In the literature, the problem is defined as the Vehicle Routing Problem (VRP), and is considered one of the most difficult optimization problems in the field of combinatorial optimization.

In practice, there are a number of limitations that significantly reduce optimization capabilities. Sometimes, finding a feasible solution is a difficult enough task due to the limited number of vehicles and limited time. Research in the field has recently focused on solving the problem of realistic vehicle routing, called Rich Vehicle Routing Problem (RVRP). The idea is to find effective methods of solving problems in the real world, where all the necessary restrictions (capacity, time windows, delivery restrictions, etc.) are respected at the same time.

Distribution companies sometimes divide customers into clusters, which can be used to significantly reduce the complexity of problem-solving. The paper describes the approach to solving RVRP problems based on two phases: automatic cluster creation and solving RVRP for each cluster. The implemented approach has been used in some of the largest distribution companies in Bosnia and Herzegovina, and contains a number of limitations and experiences collected from everyday use.

The paper consists of five sections. In this section, an introduction is given, and the motivation for implementing the algorithm is described. In the second section, a literature review is provided with a focus on the current state of the field. The third section describes the approach used to implement the solution. The firefly algorithm is described, and the method for solving the problem of customer clustering is described. A proposed algorithm for solving routing problems is also given. In the fourth section, the results of the proposed approach are described, and a comparison is made with the results of previously used approaches. The results were tested on data collected in real warehouses distribution companies in Bosnia and Herzegovina. The fifth section gives a brief conclusion, and offers guidelines for future research.

2 Literature Review

The problem of vehicle routing belongs to a class of $\mathcal{NP}$-hard problems for which no algorithm is known that will find the optimal solution in real time. Therefore, a number of heuristic and metaheuristic approaches have been implemented to solve routing problems.

The application of metaheuristic algorithms in the field of logistics is of great importance for the efficiency of Supply Chain Management (SCM). The ant colony optimization algorithm was applied to the process of order splitting in the warehouse [1], while the application of a discrete bat algorithm to solve the problem of order batching is described in [2]. The mentioned algorithms are improvements to the earlier described concept of Smart Warehouse Management System (SWMS) [3].

The problem of vehicle routing represents one of the most important problems in SCM due to the large savings that optimization brings. The field has been extremely active in recent years. Well-known approaches such as genetic algorithm [4], simulated annealing [5], or Tabu search [6] have yielded successful results in the field.

In the paper [7], different approaches for solving the RVRP are presented. In [8], the authors describe several applied metaheuristic approaches. The paper offers a comprehensive overview of problem variants and approaches to solving them in recent years. The particle swarm optimization algorithm has been implemented to solve many versions of VRP [9, 10], as well as bat algorithm [11, 12], fireworks algorithm [13], firefly algorithm (FA) [14–16], and others.

There are a number of cluster-based versions of the problem in the literature. In [17], the hybrid metaheuristic algorithms were proposed for the clustered vehicle routing problem. In the paper [18], a cluster-based optimization for the multi-depot heterogeneous fleet VRPTW is presented. In [19], an approximate two-level algorithm for solving a similar clustered CVRP is proposed.

The cluster-based approach for solving RVRP is described in [20]. The approach to solving is similar to the approach described in this paper. In the first phase, customers are divided into clusters. In the second phase, the clusters are individually optimized with the aim of reducing complexity and obtaining feasible routes. In [21], the applications of the firefly algorithm to solve clustering problems are described. The main significance of the described approach is in reducing the complexity and fulfilling additional user restrictions.

For distribution companies, customers are often repetitive, and therefore there are a number of opportunities to improve the input for future algorithm launches. Vehicles are tracked through a GPS system, and their movements are fully recorded and their accuracy is ensured [22]. Some of the improvements to vehicle routing algorithms based on GPS data analysis can be implemented [23, 24]. The application of machine learning techniques and GPS data to improve the unloading time prediction for VRP is described in [25].

3 Case Study

The paper describes the approach to solving the problem of vehicle routing in two phases: the creation of customer clusters and solving a number of routing problems. It is necessary to provide logistical data on available vehicles (ID, maximum load—mass and volume, cost per kilometer, working hours), customers (ID, latitude, longitude, ordered mass and volume, unloading time, time window, address), depot (ID, latitude, longitude), and customer-vehicle restrictions. In the next step, the distance between each two customers is calculated, as well as between customers and the depot. Due to the limited number of vehicles available, the primary goal is to deliver orders to all customers with available fleet and meet all conditions. The largest cost to the company is the case where certain customers remain undelivered.

3.1 Phase I: Clustering

The problem of automatic clustering was solved using a nature-inspired firefly algorithm (FA) [26]. The intensity of the light emitted by the fireflies affects the attraction. A lighter firefly always attracts a less bright firefly. The distance between the fireflies significantly affects the intensity of the light that the other firefly sees. The intensity of light is most often associated with the objective function.

The pseudocode for the adjusted FA to solve clustering problem is given in Algorithm 1. For successful implementation, it is necessary to determine the number of clusters, the objective function, the initial population, the distance of the two fireflies, and the movement operator.

Number of clusters. The total number of clusters was determined based on the input data. Let the number of clusters be denoted by k. The value of k is calculated by the formula

$$k = \lceil \frac{c}{v \cdot 2} \rceil + 1, \tag{1}$$

where by c the number of customers is denoted and by v the number of vehicles is denoted. The main idea is to use $2(\pm 1)$ vehicles per sector. Each cluster will be determined by a center. Therefore, it is necessary to initialize k coordinates. For each customer, the corresponding cluster is defined as the cluster with the nearest center.

Objective function. The objective function is calculated as a linear combination of the total distance of all customers from the centers of the corresponding clusters divided by the total distance from all customers to all centers, and the difference in the number of customers in the largest and smallest cluster. The goal is to minimize the value.

Initial population. For the initial population, each cluster center is chosen as one random customer. For each firefly, a set of k random customers was selected for the cluster centers. It is necessary to ensure that the same customer is not selected more than once.

Algorithm 1 Firefly Algorithm For Clustering

Input: customers - List of latitude, longitude pairs for each customer distances - Distance matrix between each pair of customers
Output: Clusters

Determine the number of clusters k.
Define the objective function $f(X)$ where $X = (x_1, x_2, \ldots x_k)$.
Generate the initial population of n fireflies x_i, where $i = \overline{1, n}$.
Formulate the light intensity I as a function of $f(X)$.
Define the light absorption coefficient γ.
Define the maximum generation number $MaxGeneration$ and the counter t.
while $t < MaxGeneration$ **do**
  for $i = 1$ to n **do**
    for $j = 1$ to n **do**
      if $I_i < I_j$ **then**
        Move firefly i towards j
      end if
      Vary attractiveness with distance r via exp $-\gamma r$
      Evaluate new solutions and update light intensity
    end for
  end for
  Rank fireflies and find the current global best g^*
end while
return g^*

Distance and the movement operator. When the firefly x_i is moving toward x_j, for each cluster center a in x_i, the closest cluster center b in x_j is calculated, and the center a is moved toward b. After that, the corresponding cluster was determined for each customer, and for each cluster the center was recalculated as the mean location (arithmetic mean). The total distance between the two fireflies is equal to the sum of the previously described distances of the centers.

Other parts of the algorithm and all necessary operators have not been modified compared to the original FA version as described [26].

3.2 Phase II: Cluster Routing

Once customers are divided into clusters, it is necessary to approach solving a few minor vehicle routing problems. Routing is done for each individual cluster, which reduces the routing problem of hundreds of customers to a few simpler routing problems of several dozen customers, which is a less complex problem, but at the same time provides quality and feasible solutions.

To begin with, the cluster of customers to which the depot belongs is determined. Let us mark this cluster with D. Customers from cluster D are used to supplement routes in other clusters, so that the vehicle capacity is used appropriately. This cluster is routed last.

For other clusters, priority is determined. The priority of the cluster can be determined in several ways, of which in practical use, the best results were given by the algorithm that first routes the regions with more customers and fewer customer-vehicle constraints. The complete pseudocode is given in Algorithm 2.

Algorithm 2 Firefly Algorithm For Clustering

Input: Clusters CL, customers, depot, vehicles, other constraints
Output: Routes

Initialize parameters n, α, β, γ.
Define a priority for each cluster CL_i, $i = \overline{1, k}$.
Sort clusters by priority (higher priority before).
Define the objective function f_GOAL.
for CL_i in CL **do**
    Get all customers C in sector CL_i and free vehicles
    Set number of vehicles to use ($i = 1$)
    while true **do**
        Create all combinations V of i free vehicles
        for v in V **do**
            Solve VRP using vehicles v for customers C
            Update best solution if found
        end for
        if best solution feasible **then**
            For each vehicle, add customers from D when possible
            break
        end if
        $i = i + 1$
    end while
end for
Try to swap vehicles to reduce cost and return best solution

Once the priorities of the cluster are determined, the complete process contains several important elements: defining the objective function, solving the VRP with a given set of vehicles, adds customers to created routes.

Objective function. The objective function minimizes the total cost. For each used vehicle, the total number of kilometers traveled is calculated, and the obtained number is multiplied by the cost per kilometer. Constraints and other realistic conditions are included as penalties on the objective function. If any of the restrictions are violated, the route is declared unfeasible. In this way, the customer's time windows and vehicle working hours are taken into account in solving the problem, as well as the vehicle load (longer distances are better to drive with a smaller load). Customer-vehicle constraints are also easily added to the objective function.

VRP solver. One of the most important elements is solving the VRP for a particular cluster. It is important to note a few facts. The VRP solver is called multiple times for different vehicle combinations. With each call, the problem is solved with a smaller number of customers (several dozen), which enables efficient and precise solving with simple and fast algorithms. Each vehicle is added as a customer, and its initial location is set to the depot location. The distances between each of the two vehicles are set to a large value. Routes are then created using the well-known 2-opt algorithm for solving the TSP. In addition to this, it is possible to use any VRP or TSP solver for smaller instances.

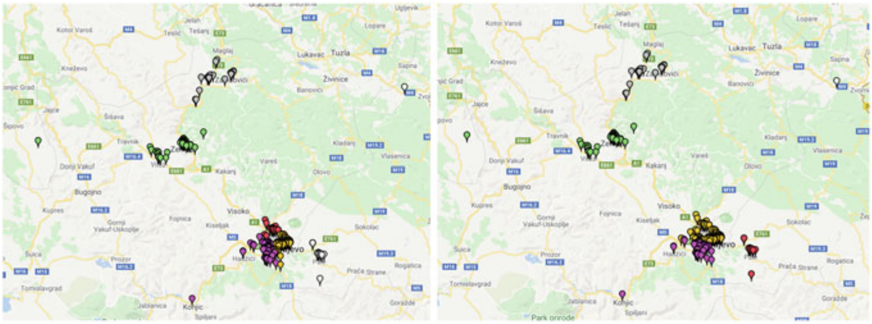
Each region can be satisfied by a smaller number of vehicles, as well as that the number of vehicles considered for a given region can be limited in advance (volume and mass of customers, vehicle capacity, etc.). Therefore, the number of observed permutations is small enough to be observed in real time, and not to significantly affect the efficiency of the algorithm.

Add customers to route. Sometimes, when creating routes within one region, the vehicle is not loaded enough. In that case, a smaller route is created, the vehicle is returned to the depot earlier and remains underused. For situations with fewer vehicles and more customers, such a route can be an insurmountable problem for finding a feasible solution. Therefore, cluster D has been singled out containing customers near the depot. These customers are used to complete the route. Of all the unvisited customers in route, it is necessary to find the customer closest to the starting and ending point of the route other than the depot. A customer who does not violate the restrictions by adding to the route is added and the process is repeated.

In addition, four algorithm improvements have been implemented. After completing all the steps, an additional vehicle check is done and the cost is checked after swapping each two vehicles. In case the total cost is reduced, the vehicles remain replaced. If the solution creates a region with a smaller number of customers or a certain region is on the way from the depot to another region, they are combined. We say that cluster A is on the way to cluster B from depot X, if the centers of cluster A, B, and the location of depot X are almost collinear points. In doing so, a deviation is allowed which is set as an algorithm parameter. Occasionally, regardless of the addition of routes and the possibility of combining regions, there are cases when the vehicle is returned to the depot earlier, and the rest of the day would be unused. In this case, all remaining routes in the region are canceled, a shorter route is saved, and the vehicle is made available for a new routing. Again, the routing process is repeated from the observed region, which opens the possibility of multiple routing of one vehicle and maximum utilization. In case none of the options is satisfied, a series of fictitious vehicles is created, and the rest of the customers are routed using those vehicles. The number of fictitious vehicles is equal to the number of clusters of remaining customers.

4 Results

As mentioned earlier, the described approach has been used in the real world, and is used by some of the largest distributor companies in Bosnia and Herzegovina. The approach has emerged through years of experience in solving routing problems. Through research, multiple approaches have been implemented. Tabu searches gave up to 10 times slower results compared to the described cluster method. At the same



(a) k-means approach (b) Firefly algorithm

Fig. 1 Real-world customers clustering algorithm

time, the solutions often violated the restrictions, which led to the subsequent change of routes and cases that were difficult to explain to the company's management.

The clustering approach was implemented subsequently. It was initially based on manual division by cities. For each cluster, the corresponding cities are determined. The approach has shown quality results in some cases; however, a major problem has arisen in complexity in the case of urban regions with a large number of customers. Subsequently, automatic approaches were implemented that overcame this problem, and showed generally quality results. Figure 1 shows approaches to creating customer clusters based on the k-means and firefly algorithm. The FA clustering showed the highest quality results in the case of urban regions.

Implemented approaches were tested for more than 30 days, and the obtained results are shown in Table 1. The firefly algorithm was run 10 times before the second step of the algorithm, and the best result was saved. Small deviations in the clustering process were observed, and the FA showed enviable stability ($n = 20$, $\beta_0 = 1$, $\gamma = 0.1$, $\alpha_0 = 1$, $maxGeneration = 100$, and $\theta = 0.97$).

All clustering approaches gave quality results; however, in some routes, the FA gave better results due to the balanced number of customers by sectors. FAVRP showed significantly better results compared to other algorithms in urban regions (e.g., a city with hundreds of customers), where the largest difference was achieved.

5 Conclusion

The paper describes the process of solving the problem of vehicle routing with a number of realistic constraints. The process consists of two phases. In the first phase, customers are divided into clusters, and the second phase consists of the process of routing the vehicle. Clusters are created using the firefly algorithm, while a number of realistic constraints are observed in the routing process in combination with the

Table 1 Comparison of firefly algorithm VRP (FAVRP), Tabu search (TS), k-means VRP (KM), and manual sectors VRP (MVRP)

Observed property	Results
Number of days	30
Restriction violation	TS for 10 days, others for none
Total time (FAVRP vs. TS)	FAVRP 5.6 times faster
Total time (FAVRP vs. KM)	FAVRP 20% faster
Total time (FAVRP vs. MVRP)	FAVRP 25% faster
Total cost (FAVRP vs. TS)	FARP with 7% lower cost for feasible routes
Total cost (FAVRP vs. KM)	FARP gave a 14% lower cost
Total cost (FAVRP vs. MVRP)	FARP gave a 10% lower cost

2-opt algorithm. The results were tested on real data of distribution companies in Bosnia and Herzegovina. The significance of the proposed approach is based on the fact that it respects a number of realistic constraints (time windows, customer-vehicle constraints, capacity, driver working hours), and allows routing of a large number of customers within a reasonable time.

In the future, the implementation of metaheuristic algorithms in the second phase of the process is planned, as well as the application of other metaheuristic approaches to solve the problem of clustering.

References

1. Zunic E, Delalic S, Tucakovic Z, Hodzic K, Besirevic A (2019) Innovative modular approach based on vehicle routing problem and ant colony optimization for order splitting in real warehouses. In: Communication papers of the 14th federated conference on computer science and information systems (FedCSIS). https://doi.org/10.15439/2019f196
2. Delalic S, Zunic E, Alihodzic A, Selmanovic E (2020) The order batching concept implemented in real smart warehouse. In: 2020 43rd international convention on information and communication technology, electronics and microelectronics (MIPRO). https://doi.org/10.23919/mipro48935.2020.9245256
3. Žunić E, Delalić S, Hodžić K, Beširević A, Hindija H (2018) Smart warehouse management system concept with implementation. In: 14th symposium on neural networks and applications (NEUREL). https://doi.org/10.1109/NEUREL.2018.8587004
4. Baker BM, Ayechew MA (2003) A genetic algorithm for the vehicle routing problem. Comput Oper Res. https://doi.org/10.1016/S0305-0548(02)00051-5
5. Chiang WC, Russell RA (1996) Simulated annealing metaheuristics for the vehicle routing problem with time windows. Ann Oper Res. https://doi.org/10.1007/BF02601637
6. Gendreau M, Hertz A, Laporte G (1994) A tabu search heuristic for the vehicle routing problem. Manag Sci. https://doi.org/10.1287/mnsc.40.10.1276
7. Caceres-Cruz J, Arias P, Guimarans D, Riera D, Juan AA (2014) Rich vehicle routing problem: survey. ACM Comput Surv (CSUR). https://doi.org/10.1145/2666003

8. Osaba E, Yang XS, Del Ser J (2020) Is the vehicle routing problem dead? An overview through bioinspired perspective and a prospect of opportunities. In: Nature-inspired computation in navigation and routing problems. https://doi.org/10.1007/978-981-15-1842-3_3
9. Ai TJ, Kachitvichyanukul V (2009) A particle swarm optimization for the vehicle routing problem with simultaneous pickup and delivery. Comput Oper Res. https://doi.org/10.1016/j.cor.2008.04.003
10. Belmecheri F, Prins C, Yalaoui F, Amodeo L (2013) Particle swarm optimization algorithm for a vehicle routing problem with heterogeneous fleet, mixed backhauls, and time windows. J Intell Manuf. https://doi.org/10.1007/s10845-012-0627-8
11. Taha A, Hachimi M, Moudden A (2015) Adapted bat algorithm for capacitated vehicle routing problem. Int Rev Comput Softw (IRECOS). https://doi.org/10.15866/irecos.v10i6.6512
12. Osaba E, Carballedo R, Yang XS, Fister I Jr, Lopez-Garcia P, Del Ser J (2018) On efficiently solving the vehicle routing problem with time windows using the bat algorithm with random reinsertion operators. In: Nature-inspired algorithms and applied optimization. https://doi.org/10.1007/978-3-319-67669-2_4
13. Yang W, Ke L (2019) An improved fireworks algorithm for the capacitated vehicle routing problem. Front Comput Sci. https://doi.org/10.1007/s11704-017-6418-9
14. Osaba E, Yang XS, Diaz F, Onieva E, Masegosa AD, Perallos A (2017) A discrete firefly algorithm to solve a rich vehicle routing problem modelling a newspaper distribution system with recycling policy. Soft Comput. https://doi.org/10.1007/s00500-016-2114-1
15. Altabeeb AM, Mohsen AM, Ghallab A (2019) An improved hybrid firefly algorithm for capacitated vehicle routing problem. Appl Soft Comput. https://doi.org/10.1016/j.asoc.2019.105728
16. Osaba E, Carballedo R, Yang XS, Diaz F (2016) An evolutionary discrete firefly algorithm with novel operators for solving the vehicle routing problem with time windows. In: Nature-inspired computation in engineering. https://doi.org/10.1007/978-3-319-30235-5_2
17. Vidal T, Battarra M, Subramanian A, Erdogan G (2015) Hybrid metaheuristics for the clustered vehicle routing problem. Comput Oper Res. https://doi.org/10.1016/j.cor.2014.10.019
18. Dondo R, Cerdá J (2007) A cluster-based optimization approach for the multi-depot heterogeneous fleet vehicle routing problem with time windows. Eur J Oper Res. https://doi.org/10.1016/j.ejor.2004.07.077
19. Expósito-Izquierdo C, Rossi A, Sevaux M (2016) A two-level solution approach to solve the clustered capacitated vehicle routing problem. Comput Ind Eng. https://doi.org/10.1016/j.cie.2015.11.022
20. Žunić E, Đonko D, Šupić H, Delalić S (2020) Cluster-based approach for successful solving real-world vehicle routing problems. In: 15th conference on computer science and information systems (FedCSIS). https://doi.org/10.15439/2020F184
21. Senthilnath J, Omkar SN, Mani V (2011) Clustering using firefly algorithm: performance study. Swarm Evol Comput. https://doi.org/10.1016/j.swevo.2011.06.003
22. Žunić E, Delalić S, Hodžić K, Tucaković Z (2019) Innovative GPS data anomaly detection algorithm inspired by QRS complex detection algorithms in ECG signals. In: EUROCON 2019—18th international conference on smart technologies. https://doi.org/10.1109/EUROCON.2019.8861619
23. Žunić E, Hindija H, Beširević A, Hodžić K, Delalić S (2018) Improving performance of vehicle routing algorithms using GPS data. In: 14th symposium on neural networks and applications (NEUREL). https://doi.org/10.1109/NEUREL.2018.8586982
24. Žunić E, Delalić S, Đonko, Dž (2020) Adaptive multi-phase approach for solving the realistic vehicle routing problems in logistics with innovative comparison method for evaluation based on real GPS data. Transp Lett. https://doi.org/10.1080/19427867.2020.1824311
25. Žunić E, Kuric A, Delalić S (2020) Improving unloading time prediction for vehicle routing problem based on GPS data. In: Position papers of the 15th federated conference on computer science and information systems (FedCSIS). https://doi.org/10.15439/2020F123
26. Yang X (2010) Nature-inspired metaheuristic algorithms, 2nd edn. ISBN: 1905986289, 9781905986286

Controlled Social Network Adaptation: Subjective Elements in an Objective Social World

Jan Treur

Abstract In this paper, the role of subjective elements and control in social network adaptation is analyzed computationally. In particular, it is analyzed: (1) how the coevolution of social contagion and bonding by homophily may be controlled by the persons involved, and (2) how subjective representation states (e.g., what they know) can play a role in this coevolution and its control. To address this, a second-order adaptive social network model is presented in which persons do have a form of control over the coevolution process, and in relation to this, their bonding depends on their subjective representation states about themselves and about each other, and social contagion depends on their subjective representation states about their connections.

Keywords Controlled social network adaptation · Bonding by homophily

1 Introduction

Social networks often do not only show dynamics *within* the network but also dynamics *of* the network, where the latter is also called network adaptation. These combined dynamics are sometimes referred to as the coevolution of the network states and the network connections. An often studied case for social networks is the coevolution of social contagion (for the dynamics of the network nodes or states) and bonding by homophily (for the dynamics of the weights of the network connections). The bonding by homophily adaptation principle expresses how 'being alike' strengthens the connection between two persons, also explained as 'birds of a feather flock together' (e.g., [12, 13]). On the other hand, social contagion makes that network states affect each other through their connections, which implies that the stronger the two persons are connected, the more they will become alike [10]. This makes circular, reciprocal causal relations between the two processes. It has been found in simulations that, as in the real world, the emerging behavior of adaptive network models

J. Treur (✉)
Social AI Group, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
e-mail: j.treur@vu.nl

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_26

based on the coevolution of these two processes often shows a form of clustering, segregation or community formation (e.g., [3, 4, 8, 14, 16, 17, 20, 21]).

Usually, as mentioned in the literature, these social processes are considered without taking into account subjective elements for the persons involved. For example, do the persons themselves actually know how far they are alike? Do they *have* to know that to let the bonding work properly? Do they know their connections? Are persons able to have some control over their bonding? Or are they just willless victims of objective social laws independent of what they know or what they want? Such subjective aspects are lacking in (computational) research on bonding by homophily as mentioned, as usually these coevolution processes are addressed exclusively from the perspective of an objective social world. Note that in social science literature works such as [7, 9, 22] from a wider perspective also the role of cognitive and cultural interpretation in social dynamics is emphasized.

In the current paper, it is assumed that such subjective elements indeed do matter and it is analyzed computationally how some of them can play their role in the coevolution process. More specifically, it is analyzed: (1) how the coevolution of social contagion and bonding by homophily may be controlled by the persons involved, and (2) how subjective states representing what they know about themselves, others and their connections play a role in this coevolution and its control. To this end, a second-order adaptive social network model has been developed in which persons have control over the coevolution process, and their bonding and social contagion depend on subjective representations of the involved persons about themselves and each other, and about their connections.

In the paper, in Sect. 2 the higher-order adaptive network-oriented modeling approach from [19] used is briefly explained. In Sect. 3 the designed second-order adaptive social network model is presented. Section 4 addresses simulation results for a case study on adaptation in tetradic relationships. Finally, Sect. 5 is a discussion section, where, among others, it is discussed how the model can predict that faking your properties can be an effective way to achieve the desired bonding.

2 Higher-Order Adaptive Network Models

In this section, the network-oriented modeling approach used is briefly introduced. Following [15, 19], a temporal-causal network model is characterized by:

- **Connectivity characteristics** Connections from a state X to a state Y and their weights $\boldsymbol{\omega}_{X,Y}$
- **Aggregation characteristics** For any node Y, some combination function $\mathbf{c}_Y(..)$ defines aggregation that is applied to the impacts $\boldsymbol{\omega}_{X_i,Y}X_i(t)$ on Y from its incoming connections from states $X_1, \ldots, X_k$
- **Timing characteristics** Each state Y has a speed factor $\boldsymbol{\eta}_Y$ defining how fast it changes for the given causal impact

The following difference (or differential) equations that are useful for simulation purposes and also for analysis of temporal-causal networks incorporate these network characteristics $\boldsymbol{\omega}_{X,Y}$, $\mathbf{c}_Y(..)$, $\boldsymbol{\eta}_Y$:

$$Y(t + \Delta t) = Y(t) + \eta_Y\left[\mathbf{c}_Y\left(\boldsymbol{\omega}_{X_1,Y} X_1(t),\ \ldots,\ \boldsymbol{\omega}_{X_k,Y} X_k(t)\right) - Y(t)\right]\Delta t \quad (1)$$

for any state Y and where $X_1, \ldots, X_k$ are the states from which it gets its incoming connections. Within the software environment described in [19] (Chap. 9), a large number > 45 of useful combination functions are included in a combination function library. The three combination functions from this library used for state Y in the introduced network model are:

- The Euclidean combination function $\mathbf{eucl}_{\boldsymbol{n},\boldsymbol{\lambda}}(V_1, \ldots, V_\mathrm{k})$ defined by

$$\mathbf{eucl}_{n,\boldsymbol{\lambda}}(V_1,\ \ldots,\ V_k) = \sqrt[n]{\frac{{V_1}^n + \cdots + {V_k}^n}{\lambda}} \quad (2)$$

where n is the order, $\boldsymbol{\lambda}$ is a scaling factor and $V_1, \ldots, V_k$ are the impacts from which the considered state Y gets incoming connections.

- The advanced logistic sum combination function $\mathbf{alogistic}_{\sigma,\tau_{log}}(V_1, \ldots, V_k)$ defined by:

$$\mathbf{alogistic}_{\boldsymbol{\sigma},\boldsymbol{\tau}_{\mathbf{log}}}(V_1, \ldots, V_k) = \left[\frac{1}{1 + \mathrm{e}^{-\sigma(V_1 + \ldots + V_k - \boldsymbol{\tau}_{\mathbf{log}})}} - \frac{1}{1 + \mathrm{e}^{\boldsymbol{\sigma}\boldsymbol{\tau}_{\mathbf{log}}}}\right] \left(1 + \mathrm{e}^{-\boldsymbol{\sigma}\boldsymbol{\tau}_{\mathbf{log}}}\right) \quad (3)$$

where $\boldsymbol{\sigma}$ is a steepness parameter, $\boldsymbol{\tau}_{\mathbf{log}}$ is a threshold parameter and $V_1, \ldots, V_k$ are the impacts from the states from which the considered state Y gets incoming connections

- The simple linear homophily combination function $\mathbf{slhomo}_{\boldsymbol{\alpha},\boldsymbol{\tau}_{\mathbf{homo}}}(V_1,\ V_2,\ W)(V_1, V_2, W)$ defined by

$$\mathbf{slhomo}_{\boldsymbol{\alpha},\boldsymbol{\tau}_{\mathbf{homo}}}(V_1,\ V_2,\ W) = W + \alpha W(1 - W)(\boldsymbol{\tau}_{\mathbf{homo}} - |V_1 - V_2|) \quad (4)$$

where $\boldsymbol{\alpha}$ is an amplification parameter, $\boldsymbol{\tau}_{\mathbf{homo}}$ is the tipping point parameter and V_1, V_2 are a person's representations of the two persons' states involved and W represents the weight of their connection.

In Sect. 3, the combination function $\mathbf{eucl}_{\boldsymbol{n},\boldsymbol{\lambda}}(\ldots)$ will be used to model social contagion and formation of internal state representations, $\mathbf{slhomo}_{\boldsymbol{\alpha},\boldsymbol{\tau}_{\mathbf{homo}}}(V_1,\ V_2,\ W)(V_1, V_2, W)$ to model bonding based on homophily by internal connection weight representations, and $\mathbf{alogistic}_{\sigma,\tau_{\mathbf{log}}}(\ldots)$ to model control of the bonding. Note that the homophily tipping point $\boldsymbol{\tau}_{\mathbf{homo}}$ is the point where the difference between the states of the two individuals (represented by $|V_1 - V_2|$) turns an increase of bonding (outcome $> W$) into a decrease (outcome $< W$), and conversely.

The above concepts enable to design network models and their dynamics in a declarative manner, based on mathematically defined functions and relations. Realistic network models are usually adaptive: often some of their network characteristics change over time. By using *self-modeling networks* (or *network reification*), a similar network-oriented conceptualization can also be applied to *adaptive* networks to obtain a declarative description using mathematically defined functions and relations for them as well; see [18, 19]. This works through the addition of new states to the network (called *reification states* or *self-model states*) which represent network characteristics by network states. If such self-model states are dynamic, they describe adaptive network characteristics. In a graphical 3D format, such self-model states are depicted at the next level (*reification level*), where the original network is at a *base level*. As an example, the weight $\boldsymbol{\omega}_{X,Y}$ of a connection from state X to state Y can be represented (at the next reification level) by a self-model state named $\mathbf{W}_{X,Y}$ (objective representation) or $\mathbf{RW}_{X,Y}$ (subjective representation). Similarly, all other network characteristics from $\boldsymbol{\omega}_{X,Y}$, $\mathbf{c}_Y(\ldots)$, $\boldsymbol{\eta}_Y$ can be made adaptive by including self-model states for them.

As a self-modeling network model is also a temporal-causal network model itself, as has been proven in [19], Chap. 10, this self-modeling construction can be easily applied iteratively to obtain multiple self-model levels. This can provide higher-order adaptive network models, and has turned out quite useful to model, for example, plasticity and metaplasticity in the form of a second-order adaptive mental network with three levels, one base level and a first- and a second-order self-model level (e.g., [1, 11, 19], Chap. 4). In the current paper, multi-level network self-modeling will be applied for higher-order adaptive social network models in particular.

3 A Network Model for Controlled Social Network Adaptation

This section presents the introduced network model for controlled social network adaptation by using subjective representations. This network model integrates three types of interacting processes:

- The social network's within-network dynamics based on social contagion
- First-order social network adaptation based on bonding by homophily
- Second-order social network adaptation to control the network adaptation

The above three types of processes have been modeled by a second-order adaptive network architecture based on multi-level self-modeling as described in Sect. 2, with connectivity as depicted in Fig. 1. In this 3D picture, of the three-plane models, one of the three types of processes is mentioned above.

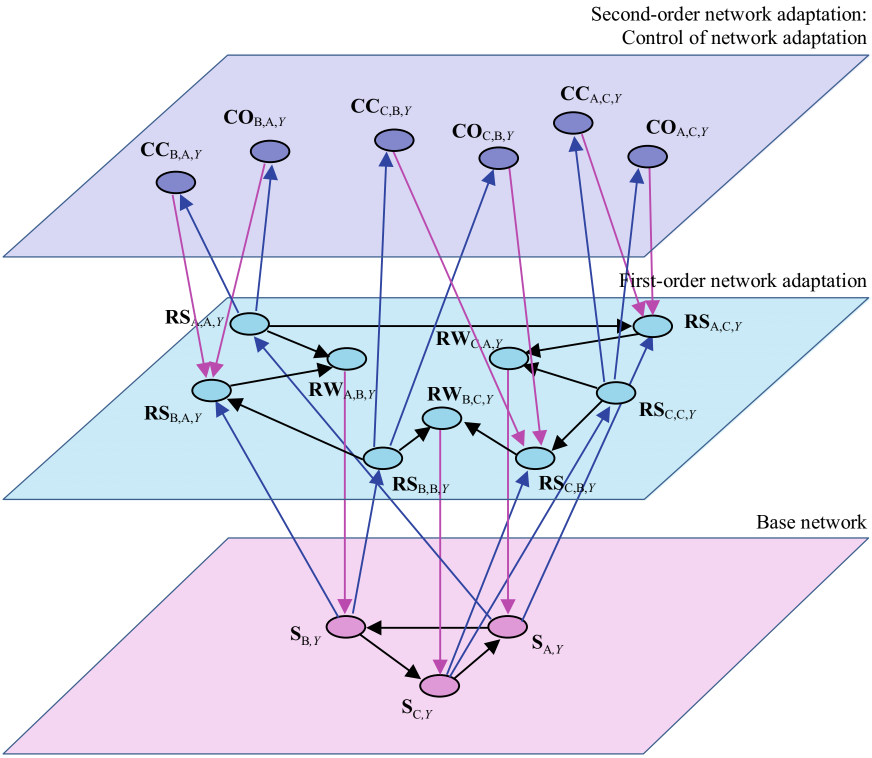


Fig. 1 Overview of the connectivity of the second-order adaptive social network model for three example persons A, B and C

The types of states and connections used at and between the three levels within this network model are shown in Tables 1 and 2. Here A and B are variables over persons and Y is a type of state of a person, for example, how much the person likes to watch Netflix series. At the base level, social contagion is modeled by connections $\mathbf{S}_{A,Y} \rightarrow \mathbf{S}_{B,Y}$. Each person has subjective internal representation states of other persons' states Y (and the state of her or himself) and of his or her connections to

Table 1 Types of states in the introduced controlled adaptive social network model

$\mathbf{S}_{B,Y}$	Objective state Y of person B
$\mathbf{RS}_{A,B,Y}$	Subjective representation of person B for state Y of person A
$\mathbf{RW}_{A,B,Y}$	Subjective representation of person A for the connection weight from person A to person B
$\mathbf{CC}_{A,B,Y}$	Control state for communication from A to B: representation of the weight of the connection from $\mathbf{RS}_{A,A,Y}$ to $\mathbf{RS}_{A,B,Y}$
$\mathbf{CO}_{A,B,Y}$	Control state for observation by B: representation of the weight of the connection from $\mathbf{S}_{A,Y}$ to $\mathbf{RS}_{A,B,Y}$ for the state Y of A observed by B

Table 2 Types of connections in the controlled adaptive social network model

Intralevel connections		
$\mathbf{S}_{A,Y} \to \mathbf{S}_{B,Y}$	Social contagion from A to B for state Y	
$\mathbf{RS}_{A,A,Y} \to \mathbf{RS}_{A,B,Y}$	Communication of $\mathbf{S}_{A,Y}$ from A to B	
$\mathbf{RS}_{A,A,Y} \to \mathbf{RW}_{A,B,Y}$	Effect of state Y of A on bonding by homophily from A to B	
$\mathbf{RS}_{B,A,Y} \to \mathbf{RW}_{A,B,Y}$	Effect of state Y of B on bonding by homophily from A to B	
Interlevel connections		
$\mathbf{S}_{A,Y} \to \mathbf{RS}_{A,B,Y}$	Observation by B of A's state Y	Upward from base level to first reification level
$\mathbf{RW}_{A,B,Y} \to \mathbf{S}_{B,Y}$	Effectuation of base connection weights for social contagion from A to B	Downward from first reification level to base level
$\mathbf{RS}_{B,B,Y} \to \mathbf{CO}_{A,B,Y}$	Observation control monitoring connections for B	Upward from first to second reification level
$\mathbf{RS}_{B,B,Y} \to \mathbf{CC}_{A,B,Y}$	Communication control monitoring connections for B	
$\mathbf{CO}_{A,B,Y} \to \mathbf{RS}_{A,B,Y}$	Effectuation of control of observation of A by B	Downward from second to first reification level
$\mathbf{CC}_{A,B,Y} \to \mathbf{RS}_{A,B,Y}$	Effectuation of control of communication from A by B	

others. This is modeled by the first-order self-model. A person B's internal representation state for person A having state Y is modeled by state representation $\mathbf{RS}_{A,B,Y}$. A person A's subjective representation of his or her connection to B is modeled by connection weight representation $\mathbf{RW}_{A,B,Y}$. There are two pathways that contribute to the formation of state representations $\mathbf{RS}_{A,B,Y}$. First, these representations can be communicated between persons. For example, if A communicates his or her subjective representation $\mathbf{RS}_{A,A,Y}$ of the own state $\mathbf{S}_{A,Y}$ to B, this is modeled by a connection $\mathbf{RS}_{A,A,Y} \to \mathbf{RS}_{A,B,Y}$. A second pathway for a person B to get information on person A's state is through observation of $\mathbf{S}_{A,Y}$ by B. This is modeled by a connection $\mathbf{S}_{A,Y} \to \mathbf{RS}_{A,B,Y}$.

As indicated, person A's representation of her or his connection to person B is modeled by $\mathbf{RW}_{A,B,Y}$. It is assumed that the adaptive change of the represented connections depends on the internal representation states $\mathbf{RS}_{A,B,Y}$. As the changes considered here are based on a homophily principle for state Y, this adaptation is supported by connections $\mathbf{RS}_{A,A,Y} \to \mathbf{RW}_{A,B,Y}$ and $\mathbf{RS}_{B,A,Y} \to \mathbf{RW}_{A,B,Y}$. The connection representations $\mathbf{RW}_{A,B,Y}$ in turn affect the social contagion within the social network, which is modeled by downward connections $\mathbf{RW}_{A,B,Y} \to \mathbf{S}_{B,Y}$.

To control the social network adaptation processes, two types of control actions are considered in particular:

- Controlling the communication of state Y from person A to person B, modeled by control states $\mathbf{CC}_{A,B,Y}$

- Controlling the observation of state Y from person A by person B, modeled by control states $\mathbf{CO}_{A,B,Y}$

Activation of a communication control state $\mathbf{CC}_{A,B,Y}$ makes that the connection $\mathbf{RS}_{A,A,Y} \rightarrow \mathbf{RS}_{A,B,Y}$ from A's state $\mathbf{RS}_{A,A,Y}$ to B's state $\mathbf{RS}_{A,B,Y}$ gets a high value (1 or close to 1) so that the transfer of information by communication happens; this is modeled by connections $\mathbf{CO}_{A,B,Y} \rightarrow \mathbf{RS}_{A,B,Y}$. This can be considered as B asking A for the information about him or herself, upon which A communicates this information. Similarly, activation of an observation control state $\mathbf{CO}_{A,B,Y}$ makes that the connection $\mathbf{S}_{A,Y} \rightarrow \mathbf{RS}_{A,B,Y}$ from A's state $\mathbf{S}_{A,Y}$ to B's state $\mathbf{RS}_{A,B,Y}$ gets a high value (1 or close to 1) so that the transfer of information by observation takes place; this is modeled by connections $\mathbf{CO}_{A,B,Y} \rightarrow \mathbf{RS}_{A,B,Y}$. As an example used in the case study in Sect. 4, the control states $\mathbf{CC}_{A,B,Y}$ and $\mathbf{CO}_{A,B,Y}$ themselves may become active depending on B's state $\mathbf{RS}_{B,B,Y}$; this is modeled by connections $\mathbf{RS}_{B,B,Y} \rightarrow \mathbf{CO}_{A,B,Y}$ and $\mathbf{RS}_{B,B,Y} \rightarrow \mathbf{CO}_{A,B,Y}$. But this can be addressed in many other ways as well, including externally determined control, for example, by enabling or allowing observation or communication (only) at specific time slots.

4 Simulation for a Tetradic Relationship Example

In this section, a simulation of an example scenario will be discussed to illustrate the introduced second-order adaptive social network model. The example scenario describes an adaptive tetradic relationship configuration with initially two couples all four of which are friends: Mark and Dion, and Ann and Jenny. After the process described in the scenario they find themselves in a slightly changed configuration, where Mark and Jenny, and Dion and Ann have the stronger connections; see Fig. 2. This adaptation process takes place because Mark and Jenny realize that they have more in common with as an example used here their preference to watch Netflix series. Similarly, Dion and Ann realize that they also have more in common, in their case disliking watching Netflix series (instead they have a preference for outdoor activities).

To specify a network model according to the approach described in [19], as discussed in Sect. 2, three types of network characteristics are to be addressed: *connectivity*, *aggregation* and *timing* characteristics. They have been specified in role matrix format as shown in the Appendix [23] and used for the simulation discussed

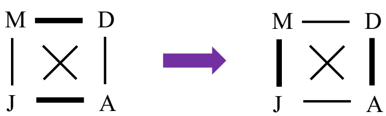


Fig. 2 Example scenario for a tetradic relationship configuration where initially M and D and J and A have strong connections and in the end M and J and D and A have the strong connections

after. For the sake of simplicity, the subscript Y (which for the example stands for a preference to watch Netflix series) has been left out here. Role matrices indicate in rows successively for all network states the factors that affect them from different roles. In role matrix **mb** (see Appendix [23]), for each state it is indicated from which other states it has incoming connections from the same or a lower level. In the same box in role matrix **mcw**, it is indicated what are the connection weights for the connected states indicated in **mb**. If the connection weights are static, their static value is indicated in matrix **mb**, but if the connection weight is adaptive, the self-model state representing this weight is indicated, as in that case at each time point this is where the (dynamic) connection weight value can be found. This can be seen for all incoming connections for the first four states X_1 to X_4, and for all incoming connections for the state representation states X_9 to X_{20}. Indicating these adaptive value representations defines the downward connections of Fig. 1. Also, the speed factors are shown in Appendix [23] (role matrix **ms**, which actually is a vector).

In the second box in Appendix [23], showing the aggregation characteristics, it can be seen which states use which combination functions (role matrix **mcfw**) and which parameter values for them (role matrix **mcfp**); also the initial values for the example simulation are shown here.

In Figs. 3 and 4 the simulation for the example scenario is shown. In Fig. 3 the coevolution of changing states and connection (self-model) representations is shown without showing the underlying personal state representations. Here the states $\mathbf{S}_A$ are slowly changing, whereas the connection representations $\mathbf{RW}_{A,B}$ are changing faster. It can be indeed seen that for Mark and Jenny both directional connection representations $\mathbf{RW}_{\mathrm{M,J}}$ and $\mathbf{RW}_{\mathrm{J,M}}$ start to increase from timepoint 5 resp. 10 on to finally end up at a value (close to) 1. Similarly, the connection representations $\mathbf{RW}_{\mathrm{D,A}}$ and $\mathbf{RW}_{\mathrm{A,D}}$ between Dion and Ann start to increase after time 5. In the same time period, the connection representations $\mathbf{RW}_{\mathrm{M,D}}$ and $\mathbf{RW}_{\mathrm{D,M}}$ between Mark and Dion and $\mathbf{RW}_{\mathrm{J,A}}$ and $\mathbf{RW}_{\mathrm{A,J}}$ between Jenny and Ann decrease to (close to) 0. All

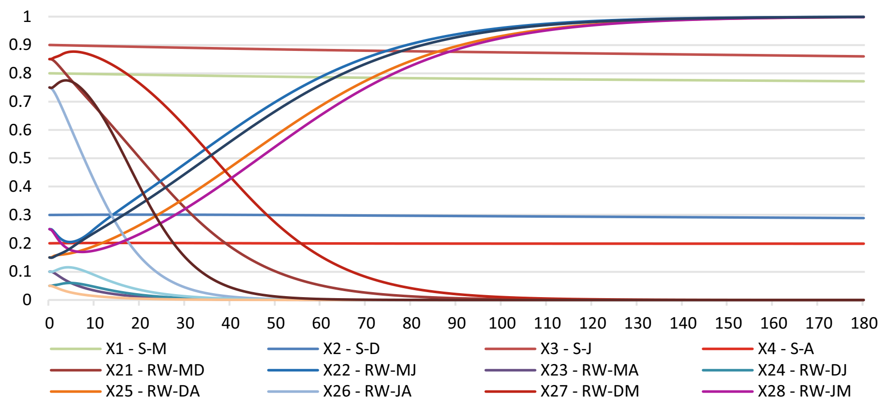


Fig. 3 Outcomes for the example scenario simulation: the changes in all relationships

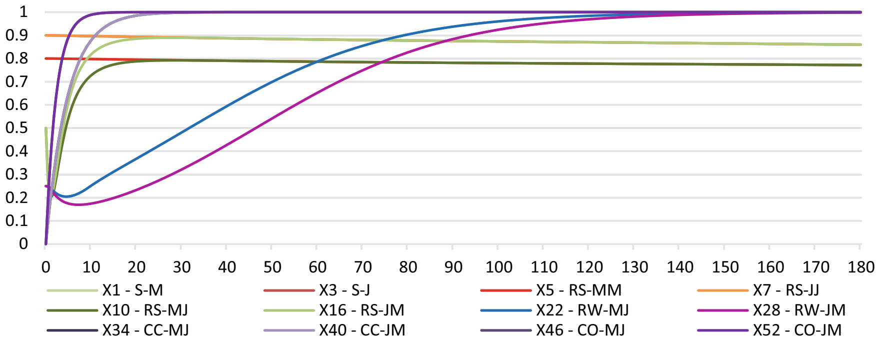


Fig. 4 The role of control and subjective states for the relationship between Mark and Jenny

these changes are a consequence of the homophily principle, as the state values $\mathbf{S}_{\mathrm{M}}$ and $\mathbf{S}_{\mathrm{J}}$ for Mark and Jenny are close to each other (0.8 and 0.9), and $\mathbf{S}_{\mathrm{D}}$ and $\mathbf{S}_{\mathrm{A}}$ for Dion and Ann also (0.2 and 0.3); note that the tipping point for similarity set was 0.25, so (only) a difference < 0.25 is strengthening a relationship. In contrast, the values for Mark and Dion differ a lot (0.8 versus 0.3), which is much higher than the tipping point of 0.25, and therefore has a decreasing effect on their relationship; the same pattern holds for Jenny and Ann.

Finally, in the right lower corner it can be seen that the other connection representations (e.g., for Jenny and Dion) were already low and still became lower because of big differences in their states. It can be noted that all connection representations converge to 0 or 1, which shows that clustering (and segregation) takes place, where the emerging clusters are Mark–Jenny and Dion–Ann, whereas the initial configuration (approximately) had clusters Mark–Dion and Jenny–Ann (also see Fig. 2).

In Fig. 4, the focus is on the development of the connections between Mark and Jenny; in particular, it zooms in on the role that is played by the control states $\mathbf{CC}_{A,B,Y}$ and $\mathbf{CO}_{A,B,Y}$ and the subjective representation states $\mathbf{RS}_{A,B,Y}$. The dark purple line that gets close to 1 before time 10 indicates the control states $\mathbf{CC}_{A,B,Y}$ for the communication between them, which makes that at that time their mutual communication channels $\mathbf{RS}_{A,A,Y} \rightarrow \mathbf{RS}_{A,B,Y}$ get weights close to 1. This implies that before time 10 they indeed both communicate to each other that they like watching Netflix series. These control states for the communication are triggered in this example scenario because each of them observes his or her own behavior and therefore they form representations $\mathbf{RS}_{A,A}$ of their own states $\mathbf{S}_{A}$ concerning watching the series. Next, around time point 20 the control states $\mathbf{CO}_{A,B,Y}$ for observation (the grey line) get close to 1, triggered in a similar way (but just a bit slower) as the control states for communication. This gives the relevant observation channel $\mathbf{S}_{A,Y} \rightarrow \mathbf{RS}_{A,B,Y}$ a weight close to 1. Due to that, mutual observation takes place.

Because of these communication and observation actions, the mutual subjective representations $\mathbf{RS}_{\mathrm{M,J},Y}$ of Jenny about Mark (the dark green line) and $\mathbf{RS}_{\mathrm{J,M},Y}$ of Mark about Jenny (the light green line) are formed and around time 20 reach levels around 0.8 (Jenny representing Mark) and 0.9 (Mark representing Jenny), respectively; these representations are close to the actual values, as are the representations $\mathbf{RS}_{A,A}$ of their own states, so all of them achieve faithful representations. Only now these subjective representations have been formed in a controlled manner, the homophily principle can start to work: the bonding works through the (subjective) representation states $\mathbf{RS}_{A,B,Y}$, not through the (objective) states $\mathbf{S}_{A,Y}$ themselves. More specifically, from the moment on that the subjective representations of Jenny about Mark and Jenny's own subjective representation about herself get closer than 0.25 (which is somewhere before time point 10 but not earlier), her self-model representation $\mathbf{RW}_{\mathrm{J,M},Y}$ of her connection to Mark (the pink line) starts to increase from 0.2 or lower to finally becoming very close to 1. Similarly, the effect of the subjective representations of Mark for Jenny and Mark's own subjective self-model representation about himself, on the subsequent increase of his representation $\mathbf{RW}_{\mathrm{M,J},Y}$ of his connection to Jenny (the blue line) can be noted. Before that point in time their connections were not increasing, but instead go slightly downward; this illustrates the effect of the control via the subjective self-model representation states on the adaptation.

5 Discussion

In this paper, computational analysis was made of the role of subjective elements and control in social network adaptation. It was analyzed: (1) how the coevolution of social contagion and bonding by homophily may be controlled by the persons involved, and (2) how subjective representation states (e.g., what they know about themselves and each other and about their connections) can play a role in this coevolution and its control. To address this, a second-order adaptive social network model was presented in which persons do have a form of control over the coevolution process, and, in relation to this, their bonding depends on their subjective representation states about themselves and about each other, and social contagion depends on their subjective representation states about their connections.

Concerning evaluation, the model behavior is as expected from the mentioned literature. Moreover, based on mathematical analysis, from formula (3) for the homophily function it can be predicted that when the model reaches equilibrium, it holds:

$$W = 0 \text{ or } W = 1 \text{ or } |V_1 - V_2| = \tau_{\mathbf{homo}}$$

This is indeed the case, as can also be seen in the case study simulation in Fig. 3 where all connection weight representations end up in 0 or 1.

Also, note that a basic design choice for the model is that the subjective representations of the connections determine the actual social contagion in the objective social world. This is based on the assumption that persons socially behave according to what they know or believe about their connections. Also, here misrepresentation can be modeled easily by introducing some deviations in the subjective bonding by homophily mechanism within the model. Then the social behavior leading to social contagion will (falsely) take place based on these misrepresentations of connections. On the other hand, it may as well be assumed that the subjective representations of the connections do not play an exclusive role in the social behavior but also a more objective form of connections may have influence. To cover this, the model can easily be extended by also adding (in parallel) a more standard objective mechanism for bonding by homophily based on the objective states and then combine (according to some chosen ratio) both the objective and subjective connection representations to jointly make social contagion work. Also, this may be worked out in more detail for a possible extended version for a journal.

The proposed computational network model where mental states are modeled as a basis for social mechanisms also roughly relates to (noncomputational) literature in social science such as [7, 9, 22] which addresses more, in general, the role of cognitive interpretation and cultural influence on social interactions. Such literature may provide inspiration to design computational network models for other situations where mental states and social dynamics interact.

Adaptation inhibition of social networks (e.g., for terrorists) is a topic addressed in [5, 6]. It can be an interesting challenge to explore how far a similar architecture for controlling social network adaptation as discussed in the current paper can be applied to these types of inhibited adaptive social networks. Other possible extensions may consider the integration of different adaptation principles, such as addressed (without control) in [2].

References

1. Abraham WC, Bear MF (1996) Metaplasticity: the plasticity of synaptic plasticity. Trend Neurosci 19(4):126–130
2. van Beukel S, Goos S, Treur J (2019) An adaptive temporal-causal network model for social networks based on the homophily and more-becomes-more principle. Neurocomputing 338:361–371
3. Blankendaal R, Parinussa S, Treur J (2016) A temporal-causal modelling approach to integrated contagion and network change in social networks. In: Proceedings of the 22nd European conference on artificial intelligence, ECAI'16 (Frontiers in artificial intelligence and applications), vol 285. IOS Press, pp 1388–1396
4. Boomgaard G, Lavitt F, Treur J (2018) Computational analysis of social contagion and homophily based on an adaptive social network model. In: Proceedings of the 10th international conference on social informatics, SocInfo'18, Lecture notes in computer science, vol 11185. Springer Publishers, pp 86–101
5. Carley KM (2002) Inhibiting adaptation. In: Proceedings of the 2002 command and control research and technology symposium, Naval Postgraduate School, Monterey, CA, pp 1–10
6. Carley KM (2006) Destabilization of covert networks. Comput Math Organ Theor 12:51–66
7. Casciaro T, Carley KM, Krackhardt D (1999) Positive affectivity and accuracy in social network perception. Motiv Emot 23(4):285–305
8. Holme P, Newman MEJ (2006) Nonequilibrium phase transition in the coevolution of networks and opinions. Phys Rev E 74(5):e056108

9. Krackhardt D (1987) Cognitive social structures. Soc Netw 9:109–134
10. Levy DA, Nail PR (1993) Contagion: a theoretical and empirical review and reconceptualization. Genet Soc Gen Psychol Monogr 119(2):233–284
11. Magerl W, Hansen N, Treede RD, Klein T (2018) The human pain system exhibits higher-order plasticity (metaplasticity). Neurobiol Learn Mem 154:112–120
12. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444
13. Pearson M, Steglich C, Snijders T (2006) Homophily and assimilation among sport-active adolescent substance users. Connections 27(1):47–63
14. Sharpanskykh A, Treur J (2014) Modelling and analysis of social contagion in dynamic networks. Neurocomputing 146:140–150
15. Treur J (2016) Network-oriented modeling: addressing complexity of cognitive, affective and social interactions. Springer Publishers, New York
16. Treur J (2017) Modelling and analysis of the dynamics of adaptive temporal-causal network models for evolving social interactions. Comput Soc Netw 4:e4
17. Treur J (2019) Mathematical analysis of the emergence of communities based on coevolution of social contagion and bonding by homophily. Appl Netw Sci 4:e1
18. Treur J (2020) Modeling higher-order adaptivity of a network by multilevel network reification. Netw Sci 8:S110–S144
19. Treur J (2020) Network-oriented modeling for adaptive networks: designing higher-order adaptive biological, mental and social network models. Springer Nature Publishing, Cham, Switzerland
20. Vazquez F (2013) Opinion dynamics on coevolving networks. In: Mukherjee A et al (eds) Dynamics on and of complex networks, volume 2, modeling and simulation in science, engineering and technology. Springer, New York, pp 89–107
21. Vazquez F, Gonzalez-Avella JC, Eguíluz VM, San Miguel M (2007) Time-scale competition leading to fragmentation and recombination transitions in the coevolution of network and states. Phys Rev E 76:e046120
22. Vaisey S, Lizardo O (2010) Can cultural worldviews influence network composition? Soc Forces 88(4):1595–1618
23. Appendix at https://www.researchgate.net/publication/343682520

Toward Understanding the Use of Social Media by a Metropolitan Municipality

W. Bvuma and M. J. Hattingh

Abstract One of the most important functions of the municipalities is service delivery. The citizens need to fully engage with the municipalities. This has been made easier with internet-enabled communication such as social media. However, there is limited empirical research available that explains the level of engagement between municipalities and the citizens. The main objective of this study is to develop a social media citizen engagement framework. The framework was derived by following a netnographic approach by studying publically available Facebook and Twitter data from the metropolitan municipality, for a period of 3 months. The framework entails the elements of Facebook and Twitter of the municipality and its staff and citizens who engage in the municipality's social media platform. It was designed to illustrate the current structure of the metropolitan municipality. The results of the study have important implications for the development of the citizen engagement model for municipalities. Future work can be done based on effectively understanding the current usage of the metropolitan municipality's social media platforms.

Keywords Citizen engagement · e-Government · Social media · Netnography

1 Introduction

Using social media to engage with the citizens in local government is becoming common, the organizational dynamics underlying adoption and use decisions follow a process like that of previous waves of new information and communication technologies [1]. Social media introduce substantial and pervasive changes to communication between organizations, communities and individuals [2]. The study examines how the metropolitan municipality was engaging with its citizens using social media

W. Bvuma (✉) · M. J. Hattingh
Department of Informatics, University of Pretoria, Pretoria 0001, South Africa

M. J. Hattingh
e-mail: winniesbb@gmail.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_27

(Facebook and Twitter). Instead of having people talking about the municipal services in the background, the municipality will have an opportunity to hear the concerns of the citizens. The municipality exercised the option of publicly engaging the citizens on social media platforms. This option could assist the municipality to understand the needs of the people. Additionally, it may enable the government to communicate services delivered to citizens in a timely manner [1]. Citizens want the municipality to listen; the municipality should then listen and give the appropriate response to the citizens [3]. The municipality can use this opportunity of citizens' engagement to understand the needs and frustration of community members [4]. The municipality can choose their own social media strategy of engagement with the citizens that suit both citizens and municipality well. Social media activities vary in terms of their function and impact, to develop a congruent social media strategy based on the appropriate balance of building blocks for their community [2]. There is a high demand in local government operations that has prompted many countries to use social media to engage with citizens to better understand all the needs of the citizens so that this process could assist the local government to prioritize the service demand required by the citizens. In many countries, e-government is also used to ensure that the citizens pay their bills online and further engage with communities online to speed up the service delivery processes [5]. This could enable the citizen to start doing some of the operations by themselves. e-Government is defined as a capacity to transform the public administration using ICTs or indeed is used to describe a new form of government built around the ICTs. This aspect is usually linked to internet use [6]. Many South African citizens are already using social media and many more are continuously joining. Social media technologies can create new ways of democratic participation, pressures for new institutional structures, and processes and frameworks for the open and transparent government on an unprecedented scale [7]. The widespread use of social media has alerted the South African government to consider using social media as a tool to communicate with communities at the local government and national level. The government took the steps to look at both the benefits and the risks associated with the use of social media. The purpose of these guidelines is to create awareness of some of the opportunities that social media presents for the government, as well as making government agencies and staff aware of how to manage the risks associated with the use of this kind of technology [7].

South African government has a mandate to deliver basic services and poverty alleviation as national priorities. The basic services could include water, electricity, solid waste management, sewerage and sanitation. Municipalities engage with citizens through traditional means as well as social media platforms. However, very little research has been done on how municipalities engage with citizens through social media. According to [8] there were 18 million South Africa social media in January 2018. Based on these figures, social media users are increasing tremendously in South Africa, and it was also for this reason that this study should be undertaken. Social media could also be utilized to "improve government transparency, participation and interaction with the public" [2]. All these opportunities of social media also resulted in this study.

Against this background, the paper reports on a study that aimed to understand: How does a large metropolitan municipality use social media platforms to engage with citizens?

2 Research Methodology

The research employed netnography [9] to observe citizen engagement with a particular municipality in South Africa. Both Facebook and Twitter posts for the specific municipality were observed for a period of three months.

The reason for the netnographic method was to ensure that audiences are reached at large. This method was most effective for reaching many participants at low-cost budget. This approach was able to assist the researchers to gather more information about the municipality's interaction with citizens on social media accounts. Municipalities' customers were not asked any questions at all, but their actions were observed online for three months.

The data were classified into different categories depending on what the municipality social media site administrator was positing, or tweeting related to the service. The researcher was not directly participating in the social media forums of the municipality, whether the researcher could have participated in the forums or not; the collection of data would not have been influenced differently. The researcher played an observing role, and this enables him to be completely unobtrusive in line with previous netnography studies [10].

Following data collection, an engagement model was developed. This model was evaluated through semi-structured interviews with municipality staff.

Salzmann-Erikson and Eriksson [11] argue that ethical consideration is an early step that should be considered in netnography. Dawson [10] argues that the nature of participant observation tends to have issues involving ethics and morals. Salzmann-Erikson and Eriksson [11] continue to say that based on common knowledge, pages that do not require passwords are considered public, and "therefore research conducted using them should not be considered as human subject research". The study did not have any aspects that can lead to the immoral behavior of the observer or cause any participant to engage in immoral activities. The rights to be anonymous by the participants were also observed.

3 Data Analysis and Findings

The study analyzed the social media posts and tweets in the large metropolitan municipality's Facebook and Twitter accounts for the period of three months starting from 1 June 2018 until 31 August 2018. In the study, the netnographic research methodology was used to answer the research question: **How does a metropolitan municipality use social media platforms to engage with citizens?** The process of answering the question was involving observing the engagement of the citizens and the municipality through the social media sites of the municipality.

The researcher grouped the 1068 posts and tweets into different topics. The data were collected in a spreadsheet and the groups were analyzed as per topic. The study divided the posts and tweets into 14 categories: Electricity, Water, Cleaning, Account enquiries, Celebrations and Public Holidays, Sports and Recreation, Other, Protests, Warnings, Festivals and Choirs, Service Delivery Updates, Adverts, Other Awareness Campaigns, and Education.

The study found that at least 98% of the time there was one-way communication on Facebook, but the same could not be said about Twitter. The municipality could reply to most of the questions tweeted by the citizens. One-way communication means that the municipality could post a problem or a request; however, when the citizens start to engage the municipality staff who posted the status under the municipality account was nowhere to be found, this was common on Facebook. There were at least 40% of two-way communications on Twitter.

3.1 Engagement Model

The engagement model is the main purpose of this research. The model summarizes how the municipality engages with its citizens or the communities that the municipality is serving. By understanding the engagement model, the government can determine its current level of engagement maturity.

Figure 1 shows the current model of the large metropolitan municipality. The model will assist the municipality to understand its current stage of maturity.

The main aim of the model is to find how the municipality engages with its citizens through social media. Various metrics have been used to analyze citizen engagement through social media, with different aims [12]. According to [13] the concept of citizen engagement has multiple meaning, depending on the kind of environment, where the term is applied; for example, the term might be used in social mass movement campaign could class action lawsuits which will represent the massive amount of people or it could be workers who are fighting for the rights at the workplace environment. Social media is growing tremendously, with the possibility that all government departments will be using social media. Understanding the current model is very important to ensure that the models should fit well in communication and media strategies. Macnamara and Zefass [14] argue that strategic focus and integration of all forms of corporate communication are supported by social media specialists. This is the reason we should give much attention to social media in the media and communication department. The study constructed the model based on the two social media platforms (Facebook and Twitter) used by the large metropolitan municipality. Based on the observation and data collected from the Facebook and Twitter platform for the period of three months from June to August 2018, on Facebook only, the municipality official homepage can start the conversation by posting or sharing the status on the Facebook page of the metropolitan municipality. The

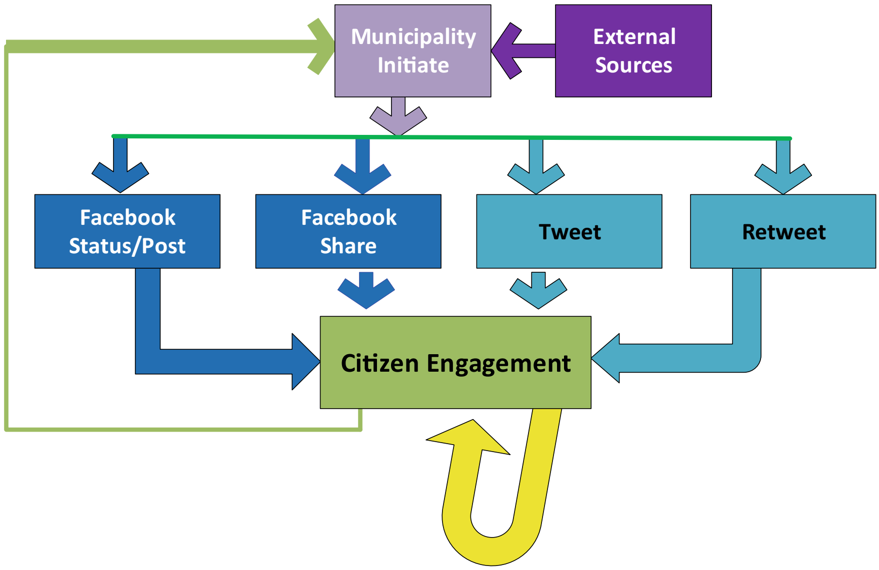


Fig. 1 Municipality social media engagement model

citizens will then engage with the municipality based on the status the municipality posted. However, there was no further engagement on Facebook from the municipal side after the initial post. In order to encourage citizens to engage, the municipality should keep the conversation going by answering the questions and giving solutions to the citizens. "A team of trustworthy people, who are authorized and competent to respond within departmental policy guidelines, is needed to ensure that a flow of conversation is maintained" [2]. After the municipality posted the status, 98% of the time the citizens will engage by commenting on the status, and only 2% of the time the status will get no comments at all. It was a very positive sign that the citizen comments in almost all the statuses posted by the municipality were different from the studies conducted in Greek municipalities by [15] where the study found that citizens communicate with the municipality in passive ways (pressing the "like" button) than by actively sharing or commenting on posts. The municipality should ensure that all the comments are answered whether good or bad. The specialized municipality "team should develop an approach for responding to negative criticism, so that serious issues can be properly escalated and are not ignored or forgotten" [2].

The initiation part done by the municipality is shown by the purple color area in the model in Fig. 1. The purple color area labeled "Municipality Initiate" is the beginning of the process on the Facebook platform. When the municipality is posting on Facebook the areas to follow in the engagement model are represented by the dark blue area labeled "Facebook Status/Post" and "Facebook share". It does not matter which dark blue process starts, but depending on the action taken first. "Facebook Status/Post" is the most common initiation of Facebook conversation. More than 99% of the posts are coming directly from the municipality posts. The less common area of initiation in Facebook is status sharing, and it was highly unlikely that the municipality will share the status of other web pages on Facebook. From the past three-month status that was collected and analyzed, only one status was shared from Facebook by the municipality. The shared page was coming from Geographical communication and Information Systems (GCIS).

Unlike on Facebook, Twitter did not only start at “municipality initiate”, it could start on either “Municipality initiate” or “External Sources”. Starting from “external sources” when an external or legal person tweeting on the municipality’s account, such tweet will eventually find the space in “municipality initiate” because it appears on the municipality Twitter home page. The social media platform that represents Twitter is shown by the light blue area labeled “Tweet” in the engagement model. The tweet area is far more common on Twitter than in a retweet. The busiest area of the model is represented by the green area labeled “Citizen Engagement”. The citizen engagement area represents the comments or retweets that are created by the citizens. This is the area where the citizens are engaged on the topic that has been initiated by the municipality whether on Facebook or Twitter. The yellow U-turn arrow shows that the citizens are replying to each other. This is the most common method of communication. In most cases, the municipality posts or tweets do not respond to the citizens but the citizens’ end up responding to each other. Based on the collected data, there is no much direct engagement between the municipality and the citizens. Engagement can be improved if the municipality reply to all the posts and tweets; without the “responses to public comments, the audience does not feel that they are engaged in a conversation, that there is little or no interaction” [2]. This is demonstrated by the number of comments per status or the number of retweets per tweet. The orange arrow that goes from the citizen’s engagement area to “municipality initiate” represents the citizens who answered the status or tweet created directly by the municipality rather than citizens replying to other citizens’ comments or tweets.

The municipality employees were given structured questionnaire questions to find out if the designed model represented their engagements with citizens over social media.

4 Conclusion

The main research question had to be answered: **How does the metropolitan municipality** use **social media platforms to engage with citizens?** The study found that the municipality’s most useful social media method of engagement was Twitter.

- Sub-research question 1: What does the municipality post on the social media platforms?

The metropolitan municipality post about service delivery-related topics, the activities that are happening within the municipality area, advertising the municipality to the citizens and service providers, sports and recreation, and education-related topics. The most common service delivery posts were issues related to electricity, water, cleaning, account enquiries, and protests.

- Sub-research question 2: What is the frequency of social media usage by the municipality?

They posted almost daily on both Facebook and Twitter. There are more posts on Twitter than on Facebook. The tweets per day can easily get to 30, for example, there were 31 tweets on 4 July 2018, and this is not a rare situation. It happened on many other days within the three months of data collected. The average tweet per day is ten. Facebook had fewer number of posts per day and did not post on many weekends while Twitter could tweet sometimes. There was a possibility that there is no tweet on any day of the week, for example on 25 July 2018, there was no official tweet from the large metropolitan municipality; however, the side remained active because the followers kept on replying to the previous tweets.

- Sub-research question 3: What is the level of engagement between the municipality and the citizens?

There was a full engagement on Twitter. It was two-way communication on Twitter. Full engagement refers to two-way communication, where both municipality and citizens comment and answer each other's questions. Most of the time there was one-way communication on Facebook, so the municipality could create a Facebook post, but when the citizens start to engage by asking questions the municipality could not completely answer the question. At 1% of the time, the municipality could answer the Facebook question. In case the residence asks a question related to the Twitter post, the municipality will continue to engage by replying or retweeting. There was fully two-way communication in Twitter and partial two ways in Facebook.

- Sub-research question 4: How do the citizens respond to the municipality's posts?

When the large metropolitan municipality tweets, the citizens will engage by either replying, retweeting or pressing the like button. Same with Facebook, when the municipality post the status, the citizens will engage by replying with the comments, sharing or liking the post. The comments from the citizen could be either relevant to the topic or not.

4.1 Contribution of the Study

The study can contribute by adding the literature to the academic research and practice. The increase in social media usage generated a large amount of data; such data need to be better used to improve the level of engagement in municipalities. Better engagement can eventually create, promote better communication and result in high quality of service delivery. The collected data might be analyzed and personalized to suit the need of a specific department or function of the municipality. The information gathered can also be analyzed differently depending on the aim of the researcher, for example, there are few comments per post or tweet. **What do the few comments tell us?** Few comments do not necessarily mean that the page was not visited, people could have a simple view and left. Most of the citizens just check the information but do not ask further for comments.

The study can also contribute to municipality employee education. A dedicated team can be appointed to analyze the post and comments from the municipality's social media sites. The analysis will be about educating the municipality employees about responding to the citizens. Social media in government might look like a small thing, but is the next very big thing that the communication department needs. The dedicated team analyzes the requests from the citizens. The dedicated team should be able to guide the employees on how to answer the public request on social media. The dedicated team must enforce the written social media policy and ensures that the municipality's posts comply with the policy.

In other words, employees who are in charge of the municipality social media account should not answer as they believe is right or say whatever they want, but should provide viewpoints that are neutral and represent the municipality, and the response should not bring much conflict to the public.

4.2 Future Research

Future research can help academics and practitioners alike to understand if social media interactions are designed effectively enough to reach the right audiences of a government agency [4].

Based on the small number of citizens participating in the municipality's social media single post or tweet, it is not yet known why there was a little comment on either Facebook or Twitter.

References

1. Lovari A, Parisi L (2015) Listening to digital publics. Investigating citizens' voices and engagement within Italian municipalities' Facebook pages. Public Relat Rev 41:205–2013
2. Kietzmann JH, Hermkens K, McCarthy Ip, Silvestre Bs (2011) Social media? Get serious! Understanding the functional building blocks of social media. Bus Horiz 54:241–251
3. Mergel I (2012) Social media in the public sector: a guide to participation, collaboration and transparency in the networked world. Wiley, Hoboken
4. Mergel I (2012) Social media in the public sector: A guide to participation, collaboration and transparency in the networked world. Wiley s.l
5. Carter L, Phillips B, Weerakkody V, Vishanth Weerakkody (2016) Citizen adoption of e-government services: exploring citizen perceptions of online services in the United States and United Kingdom. Inf Syst Manag 124–140

6. OECD (2003) The case for E-government: excerpts from the OECD report “the e-government imperative”. OECD J Budg 3(1)
7. GCIS (2011) Social media policy guidelines. South African Government Communications, pp 1–60
8. Statistics South Africa (2018) Forecast of smartphone users in South Africa. https://www.statista.com/statistics/488376/forecast-of-smartphone-users-in-south-africa/. Accessed 27 Sep 2018
9. Kozinets RV (2015) Netnography: redefined, 2nd edn. SAGE, Thousand Oaks
10. Dawson C (2002) Practical research methods: a user-friendly guide to mastering research techniques and projects. How To Books Ltd., Oxford
11. Salzmann-Erikson M, Eriksson H (2012) LiLEDDA: a six-step forum-based netnographic research method for nursing science. Nurs J 4(4):7–18
12. Haro-de-Rosario A, Sáez-Martín A, Caba-Pérez MdC (2018) Using social media to enhance citizen engagement with local government: twitter or face-book? New Media Soc 20(1):29–49
13. Lukensmeyer CJ (2012) Bringing citizen voices to the table: a guide for public managers. Josssey-Bass, San Francisco, CA, p 448
14. Macnamara J, Zerfass A (2012) Social media communication in organizations: the challenges of balancing openness, strategy, and management. Int J Strat Commun 6(4):287–308
15. Lappas G, Triantafillidou A, Deligiaouri A, Kleftodimos A (2018) Face-book content strategies and citizens’ online engagement: the case of Greek local governments. Rev Socionetwork Strat 12:1–20

A Systematic Mapping Review of PMO Frameworks

Braulio Murillo, José Antonio Pow-Sang, and Rosanna Palma

Abstract Today, project management continues to represent a challenge for organizations despite the fact that there are countless methodologies and tools to facilitate its execution. This problem is compounded when organizations run many projects in parallel. Faced with these difficulties, organizations seek to define a functional structure called project management office (PMO), to provide solutions to problems that arise. However, these initiatives are presented as challenges since it is difficult to define the PMO structure that best suits the reality of the organization. The present work executes a systematic mapping review in order to find if successful experiences exist in defining the framework or methodology for the implementation of PMOs.

Keywords Project management office · PMO · Project support office · PSO · Project management · Framework

1 Introduction

Nowadays, the management of projects continues to represent a challenge for organizations even when there are countless methodologies and tools that facilitate the management of projects in execution. The problem is compounded when organizations run many projects in parallel.

While data would suggest some improvement in the successful execution rate of IT projects over time, there are nonetheless disturbing figures around the numbers of systems abandoned or regarded as working failures [1].

B. Murillo (✉) · J. A. Pow-Sang · R. Palma
Pontifica Universidad Católica del Perú, Lima 15088, Peru
e-mail: bmurillov@pucp.edu.pe

J. A. Pow-Sang
e-mail: japowsang@pucp.pe

R. Palma
e-mail: rpalmas@pucp.edu.pe

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_28

Nevertheless, in the past decade, a number of global trends have been identified to help improve project performance. Since 2013, organizations have seen a 27% decrease in the amount of money they are wasting due to poor project performance, allowing them to move faster, produce more, and achieve greater success [2].

A previous survey conducted by the PMI [3] found that only 42% of organizations report a high level of alignment between projects and organizational strategy.

Additionally, according to the experience of the authors, the most common problems that arise in these organizations are almost always the same:

- Low planning that causes an overflow in the management of risks and scope.
- Under alignment with the organization's strategy in the selection of projects.
- Project teams almost always start "from scratch" each time they execute a new project.
- Excellence in some projects but collapse in others.
- Project managers are frustrated by the high workload and the low visibility of their responsibilities in the organization.
- Lack or low level of application of methodologies in project management.

Environmental volatility, competitive pressure for innovation, organizational complexity and technological complexity suggest a growing need for better IT project management practices in challenging contexts [4].

In view of these difficulties, the organizations direct their interest toward project management offices (PMO), since these are the calls to solve the problems outlined above. However, organizations have many difficulties when trying to define the type of PMO to implement and the functions they should have. The main challenges these initiatives must face are formally defining the role of the PMO, managing the resources, ensuring the consistent application of the defined processes and demonstrating the added value of having a PMO.

Currently, there is no defined standard or model for PMO definition that includes its structure or how to implement it [5]. Therefore, the present work executes a systematic mapping review to find if successful experiences exist in frameworks or methodology for the implementation of PMOs.

In the first part of this paper, the basic concepts are defined. Then this paper shows how the systematic mapping review was structured and the results of the review and analysis are displayed. Finally, the conclusions of the work are presented.

2 Background

According to Davies and Brady, the project management office (PMO) is an organizational entity that is created in order to standardize how projects are undertaken and to generate efficiencies through the so-called "economies of repetition" [6].

Similarly, the project management institute defines the PMO as an organizational structure that standardizes all the processes related to project management: governance, resources, methodologies, tools, and techniques. The responsibility of a PMO is providing project management support functions to the direct management of one or more projects [7].

In the case of ISO 21500, the PMO is considered as an external stakeholder to the project management, but depending on its role and functions, it could also fit within the project governance group. The PMO is called the project support office (PSO) [8].

In PRINCE2, the PMO is also called PSO and is conceived as an office that centralizes qualified resources. The PRINCE2 manual indicates that a PSO is not essential, but recognizes that it is very useful to support project managers with their administrative tasks and to ensure the proper use of PRINCE2 in all projects. The creation of the PSO is recommended for large organizations since that way we can count on personnel exclusively dedicated to providing the services associated with project management [9].

3 Performing the Systematic Mapping Review

3.1 Research Questions

The major objective of this paper was to find the most used frameworks for the implementation of PMOs for the management of technological projects. Additionally, it is important to know in what context the frameworks are applied and whether they are proposals developed by the authors themselves or are based on some well-known standards or best practices (e.g., PMI, PRINCE2, etc.).

For this purpose, the following research questions are defined:

RQ1: What are the most widely used PMO framework?
RQ2: In what industries are PMO frameworks being applied?
RQ3: In what type of organizations are PMO frameworks being applied?
RQ4: Is the PMO framework applied only to IT projects?
RQ5: In what country was applied the study?

In order to perform this systematic mapping review, general concepts based on PICOC are defined. The "comparison" criterion was not considered because this research did not compare interventions. The PICOC criteria are shown in Table 1.

Table 1 Definition of the concepts using PICOC

Criteria	Description
Population	All kind of projects
Intervention	Project management office framework, methodology or guideline
Outcomes	Cases studies where frameworks are defined to implement PMOs
Context	All contexts and empirical studies

3.2 Search Strategy

In order to find if evidence of the use of PMO frameworks exists, the following search string is defined:

("Project*") AND ("Project management office" OR "PMO" OR "Project Support Office" OR "PSO") AND ("Methodology" OR "Method" OR "Framework" OR "Guidelines" OR "Principles").

Only papers from 2015 onwards were considered. The search was performed in Scopus, IEEE Xplore, and Web of Science.

All the works obtained were reviewed by the authors in order to determine if they would be included as part of the objective study of this paper. A full review of the documents was conducted.

3.3 Data Extraction

The search for papers was conducted in May 2019. A total of 622 papers were obtained from the databases. The PRISMA 2009 flow diagram [10] was applied to make the selection of the works to be reviewed. Figure 1 shows the adapted flow diagram of the PRISMA proposal.

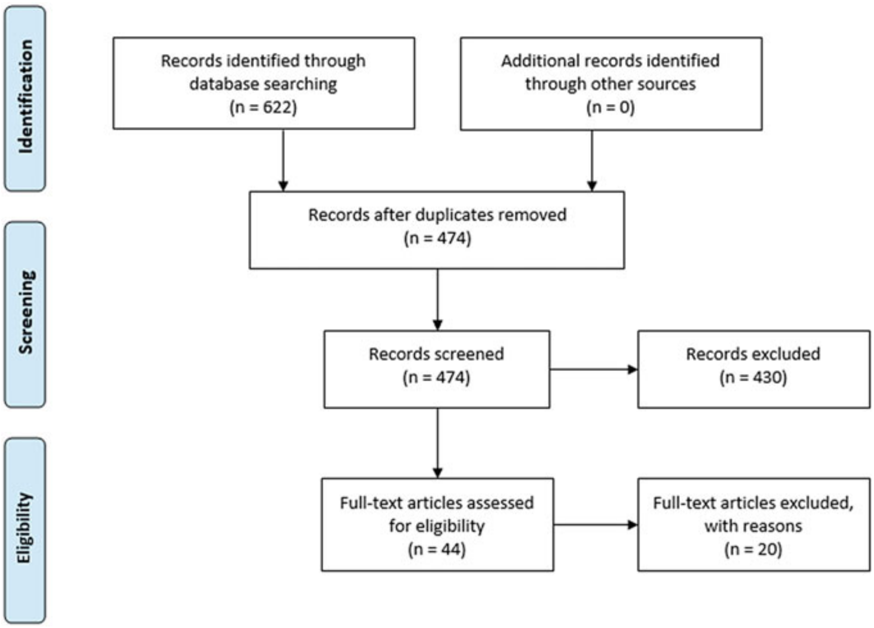


Fig. 1 PRISMA 2009 flow diagram adapted

Table 2 Results in each database

Database name	Search results	Duplicated papers	Relevant papers
IEEE Xplore	77	–	–
SCOPUS	358	29	20
WoS	187	119	4
Total	622	148	24

In the identification phase, 622 papers were obtained when applying the search string. Of these, 148 duplicate works were found. Of the remaining 474 works, 430 were excluded for the following reasons: they were not accessible, they were not in English, and their content had nothing to do with the implementation of PMO. Then, the remaining 44 works were reviewed in detail. Of these, 20 jobs were excluded because they only mentioned the PMO but did not provide any input for its implementation. Table 2 presented the results.

4 Data Analysis and Results

From reviewed papers, 71% do not define a framework for the implementation of PMO, they only suggest its use for better project management within the organization. Only 29% of the works define their own framework. Of these papers, the vast majority proposed their own frameworks based on an exhaustive review of the literature.

4.1 Criteria for Implementing PMO

According to the reviewed papers, the main criteria to be taken into account for the implementation of a PMO are shown in Table 3.

4.2 Results by Industry Domain

In Table 4, the main industry domain found in the results is presented. From the reviewed papers, it is observed that 38% were applied or developed taking projects of any type of industry domain. The other group of papers (13%) was applied in the public sector. 8% of the jobs were found for each of the following sectors: banking/finance, military, and IT. Finally, the remaining works were applied in the context of other industry domain.

Table 3 PMO implementation criteria by article

ID	Roles and responsibilities	Functions	Types	Structure	Process	Governance	Performance indicator	Agile	Lesson learned	Benefits	Maturity model
[11]							X				
[12]		X	X								
[13]	X										
[14]		X						X			
[15]									X		
[16]	X	X								X	
[17]		X									
[18]										X	
[19]	X		X		X		X				
[20]					X	X					
[21]				X	X						
[22]	X	X									
[23]			X								
[24]	X										
[25]		X	X		X			X			
[26]		X								X	
[27]											X
[28]	X			X							
[29]	X										
[30]	X		X								
[31]	X	X									

(continued)

Table 3 (continued)

ID	Roles and responsibilities	Functions	Types	Structure	Process	Governance	Performance indicator	Agile	Lesson learned	Benefits	Maturity model
[32]	X										
[33]		X									
[34]			X								

Table 4 Frequency of main industry domain

Industry domain	Number of papers	Percentage (%)
Any type of industry	9	38
Automotive	1	4
Banking/finance	2	8
Construction	1	4
Education	1	4
Governmental	3	13
Health	1	4
IT services	2	8
Military	2	8
Mining	1	4
Non-governmental	1	4
Total	24	100

4.3 Results by Type of Organization

From the reviewed works, it is observed that 50% have proposed to apply them in any type of organization. The other half was applied in large organizations.

4.4 Results by Origin

From the results shown in Fig. 2, it is noted that the largest amount of papers (20%) were produced in Brazil. 16% of the following papers were produced in Portugal. In Canada, 12% of the papers reviewed were prepared.

4.5 Results by PMO Type of Deliveries

The vast majority of studies (75%) were applied for any type of project. The remaining review papers were applied for the IT field.

5 Conclusion

One of the main reasons for this research was to find which framework proposals to implement a PMO can be found in the literature. Furthermore, another objective

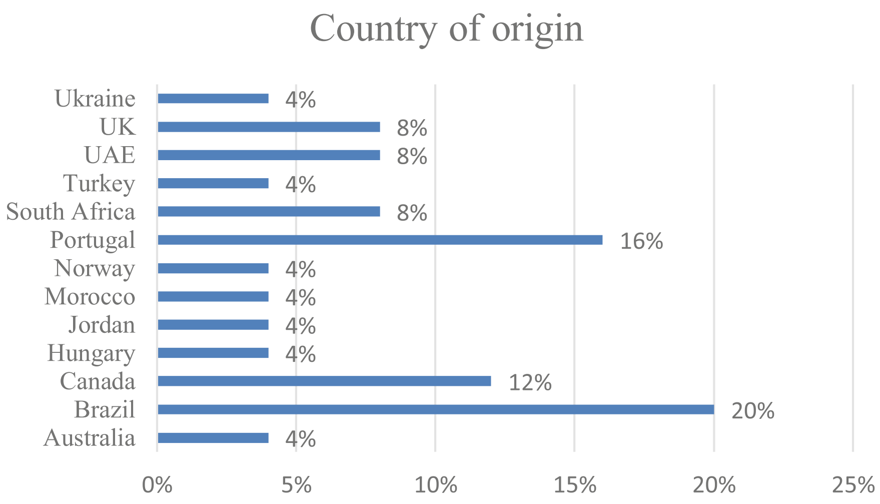


Fig. 2 Frequency of country of origin

was to identify if these frameworks were based on recommendations or standards of special institutions or organizations.

The work carried out shows that there are still very few experiences where complete frameworks for implementing PMOs are shown. Of the few papers reviewed, the vast majority of framework proposals are based on case studies developed by the authors themselves and literature review. Another important point to note is that only a quarter of the work reviewed corresponded to the implementation of PMOs aimed at managing IT projects.

The authors of this research agree with that reported by other authors in the literature, where it is indicated that there is no framework that is most used because the proposals are rigid and are not easily adapted in organizations. Finally, they recommend defining a proposal of a flexible framework that includes the best practices and standards and can be applied to different contexts and IT projects.

References

1. Letavec C, Bolles D (2011) The PMOSIG's program management office handbook: strategic and tactical insights for improving results. J. Ross Publishing, Plantation
2. Project Management Institute (2019) https://www.pmi.org/-/media/pmi/documents/public/pdf/learning/thought-leadership/pulse/pulse-of-the-profession-2018.pdf. Accessed 21 Nov 2019
3. Project Management Institute (2019) https://www.pmi.org/~/media/PDF/Business-Solutions/PMO_StrategyImplement_casestudy_v4.ashx. Accessed 21 Nov 2019
4. Sauer C, Gemino A, Reich B (2007) The impact of size and volatility on IT project performance. Commun ACM 50(11):79–84
5. Salameh H (2014) A framework to establish a project management office. Eur J Bus Manag 6(9):19–26

6. Davies A, Brady T (2000) Organisational capabilities and learning in complex product systems: towards repeatable solutions. Res Policy 29(7–8):931–953
7. Project Management Institute (2017) A guide to the project management body of knowledge (PMBOK guide), 6th edn. Newton Square
8. International Standards Office (2012) ISO 21500:2012. Guidance on project management. ISO, Geneva
9. AXELOS (2009) Managing successful projects with PRINCE2 2009 edition, Stationery Office Limited, London
10. Liberati A et al (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 62(10):e1–e34
11. Viglioni T, Cunha J, Moura H (2016) A performance evaluation model for project management office based on a multicriteria approach. Procedia Comput Sci 100:955–962
12. Alqubaisi S (2015) Assisting national governments in governance, risk, and compliance: implementation of project management offices. In: Proceedings of the 26th international business information management association conference—innovation management and sustainable economic competitive advantage: from regional development to global growth, IBIMA 2015, pp 2152–2162
13. Sandhu M, Al Ameri T, Wikström K (2019) Benchmarking the strategic roles of the project management office (PMO) when developing business ecosystems. Benchmarking Int J 26(2):452–469
14. Pinto J, Ribeiro P (2018) Characterization of an agile coordination office for IST companies. Procedia Comput Sci 138:859–866
15. Nakamura P (2016) Complex project management. Adv Transdiscipl Eng 4:3–11
16. Philbin S (2016) Exploring the project management office (PMO)—role, structure and processes. In: Proceedings of the international annual conference of the American society for engineering management, American Society for Engineering Management (ASEM), pp 1–11
17. Bredillet C, Tywoniak S, Tootoonchy M (2015) Grasping the dynamics of co-evolution between PMO and PfM: a box-changing multilevel exploratory research grounded in a routine perspective. J Mod Proj Manag 2(3)
18. Lavoie-Tremblay M et al (2017) Innovation in health service management: adoption of project management offices to support major health care transformation. J Nurs Manag 25(8):657–665
19. Szalay I, Kovács Á, Sebestyén Z (2017) Integrated framework for project management office evaluation. Procedia Eng 196:578–584
20. Tsaturyan T, Müller R (2015) Integration and governance of multiple project management offices (PMOs) at large organizations. Int J Proj Manag 33(5):1098–1110
21. Fesenko T, Shahov A, Fesenko G (2017) Modeling of maturity of gender-oriented project management office. East Eur J Inter Technol 5(3(89))
22. Oliveira C, Tereso A, Fernandes G (2017) PMO conceptualization for engineering and construction businesses. Procedia Comput Sci 121:592–599
23. Monteiro A, Santos V, Varajão J (2016) Project management office models—a review. Procedía Comput Sci 100:1085–1094
24. Aubry M (2015) Project management office transformations: direct and moderating effects that enhance performance and maturity. Proj Manag J 46(5):19–45
25. Hamad R, Al Fayoumi M (2018) Scalable agile transformation process (SATP) to convert waterfall project management office into agile project management office. In: 2018 international Arab conference on information technology (ACIT), pp 1–8
26. Van der Linde J, Steyn H (2016) The effect of a project management office on project and organisational performance: a case study. S Afr J Ind Eng 27(1):151–161
27. Tahri H, Kaitouni O (2017) The new project management maturity mixed model (P4M) and the OPM3: case of a PMO implementation. In: Proceedings of the international conference on industrial engineering and operations management, pp 2048–2057

28. Paton S, Andrew B (2019) The role of the project management office (PMO) in product lifecycle management: a case study in the defence industry. Int J Prod Econ 208:43–52
29. Tshuma B, Steyn H, Van Waveren C (2018) The role played by PMOs in the transfer of knowledge between projects: a conceptual framework. S Afr J Ind Eng 29(2):127–140
30. Fernandes G et al (2018) The roles of a programme and project management office to support collaborative university—industry R&D. Total Qual Manag Bus Excel 1–26
31. Güngör D, Gözlü S (2017) Investigating the relationship between activities of project management offices and project stakeholder satisfaction. Int J Inf Technol Proj Manag (IJITPM) 8(2):34–49
32. Terlizzi M, Albertin A (2017) IT benefits management in financial institutions: practices and barriers. Int J Project Manag 35(5):763–782
33. Lacruz A, Cunha E (2018) Project management office in non-governmental organizations: an ex post facto study. Rev Gest 25(2):212–227
34. De Oliveira J et al (2017) Project offices and the federal universities: a study on project management in the context of higher education institutions. Rev Gest Proj 8(3):18–28

Symmetric Splitting Methods for Wave Equations: First Approaches

Jürgen Geiser and Abdullah Mujahid

Abstract This article consider symmetric splitting methods for solving wave equations. We propose novel splitting methods, which are unconditionally stable, and apply diagonalization techniques to gain efficient solvers. The benefits to standard methods are discussed and we apply the methods to multidimensional wave equations. We are motivated by fast exponential computations based on matrix decompositions. We reformulate wave equations as system of coupled initial value boundary problems by discretizing only in space and use fast exponential of the resulting matrix. Therefore, we could solve fast systems of Cauchy initial value boundary problems. This process results in symmetric splitting of higher dimensional partial differential equations (PDEs), which are decoupled into a system of symmetric first-order ordinary differential equations (ODEs).

Keywords Symmetric splitting methods · Diagonalization · Nilpotent matrices · Efficient solvers · Lie-Trotter splitting · Strang splitting methods · Stability analysis · Wave equation

1 Introduction

The motivation arose to design a higher dimensional wave equation solver, which is efficient and unconditional stable. Such solvers are used to simulate delicate model electromagnetic problems, see [7]. We propose a novel method, which is based on

J. Geiser (✉)
Ruhr University of Bochum, The Institute of Theoretical Electrical Engineering, Universitätsstrasse 150, 44801 Bochum, Germany
e-mail: juergen.geiser@ruhr-uni-bochum.de
URL: https://homepage.ruhr-uni-bochum.de/Juergen.Geiser/

A. Mujahid
Computational Engineering, Ruhr University of Bochum, Universitätsstrasse 150, 44801 Bochum, Germany

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_29

exponential splitting methods. The operators are decomposed into simple diagonal and outer-diagonal operators, which can be computed very fast till the typical machine precision.

We have the following outline of the paper: The mathematical models of the wave equations are introduced in Sect. 2. In Sect. 3, we give an overview of the symmetric operator-splitting methods with respect to the wave equation and discuss the underlying numerical approaches. In Sect. 4, we present the numerical experiments, and finally, we discuss our future works in the area of high-dimensional splitting approaches and future applications.

2 Mathematical Model

The modeling equations, which are second-order hyperbolic PDEs can be applied in many directions of wave propagation, e.g., propagation of multidimensional waves, i.e., sound waves, light waves, and water waves. For further propagation of electromagnetic fields, which are important in antenna simulations or particle simulations, see [4, 7].

For the first approach, we consider wave equations with constant phase velocity coefficients, which are applied to models in E- and B-field simulations, e.g., Maxwell equation in vacuum. The linear wave equation with constant phase velocity coefficients is given as a initial-boundary value problem as

$$\frac{\partial^2 c}{\partial t^2} = D_1 \frac{\partial^2 c}{\partial x_1^2} + \cdots + D_d \frac{\partial^2 c}{\partial x_d^2} \quad \text{on } \Omega \times [0, T], \tag{1}$$

$$c(x, 0) = c_0(x), \ \text{ and } \frac{\partial c}{\partial t}(x, 0) = c_1(x) \quad \text{on } \Omega,$$

$$c(x, t) = c_2(x, t) \ \text{ on } \partial\Omega \times [0, T], ,$$

where $D_i \in \mathbb{R}^+$ are constant for $i = 1, \ldots, d$. The initial conditions are c_0, c_1 and the boundary condition is c_2, e.g., the Dirichlet boundary condition.

For the comparison to the numerical approaches, we can derive an analytical solution, which is given as

$$c(x, t) = \sin(\frac{1}{\sqrt{D_1}} \pi x_1) \cdot \ldots \cdot \sin(\frac{1}{\sqrt{D_d}} \pi x_d) \cdot \cos(\sqrt{d} \pi t), \tag{2}$$

where $x = (x_1, \ldots, x_d)^t$ and d is the spatial dimension, e.g., $d = 1, 2, 3$.

In the following, we discuss a novel numerical scheme to solve the wave equation with efficient and fast solver ideas, which are based on splitting approaches, see [4].

3 Splitting Methods

Splitting methods of PDEs are nowadays very important and reduce the computational time, see [1]. The underlying idea is to decompose fully coupled PDE into simpler solvable decoupled PDE equation parts, see [8, 9]. In combination with physical behaviors, they can be designed with additional conservation properties, e.g., symplectic splitting schemes, [5, 14].

To obtain an efficient and high accurate solver, we deal for our applications with two different splitting ideas:

– the first one (exponential splitting) is used for splitting into fast solvable operator equations, which are based on solving fast exponential matrices, see [4, 10, 12],
– while the second one (symmetric splitting) is used for decomposing the matrix-operators into diagonal and nilpotent matrices, which could be fast computed as matrix–vector multiplications, see [6]; further, we can decompose high-dimensional PDE into a symmetric system of first order ODEs, see [13].

3.1 Exponential Splitting Methods

The exponential splitting methods are widely applied in parabolic and hyperbolic PDEs, see [4, 9]. We could reformulate the semi-discretized PDEs, where we assume to embed the boundary conditions, as a Cauchy problem, which is a system of linear ODEs. Such a system can be solved with exponential functions and the methods are unconditional stable for the time evolution, see [9]. In a natural way, we decompose a fully semi-discretized PDE into simpler semi-discretized PDE parts, which can be done, for example:

$$\frac{dc(t)}{dt} = A_{\text{full}}\, c(t), \text{ for } t \in (t^n, T), \tag{3}$$

$$\frac{dc(t)}{dt} = \sum_{i=1}^{m} A\, c(t), \quad \text{for } t \in (t^n, T), \tag{4}$$

$$c(t^n) = c^n, \text{ (initial condition)}, \tag{5}$$

where $t^n, T \in \mathbb{R}^+$ and $t^n \leq T$. The operator A_{full}, which concludes of the semi-discretization, e.g., with finite-difference or finite-element methods, is decoupled into the simpler operators, e.g., $A_1, \ldots, A_m$, see [10].

The solution of the subproblems are given as

$$c_i(t^{n+1}) = \phi_{A_i}(\Delta t)\, c_i(t^n) = \exp(\tau A_i)\, c_i(t^n), \; i = 1, \ldots, m, \tag{6}$$

where $\tau = t^{n+1} - t^n$ is the uniform time step and $t^{n+1} \leq T$.

We have the following splitting approaches:

– Lie–Trotter splitting:

$$\chi(\Delta t) = \phi_{A_m}(\Delta t) \dots \phi_{A_1}(\Delta t), \tag{7}$$

and

– Strang splitting:

$$\chi(\Delta t/2) \circ \chi^*(\Delta t/2) \text{ or } \chi^*(\Delta t/2) \circ \chi(\Delta t/2), \tag{8}$$

where $\chi^*(\Delta t) = \chi^{-1}(\Delta t) = \phi_{A_1}(\Delta t) \dots \phi_{A_m}(\Delta t)$ is the adjoint map.

3.2 Symmetric Splitting Method

The symmetric splitting method is a novel splitting idea, which symmetric decomposed fast computable exponential matrices, see [6]. The exponential matrices are obtained by exponential splitting methods, see the previous paragraph.

In the following, we discuss the different steps of the symmetric splitting for the wave equation:

– Symmetric splitting of the advection equation, which is one part of symmetric splitting of the wave equation.
– Symmetric splitting of the wave equationsis based on transformation of a system of advection equation, which is applied to the symmetric splitting.
– Decomposition into decoupled or coupled options, where we assume different boundary conditions.

3.2.1 Symmetric Splitting Method of the Advection Equation

The fundamental part of the symmetric splitting is the underlying advection equation, which is given as

$$\frac{\partial u}{\partial t} = -v\frac{\partial u}{\partial x}, \quad v > 0, (x,t) \in \Omega \times [0,T], \tag{9}$$

$$u(x,0) = u_0(x), \ x \in \Omega, \text{ initial condition}, \tag{10}$$

$$u(x,t) = 0, \ (x,t) \in \partial\Omega \times [0,T], \text{ boundary condition}, \tag{11}$$

We discretize with the splitting approach:

$$U((n+1)\Delta t) = \exp(-\nu)\ E\ U(n\Delta t), \tag{12}$$

with $\nu = v\frac{\Delta t}{\Delta x}$ and the matrix E is given as:

$$E_{ij} = \begin{cases} \frac{\nu^{i-j}}{(i-j)!} & j \le i, \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

We have the time-steps $n = 0, \ldots, N-1$, while we start with U_0 as initial condition.

The same can be done with $v \le 0$, then we have to transpose E with the value $|v|$ and also in the diagonal, we have to apply $|v|$.

3.2.2 Symmetric Splitting Method of the Wave Equation

In the following, we deal with the wave equation, which results of two coupled advection equations and given as

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \, (x, t) \in \Omega \times [0, T], \quad (14)$$

$$u(x, 0) = u_0(x), \quad (15)$$

$$u_t(x, 0) = 0, \quad (16)$$

we assume to deal with Dirichlet boundary conditions as

$$u = u_D = 0 \text{ on } \partial\Omega.$$

Based on our assumptions, we could the shift the wave equation (14) into a system of first-order hyperbolic differential equations, where we have $v = c^{-1}u_t$ and $w = u_x$:

$$\partial_t \begin{bmatrix} v \\ w \end{bmatrix} = \partial_x \begin{bmatrix} 0 & c \\ c & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix}, \quad (17)$$

$$u(x, 0) = \sin(\pi x), \quad (18)$$

$$u_t(x, 0) = 0, \quad (19)$$

and with the Dirichlet boundary conditions of $v(x, t) = 0$ for $x \in \partial\Omega$, which is related to the Dirichlet BC of $u(x, t) = 0$ for $x \in \partial\Omega$.

Then, we diagonalize the coupled PDEs (17) and obtain

$$\partial_t \begin{bmatrix} p \\ q \end{bmatrix} = \partial_x \begin{bmatrix} c & 0 \\ 0 & -c \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix}, \quad (20)$$

$$p(x, 0) = \frac{1}{\sqrt{2}}(c^{-1}u_t(x, 0) + u_x(x, 0)), \quad (21)$$

$$q(x, 0) = \frac{1}{\sqrt{2}}(c^{-1}u_t(x, 0) - u_x(x, 0)), \quad (22)$$

$$q(0, t) = -p(0, t), \quad (23)$$

$$p(2,t) = -q(2,t), \tag{24}$$

where we have the transformation matrix:

$$\begin{bmatrix} v \\ w \end{bmatrix} = Q \begin{bmatrix} p \\ q \end{bmatrix}, \tag{25}$$

where $Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $QQ^T = I$.

The solutions for p and q are given as

$$p(x,t) = \frac{1}{\sqrt{2}}(c^{-1}u_t(x,t) + u_x(x,t)), \tag{26}$$

$$q(x,t) = \frac{1}{\sqrt{2}}(c^{-1}u_t(x,t) - u_x(x,t)), \tag{27}$$

with the Dirichtlet BC $p(x,t) + q(x,t) = 0$ for $x \in \partial\Omega$. Then, we have

$$p(x,t) = \frac{1}{\sqrt{2}}(-\pi \sin(c\pi t)\sin(\pi x) + \pi \cos(\pi x)\cos(c\pi t)), \tag{28}$$

$$q(x,t) = \frac{1}{\sqrt{2}}(-\pi \sin(c\pi t)\sin(\pi x) - \pi \cos(\pi x)\cos(c\pi t)). \tag{29}$$

The initial conditions are given as

$$p(x,0) = \frac{1}{\sqrt{2}}(c^{-1}u_t(x,0) + u_x(x,0)), \tag{30}$$

$$q(x,0) = \frac{1}{\sqrt{2}}(c^{-1}u_t(x,0) - u_x(x,0)). \tag{31}$$

for the time derivations of the initial conditions:

$$p_t(x,0) = \frac{1}{\sqrt{2}}(-\pi^2 c)\sin(\pi x), \tag{32}$$

$$q_t(x,0) = \frac{1}{\sqrt{2}}(-\pi^2 c)\sin(\pi x), \tag{33}$$

and for the boundary conditions at $x = 0$ and $x = 2$, we have

$$p(0,t) = -q(0,t) = \frac{1}{\sqrt{2}}\pi \cos(c\pi t), \tag{34}$$

$$p(2,t) = -q(2,t) = \frac{1}{\sqrt{2}}\pi \cos(c\pi t). \tag{35}$$

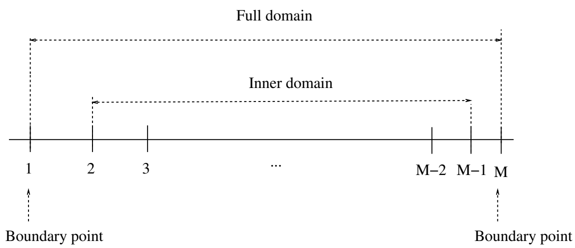


Fig. 1 The spatial domain in 1D with the notations

In Fig. 1, we have the 1D spatial-grid notations.

3.2.3 Decoupling/Coupling of the Parts p and q with Exact BCs

We discuss two different algorithms, which reduce the wave equations into a fast computation of diagonal and nilpotent matrices.

In the following Algorithm 1, we have two versions related to decouple/couple the p and q boundary conditions:

– decoupled version, here we assume analytical boundary conditions, such that we could skip the parts 3 and 5 of Algorithm 1 and
– coupled version, here we assume the realistic boundary conditions, such that we have to apply coupling in the boundary as is done in parts 3 and 5 of Algorithm 1).

Algorithm 1 We solve the pairs of advection equation in the decoupled or coupled version as

1. We initialize with $p(0) = (p(0)_1, \ldots, p(0)_M)^t, q(0) = (q(0)_1, \ldots, q(0)_M)^t$ and $u(0) = (u(0)_1, \ldots, u(0)_M)^t$
   We have $n = 0$. Further, we have the inner domain with the vectors $P = (p_2, \ldots, p_{M-1})^t$, $Q = (q_2, \ldots, q_{M-1})^t$ and the boundary points p_1, p_M and q_1, q_M.
2. We solve the parts for p (inner domain which is decoupled of q)

$$\frac{\partial p}{\partial t} = c\frac{\partial p}{\partial x}, \quad c > 0, (x,t) \in \Omega \times [0,T], \tag{36}$$

$$p(x,0) = p_0(x),\ x \in \Omega,\ \text{initial condition,} \tag{37}$$

$$p(x,t) = -q(x,t) + \frac{\sqrt{2}}{c}u_t(x,t),\ (x,t) \in \partial\Omega \times [0,T],\ \text{boundary condition,}$$

   where $p_0(x) = \frac{1}{\sqrt{2}}(c^{-1}u_t(x,0) + u_x(x,0))$.

   – For the decoupled algorithm, we use analytically pre-computed values for p and q avoiding the non-homogeneous equations. Such that the equations relating p and q are fully decoupled and we goto step 4.

– For the coupled algorithm, we apply the boundary condition of $p(x,t) = -q(x,t) + \frac{\sqrt{2}}{c} u_t(x,t)$, such that the last component of P is a non-homogeneous equation and we goto step 3.

We discretize with the splitting approach:

$$P((n+1)\Delta t) = \exp(-\nu)\ E^t\ P(n\Delta t), \tag{38}$$

with $\nu = c\frac{\Delta t}{\Delta x}$ and the matrix E is given as

$$E_{ij} = \begin{cases} \frac{\nu^{i-j}}{(i-j)!} & j \leq i, \\ 0 & \text{otherwise} \end{cases} \tag{39}$$

3. Coupling part (we couple with the BC, means the outer domain):
We have to solve the last row with the RHS means, we have

$$\frac{\partial p_{M-1}(t)}{\partial t} = \frac{c}{\Delta x}(-p_{M-1}(t) + p_M(t)). \tag{40}$$

We solve via the variation of constants in the interval $t \in [t^n, t^{n+1}]$ and obtain:

$$p_{M-1}(t^{n+1}) = \exp(-\frac{c}{\Delta x}\Delta t) p_{M-1}(t^n) + \int_{t^n}^{t^{n+1}} \exp(-\frac{c}{\Delta x}(t^{n+1} - s)) p_M(s) ds, \tag{41}$$

where we evaluate the integral with higher order numerical integration, see [2, 11].

4. Part for q (inner domain which is decoupled of p)

$$\frac{\partial q}{\partial t} = -c\frac{\partial q}{\partial x}, \quad c > 0, (x,t) \in \Omega \times [0,T], \tag{42}$$

$$q(x,0) = q_0(x),\ x \in \Omega,\ \text{initial condition}, \tag{43}$$

$$q(x,t) = -p(x,t) + \frac{\sqrt{2}}{c} u_t(x,t),\ (x,t) \in \partial\Omega \times [0,T],\ \text{boundary condition},$$

where $q_0(x) = \frac{1}{\sqrt{2}}(c^{-1} u_t(x,0) - u_x(x,0))$.

– For the decoupled algorithm, we use analytically pre-computed values for p and q avoiding the non-homogeneous equations. Such that the equations relating p and q are fully decoupled and we goto step 6.
– For the coupled algorithm, we assume to have to apply the boundary condition of $q(x,t) = -p(x,t) + \frac{\sqrt{2}}{c} u_t(x,t)$, such that the first component of Q is non-homogeneous equation and we goto step 5.

We discretize with the splitting approach:

$$Q((n+1)\Delta t) = \exp(-\nu)\ E\ Q(n\Delta t), \tag{44}$$

with $\nu = c\frac{\Delta t}{\Delta x}$ and the matrix E is given as

$$E_{ij} = \begin{cases} \frac{\nu^{i-j}}{(i-j)!} & j \leq i, \\ 0 & \text{otherwise} \end{cases} \tag{45}$$

5. Coupling part (we couple with the BC, means the outer domain): We have to solve the first row with the RHS means, we have

$$\frac{\partial q_2(t)}{\partial t} = -\frac{c}{\Delta x}(q_2(t) - q_1(t)). \tag{46}$$

We solve via the variation of constants in the interval $t \in [t^n, t^{n+1}]$ and obtain

$$q_2(t^{n+1}) = \exp(-\frac{c}{\Delta x}\Delta t) q_1(t^n) + \int_{t^n}^{t^{n+1}} \exp(-\frac{c}{\Delta x}(t^{n+1} - s)) q_1(s) ds, \tag{47}$$

where we evaluate the integral with higher order numerical integration, see [2, 11].

6. Then, we have the solution with

$$u(x,t) = u(x,0) + \frac{c}{\sqrt{2}} \int_0^t (p(x,s) + q(x,s)\ ds, \tag{48}$$

which is applied in the discretized form as

$$u((n+1)\Delta t) = u(n\Delta t) + \frac{c\ \Delta t}{\sqrt{2}}(p((n+1)\Delta t) + q((n+1)\Delta t), \tag{49}$$

7. If $n = N + 1$, we are done, if not we apply $n = n + 1$ and goto Step 2.

4 Numerical Examples

In the numerical examples, we apply a first approach of the one-dimensional wave equation, which is given as

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, (x,t) \in \Omega \times [0,T], \tag{50}$$

$$u(x,0) = \sin(\pi x)\ x \in \Omega, \tag{51}$$

$$u_t(x,0) = 0, \tag{52}$$

where $\Omega = [0, 2]$, $T = 1, 10, 100$.

We test the following two different Algorithms:

- Algorithm (decoupled) 1: Here, we assume to deal with analytical boundary conditions, which are known in previous and we could fully decouple into two advection equations, means we could ignore the non-homogeneous part.
- Algorithm (coupled) 1: Here, we have the given boundary conditions, such that we have to apply a coupling between the two advection equations, which means we have an additional RHS.

Error in following $\mathcal{L}_2$ norm is used to evaluate the numerical methods.
$err_{L_2}(t) = \sum_i^I \Delta x \; ||u_{exact}(x_i, t) - u_{num}(x_i, t)||$.

We study the convergence of the numerical scheme with varying spatial discretization. The errors are tabulated for $t = 1, 10, 100$ and $\Delta x = 0.05, 0.025, 0.0125,$ 0.00625.

In Table 1, we compare the energy conservation of the different algorithms.

Remark 1 In a first test example, we see the benefit of the symmetrical splitting approach, while we could reduce the computational time by solving 1D advection equation. The second algorithm is more realistic, while it assumes the coupled systems based on the boundary conditions. Here, we see that the L_1-error is reduced and also the energy computation is more precise with smaller time and spatial steps.

Remark 2 For the computation, the exponential matrix $E = exp(A)$ is in general a triangular matrix for a bidiagonal A, it's entries for numerical purposes are sparse. After 3–4 diagonals, the entries become less than typical machine epsilon, means around $1.0e^{-16}$,(for small time steps, $\frac{\nu^{(a-b)}}{(a-b)!}$, it is nearly one over factorial $(a - b)$, so for numerical implementation, it can be simplified as sparse matrix vector product. Though the framework allows for an unconditionally stable scheme, it is still a sparse matrix–vector product (Fig. 2).

5 Conclusions and Discussions

In the paper, we presented novel symmetric splitting methods for solving wave equations. Based on an unconditionally stable scheme and preserving the physical damp-

Table 1 Comparison of the energy conservation of decoupled and coupled algorithms

Δx	Algorithm 1 (decoupled)			Algorithm 1 (coupled)		
	$t = 1$	$t = 10$	$t = 100$	$t = 1$	$t = 10$	$t = 100$
0.05	60.92	56.785	56.785	92.897	106.9	106.9
0.025	143.3	128.87	128.87	216.63	255.92	255.92
0.0125	317.84	283.31	283.31	470.13	566.5	566.5
0.00625	672.01	598.26	598.26	980.96	1194.9	1194.9

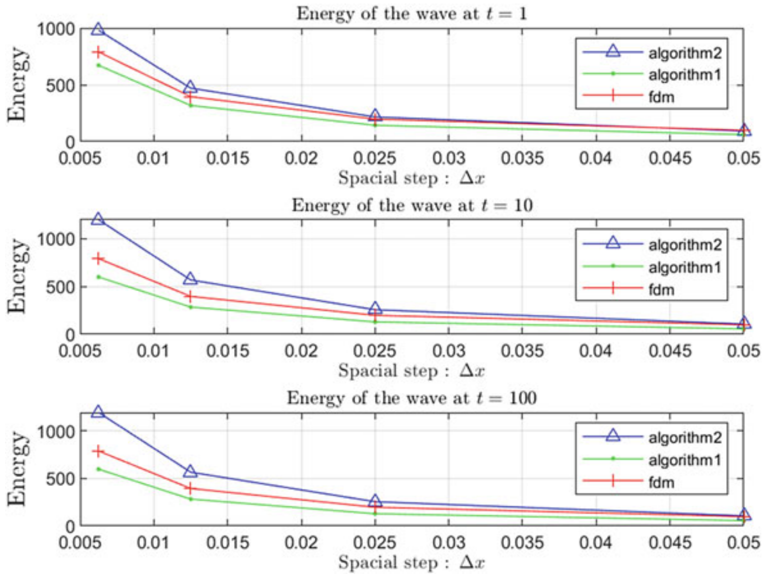


Fig. 2 Energy of the wave: comparison

ing, we could obtain results, which are in the region of standard finite-difference thuds (FDMs). For higher dimensional problems, the new symmetric splitting method might be a promising method. Here, we could decompose into 1D PDE problems based on splitting approaches, e.g., locally 1D (LOD). Then the locally 1D problems can be computed with the fast symmetric splitting methods in parallel. Here, the symmetric splitting approaches can be accelerated based on the fast computations of shifting matrices instead of vector-matrix multiplications. Therefore, we have an attractive alternative approach for numerical methods for wave equations.

References

1. Geiser J (2007) Iterative operator-splitting methods with higher order time-integration methods and applications for parabolic partial differential equations. J Comput Appl Math (accepted)
2. Geiser J (2011) Computing exponential for iterative splitting methods: algorithms and applications. J Appl Math **2011**(Article ID 193781):27
3. Geiser J, Noack L (2007) Iterative operator-splitting methods for wave-equations with stability results and numerical examples, Preprint 2007–10 of Humboldt University of Berlin. Department of Mathematics, Germany
4. Geiser J (2016) Multicomponent and multiscale systems: theory, methods, and applications in engineering. Springer, Cham, Heidelberg, New York, Dordrecht, London
5. Hairer E, Lubich C, Wanner G (2002) Geometric numerical integration. SCM, Springer, Berlin, Heidelberg, New York, No. 31
6. Hosseini R, Tatari M (2019) Some splitting methods for hyperbolic PDEs. Appl Numer Math (available online)

7. Jackson JD (1998) Classical electrodynamics, 3rd edn. Wiley
8. Karlsen KH, Lie K-A, Natvig JR, Nordhaug HF, Dahle HK (2001) Operator splitting methods for systems of convection-diffusion equations: nonlinear error mechanisms and correction strategies. J Comput Phys 173(2):636–663
9. McLachlan RI, Quispel GRW (2002) Splitting methods. Acta Numer 341–434
10. Sportisse B (2000) An analysis of operator splitting techniques in the stiff case. J Comput Phys 161:140–168
11. Stoer J, Bulirsch R (1980) Introduction to numerical analysis. Springer, New York
12. Strang G (1968) On the construction and comparison of difference schemes. SIAM J Numer Anal 5:506–517
13. Strang G, MacNamara S (2014) Functions of difference matrices are Toeplitz plus Hankel. SIAM Rev 56(3):525–546
14. Yoshida H (1990) Construction of higher order symplectic integrators. Phys Lett A 150(5, 6, 7)

A Theory and Calculation of Zero-Point Energy and Pressure, also Resulting in a New Value for the Hubble Constant

S. W. Mason

Abstract Within this publication, the Dirac equation is modified to take account of the electromagnetic field of sub-atomic particles, as well as their relativistic spatio-temporal spin states. This procedure therefore results in what turns out to be two asymmetric and different, positive and negative energies, rather than the equally sized identical quantities of opposite signs that Dirac obtained. The difference between each of these slightly different positive and negative energies thus ultimately yields a value that is very similar to that quoted for zero-point energy or dark energy, otherwise equally termed vacuum energy, and a final calculation for a predicted, outward-directed vacuum pressure of approximately -0.6×10^{-8} ergs/cm^2—possibly explaining the expansion of the Universe—is also made. Finally, using this paper's own theoretically calculated magnitude for a vacuum energy density of 0.5963×10^{-8} ergs/cm^3, it is shown that a possible value for the Hubble constant is $71.55587\,\text{km}/(\text{s Mpc})$.

Keywords Dirac equation · Modified Dirac equation · Pauli matrices · Zero-point energy · Dark energy · Vacuum energy · Vacuum pressure · Hubble constant

1 Introduction

Ever since its initial postulation in 1911, by Max Planck, zero-point energy, often termed dark energy or vacuum energy, has eluded our understanding of both its nature and existence, merely acting as little more than a place-holder for something else unknown that is clearly present. In order to attempt to ascertain more of an understanding of what this energy is, it therefore made sense to adopt the energy-yielding Dirac equation in order to try and not only prove its theoretical existence but also calculate its numerical value as well, regardless of the name attributed to it.

It is well known that the solution to the Dirac equation yields *two* energy values of identical magnitude, but of *opposite* signs, as most textbooks on the subject indicate.

S. W. Mason (✉)
Leeds, UK

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_30

This has apparently been somewhat of a conundrum for many people, although it is now recognised that the positive energy within the Universe is also counterbalanced by an equal and opposite negative energy. However, it is the contention of this paper that these two energies are not exactly equal, and that the negative energy resulting from the modified Dirac equation featured in this work is in fact slightly greater in magnitude than that of its positive counterpart. This results from an additional term for each particle's external electromagnetic field, defined by the Pauli matrix, σ_z, which thus affects their spin along the z-axis, as well as the presence of the term γ^μ, the latter of which affects the relativistic nature of such particles. In introducing the Pauli matrix, σ_z, as a further term added to γ^μ, it is thus shown that there are not only two slightly differing energy magnitudes of opposite sign, but that the residual of their sum falls within the experimental and theoretical limits for both zero-point-energy and dark energy, otherwise known as vacuum energy. Close to the end of this paper's findings, it is also suggested that this vacuum energy produces a negative energy pressure of -0.59522×10^{-8} ergs/cm^2, directed outwards from each face of a 1 cm^3 cube, due to the minus sign, which could potentially explain the reason behind the expansion of the Universe, should it turn out that this pressure is behind the force responsible for this expansion. Whilst this energy pressure, and the *residual energy density* of -0.59522×10^{-8} ergs/cm^3, are obtained by considering a combined quantity of 10^6 particles (involving equal numbers of electrons and protons) per cubic meter in interstellar space, it is further suggested that, given the claim that vacuum energy does not comprise ordinary matter, the scientific community might be further encouraged to search with even more vigour for the true nature of what this energy is, possibly spurred on by the values obtained here. At the very end of this paper, a further, final calculation is made to determine the theoretical value of the Hubble constant using the magnitude of the *residual energy density* of 0.5963×10^{-8} ergs/cm^3 obtained in Eq. (60), leading to $H = 71.55587$ km/(s Mpc) in Eq. (65). If the theory in this publication is correct, and if it can be utilised, then the scientific implications for Engineering, Information and Communications Technology (ICT), and other technologies are immense, along with allowing us to have a better Cosmological understanding of the Universe, its creation, and structure.

Manuscript Organisation: None. The author, S. W. Mason, is not a member of any university, other institution, organisation, or business, is not salaried, and works, voluntarily, from home for zero payment and profit.

2 Results and Discussion

Since previous research, such as [9], has shown that zero-point, dark, or vacuum energy exists, whilst simultaneously failing to identify what it actually is so far, it made sense to attempt to not only theoretically prove its existence but also to calculate its numerical value. It was hoped therefore that by adopting and solving a modified version of the Dirac equation, the original version of which yields two identical energies of opposite sign, at least one of the newly obtained energies might be that of the known value of zero-point energy, again otherwise known as vacuum or dark energy. This consequently set forth the motivation for the following theory and analysis, as illustrated in Fig. 1, in the hope of contributing to, and furthering, existing knowledge on this elusive subject.

It is suggested therefore that the Dirac equation

$$i\hbar\gamma^{\mu}\partial_{\mu}\psi(x) - mc\psi(x) = 0 \tag{1}$$

can be modified for an oscillating, 'hidden', *zero-point* energy field, by considering not only the *relativistic spatio-temporal spin states* of a sub-atomic particle, defined by γ^{μ}, but also the *influence that an external electromagnetic field* has on that particle's *spin*, defined by σ_z. This latter quantity is defined by the 2×2 Pauli Matrix along the z-axis, as

$$\sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{2}$$

where

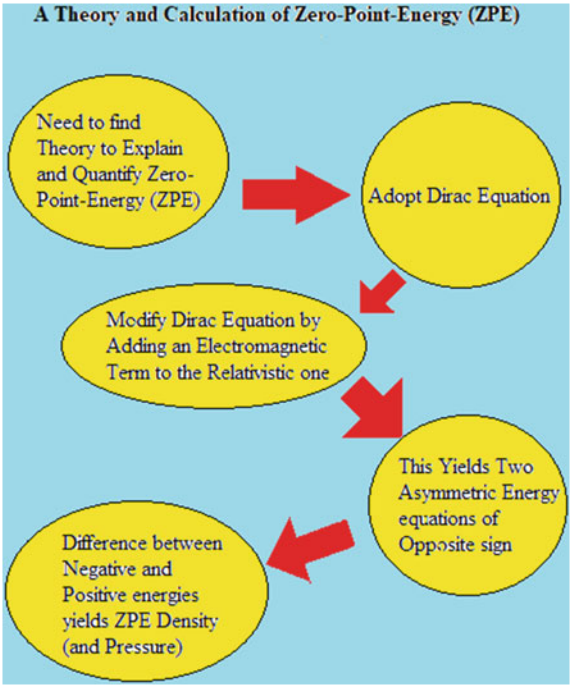


Fig. 1 Object process diagram/infographic illustrating the process by which the theory was developed resulting in the values obtained

$$\partial_\mu \psi(x) = \frac{\partial \psi(x)}{\partial x^\mu} \tag{3}$$

and the spinor, $\psi(x)$, is represented as

$$\psi(x) = \begin{bmatrix} \psi^1(x) \\ \psi^2(x) \\ \psi^3(x) \\ \psi^4(x) \end{bmatrix} \tag{4}$$

Also, $\gamma^\mu = \{\gamma^0,\ \gamma^1, \gamma^2, \gamma^3\}$, such that

$$\gamma^0 = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \tag{5}$$

and

$$\gamma^1 = \begin{bmatrix} 0 & \sigma_1 \\ -\sigma_1 & 0 \end{bmatrix},\ \gamma^2 = \begin{bmatrix} 0 & \sigma_2 \\ -\sigma_2 & 0 \end{bmatrix},\ \gamma^3 = \begin{bmatrix} 0 & \sigma_3 \\ -\sigma_3 & 0 \end{bmatrix} \tag{6}$$

In order to attempt to determine an expression for such a zero-point energy, we combine the *relativistic* nature of a particle's *spin*, γ^μ, by adding it to the *spin* defined by the Pauli Matrix under an *external electromagnetic field* along the z-axis, $\frac{1}{2}\sigma_z I$, to obtain $\left(\gamma^\mu + \frac{1}{2}\sigma_z I\right)$, where $\frac{1}{2}\sigma_z I$ is a two-component spinor, [8], and I is the identity matrix:

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{7}$$

We may therefore re-write the Dirac equation, Eq. (1), as

$$i\left(\gamma^\mu + \frac{1}{2}\sigma_z I\right)\partial_\mu \psi(x) - \frac{mc}{\hbar}\psi(x) = 0 \tag{8}$$

or, setting $\beta = \frac{mc}{\hbar}$ for greater simplicity of calculation, we have

$$i\left(\gamma^\mu + \frac{1}{2}\sigma_z I\right)\partial_\mu \psi(x) - \beta\psi(x) = 0 \tag{9}$$

Here, the factor of $\frac{1}{2}$ in the term $\frac{1}{2}\sigma_z I$, on the left-hand side of Eq. (9), represents the half-spin nature of each particle, [1, 8], aligned along the z-axis.

Since we can also write $\beta\psi(x)$ as $\beta I\psi(x)$, Eq. (9) now becomes

$$i\left(\gamma^{\mu}+\frac{1}{2}\sigma_z I\right)\partial_{\mu}\psi(x)-\beta I\psi(x)=0 \tag{10}$$

Rather than attempting to solve Eq. (10) in terms of a set of 4×4 matrices, we may instead simplify it by now converting it into a 2×2 component system by writing the spinor, $\psi(x)$, as

$$\psi(x)=\begin{bmatrix}\psi^1(x)\\ \psi^2(x)\\ \psi^3(x)\\ \psi^4(x)\end{bmatrix}=\begin{bmatrix}u\\ v\end{bmatrix} \tag{11}$$

and we note that the *Fourier transform* of the *Dirac field*, $\psi(k)$, is

$$\psi(x)=\int\frac{d^4k}{(2\pi)^4}\psi(k)e^{-ik_{\mu}x^{\mu}} \tag{12}$$

where k represents the momentum space into which we are hoping to transform the system and thus determine expressions for the energy of the particle or for a more general particle field.

First, expanding Eq. (10) to yield

$$i\gamma^{\mu}\partial_{\mu}\psi(x)-\beta I\psi(x)+\frac{1}{2}i\sigma_z I\partial_{\mu}\psi(x)=0 \tag{13}$$

we can re-write Eq. (13) in the following form:

$$i\gamma^{\mu}\partial_{\mu}\int\frac{d^4k}{(2\pi)^4}\psi(k)e^{-ik_{\mu}x^{\mu}}-\beta I\int\frac{d^4k}{(2\pi)^4}\psi(k)e^{-ik_{\mu}x^{\mu}} \\ +\frac{1}{2}i\sigma_z I\partial_{\mu}\int\frac{d^4k}{(2\pi)^4}\psi(k)e^{-ik_{\mu}x^{\mu}}=0 \tag{14}$$

Hence,

$$\int\frac{d^4k}{(2\pi)^4}\psi(k)(i\gamma^{\nu})\partial_{\nu}(e^{-ik_{\mu}x^{\mu}})-\int\frac{d^4k}{(2\pi)^4}\beta I\psi(k)e^{-ik_{\mu}x^{\mu}} \\ +\frac{1}{2}\int\frac{d^4k}{(2\pi)^4}\psi(k)(i\sigma_z I)\partial_{\nu}(e^{-ik_{\mu}x^{\mu}})=0 \tag{15}$$

Since the partial derivative component, $\partial_{\nu}(e^{-ik_{\mu}x^{\mu}})=-ik_{\nu}e^{-ik_{\mu}x^{\mu}}$, represents a *rotation* with respect to the coordinate axes (in which the 'coefficient' of the above derivative, $-ik_{\mu}$, actually becomes $-ik_{\nu}$), then we can re-write Eq. (15) as

$$\int \frac{d^4k}{(2\pi)^4}\psi(k)(i\gamma^{\nu})(-ik_{\nu})(e^{-ik_\mu x^\mu}) - \int \frac{d^4k}{(2\pi)^4}\beta I\psi(k)e^{-ik_\mu x^\mu} + \frac{1}{2}\int \frac{d^4k}{(2\pi)^4}\psi(k)(i\sigma_z I)(-ik_{\nu})(e^{-ik_\mu x^\mu}) = 0 \tag{16}$$

which simplifies to

$$\int \frac{d^4k}{(2\pi)^4}\psi(k)\gamma^{\nu}k_{\nu}e^{-ik_\mu x^\mu} - \int \frac{d^4k}{(2\pi)^4}\beta I\psi(k)e^{-ik_\mu x^\mu} + \frac{1}{2}\int \frac{d^4k}{(2\pi)^4}\psi(k)(k_{\nu}\sigma_z I)e^{-ik_\mu x^\mu} = 0 \tag{17}$$

Factorising out the integral, $\int \frac{d^4k}{(2\pi)^4}e^{-ik_\mu x^\mu}$, we get

$$\int \frac{d^4k}{(2\pi)^4}\left\{\gamma^{\nu}k_{\nu}\psi(k) - \beta I\psi(k) + \frac{1}{2}k_{\nu}\sigma_z I\psi(k)\right\}e^{-ik_\mu x^\mu} = 0 \tag{18}$$

Thus,

$$\gamma^{\nu}k_{\nu}\psi(k) - \beta I\psi(k) + \frac{1}{2}k_{\nu}\sigma_z I\psi(k) = 0 \tag{19}$$

Since I is the 2×2 identity matrix for each half-spin state, and since $\gamma^{\mu}k_{\mu} = \begin{bmatrix} k_0 & k_\mu\sigma_j \\ -k_\mu\sigma_j & -k_0 \end{bmatrix}$, indicating that $\gamma^{\nu}k_{\nu} = \begin{bmatrix} k_0 & -k_\nu\sigma_j \\ k_\nu\sigma_j & -k_0 \end{bmatrix}$ for a rotation, we can rewrite Eq. (19) in the matrix form:

$$\begin{bmatrix} k_0 & -k_\nu\sigma_j \\ k_\nu\sigma_j & -k_0 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} - \beta\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \frac{1}{2}k_\nu\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = 0 \tag{20}$$

so that

$$\begin{bmatrix} k_0 & -k_\nu\sigma_j \\ k_\nu\sigma_j & -k_0 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} -\beta & 0 \\ 0 & -\beta \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} \frac{1}{2}k_\nu & 0 \\ 0 & -\frac{1}{2}k_\nu \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = 0 \tag{21}$$

Henceforth, we have

$$\begin{bmatrix} k_0 + \frac{1}{2}k_\nu - \beta & -k_\nu\sigma_j \\ k_\nu\sigma_j & -k_0 - \frac{1}{2}k_\nu - \beta \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = 0 \tag{22}$$

or

$$\begin{bmatrix} k_0 + \frac{1}{2}k_\nu - \beta & -k_v\sigma_j \\ k_v\sigma_j & -(k_0 + \frac{1}{2}k_\nu + \beta) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = 0 \tag{23}$$

Temporarily letting $a = k_0 + \frac{1}{2}k_\nu$ and defining the resulting matrix as

$$K = \begin{bmatrix} a - \beta & -k_v\sigma_j \\ k_v\sigma_j & -(a + \beta) \end{bmatrix} \tag{24}$$

such that

$$\begin{bmatrix} a - \beta & -k_v\sigma_j \\ k_v\sigma_j & -(a + \beta) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = 0 \tag{25}$$

where

$$\det(K) = \det\begin{bmatrix} a - \beta & -k_v\sigma_j \\ k_v\sigma_j & -(a + \beta) \end{bmatrix} = -(a - \beta)(a + \beta) + \left(k_v\sigma_j\right)^2 \tag{26}$$

we find that

$$\det(K) = -a^2 + \beta^2 + \left(k_v\sigma_j\right)^2 \tag{27}$$

Since $\sigma_j^2 = I$, we therefore know that $\left(k_v\sigma_j\right)^2 = k_v^2$. Thus, resubstituting $a = k_0 + \frac{1}{2}k_\nu$, we may now write

$$\det(K) = -\left(k_0 + \frac{1}{2}k_\nu\right)^2 + \beta^2 + k_v^2 \tag{28}$$

But, for a solution to Eq. (28) to exist,

$$\det(K) = 0 \tag{29}$$

So

$$-\left(k_0 + \frac{1}{2}k_\nu\right)^2 + \beta^2 + k_v^2 = 0 \tag{30}$$

or

$$\left(k_0 + \frac{1}{2}k_\nu\right)^2 = \beta^2 + k_v^2 \tag{31}$$

such that

$$k_0 + \frac{1}{2}k_\nu = \pm\sqrt{\beta^2 + k_\nu^2} \tag{32}$$

Therefore,

$$k_0 = -\frac{1}{2}k_\nu \pm \sqrt{\beta^2 + k_\nu^2} \tag{33}$$

Resubstituting $\beta = \frac{mc}{\hbar}$ into Eq. (33), and then multiplying both sides of it by c, we have

$$k_0 c = -\frac{1}{2}k_\nu c \pm c\sqrt{\left(\frac{mc}{\hbar}\right)^2 + k_\nu^2} \tag{34}$$

and hence

$$\hbar k_0 c = -\frac{1}{2}\hbar k_\nu c \pm \sqrt{\left(mc^2\right)^2 + (\hbar k_\nu c)^2} \tag{35}$$

Since $p = \hbar k$, where p represents the actual momentum, and $E_0 = \hbar k_0 c$ represents the energy, then

$$E_0 = -\frac{1}{2}pc \pm \sqrt{\left(mc^2\right)^2 + (pc)^2} \tag{36}$$

This means that there are two energies, which are

$$E_{+ve} = -\frac{1}{2}pc + \sqrt{\left(mc^2\right)^2 + (pc)^2} \tag{37}$$

and

$$E_{-ve} = -\frac{1}{2}pc - \sqrt{\left(mc^2\right)^2 + (pc)^2} \tag{38}$$

meaning that

$$E_{-ve} < 0 < E_{+ve} \tag{39}$$

Now, the magnitudes of each energy are such that

$$|E_{+ve}| < |E_{-ve}| \tag{40}$$

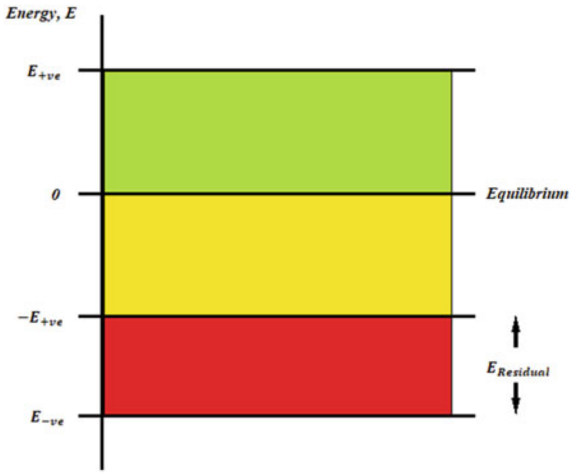


Fig. 2 The residual energy, $E_{\mathrm{Residual}} = -(|E_{-ve}| - |E_{+ve}|)$, is the section in red illustrating the negative difference between E_{-ve} and E_{+ve}

so that the magnitude of the negative energy

$$E_{-ve} = -\frac{1}{2}pc - \sqrt{\left(mc^2\right)^2 + (pc)^2} \tag{41}$$

prevails over the positive energy, $E_{+ve} = -\frac{1}{2}pc + \sqrt{\left(mc^2\right)^2 + (pc)^2}$, by way of a resultant negative energy that would pervade the very fabric of the cosmos, as Fig. 2 illustrates. This energy could thus be responsible for the dark energy that is believed to be potentially causing the accelerated expansion of the Universe.

In the first instance, we will assume that space is pervaded mainly by electrons although, later, we will also consider protons as well. To calculate the value of zero-point energy, or even dark energy, we first substitute the relevant values for the electron into Eqs. (37) and (38) and calculate the difference, where the mass of the electron is $m_e = 9.1093897 \times 10^{-31}$ kg and the speed of light, $c = 299792458$ m/s. The de Broglie momentum of the electron, p_e, is then calculated from the relationship:

$$p_e = m_e v_e = m_e \alpha c \tag{42}$$

where v_e is the electron's velocity and α is the fine structure constant equal to 7.297351×10^{-3}. Thus, the electron's zero-point momentum, p_e, is

$$p_e \cong 1.9929 \times 10^{-24}\,\mathrm{kg\,ms^{-1}} \tag{43}$$

Substituting the above momentum and the previously listed values into Eq. (36), we obtain

$$E = -\frac{5.9743 \times 10^{-16}}{2} \pm \sqrt{\left(8.1871 \times 10^{-14}\right)^2 + \left(5.9743 \times 10^{-16}\right)^2} \qquad (44)$$

or

$$E = -2.9872 \times 10^{-16} \pm \sqrt{\left(8.1871 \times 10^{-14}\right)^2 + \left(5.9743 \times 10^{-16}\right)^2} \qquad (45)$$

where all displayed values are written to five significant figures for brevity and clarity.

We thus see that Eq. (45) yields the two values (in this case quoted to a maximum of 10 significant figures for the purposes of deducing a very small difference) of

$$E_{+ve} = 8.15745704 \times 10^{-14}\,\mathrm{J} \qquad (46)$$

per electron, and

$$E_{-ve} = -8.217201264 \times 10^{-14}\,\mathrm{J} \qquad (47)$$

per electron.

The residual energy that prevails is therefore

$$E_{\mathrm{Residual}} = -(|E_{-ve}| - |E_{+ve}|) \qquad (48)$$

such that

$$E_{\mathrm{Residual}} = -\left(8.217199475 \times 10^{-14} - 8.157456475 \times 10^{-14}\right) \qquad (49)$$

leading to

$$E_{\mathrm{Residual}} = -0.59744 \times 10^{-15}\,\mathrm{J} \qquad (50)$$

per electron, again to five significant figures.

If we take there to be an *average* of 1×10^6 separate particles per cubic metre in interstellar space [3–6, 10, 11], and *initially assume* that each of these particles is an electron, then the whole, total residual energy for all 10^6 electrons/m^3 is

$$E_{\mathrm{Residual}} = -0.59744 \times 10^{-15} \times 1 \times 10^6\,\mathrm{J/m^3} \qquad (51)$$

or

$$E_{\mathrm{Residual}} = -0.59744 \times 10^{-9}\,\mathrm{J/m^3} \qquad (52)$$

Alternatively, the result in Eq. (52) may equally be written as

$$E_{\text{Residual}} = -0.59744 \times 10^{-15}\,\text{J/cm}^3 \tag{53}$$

or, since 1 erg $= 10^{-7}$ Joules , and 1 m$^3 = 10^6$ cm^3, we may write the above residual energy in the form:

$$E_{Residual} = -0.59744 \times 10^{-8}\,\text{ergs/cm}^3 \tag{54}$$

However, as implied above, this does assume that *only* electrons are present in the cosmos, not protons, so a further calculation will be made in *a More Realistic Scenario*, below, in which the ground state of the composite hydrogen atom is considered, together with equal numbers of electrons and protons post ionisation.

Because of the negative sign associated with this final, theoretically calculated value of zero-point energy in Eq. (54), the implication is that the zero-point-energy field is thus producing a surplus of *fermions*, such as electrons and protons, or at least something that is mimicking them. This could in fact explain the presence of the mysterious *dark matter*, gravitationally responsible for the faster than expected stellar velocities with which stars orbit their galactic centres, predicted by scientists to exist close to this value [2].

As an indication of how accurate Eq. (54) is, we note that Beck and Mackey, [2], experimentally verified the magnitude of dark energy to be $\rho_{\text{dark}} = (3.9 \pm 0.4)\,\text{GeV/m}^3$ which, after conversion, becomes an average of $\rho_{\text{dark}} = 0.62485 \times 10^{-8}\,\text{ergs/cm}^3$. Gründler [7] also calculates a vacuum energy density value of $5.4 \times 10^{-10}\,\text{J/m}^3$, or $0.54 \times 10^{-8}\,\text{ergs/cm}^3$. It is also well known in Cosmology that dark energy is often measured in terms of mass per cubic centimetre, at 7×10^{-30} g/cm^3, or 7×10^{-33} kg/cm^3, as highlighted by Steinhardt and Turok, [12], and, given that Einstein's energy equation is $E = mc^2$, we see that this latter mass density of 7×10^{-33} kg/cm^3 easily converts into $0.6291 \times 10^{-15}\,\text{J/cm}^3$, otherwise written as $0.6291 \times 10^{-8}\,\text{ergs/cm}^3$. This value is very close to Beck and Mackey's [2], along with the magnitude of the value obtained in this research, if we assume that the number of combined but separate interstellar particles per cubic metre of 10^6 is adopted.

In particular, since the *maximum* of this energy density, as Beck and Mackey [2] state, is $4.3\,\text{GeV/m}^3$, which, by the same method of conversion, equates to $\rho_{Max} = 0.68894 \times 10^{-8}\,\text{ergs/cm}^3$, with the *minimum* energy being $3.5\,\text{GeV/m}^3$, or $\rho_{Min} = 0.560762 \times 10^{-8}\,\text{ergs/cm}^3$, then the *average value* of the cosmic dark energy density is $\bar{\rho} = \frac{1}{2}\left(0.68894 \times 10^{-8} + 0.560762 \times 10^{-8}\right)\text{ergs/cm}^3 = 0.62489 \times 10^{-8}\,\text{ergs/cm}^3$. Thus, the upshot of this would therefore seem to vindicate the theory in this publication, and vice versa, in which the *upper* and *lower* bounds of the dark energy calculated by Beck and Mackey [2] seem to very accurately prove correct the magnitude of the zero-point energy (or dark or vacuum energy) calculated in Eq. (54), namely $0.59744 \times 10^{-8}\,\text{ergs/cm}^3$, which lies perfectly within the stated limits of $(3.9 \pm 0.4)\,\text{GeV/m}^3$.

3 A More Realistic Scenario: Considering 5×10^5 Free Electrons Per m^3 and 5×10^5 Free Protons Per m^3, *Separately*, Totalling 10^6 particles/m^3 in All

A *word of caution* here indicates that since we have already established that there are 10^6 individual *particles* per cubic meter (1 particle per cm^3) of interstellar space, in total, [3–6, 10, 11], then we might also perhaps more wisely, and more realistically, assume that there are approximately 5×10^5 free electrons and 5×10^5 free protons present, separately, post ionisation of each individual hydrogen atom. We note here that although 5×10^5 free electrons and 5×10^5 free protons are actually equal in sum to 10^6 separate particles, post ionisation, on recombining, they actually form 5×10^5 hydrogen atoms, although we could make the assumption that there are also 10^6 hydrogen atoms per cubic metre in certain regions of space as well. Re-working Eq. (50) for only 5×10^5 electrons/m^3 yields

$$E_{\text{Electron}} = -0.59744 \times 10^{-15} \times 5 \times 10^5 \text{ J/m}^3 \tag{55}$$

or

$$E_{\text{Electron}} = -0.2987 \times 10^{-15} \text{ J/cm}^3 \tag{56}$$

leading to

$$E_{\text{Electron}} = -0.2987 \times 10^{-8} \text{ ergs/cm}^3 \tag{57}$$

for any cubic centimetre of interstellar space solely containing *electrons*.

This means that, for the *proton*, the difference between the *hydrogen's* ground state of -13.6 eV, or -2.17896×10^{-18} J, which equates to $-0.108948 \times 10^{-10}$ ergs/cm^3 for an *assumed* total of 5×10^5 hydrogen atoms per cubic metre (as opposed to 10^6 atoms/m^3), and the zero-point energy of the electron, -0.2987×10^{-8} ergs/cm^3, as deduced in Eq. (57), after the electron has been ejected from the hydrogen atom, is therefore

$$E_{\text{Proton}} = -\left(0.2987 \times 10^{-8} - 0.1089 \times 10^{-10}\right) \text{ergs/cm}^3 \tag{58}$$

or

$$E_{\text{Proton}} = -0.2976 \times 10^{-8} \text{ ergs/cm}^3 \tag{59}$$

Consequently, the total residual energy is possibly more accurately defined by the sum of Eqs. (57) and (59) as

$$E_{\text{Residual}} = -0.5963 \times 10^{-8}\,\text{ergs/cm}^3 \tag{60}$$

which not only justifies the result in Eq. (54) but also is again similarly consistent with the total zero-point energy, or dark energy, otherwise known as vacuum energy, as illustrated by Beck, Mackey, Steinhart, and Turok, [2, 12], contained within the Universe.

Using this perhaps more accurate result, if we assume that from the particle density of 10^6 particles per cubic meter, or 1 particle/cm^3, we can thus determine that there exists 1 particle per unit length of the cube (and hence 1 particle incident upon every square centimetre of its faces), and further assume that the overall pressure of zero-point energy, or dark energy, per unit area of each side of the cube is therefore

$$E_{\text{Pressure}} = -0.5963 \times 10^{-8}\,\text{ergs/cm}^2 \tag{61}$$

which is directed outwards, from the *area* of *each side* of every *cubic centimetre volume* of space, into the cosmos, as suggested by the negative sign in Eq. (61). One can therefore assert that it is this *constant vacuum energy pressure* of -0.5963×10^{-8} ergs/cm^2 that is always *exerting a force* upon the fabric of the space–time continuum, which is responsible for causing the Universe to expand. The fact that this appears to be due purely to free electrons and protons of regular matter, when dark energy is supposed to consist of matter that is, as yet, undetermined, is a conundrum that needs further scientific investigation. A suggestion would be that for every identical quantity of free electrons and protons of regular matter that exists in interstellar space, there is an equal number of sister particles—comprising dark matter—which are responsible for this dark energy reservoir and pressure thus resulting in the experimentally observed rate of cosmological expansion, and which could also have been the initial origin of the Big Bang itself.

4 Calculation of the Hubble Constant

If one adopts the formula quoted by Gründler [7], which is

$$\tau_{00}^{\text{vacuum}} = \frac{3H^2\Omega_\Lambda c^2}{8\pi G} \tag{62}$$

we can calculate a theoretical value for the Hubble constant, H, using Eq. (60). Rearranging and taking the square root of each side of Eq. (62), we have

$$H = \sqrt{\frac{8\pi G\tau_{00}^{\text{vacuum}}}{3\Omega_\Lambda c^2}} \tag{63}$$

where we take $\tau_{00}^{\text{vacuum}} = |E_{Residual}| = 0.5963 \times 10^{-8}\,\text{ergs/cm}^3 = 5.963 \times 10^{-10}\,\text{J/m}^3$, Ω_Λ = dark energy parameter = 0.69, c = speed - of - light in vacuo = 299792458 m/s, and G = Gravitational constant = $6.6743 \times 10^{-11}\,\text{m}^3\,\text{kg}^{-1}\,\text{s}^{-2}$.

Substituting each of the above values into Eq. (63), we obtain

$$H = 2.31873 \times 10^{-18}\,\text{s}^{-1} \tag{64}$$

However, since 3.086×10^{19} km is equal to one megaparsec (1 Mpc), we see that $2.31873 \times 10^{-18}\,\text{s}^{-1} = 71.55587\,\text{s}^{-1}\,\text{km}/(3.086 \times 10^{19}\,\text{km}) = 71.55587\,\text{s}^{-1}\,\text{km/(Mpc)}$, or

$$H = 71.55587\,\text{km/(s Mpc)} \tag{65}$$

the latter of which represents the more conventional *rate of expansion* of the Universe.

5 Conclusion

Within the theory laid out in this publication, the Dirac equation has been modified, thus resulting in two asymmetric positive and negative energies of differing magnitudes, respectively. This was achieved by adding the spinor, $\frac{1}{2}\sigma_z I$, representing the spin due to an external electromagnetic field, to the relativistic term, γ^μ, and solving this modified Dirac equation from that point onwards. The two energies that result, of the form $E_0 = -\frac{1}{2}pc \pm \sqrt{\left(mc^2\right)^2 + (pc)^2}$, such that the magnitude of the negative energy is larger than the magnitude of the positive one, tell us that there is a residual, negative energy that offsets the balance of what would normally be a state of equilibrium between the otherwise equal and identical positive and negative energies that result from the more regular solution of the *un-modified* Dirac equation. In short, when the *difference* between the *negative* energy and the smaller, *positive* one is calculated, the resulting zero-point energy, dark energy, or vacuum energy, density comes out to be approximately $-0.6 \times 10^{-8}\,\text{ergs/cm}^3$, with a theorised pressure of $-0.6 \times 10^{-8}\,\text{ergs/cm}^2$ upon each *interior* face of a cubic centimetre of particles. It is therefore suggested that, if this latter value is accurate, then the *force* due to the sub-atomic *pressure* created by this *outward seeking energy* could be responsible for the expansion of the Universe, as observed by astronomers and cosmologists. This is because the residual, *surplus* vacuum energy calculated in this paper appears to be a *constantly* produced quantity—directly in contravention of the law of conservation of energy—since it is always being created from the stated asymmetric difference between positive and negative energies within each 1 cm^3 volume of interstellar space. Finally, this theoretically and experimentally verified excess of vacuum energy could also have been responsible for the creation of the Universe in the first place, if we assume that it was present at the time at which the Big Bang first occurred 13.9 billion years ago, and the magnitude of the theoretical value obtained in this paper leads directly to a Hubble constant of 71.55587 km/(s Mpc).

As suggested at the end of Introduction in Sect. 1, if the theory presented here is correct, and if it can be utilised in scientifically practical ways, then the potential Engineering, ICT, and other technological advancements that could be made are huge, together with providing mankind with a fuller understanding of the creation, structure, existence, and life span, of our Universe.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. I am an independent researcher, without any affiliation to any organisation or institution. I am the sole author of this work.

Declarations of Interest None.

References

1. Ahluwalia-Khalilova DV, Grumiller D (2005) Dark matter: a spin one-half fermion field with mass dimension one? Phys Rev D 72:067701-1–067701-4
2. Beck C, Mackey MC (2005) Could dark energy be measured in the lab? Phys Lett B 605:295–300
3. Bohlin RC, Savage BD, Drake JF (1977) A survey of interstellar HI from Lα absorption measurements II, NASA (NASA-TM-78044) pre-print, Unclas, 55994, pp 1–25
4. Boulares A, Cox DP (1990) Galactic hydrostatic equilibrium with magnetic tension and cosmic-ray diffusion. Astrophys J 365:544–558
5. European Space Agency, ESAR (2019) The interstellar medium, CESAR's booklet, pp 1–6. http://cesar.esa.int/upload/201807/ism_booklet.pdf
6. Ferrière KM (2001) The interstellar environment of our galaxy. Rev Mod Phys 73:1031–1066
7. Gründler G (2017) The zero-point energy of elementary quantum fields. Astrophysical Institute Neunhof, Nürnberg, Germany. arXiv:1711.03877v1 [physics.gen-ph]
8. Ning W (2006) Non-relativistic limit of Dirac equations in gravitational field and quantum effects of gravity. Commun Theor Phys (Beijing, China) 45:452–456
9. Perlmutter S (2003) Supernovae, dark energy, and the accelerating universe. Phys Today 56(4):53. https://doi.org/10.1063/1.1580050
10. Ridpath I (1997) Oxford dictionary of astronomy. Oxford University Press, p 240 (Entry: Interstellar Matter, ISM)
11. Spitzer L Jr (1942) The dynamics of the interstellar medium, III. Galactic distribution. Astrophys J 95(3):329–344
12. Steinhardt PJ, Turok N (2006) Why the cosmological constant is small and positive. Science 312:1180–1183

Implementation-Free Forensic Watermarking for Adaptive Streaming with A/B Watermarking

Hannes Mareen, Glenn Van Wallendael, and Peter Lambert

Abstract Forensic watermarking enables the identification of digital pirates after they illegally re-distribute copyright-protected videos. For adaptive streaming, these methods are best combined with A/B watermarking, in which two watermarked versions are created for each video segment, and subsequently mixed in order to create a large number of uniquely watermarked videos. Although good video quality and low bitrate are key characteristics of a good watermarking system, existing methods objectively lower the compression efficiency. Additionally, they often require complex implementations. Therefore, this paper proposes an implementation-free, rate-distortion-preserving watermarking technique to be used with the scalable A/B watermarking concept. Even though the embedding is performed during compression, it does not change the existing video encoder implementation. Instead, it only changes the target bitrate parameter in order to create different compression artifacts. These artifacts represent the watermark but are not noticeable due to high-quality coding. As such, the rate-distortion performance is nearly equal to that of ordinary, unwatermarked compression (i.e., a BD-rate of 0.02% and −0.10% when applied with H.264/AVC and H.265/HEVC encoders, respectively). Furthermore, the robustness is equal or better than state-of-the-art methods with comparable embedding complexities. More specifically, in case of recompression attacks, nonzero false negative rates are only reported when a watermarked video is initially compressed with a high quality and degraded to a very low quality. Consequently, the proposed scheme can be used in practice by adaptive-streaming platforms without a quality decrease, bitrate increase, or implementation overhead.

H. Mareen (✉) · G. Van Wallendael · P. Lambert
IDLab, Department of Electronics and Information Systems, Ghent University – imec, Ghent, Belgium
e-mail: hannes.mareen@ugent.be
URL: http://media.idlab.ugent.be/

G. Van Wallendael
e-mail: glenn.vanwallendael@ugent.be

P. Lambert
e-mail: peter.lambert@ugent.be

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_31

Keywords Forensic watermarking · Compression efficiency · Imperceptibility · Rate-distortion performance · Implementation free

1 Introduction

Digital piracy is a common problem for copyright holders that distribute their content on video-on-demand platforms. In order to prevent direct access to the streamed media by clients, those platforms often apply encryption-based Digital Rights Management (DRM) security measures. However, these tools require extra resources on the clients' devices and do not fully prevent content from being pirated after its decryption [1]. Therefore, forensic watermarking methods are used to enable the identification of digital pirates after illegal video redistribution, i.e., to perform traitor tracing or active fingerprinting. That is, every user receives a uniquely watermarked version of the protected video, i.e., the watermarked video is linked to the receiver's identifier (ID). When a malicious user leaks their version of the video, the watermark associated to the culprit can be detected.

Video-on-demand applications typically serve a large number of users. In order to distribute videos to many users with varying internet and device conditions, adaptive-streaming protocols are typically used. More specifically, client devices dynamically react to varying conditions by streaming content representations with other bitrates and resolutions. In typical adaptive-streaming applications, the streamed videos are encoded once and then distributed to a large number of users. Hence, the visual quality and bitrate are of utmost importance. For example, popular streaming platforms were downgrading the video bitrates by 25% to avoid network congestion in March 2020, when Europe went into lockdown due to the coronavirus outbreak [2]. Moreover, because the users may stream videos from a large variety of devices and under different conditions, the videos are accessible in a variety of content representations, potentially compressed with different video encoders. For example, old devices can be limited to the older, low-complexity H.264/Advanced Video Coding (AVC) standard, whereas modern devices can use the newer, more efficient H.265/High-Efficiency Video Coding (HEVC) standard.

Watermarking in adaptive-streaming applications is not straightforward because the server must provide each user with a uniquely watermarked video. This is not practical when the number of users is large because the watermarking algorithm is too complex for real-time embedding, or because it requires an unpractical increase in storage space. Moreover, it disrupts file caching mechanisms in the Content Delivery Network (CDN). A solution is to distribute the same video to all users and perform watermark embedding at the client device [3, 4]. However, this requires extra DRM software to be installed on the client devices, as well as extra client-side resources to embed the watermark in real time during video decoding. Additionally, it decreases the security of the system since malicious users may get access to the unwatermarked video that is temporarily stored on their device. As an alternative, the A/B watermarking or two-step watermarking framework is often applied, in which only a small

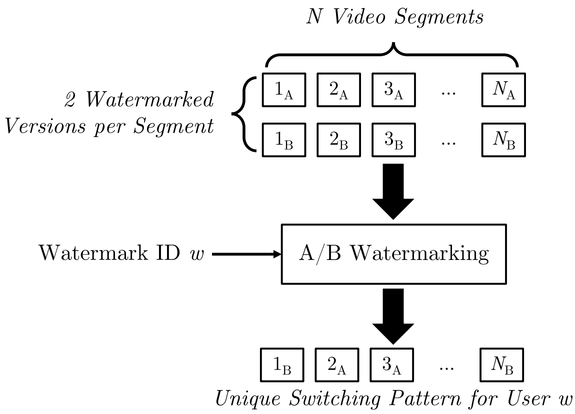


Fig. 1 Example of A/B watermarking: only two watermarked versions are created for each of the N video segments. By mixing these segments according to a watermark ID w, every user receives a unique switching pattern as a watermarked video

number of watermarked versions are created for every short video segment, using an existing watermarking method [1]. In other words, every segment is not only available in various content representations (e.g., bitrates, resolutions) but also in different watermarked versions. For example, only two watermarked variants (i.e., version A and version B) can be created for each segment. Before sending the video to a user, the watermarked segments are mixed to create a unique switching pattern that identifies the user. This example is illustrated in Fig. 1. As such, A/B watermarking is routinely used in adaptive bitrate applications where the number of users is high [5].

In the last two decades, many watermarking techniques have been created that can be used in combination with the A/B watermarking concept [6, 7]. An important requirement of these techniques is that they are imperceptible while not causing a significant bitrate overhead. In other words, the compression efficiency or rate-distortion (RD) performance may not be influenced significantly. Moreover, the watermarks should be robust to manipulations that try to destroy them. Since a more perceptible watermark is usually harder to delete, there is a trade-off between imperceptibility and robustness. For example, video encoders are made to remove imperceptible information in order to achieve stronger compression. Thus, video compression subsequent to watermarking is already an (unintentional) attack on the security system.

In order to balance the two requirements, watermarks are often embedded in a transform domain because they have interesting properties regarding perceptibility and robustness. For example, the discrete cosine transform (DCT) and wavelet transform domains are commonly used [8–14]. That is because the human visual system is more sensitive to low-frequency DCT coefficients changes, such that a watermark embedded in those coefficients is perceptible, but hard to remove without decreasing the visual quality even more. In other words, the trade-off between robustness and imperceptibility is controlled by changing either low- or high-frequency coefficients. In order to further balance imperceptibility and robustness, visual masking is commonly applied during embedding [11, 15, 16]. In this technique, the watermark is masked such that it is more strongly embedded in regions with low noise sensitivity than in those where it is more perceptible. Most importantly, although existing techniques are relatively imperceptible, they still objectively decrease the visual quality and/or increase the bitrate of the compressed video.

In an effort to minimize both the bitrate overhead and quality decrease, our previous work proposed a forensic watermarking method that does not affect the RD performance [17]. That is, the watermarking step is incorporated in the video compression process by varying the Quantization Parameters (QPs) of Coding Transform Units (CTUs) during encoding. Because the rest of the video encoder is left unchanged, the compression efficiency is preserved. Although this effectively minimizes the RD overhead, it requires the video encoder implementation to be adapted, which is not always possible or practical (e.g., due to proprietary software). Additionally, it is expensive when multiple video codecs are used, which is often the case in adaptive-streaming applications.

As a solution, this paper proposes a forensic watermarking method that preserves the RD performance and that does not change the existing video encoder implementation. Instead, only the input parameters of the video encoder are changed in order to create a watermarked video. More specifically, the target bitrate is changed. Because of the flexibility of modern video encoders and their rate-control algorithms, a different target bitrate results in different coding decisions, which in their turn result in different compression artifacts. Although these artifacts are different, the RD performance of the compressed watermarked videos is quasi-equal.

The remainder of this paper is structured as follows. First, Sect. 2 summarizes background information of video coding standards. Then, the forensic watermarking method is proposed in Sect. 3. Subsequently, Sect. 4 discusses the experimental results. Finally, the paper is concluded in Sect. 5.

2 Video Coding

Video compression standards such as H.264/AVC, H.265/HEVC, and the recently standardized H.266/Versatile Video Coding (VVC) transform an uncompressed video into a compressed bitstream. The compressed video consists of two main components: the coding information and the residual signal. The coding information involves the block and prediction structure. For example, a frame is typically partitioned into blocks of pixels of various sizes, and every block is predicted using motion vectors and intra-prediction modes. Usually, this prediction is not perfect. Therefore, the residual signal corrects the prediction errors. The residual signal is transformed and quantized with a configurable QP, which results in less bits to store, but also introduces compression artifacts. The QP is typically determined automatically by a rate-control algorithm or varied periodically by a Constant Rate Factor (CRF).

Modern video codecs have a large number of flexible tools and potential coding decisions, which rate-control algorithms use to achieve a target bitrate [18]. The QP is the most effective coding decision to be varied by rate-control algorithms, as it controls the granularity of transformed coefficients of the residual signal. Moreover, they can also change (the granularity of) other coding decisions such as block partitioning and motion vectors. In adaptive-streaming applications, this is typically done using a two-pass encoding [19]. In a first pass, the content is analyzed and the result is stored in a log file. Then, in the second pass, these results are used to achieve good encoding quality as well as accurate rate allocation.

3 Proposed Watermarking Method

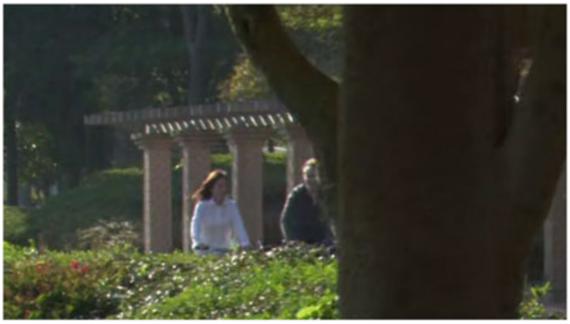
The proposed forensic watermarking method is created to be used in combination with A/B watermarking. Thus, only a few watermarked versions are created for each video segment. First, the embedding algorithm is described in Sect. 3.1, followed by the detection method in Sect. 3.2.

3.1 Watermark Embedding

The watermark is embedded during video compression, but the proposed algorithm does not require changes to the video encoder. Instead, it only changes the video encoder's input parameter that sets the target bitrate. More specifically, in order to embed watermark ID w, the default target bitrate t is changed to $t + w$ kilobit per second (kbit/s), where w is a signed integer.

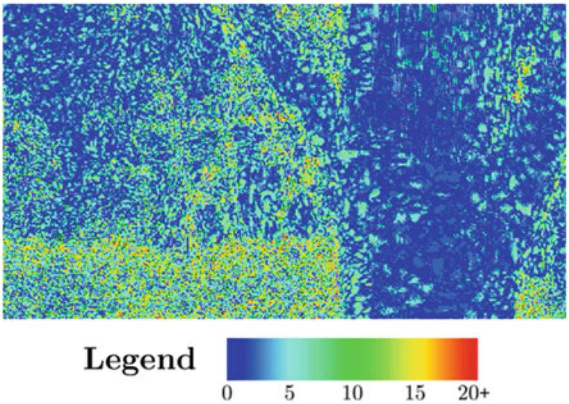
Changing the target bitrate has an effect on coding decisions made by the video encoder's rate-control algorithm, as explained in Sect. 2. Changing coding decisions such as the QP introduces different compression artifacts, which are used as a watermark representation. This effect is illustrated in Fig. 2. More specifically, Fig. 2a shows a crop of a compressed frame of the *ParkScene*-sequence, using a two-pass encoding with CRF $= 27$ and a target bitrate of $t = 3600$ kbit/s, using the *x265*-encoder. Figure 2b shows the same crop, watermarked and compressed using the same parameters, except for the target bitrate which was changed with 1 kbit/s, to $t - 1 = 3599$ kbit/s. Although both versions are encoded in a similar way and have a near-equal RD performance, there are many differences between them, as visualized in Fig. 2c.

In our proposed method, the target bitrates of each two watermarked videos differ minimally with 1 kbit/s. This value is sufficiently large to cause the creation of different compression artifacts in typical rate-control algorithms, as validated in the experiments in Sect. 4.3. Additionally, when the watermark ID is a small number, the difference in target bitrate is low, such that the output quality and bitrate are not changed significantly. In any case, the RD performance is preserved because the video compression operates as usual. Moreover, since A/B watermarking requires only few watermarked versions for each segment, the watermark ID always remains small.



(a) Compressed (3600 kbit/s).

(b) Watermarked (3599 kbit/s).



(c) Visualization of absolute pixel differences.

Fig. 2 Cropped example of **a** a compressed frame, **b** a watermarked frame, and **c** the differences between the compressed and watermarked frames. The crop has a resolution of 500 × 280 pixels, and a pixel is represented by an 8-bit integer between 0 and 255. Although the watermarked and compressed frames have an approximately equal quality and bitrate, they exhibit many differences

The embedding complexity is similar to other uncompressed domain watermarking algorithms and methods that perform watermarking during video compression. That is, they perform a separate video encoding for each watermarked version of the video. Since the proposed method should be used in combination with A/B watermarking, this high embedding complexity is not a problem, since only few watermarked versions are created for each video segment.

3.2 Watermark Detection

The compression artifacts that represent the watermark are detected by using a correlation-based technique [17, 20, 21]. More specifically, the observed video $\boldsymbol{o}$ is first compared to all distributed watermarked versions $\boldsymbol{w_i}, i \in \{1, 2, \ldots, N\}$ by calculating the root mean squared error (RMSE). It should be noted that the number of distributed watermarked versions N is very low when used with A/B watermarking. The RMSE is defined in (1), in which p represents the number of pixels of the video.

$$\mathrm{RMSE}(\boldsymbol{o}, \boldsymbol{w_i}) = \frac{1}{p}\sqrt{\sum_{j}^{p}(\boldsymbol{o}_j - \boldsymbol{w}_{i,j})^2} \tag{1}$$

When the observed video does not contain any additional distortions, the RMSE corresponding to one of the watermarked videos is zero, whereas it is nonzero for the other watermarked videos. Nevertheless, when attacks are performed prior to leaking the video, the RMSEs corresponding to all videos are nonzero, due to the distortions introduced by the attack. Because attacks increase the RMSE, it is not possible to perform robust watermark detection by simply comparing the RMSE value to a threshold.

Although a single RMSE value cannot indicate watermark presence, the collection of RMSE values of all watermarked videos can. More specifically, the RMSE value corresponding to the present watermarks is significantly smaller than the values from absent watermarks. For this reason, we convert every RMSE value into a z-score, which represents the number of standard deviations that an RMSE value differs from the mean of the distribution of RMSE values corresponding to absent watermarks. To put it in a simpler way, it performs outlier detection [22, 23]. The z-score is defined in (2). In the equation, r_i is the RMSE value corresponding to watermark w_i, and μ_A and σ_A are the mean and standard deviation of the distribution of RMSE values corresponding to absent watermarks, respectively. For simplicity, we assume that all watermarks are absent except for the one corresponding to the lowest RMSE value. Alternatively, the RMSE values corresponding to watermarks that were never distributed can be used as an estimate for all absent watermarks.

$$z(r_i, \mu_\mathrm{A}, \sigma_\mathrm{A}) = \frac{r_i - \mu_\mathrm{A}}{\sigma_\mathrm{A}} \tag{2}$$

The z-score calculation normalizes the RMSE values corresponding to absent watermarks, such that the collection of z-scores corresponding to absent watermarks has a zero mean with unit variance. In contrast, the z-score corresponding to the present watermark has a much lower (negative) value. As such, to detect a watermark's presence, we can compare the corresponding z-score to a threshold. To estimate the threshold T for a certain false positive (FP) probability P_{fp}, the Gaussian method defined in (3) is used [24, 25].

$$T = \sqrt{2}\mathrm{erfc}^{-1}(2P_{\mathrm{fp}}) \tag{3}$$

In the equation, erfc^{-1} is the inverse of the complementary error function. For example, an FP probability of $P_{\mathrm{fp}} = 10^{-6}$ results in a threshold $T \approx -4.8$. Finally, a watermark's presence is detected when the corresponding z-score does not exceed this threshold.

Note that the detected watermark is a zero-bit watermark. That is to say, it cannot embed or extract a multi-bit payload bit by bit from a single video segment. Instead, we can only detect the watermark's presence or absence. Then, by linking a zero-bit watermark with the user to which the corresponding video was distributed, the leaker is detected. Additionally, a multi-bit payload can optionally be embedded by combining the proposed method with A/B watermarking. In that case, every segment can represent one bit of information, assuming only two watermarked versions per segment are used. In this way, by using multiple segments in the entire video, a multi-bit watermark is embedded.

The proposed method requires all distributed watermarked versions of the video to detect the leaked watermark. Even so, as mentioned before, it should be stressed that there are only few watermarked versions when used in combination with A/B watermarking (e.g., only two). Nonetheless, the method is non-blind. Non-blindness has the advantage that registration techniques can be applied to undo spatial and temporal synchronization attacks, increasing the robustness of the method. Moreover, the complexity of extraction increases linearly with the number of distributed watermarked videos. Again, although the order of the asymptotic complexity may be larger than other existing methods, it should be stressed that there are only few distributed watermarked videos in A/B watermarking. Hence, in that case, the total computational complexity is still very low.

Finally, one may note that when only two watermarked versions of a video segment are created in A/B watermarking, the proposed detection method does not work as-is. That is because there is only a single absent watermark, resulting in the standard deviation σ_{A} of zero, i.e., a z-score of (negative) infinity. As a solution, few non-distributed watermarked videos can be created as extra watermarked videos for accurate detection. These limitations are the cost of an implementation-free, RD-preserving watermarking algorithm.

4 Evaluation

This section experimentally analyzes the proposed method. First, Sect. 4.1 describes the setup that was used during the experiments. Then, Sect. 4.2 analyzes the compression efficiency, i.e., the visual quality and bitrate. Lastly, the robustness is evaluated in Sect. 4.3.

4.1 Experimental Setup

We evaluated the proposed method with five 10-s sequences that have a resolution of 1920×1080 pixels: BQTerrace, Cactus, ParkJoy, Kimono1, and ParkScene [26]. These sequences consist of 600, 500, 500, 240, and 240 frames, respectively.

The watermarking method was used in combination with two existing encoders, namely the *x264*-encoder (version 0.157) and the *x265*-encoder (version 3.2), which use the H.264/AVC and H.265/HEVC standard, respectively. These open-source implementations allow a minimum target bitrate change of 1 kbit/s. The rate-control algorithms' buffer sizes were set to the same value as the target bitrate, and no parallelization was enabled such that the encoder behavior was deterministic and reproducible. To adapt the encoder settings to typical adaptive-streaming applications, a two-pass encoding was performed. The first pass is equal for all watermarked versions of a video, in which a video is encoded with a fixed CRF value of 22, 27, 32, or 37 (further represented as CRF_w). Then, the average bitrate of the first pass encoding is calculated and rounded to the nearest 100 kbit/s. In the second pass, this rounded bitrate is used as the default target bitrate t, as well as for the video buffering verifier (VBV) size (i.e., the buffer is approximately one second). Lastly, an intra-period of 10 s was set. This means that only the first frame of each sequence is an intra-frame, whereas all subsequent frames are inter-frames. The other encoding parameters were left to their default values. For each sequence and each CRF_w, 20 watermarked videos were created with IDs $-9, -8, \ldots, 9, 10$. In other words, the default target bitrate was changed with a maximum of 10 kbit/s.

We compare the experimental results to various recent state-of-the-art watermarking methods with similar embedding complexities as the proposed technique. In other words, systems in which each watermarked video is separately compressed. First, both the results of compression efficiency and robustness are compared to our previous work, i.e., the RD-preserving method by Mareen et al. [17]. Although this work is also RD preserving, its implementation requires complex modifications to the video encoders. Second, we compare the results against the uncompressed domain watermarking method called Symmetric Dynamic Level Detection (SDLD), by Asikuzzaman et al. [11]. For this method, the parameters are equal to those in the original method. Third, we compare with an adapted version [9] of the non-blind uncompressed domain Spread-Spectrum (SS) watermarking method by Cox et al. [8]. That is, the watermark length, the scale factor, and the number of skipped coefficients are

1000, 0.1, and 1000, respectively. The length and scale factor are the same that were used for the evaluation of the originally proposed method [8]. Additionally, the number of skipped coefficients is chosen to be as large as the length of the watermark, as was done in the evaluation of the adaptation by Barni et al. [9]. Since the watermarks of the SDLD and SS method are embedded in the uncompressed domain, we encoded the watermarked videos with the same four CRF_w values as this paper. Lastly, we compare the compression efficiency to the scheme by Meerwald and Uhl [12] and the system proposed by Buhari et al. [14].

4.2 Compression Efficiency: Visual Quality and Bitrate

The imperceptibility or visual quality is often measured using the Peak Signal-to-Noise Ratio (PSNR). However, care should be taken when interpreting reported PSNR values from watermarking methods because a high PSNR is not necessarily good when the bitrate increase is also high. Therefore, we measure the visual quality jointly with the bitrate increase, using the Bjøntegaard-Delta rate (BD-rate) [27]. The BD-rate estimates the difference in average bitrate that two encoders need to create a video of the same quality. More specifically, the BD-rate is a percentage that represents the bitrate overhead to compress a watermarked video at the same quality as the unwatermarked encoding. A BD-rate of 0% indicates that the relationship between the quality and bitrate is equal for both the watermarked and unwatermarked encoding. For the BD-rate calculation in this paper, we used the PSNR and bitrate of 4 quality levels (CRF_w of 22, 27, 32, and 37). These values were calculated for an encoding using the default target bitrates as the unwatermarked encodings, and for all 20 modified target bitrates (per sequence, per default target bitrate) as the watermarked encodings. As such, the unwatermarked encodings can be compared to the watermarked encodings.

Table 1 presents the obtained BD-rates of the proposed and state-of-the-art techniques. The results prove that the proposed method approximately preserves the compression efficiency since the BD-rates are very close to zero, namely 0.02% and

Table 1 Compression efficiency results and comparison with state of the art

Method	BD-rate (%)
Asikuzzaman et al. [11]	8.72
Cox et al. [8]	8.71
Meerwald and Uhl [12]	8.23
Buhari et al. [14]	1.38
Mareen et al. [17]	0.03
Proposed (*x264*)	0.02
Proposed (*x265*)	−0.10

-0.10% for the *x264* and *x265*-encoder, respectively. In other words, these results report a negligible loss in RD performance when using the *x264*-encoder, and a small but negligible increase in compression efficiency when using the *x265*-encoder. These close-to-zero values are as expected since the watermarked encodings are using an unmodified video encoder from which they only change an input parameter. Our previous rate-distortion-preserving work [17] reported a similar BD-rate of 0.03%, but requires a complex implementation integrated with the used video encoders. Lastly, other state-of-the-art algorithms report much larger BD-rates of up to 9%. Thus, the state of the art is outperformed by the proposed method in terms of compression efficiency.

4.3 Robustness

The robustness is evaluated against recompression attacks in this section. The recompression attacks were performed by re-encoding all watermarked videos with the *x265*-encoder, using 6 different CRF values: 22, 27, 32, 37, 42, and 47 (further represented as CRF_a). In other words, $5 \cdot 20 \cdot 4 \cdot 6 = 2400$ recompressions were done in total, i.e., for 5 tested sequences, 20 watermark IDs, 4 CRF_w values, and 6 CRF_a values.

For detection, the threshold was set to $T = -4.7534$, which corresponds to a FP probability of $P_{\text{fp}} = 10^{-6}$. For this reason, the false positive rates (FPR) are not further presented in this paper. For the robustness evaluation, we use the false negative rate (FNR) defined in (4). In the equation, a false negative (FN) detection is when the present watermark is not detected in the attacked video. Finally, it is important to note that a smaller FNR is better.

$$\text{FNR} = \frac{\#\text{FN Detections}}{\text{Total \# Detections}} \tag{4}$$

Table 2 shows the calculated FNR values for the proposed method, in addition to the results of the state-of-the-art methods by Asikuzzaman et al. [11], by Cox et al. [8], and our previous RD-preserving work [17]. The methods by Meerwald and Uhl [12] and Buhari et al. [14] are not present in the comparison because their implementations were not available, and the robustness results for the recompression attack is not present in scientific literature. The FP probability of the state-of-the-art methods was set also to $P_{\text{fp}} = 10^{-6}$.

For the proposed technique, the FNRs for almost all CRF_w and CRF_a values are zero, regardless of the used encoder. Only watermarks initially encoded with a higher quality (i.e., a low CRF_w) result in a nonzero FNR after recompression to a low quality (i.e., $\text{CRF}_a \geq 42$). In contrast, when the quality of the initial encoding is medium to high (i.e., $\text{CRF}_w = 27$ or $\text{CRF}_w = 32$), the watermarked videos are robust for recompression with $\text{CRF}_a \leq 42$. These results are similar to the method by Mareen et al. [17]. It should be noted that recompression with $\text{CRF}_a \geq 42$ severely

Table 2 Robustness results and comparison with state of the art

CRF_w	False negative rate (%)																	
	$CRF_a = 22$	27	32	37	42	47	22	27	32	37	42	47	22	27	32	37	42	47
	Asikuzzaman et al. [11]						Cox et al. [8]						Mareen et al. [23]					
22	0	0	0	20	100	100	0	0	0	0	100	100	0	0	0	0	0	37
27	0	0	0	20	100	100	0	0	0	0	100	100	0	0	0	0	0	19
32	0	0	0	20	100	100	0	0	0	0	100	100	0	0	0	0	0	0
37	20	20	20	40	100	100	0	0	0	65	100	100	0	0	0	0	0	0
	Proposed (*x264*)						Proposed (*x265*)											
22	0	0	0	0	1	39	0	0	0	0	19	20						
27	0	0	0	0	0	6	0	0	0	0	0	3						
32	0	0	0	0	0	1	0	0	0	0	0	0						
37	0	0	0	0	0	0	0	0	0	0	0	0						

Fig. 3 Attacked version of the watermarked cropped frame in Fig. 2b, recompressed with $\text{CRF}_a = 42$. Although the attacked frame has a low quality, the watermark's presence is still detected

decreases the quality, to a point where users do not enjoy watching it anymore. For example, Fig. 3 shows an example of a cropped attacked frame, recompressed with $\text{CRF}_a = 42$, which shows many disturbing artifacts.

For the state-of-the-art methods by Asikuzzaman et al. [11] and by Cox et al. [8], one can observe worse robustness results. That is, the method by Asikuzzaman et al. reports nonzero FNRs for all $\text{CRF}_a \geq 37$ and for $\text{CRF}_w = 37$. Similarly, the method by Cox et al. is not robust for large CRF_a values. Most interestingly, traditional state-of-the-art methods perform worse when initially compressed with a high CRF_w, whereas the proposed method achieves better results in that case. That is because the initial encoding indirectly creates the zero-bit watermark. Since a high CRF_w creates more perceptible compression artifacts, the watermark is more robust than when compressed with a low CRF_w.

5 Conclusion

Forensic watermarking in adaptive streaming is often performed using the A/B watermarking concept, which can be combined with existing watermarking methods. Because visual quality and the compressed video's bitrate are of utmost importance in adaptive-streaming applications, this paper proposed a RD-preserving forensic watermarking method which is intended to be used in combination with A/B watermarking. Additionally, although the proposed method integrates the watermarking step with video compression, it does not require changes to existing video encoders. Instead, it only changes the input parameters of the rate-control algorithm of the video encoder. In this way, the proposed scheme does not require a complex implementation, and can hence be used in combination with various video codecs without implementation modification overhead.

The results prove that our proposed approach has a negligible impact on the RD performance, i.e., on the relationship between visual quality and bitrate. More specifically, the BD-rate is only 0.02% and -0.10% when combining the proposed method with an H.264/AVC and H.265/HEVC-encoder, respectively. Moreover, the watermarks are robust against attacks that severely lower the quality of the video. For example, when the video is initially compressed with a medium-to-high quality, the system is robust against recompression with $\mathrm{CRF}_a \leq 42$. These robustness results are similar or better than state-of-the-art methods with a comparable embedding complexity. In conclusion, when combined with A/B watermarking, the proposed scheme provides a practical, implementation-free, RD-preserving forensic watermarking solution for adaptive-streaming platforms.

Acknowledgements This research was supported by the Research Foundation—Flanders (FWO) under Grant 1S55218N. Additionally, it was supported by IDLab (Ghent University—imec), Flanders Innovation & Entrepreneurship (VLAIO), and the European Union. Moreover, the computational resources (STEVIN Supercomputer Infrastructure) and services used for the evaluation of our watermarking approach were kindly provided by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation, and the Flemish Government department EWI. Furthermore, we would like to thank Martijn Courteaux for his valuable input during brainstorming sessions.

References

1. Jarnikov D, Hietbrink E, Arana M, Doumene JM (2014) A watermarking system for adaptive streaming. In: Proceedings of the IEEE international conference on consumer electronics (ICCE), pp 375–377, Jan 2014
2. Mathew R, Cushing C (2020) Netflix to slash traffic across Europe to relieve virus strain on internet providers. https://www.reuters.com/article/idUSKBN21906P. Posted 22 Mar 2020, Accessed 17 Aug 2020
3. Kim KS, Lee HY, Im DH, Lee HK (2008) Practical, real-time, and robust watermarking on the spatial domain for high-definition video contents. IEICE Trans Inf Syst E91-D(5):1359–1368
4. Piva A, Bianchi T, Rosa AD (2010) Secure client-side ST-DM watermark embedding. IEEE Trans Inf Forensics Secur 5(1):13–26
5. Fautier T (2020) Ultra HD Forum guidelines, v2.2. Technical report, Ultra HD Forum
6. Liu J, He X (2005) A review study on digital watermarking. In: Proceedings of the international conference on information communication technologies, ICICT 2005, pp 337–341
7. Asikuzzaman M, Pickering MR (2018) An overview of digital video watermarking. IEEE Trans Circuits Syst Video Technol 28(9):2131–2153
8. Cox IJ, Kilian J, Leighton FT, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. IEEE Trans Image Process 6(12):1673–1687
9. Barni M, Bartolini F, Cappellini V, Piva A (1998) A DCT-domain system for robust image watermarking. Signal Process 66(3):357–372
10. Islam SMM, Debnath R, Hossain SKA (2007) DWT based digital watermarking technique and its robustness on image rotation, scaling, JPEG compression, cropping and multiple watermarking. Proceedings of the international conference on information communication technologies, ICICT 2007, pp 246–249

11. Asikuzzaman M, Alam MJ, Lambert AJ, Pickering MR (2014) Imperceptible and robust blind video watermarking using chrominance embedding: a set of approaches in the DT CWT domain. IEEE Trans Inf Forensics Secur 9(9):1502–1517
12. Meerwald P, Uhl A (2011) Robust watermarking of H.264/SVC-encoded video: quality and resolution scalability. In: Proceedings of the international conference on digital watermarking, IWDW'10. Springer, Berlin, Heidelberg, pp 159–169
13. Aparna J, Ayyappan S(2015) Image watermarking using Diffie Hellman key exchange algorithm. Procedia Comput Sci 46:1684–1691. Proceedings of the international conference on information communication technologies, ICICT 2014
14. Buhari AM, Ling HC, Baskaran VM, Wong K (2016) Fast watermarking scheme for real-time spatial scalable video coding. Signal Process Image Commun 47:86–95
15. Coria LE, Pickering MR, Nasiopoulos P, Ward RK (2008) A video watermarking scheme based on the dual-tree complex wavelet transform. IEEE Trans Inf Forensics Secur 3(3):466–474
16. Chen W, Shahid Z, Stütz T, Autrusseau F, Le Callet P (2014) Robust drift-free bit-rate preserving H.264 watermarking. Multimed Syst 20(2):179–193
17. Mareen H, Courteaux M, De Praeter J, Asikuzzaman M, Van Wallendael G, Lambert P (2020) Rate-distortion-preserving forensic watermarking using quantization parameter variation. IEEE Access 8:63700–63709
18. Li B, Li H, Li L, Zhang J (2014) λ domain rate control algorithm for high efficiency video coding. IEEE Trans Image Process 23(9):3841–3854
19. Zupancic I, Naccari M, Mrak M, Izquierdo E (2017) Two-pass rate control for improved quality of experience in UHDTV delivery. IEEE J Sel Top Signal Process 11(1):167–179
20. Cox I, Miller M, Bloom J, Fridrich J, Kalker T (2008) Digital watermarking and steganography, 2nd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
21. Mareen H, De Praeter J, Van Wallendael G, Lambert P (2019) Traitor tracing after visible watermark removal. In: Yoo CD, Shi YQ, Kim HJ, Piva A, Kim G (eds) Digital forensics and watermarking, IWDW'18. Springer International Publishing, Cham, pp 110–123
22. Urvoy M, Autrusseau F (2014) Application of Grubbs' test for outliers to the detection of watermarks. In: Proceedings of the ACM workshop on information hiding multimedia security (IH&MMSec). Association for Computing Machinery, New York, NY, USA, pp 49–60
23. Mareen H, De Praeter J, Van Wallendael G, Lambert P (2018) A novel video watermarking approach based on implicit distortions. IEEE Trans Consum Electron 64(3):250–258
24. Kalker T, Linnartz J, Depovere G, Maes M (1998) On the reliability of detecting electronic watermarks in digital images. In: Proceedings of the European Signal Processing Conference (EUSIPCO), pp 1–4, Sept 1998
25. Miller ML, Bloom JA (1999) Computing the probability of false watermark detection. In: Proceedings of the international workshop on information hiding, pp 146–158
26. Bossen F (2013) Common test conditions and software reference configurations. Technical report, JCTVC-L1100, ITU-T Joint Collaborative Team on Video Coding (JCT-VC), Jan 2013
27. Bjøntegaard G (2001) Calculation of average PSNR differences between RD-curves. Technical report, VCEG-M33, ITU-T Video Coding Experts Group (VCEG), Apr 2001

Study of Data and Internet Propagation in Electrical Networks Versus Unshielded Twisted Pair (UTP) Cabling for the Extension of Home Wi-Fi Network Coverage

Yessica Herran, Richard García, Fabian Blanco, and Fredys Simanca

Abstract The advances in technology bring about the need to interact with and control the devices used in daily life. This process is so accelerated that the traditional methods of communication become obsolete and not very aesthetic, and this factor demands the experts to create instruments that allow a connection to the virtual world, regardless of where the person is. This paper will examine the feasibility of implementing a communication system different from the usual propagation of UTP cabling, as well as the use of new technologies such as the power line communications network extension system, where an analysis of the performance of the channel, the susceptibility of the medium to external factors, and the versatility in the process of installation in the infrastructure was carried out. The aim is to obtain an easy means of connecting with the various electronic devices that can be used at home and/or office, reducing the time of interaction with the environment.

Keywords UTP cable · Internet of things · Power line communications

1 Introduction

Communication technologies allow us to transmit information through two media: the guided ones such as coaxial cable, twisted pair, and optical fiber, and the unguided ones such as air and vacuum, which are the fundamental piece for what is called today IoT.

The guided media are the most reliable channels for transmitting information, because they may be less susceptible to interference, compared to the non-guided media which may have a greater effect on the channel, due to environmental factors or other broadcasting sources. Taking advantage of these properties, a study will be conducted on the propagation behavior of the twisted pair, and of another little-used

Y. Herran · R. García · F. Blanco · F. Simanca (✉)
Universidad Cooperativa de Colombia, Bogotá, Colombia
e-mail: fredys.simanca@campusucc.edu.co

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_32

medium such as copper wire from low-voltage power lines, which is found in the electrical distribution system of homes or residences.

Power line communications (PLC) is a technology that allows the transmission of broadband data over high, medium, and low-voltage networks. It appeared in the 1950s, allowing the control of remote relays with one-way communications. In the mid-1980s, research was initiated for the use of electrical networks and their broadband transmission, but in only one direction. Later, by 1997 and 2000, the first high-speed transmission tests were carried out in the United Kingdom and France [1].

Among the interferences that affect this technology are background noise, color noise, narrowband noise, and impulsive noise, typical of the transmission medium. This technology considers different standards, each one improving the susceptibility to these interferences, as well as the transmission speed. This paper is divided into three phases. The first phase describes the technical characteristics of the HomePlug standard, Ethernet, and wireless networks, respectively, developed for the transmission of information by electrical networks and twisted-pair cables. The second phase describes the methods used and the results obtained in the comparison of the two technologies, and the closure of the paper establishes the conclusions and recommendations.

2 Standards Applied

Each technology implemented in a data propagation system includes standards that are used as a model to evaluate security, coupling, performance, efficiency, and compatibility between media [2]. The following describes the standards applied by both technologies to analyze BPL and UTP.

The HomePlug standard makes use of transmission techniques such as *OFDM modulation*, which allows dividing the available spectrum into multiple carriers, thus being resistant to RF interference, using CSMA/CA for data transmission and TDMA for accessing the medium [3]. For the year 2001 *HomePlug Alliance* developed the *Homeplug 1.0* which used speeds up to 14 Mbps being functional for basic internet services. It sent electronic mails and was one of the first ones to support *Plug & Play* technology [4]. Later, in the year 2005 emerged a breakthrough with the AV version increasing its speeds to 150 Mbps in a range of 2–30 MHz, being suitable for services such as HDTV known as high definition television [5] and providing connection-oriented QoS. In the year 2012, the AV2 version emerges with a performance of up to 1000 Mbps that stands out for using *MIMO technology* with multiple-input and multiple-output routes used in buildings with old power distribution systems [6].

The IEEE standards are defined for the *LAN*, where typologies and rules of access to the network are defined. In the IEEE 802, three fundamental complements are defined: LLC (logical link control) for managing the logical link, MAC (medium access control) for controlling and managing the physical access, and PHY for performing tasks of encoding and decoding bits [7].

The 802.3u standard is one of the variants of the 802.3 standards and applies to networks that can transmit information at 100 Mbps and can simultaneously be integrated into 10 Mbps networks, using the CSMA/CS access protocol called multiple access with carrier detection and collision detection [8].

The IEEE 802.11 is an IEEE standard for wireless networking technology, using the frequency bands of 2.4, 3.3, and 5 GHz and making use of the CSMA/CA protocol [9]. In its versions present is the IEEE 802.11b, which defines a high-speed physical layer of up to 11 Mbps in the frequency of 2.4 GHz, and uses HR-DSSS modulation techniques. In [10], the IEEE 802.11g, which presented the advance over the 802.11b standard, uses OFDM modulation techniques, reaching speeds of 22–55 Mbps in the 2.4 GHz band [11]. Then, the IEEE 802.11n standard made it possible to take advantage of the robust cabling network infrastructure by creating multiple-input and multiple-output (MIMO) antennas [12].

IEEE 1901 defines detailed mechanisms for interoperability between different devices such as *broadband over powerline* (BPL) as well as to ensure the desired bandwidth and the quality of the service. Among the resources defined in the standard, it relates the modulation techniques such as *orthogonal frequency division multiplexing* (OFDM) to be very robust against the multipath [13], and its operation is in three phases of transmission: generation, propagation, and distribution, in order to take advantage of the different channels in a frequency range of 2–30 MHz [14].

3 Analysis and Results

This section describes the steps taken to make the comparison between BPL technology devices and conventional Ethernet. In the transmission on power lines, reference equipment DHP-208AV and DHP-W220AV from the manufacturer D-Link are used to make up the kit for BPL technology with the specifications described in Table 1.

In conventional Ethernet communication, the reference device TD-W8951ND from the manufacturer TP-LINK with the specifications described in Table 2 is used:

The device specifications are presented in order to validate its installation versatility, channel stability, and performance.

3.1 Versatility of Sub-chapter Installation

The proposed connection scheme for GLP technology is shown in Fig. 1. First, the connection of the device 208AV to the modem to establish the communication, then to the pc by Ethernet cable to perform configurations on the device and the power outlet for the spread of the data. In the other area, the device 220AV performs the reception of data, allowing for the connection of devices via wireless, taking for the latter the option of creating a separate network or clone the current one in order to extend

Table 1 General specifications of GLP devices

Device	Standards		Encryption	Transmission speeds
DHP-208AV	802.3	1901	Coding 128—AES on power line	>200 Mbps (PHY) over power line
	802.3u	HomePlug		100 Mbps over Ethernet
DHP-W220AV	802.3	1901	Coding 128—AES on power line	>150 Mbps WLAN with working frequencies of 2.4–2.976 GHz
	802.3u	HomePlug	WEP/WPA/WPA2	>200 Mbps (PHY) over power line
	802.11b	802.11g		100 Mbps over Ethernet
	802.11n			

Source DHP-W221AV—Kit Powerline N150—D-Link Latinamerica, n.d. [15]

Table 2 General specifications of the Ethernet device

Device	Standards		Encryption	Transmission speeds
TD-W8951ND	802.3	802.11b	Codification 64/128-AES	150 Mbps WLAN
	802.3u	802.11g	WEP/WPA-PSK/WPA2-PSK	10/100 Mbps over Ethernet

Source (TD-W8951ND|Módem Router Inalámbrico N ADSL2 + de 150 Mbps|TP-Link Colombia, n.d.) [16]

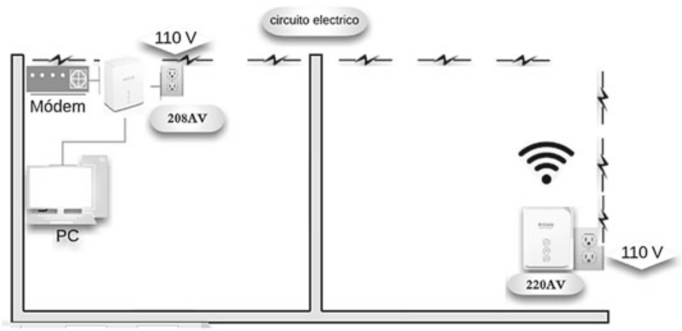


Fig. 1 Connection scheme for GLP technology devices

the coverage in areas of low reception. Unlike BPL technology, the conventional modem must have power line availability and a communication channel from the service provider's equipment to be able to access the internet, which leads to a more complex process.

3.2 Channel Stability and Performance

The standard 802.3u defines the PMA modulation as the best method of propagation in twisted pair, offering higher transmission speeds at a lower cost. It happens in a different way with the GLP devices susceptible to external and internal sources of the medium. The major effects are mainly produced by elements in inductive, capacitive, false connections, and deterioration of the environment, and therefore, they will be represented as background noise such as color noise (different interconnected sources of energy), the noise of narrowband (induction of broadcasting signals), and impulsive noises such as disconnection of electronic devices [17]. As an additional factor, there is power loss due to the topology adopted by the system, which would be defined as a bus topology in which multipath is generated throughout the power line [18]. In fact, for correcting this type of affectations, the Homeplug standard defines the OFDM modulation technique as the best propagation option correcting most of the affectations of the medium. The proposed connection scheme for GLP technology is shown in Fig. 2.

In the BPL technology, hostile scenarios are reflected and, in the same way, it was verified that to preserve the stability in the communication channels it must ensure the connection scheme proposed in Fig. 3, which finally allows to obtain the highest performance in the devices both in the physical and wireless environment.

Therefore, if the connection is modified to a scheme where active devices are intermediate or which, due to their operating principle, force an opening of the electrical circuit, an instability will be evident since the electrical line behaves like an information bus. Figure 4 presents a connection scheme where the case of involvement for the BPL communication channel is represented.

It should be borne in mind that most of the effects generated by inductive elements directly affect the power stage of the transmitter/receiver equipment, due to the abrupt variation in current.

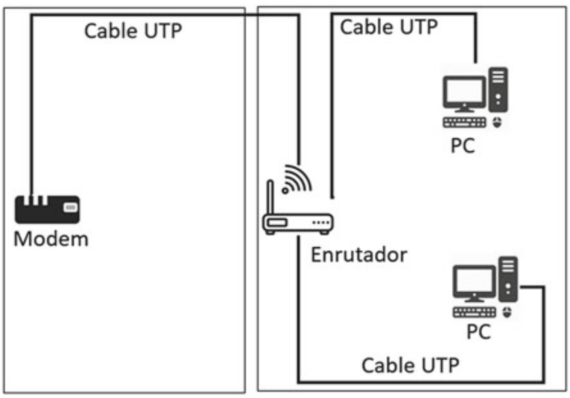


Fig. 2 Connection scheme for conventional Ethernet modem-to-modem technology

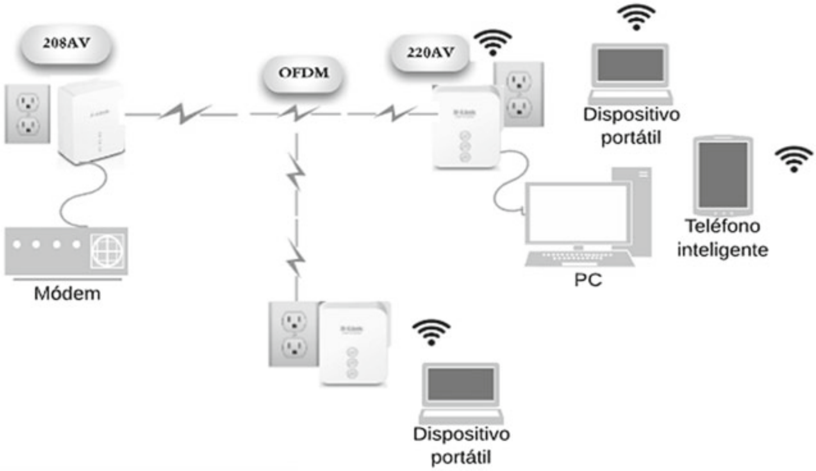


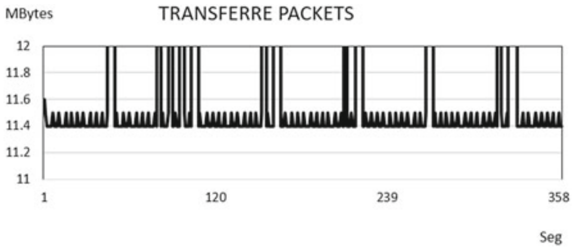
Fig. 3 BPL connection with optimum performance

Fig. 4 BPL scheme with energy meter

3.3 Transmission Speeds

In a first impression, it is said that GLP technology is inferior to Ethernet technology, because the transmission medium is not only dedicated to the propagation of information and, as is well known, it is multiservice, contrary to twisted pair which is defined for a communication service. In order to check the channel capacity and evaluate if the GLP technology is optimal, the Iperf software is used. It has an easy-to-use interface and is developed in such a way that it allows to obtain real-time propagation results by creating end-to-end TCP/UDP flows. The sampling obtained was carried out with UDP flows, since this allows to flood the medium without restriction to determine the speed with duration times of 360 s. The first system tested is the BPL where 11.5 Mbyte second-to-second packet shipments are recorded, as shown in Fig. 5, with intervals outside the threshold and mainly generated by the modulation technique.

Fig. 5 Transferred BPL packages



In parallel, transmission speeds of 95.6 Mbps are recorded, with out-of-range intervals newly generated by the propagation method, as shown in Fig. 6. Every registered variation coincides with the transmission speed and the transferred packet; the total result in transferred packets is 17.3 Gbytes with a transfer speed of 414 Mbps.

These intervals represent the relayed packets that arrive over another channel which is reflected as an increase in the speed achieved, so the modulation method influences the data collection.

As for transmission with conventional Ethernet technology, the same configuration of 360 s without bandwidth limit is used with a UDP stream, where a log of packets of 11.4 Mbytes is transferred constantly throughout the test, the information of which is shown in Fig. 7. In parallel, the transmission rate is sampled with a constant value of 95.9 Mbps and is graphed in Fig. 8.

The results obtained from the sampling make it possible to determine that the modulation technique has a great influence on the reception of the end-to-end data, since it is constant in one of them, while on the other it can present variations

Fig. 6 BPL transmission speed

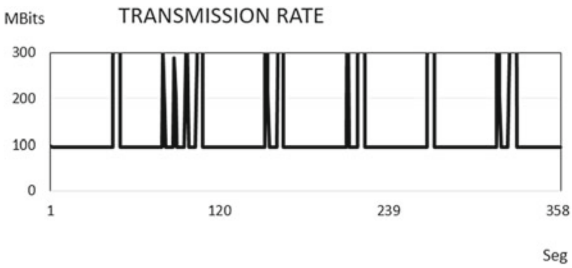


Fig. 7 Packets transferred on conventional Ethernet technology

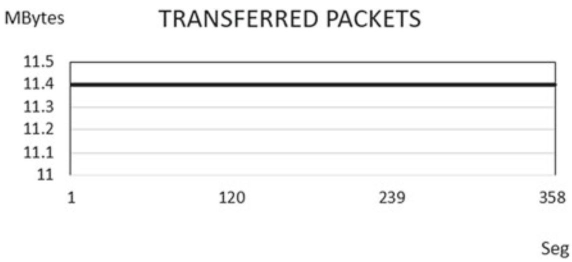


Fig. 8 Transmission speed in conventional Ethernet technology

over time in order to correct transmission errors. Another of the environments to be evaluated is the transfer of the wireless service, so the same simulation techniques were used as in the physical sampling, with results for the GLP technology of better performance than the conventional Ethernet technology as can be observed in the following analysis.

In the BPL technology through wireless network service, an average of 6.60 MB of transmitted packets is recorded, although transmission peaks are obtained in similarity with the physical medium, and this information is presented in Fig. 9. As a result, the transmission rate presents a reading of 56.9 Mbps as an average value. One of the particularities identified in this test model for the wireless service is that it shows the constant variation that the communication undergoes, and does not retain a transmission pattern as in the physical medium. This record can be seen in Fig. 10. In the tests carried out with the wireless connection typical of the conventional Ethernet

Fig. 9 Packets transferred via BPL wireless service

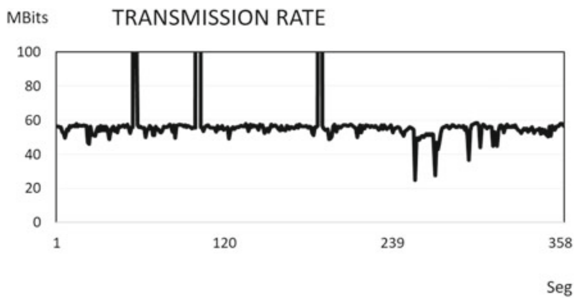


Fig. 10 Transmission rates

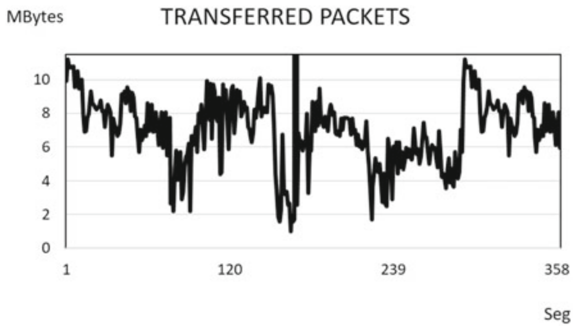


Fig. 11 Packets transferred via wireless service on Ethernet technology

Fig. 12 Transmission rates via wireless service on Ethernet technology

technology, a more erratic behavior was tracked, since constant data was not achieved and with a final result of 2.51 Gbytes in transferred packets as recorded in Fig. 11.

Therefore, the transmission rate is also recorded with random behavior as shown in Fig. 12, showing a final average of 59.9 Mbps. The tests with the wireless service were carried out with a distance of 3 m to the receiving device in five repetitions.

The use of these tools makes it possible to observe the behavior of the data propagated in the physical and wireless environment of both technologies.

4 Conclusions

The results collected from each technology demonstrate that they are technically similar no matter what medium is being used, even if it is emphasized that GLP technology has a great breakthrough in propagating information by non-communication means, having robust components in very small units that can be safely deployed for the growth of the local network. In the last aspect of comparison of the achieved speeds, technologies manage to have a great similarity with links averages of 95.6 Mbps which is the speed that achieves the network card of the computer used, as stated in the IEEE 802.3 standard. A discrepancy between the two technologies is that the BLP system exceeds the speed of transmission, but this is due to the method that spreads the information and the corrections that the system performs. Although the physical links are reliable, wireless technology is not left behind. By having certifications in the IEEE 802.11 standards, the technologies guarantee the transport of the information end-to-end without having a low performance, which is demonstrated with the network tests at the moment of transporting information, although greater variations of one technology with respect to the other are identified, being of better performance the wireless service in the BPL devices than in the conventional modems, demonstrating that the service of the BPL system is reliable and of high capacity for the wireless networks.

In this way, it is verified that the BPL systems are a usable technology for the extension of physical and wireless networks, considering that the equipment used is of medium range with respect to others that are in the market. Finally, the BPL devices were used for six months without interruption in any of its services and are still used today for the communication of information.

References

1. Serna V (2011) Comunicaciones a través de la red eléctrica–PLC. Comunicaciones 62–65
2. Moreno Castro G, Amaya Fernández F (2012) Sistemas de comunicación BPL—Hibridos y su aplicación en áreas rurales. Rev Telecomun Inform 2(3):41–54
3. Ferreira JC, Acuna GG, Barrera JAV (2016) Evaluación del rendimiento de una red LAN sobre power line communications para la transmisión de VOIP, pp 83–95
4. HP Alliance (2001) Homeplug 1.0 technology white paper, p 9
5. P Alliance (2005) HomePlug AV white paper. HomePlug Powerline Alliance Inc, pp 1–11
6. Yonge L, Abad J, Afkhamie K, Guerrieri L, Katar S, Lioe H, Pagani P, Riva R, Schneider DM, Schwager A (2013) An overview of the HomePlug AV2 technology
7. Navarro García S (2005) Algoritmos cross-layer para la optimización de las prestaciones del tcp en redes wireless Ad-Hoc 165:165
8. Carrión Morales MF (2006) Diseño De La Red De Comunicaciones De La Mutualista Pichincha Para La Ciudad De Quito Basado En Tecnologías Ethernet De Alta Velocidad
9. Akbar MS, Khan MS, Khaliq KA, Qayyum A, Yousaf M (2014) Evaluation of IEEE 802.11n for multimedia application in VANET. Procedia Comput Sci 32:953–958
10. Soto Sanchez OA (2012) Comparación de la eficiencia volumétrica entre redes inalámbricas WiFi y WiMAX, pp 33–54
11. Anguís Horno JJ (2008) Redes de área local inalámbricas: diseño de la WLAN de wheelers lane technology college, p 272
12. Perahia E (2008) IEEE 802.11n development: history, process, and technology. IEEE Commun Mag 46(7):48–55
13. Peralta A (2010) BPL (Broadband Power Line): una alternativa para brindar el servicio de internet de alta velocidad 4:25–34
14. Goldfisher S, Tanabe S (2010) ITU S TANDARDS IEEE 1901 access system : an overview of its uniqueness and motivation. IEEE Xplore 48(10):150–157
15. DHP-W221AV (n.d.) Kit Powerline N150—D-Link Latinamerica
16. TD-W8951ND (n.d.) Módem Router Inalámbrico N ADSL2+ de 150Mbps|TP-Link Colombia
17. Apaza Mamani V (2017) Estudio del canal plc (power line communications) para envío de datos a bajo costo en redes eléctricas domiciliarias, p 92
18. Gallardo Moreno JP (2006) Estudio de un sistema OFDM-CDMA

A Study on the Generalized Normalization Transformation Activation Function in Deep Learning Based Image Compression

Qiang Duan, Xue Li, Qingshan Yin, Luoluo Feng, Jing Zhao, Yijin Teng, Xiaohui Duan, Yanhan Zhao, Ming Gao, Jianhua Wang, Wei Cai, and Rui Li

Abstract Image compression has a long history, being widely applied in numerous applications. Yet, the commonly applied approach is based on traditional signal processing, such as JPEG. The lack of learning process in the approach limits the performance of those methods. Deep learning, recently, shows great performance in image compression, whose success may be attributed to various factors. Activation function is known as one of these influential factors. The present work is devoted to analyze the different effects of various activation functions, and the experimental results suggest that the generalized divisive normalization (GDN) is probably the best activation function in deep image approach-based image compression.

Keywords Image compression · Deep learning · Convolutional neural networks · Auto-encoder · Activation function

---

Authors Qiang Duan, Xue Li and Qingshan Yin contributed equally to this study.

Q. Duan · X. Li · Q. Yin · L. Feng · J. Wang · R. Li (✉)
Inspur Technology Center, Inspur Group, Jinan, People's Republic of China
e-mail: lirui01@inspur.com

J. Wang
e-mail: jhwang@inspur.com

Shandong Institute of New Generation Information Industry Technology Co., Ltd., Jinan, People's Republic of China

J. Zhao
Shandong Yingxin Computer Technology Co., Ltd., Jinan, People's Republic of China

Y. Teng
Department of Program Marketing, Inspur Group, Jinan, People's Republic of China

X. Duan · Y. Zhao · W. Cai
Department of Enterprise Development, Inspur Group, Jinan, People's Republic of China
e-mail: caiwei@inspur.com

M. Gao
Baidu Knows Business Department, Baidu Science and Technology Park, No. 1036, Langchao Road, Jinan Hi-Tech Development Zone, Jinan, People's Republic of China

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_33

1 Introduction

Nowadays, more than 80% of information on the internet is transmitted through images and videos, and it will continue to grow in the future. Therefore, an efficient and high-quality image compression method is essential. Image compression is a process of data compression that encodes the original image with few bits. The objective of this process is to reduce the cost of storage and to transmit image data efficiently. Image processing is closely related to many computer vision tasks such as object classification and object detection. An efficient image compression method is helpful for image processing and model compression. There are two kinds of image compression, namely, lossless and lossy. In lossless compression, the original image can be reconstructed without losing information. On the contrary, some information can be ignored during the compression. At the time being, lossy image compression becomes more and more necessary in saving bandwidth and storage. In the last couple of years, image compression algorithms rely on hand-crafted modules, the codec usually consists of three parts: encoder, quantizer, and decoder. The traditional image encoding standards, e.g., JPEG, are not end-to-end optimized, in contrast to deep learning based-method.

In recent years, deep learning has achieved significant success in computer vision fields such as image classification, object detection, and face recognition. As for image compression, deep convolutional neural networks (CNNs) are expected to be more powerful than the traditional image compression method for the following reasons. First, CNNs can extract more accurate feature maps from a given image than the traditional method. Second, flexible nonlinear analysis and synthesis transformations can be easily applied by stacking multiple convolutional layers. Finally, deep learning enables encoder and decoder optimized in an end-to-end manner. Some recent works have already shown some evidence [4]. However, most of the works focus on the design of network architectures. Having done a large number of experiments, we found that the activation function has a significant impact on the quality of the image after compression.

In this paper, we present a CNN-based image compression framework to show that the Generalized Normalization Transformation (GDN) [2] is likely the best activation function in most cases for image compression. In our approach, the feature map of the input image is extracted by a stack of multiple layers in CNNs. Next, the feature map is quantized, and entropy coding uses a lossless encoding to the bitstream. After that, the compact bitstream is used to obtain the reconstruction of the original image through entropy decoding, inverse quantization, and deconvolution.

Experimental results show that the GDN activation function outperforms the Rectified Linear Unit (ReLU), leaky ReLU, Gaussian Error Linear Units (GeLU), and Scaled Exponential Linear Unit (SeLU) in both Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity Index Metrics (MS-SSIM) across a large range of bitrates. For example, using GDN activation function can be 3.0 dB better in PSNR than using ReLU activation function for some testing images. Moreover, images compressed by GDN activation function look smooth, which are more in line with the characteristics of the human visual system with less blurring, ringing, and blocking.

2 Related Work

Image compression is usually based on traditional signal processing-based methods, in which quantization and entropy coding are designed manually. Deep learning-based image compression approaches are capable of automatically discovering and utilizing patterns hidden in data, through constructing a neural network structure that chooses the appropriate activation and loss function. Therefore, such a method shows better effects in the compression.

Deep learning-based image compression has recently become an active research area. A variety of approaches and frameworks for this task have been proposed. At present, the popular DNN architectures utilized for image compression are auto-encoders [3, 8] and recurrent neural networks (RNNs) [10].

To make entropy coding more efficient, Toderici et al. [10] utilized Long Short-Term Memory (LSTM) to extract binary representation. Moreover, LSTM convolution was applied to estimate the probability distribution of image information in the entropy encoding section. In 2016, Ball et al. [3] proposed an end-to-end optimized image compression framework that involved a nonlinear analysis transformation, a uniform quantizer, and a nonlinear synthesis transformation.

At the same time, scientists have made efforts to combine framework structures of image compression with other computer vision tasks. For example, in [7], the computer vision tasks, such as image classification and semantic segmentation, analyzed the feature maps generated from image compression approaches. Meanwhile, the synthesis model from the frameworks of semantic segmentation was also used to reconstruct background regions such as streets and trees from the label map [1]. In low bitrates, this approach has the ability to generate higher quality synthetic images.

3 Methods

The background knowledge is first introduced as follows:

3.1 Introduction to Deep Learning-Based Image Compression

Our proposed auto-encoder framework is shown in Fig. 1. Both encoder and decoder have convolutional layer, deconvolutional layer, and GDN/IGDN (inverse

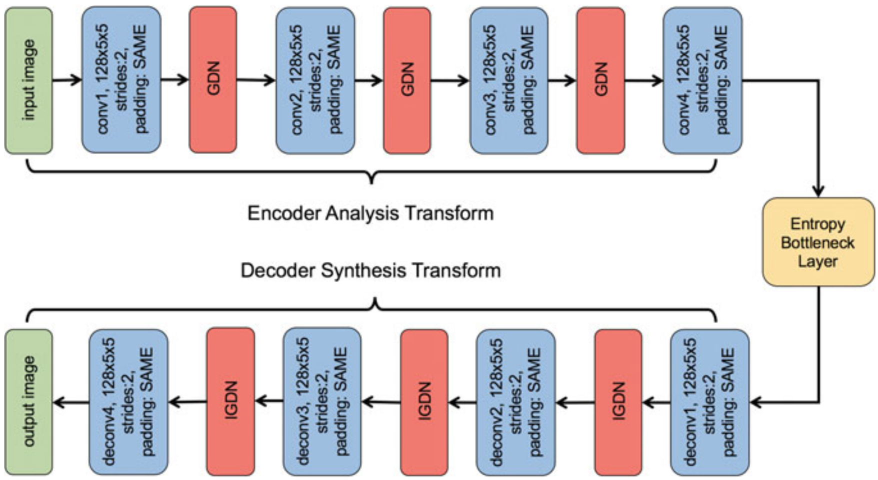


Fig. 1 The network architecture of used deep learning image compression

GDN) layer. The GDN/IGDN implement a type of local divisive normalization transformation [3].

In order to effectively perform the entropy encoding operation, we first use a quantization operation to convert float data into integer data. Assuming the feature maps after encoding are powerful enough, we take round quantization function as quantization representing values, which is $\hat{y}_i = round(y_i)$. The marginal density of $\hat{y}_i$ can be represented as a discrete probability mass with weights equal to the probability mass function [1]:

$$P_{\hat{y}_i}(\hat{y}_i = n) = \int_{n-0.5}^{n+0.5} P_{\hat{y}_i}(x)dx. \tag{1}$$

However, since the derivatives of the round quantization function are zero almost everywhere, the gradient descent of round quantization function is ineffective. In the stage of training, we replace the quantizer with an additive uniform noise, that is $\hat{y}_i = y_i + \epsilon$, where ϵ is a random noise. It is obvious that the entropy of $\hat{y}$ can be used as an approximation of the entropy of $\hat{y}$. Hence, in the phases of validation and testing, we use the function $\hat{y} = round(y_i)$ as the quantization function. Entropy coding calculates the number of bits required to represent a particular image given a quantization level. Currently, entropy coding algorithms include Range coding, Arithmetic Coding, Huffman coding, etc. Huffman coding is adopted in this paper. The process of Huffman coding and decoding is lossless, and the quantized feature maps are mapped into a bitstream using variable-length codes.

As described in [3], we optimized the set of parameters over the DIV2K database [5] which consists 800 train and 100 validation images. We used all of those images

Table 1 List of compared activation functions

$\mathrm{SeLU}(x) = \lambda \begin{cases} \alpha(e^x - 1) & x < 0 \\ x & x \geq 0 \end{cases}$	$\mathrm{GeLU}(x) = 0.5x \left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$
$\mathrm{ReLU}(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$	$\text{leaky ReLU}(x) = \begin{cases} x & x > 0 \\ \alpha x & x \leq 0 \end{cases}$

to train modified models and conduct various experiments using identical neural network structure and test dataset, with various hyperparameters and a series of activation functions. One important hyperparameter is λ that controls the rate-distortion trade-off through the loss function $loss = bitrate + \lambda * distortion$. The bitrate indicates bits-per-pixel that measures the average number of bits required for one pixel. The distortion measures the cost of representing a pixel x by an approximated $\hat{x}$, and the λ is a weighting factor for end-to-end training. The larger λ is set, the better image decoding quality and the higher bits-per-pixel (BPP) are obtained are vice versa. Training step was set to 100000 for all experiments. The activation functions for different models include ReLU, leaky-ReLU, ReLU + GDN, SeLU, SeLU + GDN, GeLU, and GDN.

3.2 Activation Functions in Image Compression

To introduce nonlinearity into the approach, deep learning employs many different activation functions. Currently, the popular activation functions in deep learning include ReLU, Sigmoid, Tanh, etc. Besides, GDN, a form of gain control in which responses are divided by pooled activity of neighbors, has been preferred to use in image compression due to the nonlinear properties of sensory neurons [2]. A commonly used form for GDN is

$$y_i = \gamma \frac{x_i^{\alpha}}{\beta^{\alpha} + \Sigma_j x_j^{\alpha}}, \tag{2}$$

where $\theta = \{\alpha, \beta, \gamma\}$ are parameters. GDN adjusts responses to be within a desired operating range while keeping their relative values unchanged. As a special case of the model, the optimal estimator for the Gaussian vector can be proved to be a modified form of divisive normalization using a weighted $L2$-norm. In this paper, we also fix the value of α to 2 as commonly suggested in other works (Table 1).

The network structure is shown in Fig. 1. It has a minor modification to the described structure in the work of [3, 9]. Note that the small convolutional kernels can have the same receptive field as large convolutional kernels. For example, a 5×5 convolutional layer can be replaced by two 3×3 convolutional layers. Therefore, we used two 5×5 convolutional layers to replace the original 9×9 convolutional layer, positioning them at the first layer of analysis transform and the last layer of synthesis transform, respectively. The reason why we did not directly use 3 × 3 convolutional layers is that the further decomposing will introduce more layers and result in a rather deep network structure because decomposing larger filters into smaller ones need more layers for computation. As Zhou et al. [11] suggested, a deep convolution auto-encoder is not good at capturing the long-range global dependencies between features without residual-like block. Hence, we used four 5 × 5 convolutional layers with 128 kernels and strides 2 for the encoder part, which down-sampling the size of images to 1/16 with 128 channels. Then, we used another four 5 × 5 deconvolutional layers with 128 kernels (3 kernels at output layer) and strides 2 for the decoder part, which up-sampled the features to their original size with three channels.

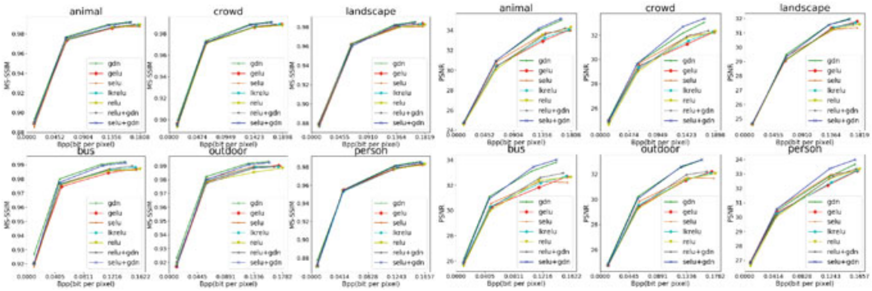
We also try to study the model behavior when applied to six different scenes, including animal, crowd, landscape, monitor, outdoor, and person. There are two kinds of evaluation metrics used to measure the quality of resulting (decoded) image, i.e., PSNR and MS-SSIM. The horizontal axis has four kinds of BPP derived from four values of λ (0.001, 0.01, 0.05, 0.1). As mentioned above, a higher λ generates better quality and higher BPP. Bits per pixel (BPP) is also called bitrate, which means the required bits for each pixel of images. It represents the size of encoded image.

4 Experimental Results

We empirically study the effects of various activation functions in six different scenarios. The differences are revealed in visual effects and evaluation metrics.

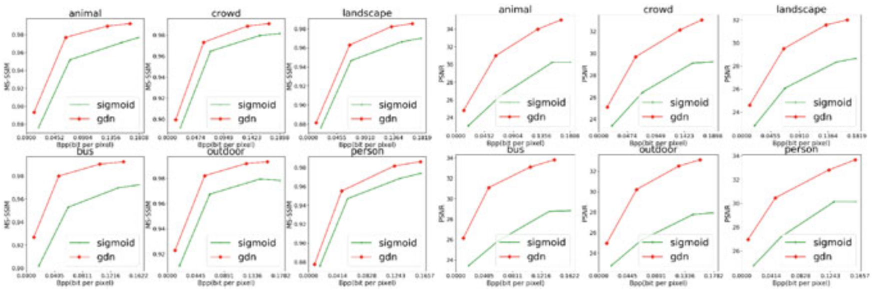
Figure 2a demonstrates the decoding image quality of different models measured by MS-SSIM. Generally, with the bitrate increasing, MS-SSIM of different models are also increasing but the growth is slowing down. In addition, GDN shows the dominant performance over all scenes comparing to rest activation functions. Only SeLU + GDN is as good as GDN over some scenes by the cost of extra computation. Specifically, over the animal, crowd, landscape, and people scenes, the separation among different activation functions is smaller than the monitor and outdoor scenes. ReLU, SeLU, GeLU, leaky ReLU, and ReLU + GDN are almost overlapping over those four scenes, which means their performance is similar. However, the lines of GDN and SeLU + GDN show a distinctive trend that they are always above the others. Furthermore, since the horizontal axis represents bitrate, we can see that with the same λ, GDN, and SeLU + GDN have less bitrate, which indicates their compression ability is stronger.

The situation is slightly different while using PSNR metric. In Fig. 2b, SeLU + GDN performs better than GDN over most scenes and the separation is distinguishable. According to [6], the normalization occurs inside the SeLU activation function, which is absent from original auto-encoder. Therefore, it can deal with gradient issues during backpropagation. It may be the reason SeLU + GDN performing slightly better than GDN, and SeLU performing relatively better than ReLU, leaky



(a) (b)

Fig. 2 MS-SSIM and PSNR comparison of different activation functions based on six images



(a) (b)

Fig. 3 MS-SSIM and PSNR comparison between GDN and Sigmoid using six scenarios. Bitrate comparison of using various activation functions

Table 2 Bitrate comparison of using various activation functions

	Animal	Crowd	Landscape	Bus	Outdoor	Person
GDN	**0.129**	**0.135**	**0.133**	**0.106**	**0.123**	0.122
SeLU + GDN	0.130	0.136	0.133	0.108	0.125	0.122
SeLU	0.140	0.147	0.141	0.169	0.135	0.124
ReLU	0.136	0.141	0.137	0.116	0.132	0.121
leaky ReLU	0.139	0.145	0.140	0.118	0.135	0.124
ReLU + GDN	0.140	0.142	0.139	0.118	0.133	0.127
GeLU	0.136	0.143	0.137	0.116	0.132	0.120

ReLU, etc. Empirically, the performance gap will become more and more distinct when the training step increases.

We have presented the effectiveness of various activation functions in image compression. The results show that the GDN has excellent performance on encoding bitrate and decoding quality (the lower the better) as indicated by Table 2. Lower bitrate requires less transmission bandwidth for communication. As for the remaining activation functions, the SeLU + GDN has the closest performance to GDN, whose encoding bitrate is slightly worse and decoding quality is sometimes even better than GDN as depicted in Fig. 2.

As a classic activation function, Sigmoid is frequently used. Thus, whether GDN is still better than Sigmoid $\left(\frac{1}{1+e^{-x}}\right)$ in image compression deserves some experiments. As Fig. 3 indicates, using GDN yields a much better PSNR value than Sigmoid, and it is similar for the measure of MS-SSIM that is neglected for plotting due to page limit.

5 Conclusion

In this study, we focus on comparing the effectiveness of various activation functions used in the convolutional neural networks. The evaluation is investigated via six distinct images based on PSNR and MS-SSIM metrics. Experimental results suggest that the GDN is, in general, a great choice in deep learning-based image compression. In specific, using GDN yields the best MS-SSIM. Combined with SeLU, GDN shows even slightly better performance than applying GDN alone in terms of MS-SSIM according to the experimental results. Therefore, both measures indicate that the use of GDN in deep image compression should provide better compression results. The underlying reason might partly because of the normalization in GDN. For future work, it might be a good idea to study the effectiveness of GDN in deep learning-based video compression. The current research work only sheds some light on image compression, to inspire the community.

References

1. Agustsson E, Tschannen M, Mentzer F, Timofte R, Van Gool L (2019) Generative adversarial networks for extreme learned image compression. In: Proceedings of the IEEE international conference on computer vision, pp 221–231
2. Ballé J, Laparra V, Simoncelli EP (2016) Density modeling of images using a generalized normalization transformation. In: 4th international conference on learning representations, ICLR
3. Ballé J, Laparra V, Simoncelli EP (2016) End-to-end optimized image compression. arXiv:1611.01704
4. Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. Springer, pp 184–199

5. Ignatov A, Timofte R et al (2019) PIRM challenge on perceptual image enhancement on smartphones: report. In: European conference on computer vision (ECCV) workshops, Jan 2019
6. Klambauer G, Unterthiner T, Mayr A, Hochreiter S (2017) Self-normalizing neural networks. In: Advances in neural information processing systems, pp 971–980
7. Luo S, Yang Y, Yin Y, Shen C, Zhao Y, Song M (2018) DeepSIC: deep semantic image compression. In: International conference on neural information processing. Springer, pp 96–106
8. Minnen D, Ballé J, Toderici GD (2018) Joint autoregressive and hierarchical priors for learned image compression. In: Advances in neural information processing systems, pp 10771–10780
9. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
10. Toderici G, O'Malley SM, Hwang SJ, Vincent D, Minnen D, Baluja S, Covell M, Sukthankar R (2016) Variable rate image compression with recurrent neural networks. In: The international conference on learning representations
11. Zhou L, Sun Z, Wu X, Wu J (2019) End-to-end optimized image compression with attention mechanism. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops

Sentiment Analysis of Movie Reviews Using Machine Learning Techniques

Duc Duy Tran, Thi Thanh Sang Nguyen, and Tran Hoang Chau Dao

Abstract Sentiment analysis is the interpretation and classification of emotions and opinions from the text. The scale of emotions and opinions can vary from positive to negative and maybe neutral. Customer sentiment analysis helps businesses to point out the public's thoughts and feelings about their products, brands, or services in online conversations and feedback. Natural language processing and text classification are crucial for sentiment analysis. That means we can predict or classify customers' opinions given their comments. In this paper, we do sentiment analysis in the two different movie review datasets using various machine learning techniques including decision tree, naïve Bayes, support vector machine, blending, voting, and recurrent neural networks (RNN). We propose a few frameworks of sentiment classification using these techniques on the given datasets. Several experiments are conducted to evaluate them and compared with an outstanding natural language processing tool (Stanford CoreNLP) at present. The experimental results have shown our proposals can achieve higher performance, especially, the voting and RNN-based classification models can result in better predictions.

Keywords Sentiment analysis · Opinion mining · Classification · Blending · Voting · Stanford CoreNLP

D. D. Tran · T. T. S. Nguyen (✉) · T. H. C. Dao
School of Computer Science and Engineering, International University, VNU-HCMC, Ho Chi Minh City, Vietnam
e-mail: nttsang@hcmiu.edu.vn

D. D. Tran
e-mail: ITITIU14127@student.hcmiu.edu.vn

T. H. C. Dao
e-mail: dthchau@hcmiu.edu.vn

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_34

1 Introduction

Nowadays, social media platforms, such as Facebook, YouTube, Flickr, Instagram, and Pinterest allow customers to share their opinions, comments, views, feelings, and judgments about various topics, ranging from education to entertainment. These platforms contain a large amount of data in many different forms like tweets, blogs, pictures, status updates, posts, icons, etc. Sentiment analysis, also known as opinion mining, is a text analysis method that detects "polarity" between positive and negative opinions in a document, paragraph, sentence, or clause [1]. Positive states can be happiness, excitement, joy, hope, inspiration, and so on. In contrast, negative states can be defined as sorrow, hatred, anger, grief, etc. Opinion mining deals with the complexity of slang words, misspellings, short forms, repeated characters, linguistic variation, and continually created emoticons. Sentiment score is determined based on individual or compound words, and sentence structure. In addition, the explosion of data makes it almost impossible to manually analyze all user-generated content. Therefore, machine learning methods are necessary to handle these issues. The following will present related works in Sect. 2, our research methodology in Sect. 3, experimental results in Sect. 4, and some conclusions in Sect. 5.

2 Related Works

2.1 Opinion Mining

As known, opinion mining or sentiment analysis is a branch of data mining that applies text mining techniques to analyze customer reviews for the purpose of marketing. Song and Chambers [2] provided numerous techniques of sentiment analysis, classification, and prediction, such as Ling Pipe, CoreNLP, and SentiWordnet, in which a dataset of 2000 comments was used to test the models. The research rather focused on sentiment analysis at the sentence level than the document level. Classification algorithms play an important role in text mining. For example, support vector machine was used to build a sentiment system [3] based on the Weka library [4] for English comments in various categories, such as shoes, handbags, rings, and dresses. Their datasets were taken from Amazon and divided into a trainset (300 comments) and a test set (400 comments). FastVector was applied to create a specific dataset for sentiment analysis and a confusion matrix was used to compare the machine learning method with the truth sentiment. The results of this system showed that the precision, recall, and F1 score of positive comments were 89.3, 95, and 92.1%, respectively; those of negative comments were 97.1, 78.5, and 81.2%, respectively; and neutral comments were 76.7, 86.2, and 81.2%, respectively.

By discovering Weka [4], we can improve the prediction accuracy by combining the results of multiple algorithms together [5]. These ensemble methods can be boosting, bagging, and blending. The accuracy of boosting is often higher than bagging. Moreover, deep learning methods have emerged to contribute significantly to data mining. Lang et al. [6] mentioned the idea of building a network consisting of a long short-term memory (LSTM) [7] layer and a recurrent neural network output layer (RnnOutputLayer), which would be of great help in text mining. The dataset can be loaded and mapped to a sequence of vectors in the embedding space that is defined by the Google News model. These data processing techniques are supported in WekaDeeplearning4j [6], which is available for use.

2.2 Stanford CoreNLP

Stanford CoreNLP [1], a Java natural language analysis library, makes the job of linguistic analysis easier by integrating all natural language processing (NLP) tools, including the part-of-speech (POS) tagger, the named entity recognizer (NER), the parser, the co-reference resolution system, and the sentiment analysis. For sentiment analysis, the Stanford team grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences and presents new challenges for sentiment compositionality [8]. To address them, they introduce the recursive neural tensor network. While most sentiment prediction systems work just by looking at words in isolation and ignoring the order of the words in a sentence, their new deep learning model builds up a representation of phrases by word vectors to determine positive and negative points of longer word groups.

2.3 Machine Learning Methods

Classification algorithms in machine learning play an important role in classifying text data objects into different sentiment classes. Therefore, this study considers some popular classification algorithms [9] from basic to advanced, such as naïve Bayes, support vector machine (SVM), decision trees, and recurrent neural networks (RNN). Moreover, some ensemble methods, such as blending and voting, are considered to leverage the classification performance of assembling algorithms.

Naïve Bayes is well-known because of its simplicity and outperforming compared to some other more sophisticated classification techniques [9]. It is used to calculate the probability of each word in the text/sentence to estimate the probability of a sentence belonging to a class. The SVM [9] is one of the most popular and successful binary supervised classification methods. However, it has got the problem of quadratic programming when training the classifier model. Therefore, sequential minimal optimization was introduced to solve this problem [10]. In this study, we use SMO instead of SVM. Decision trees [9] construct a tree structure in a top-down. They are good to select the best attributes for classification but quite time-consuming. In Weka [5], J48 is implemented for decision trees. More advanced, we have RNNs which are a class of neural networks, mostly used in the fields of NLP and speech recognition [7].

Besides, the ensemble methods of blending and voting [6] are useful for utilizing the advantages of single algorithms and make more effective learning models. In blending, multiple different algorithms are prepared on the training data and a meta classifier is prepared to learn how to take the predictions of each classifier and make accurate predictions on unseen data. On the other hand, the voting classifier [6] can be trained by several different machine learning algorithms. Each sub-model (algorithm) makes predictions which are then compared and combined in some way, such as, taking the mean or the mode of the predictions, allowing each sub-model to vote on what the outcome should be.

3 Methodology

This project was implemented in Java and uses the Weka 3.8 library [5] for the naïve Bayes, decision tree, SMO, logistic regression, blending, and voting classifiers. Besides, RNNs and SentimentCoreNLP models were developed by using Wekadeeplearning4j [7] and Stanford CoreNLP [8], respectively.

3.1 The Proposed Frameworks of Classification

Figures 1, 2, and 3 depict the proposed frameworks of classification and evaluation

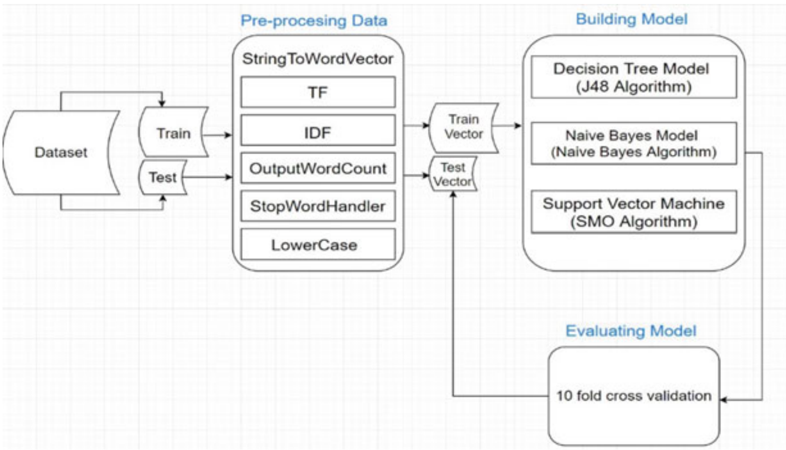


Fig. 1 Framework of classifier models using decision tree, naïve Bayes, SMO

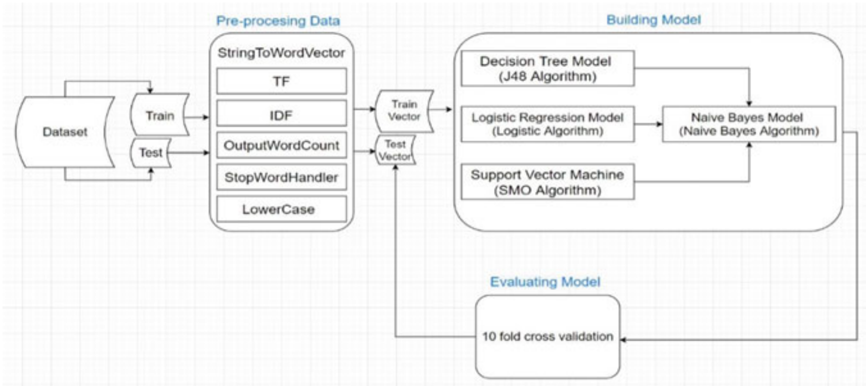


Fig. 2 Framework of blending classifier model

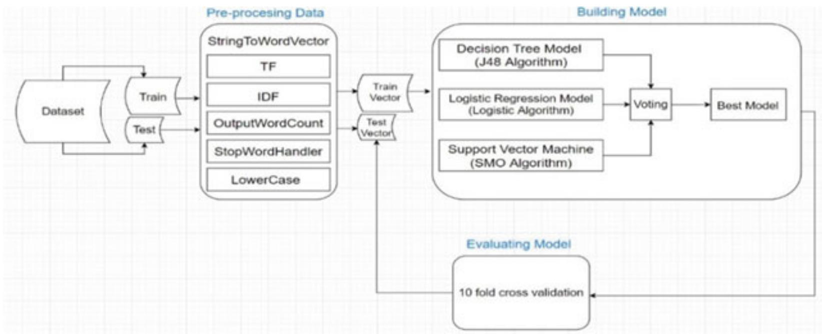


Fig. 3 Framework of voting classifier model

for sentiment analysis. There are three processing units: pre-processing data, building model, and evaluating model. These processing units are described as follows.

Step 1: Pre-processing (cleaning) the data, we do:

- Vectorizing string data (StringToWordVector): an unsupervised filter that converts string attributes into a set of numeric attributes representing word occurrence in the text strings. This job is performed by calculating term frequency–inverse document frequency (TF-IDF) for words in text data to construct the term-document matrix.
- Controlling the number of words (output) during the training process, namely, OutputWordCount.
- Removing stop words, e.g., “a”, “the”, “an”, “is”, “are”, etc. (StopWordHandler)
- Converting all the text data to lowercase to minimize the corpus size (LowerCase).

Step 2: Building the classifier models. The first framework (Fig. 1) includes the classification algorithms of decision tree, naïve Bayes, and SMO. The second framework (Fig. 2) applies the blending method, in that, the algorithms used in the base classifiers are decision tree, regression logistic, and SMO; and the meta classifier is naïve Bayes. The third framework (Fig. 3) uses the voting method, in that, the used classification algorithms are decision tree, naïve Bayes, and SMO.

Step 3: Evaluating the classifier models using the tenfold cross-validation.

3.2 The Proposed Framework of RNN-Based Classification

In this proposed framework, the RNN-based classification algorithm was built. To implement the RNNs, Weka Deeplearning4j and the Google News models are utilized. Figure 4 depicts the framework of the RNN-based classifier model and classifier evaluation in three steps.

Step 1: Pre-processing (vectorizing) data. The RNN pre-processing configuration is divided into four layers. The first input layer configuration is used to refine and vectorize the text data. In this layer, the GoogleNews Embeddings is used to map the dataset to a sequence of vectors in the embedding space that is defined by the Google News model. The second hidden layer uses LSTM [1] for text analysis. In this layer, the used activation function is the tanh function. The third output layer was the RNN output layer. The last layer is the process of gradient normalization.

Step 2: Building the classifier model. Some parameters of RNN are set, e.g., seed, the number of epochs, and backpropagation through time.

Step 3: Evaluating the model using the tenfold cross-validation.

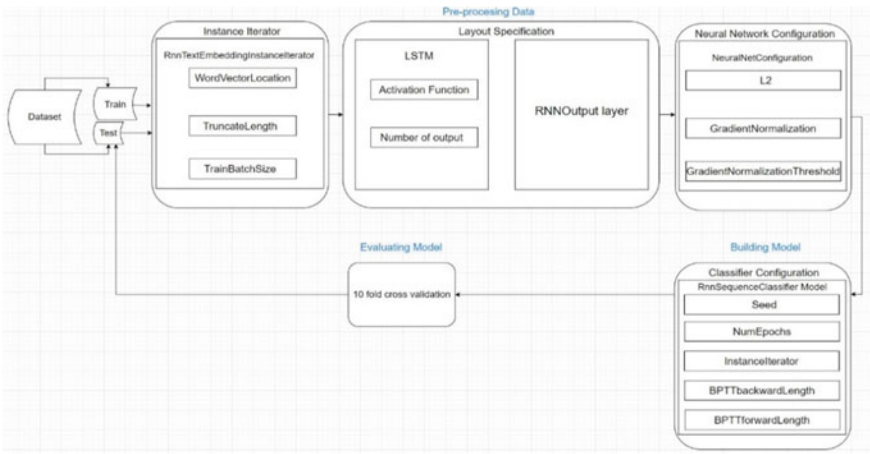


Fig. 4 Framework of RNN-based classifier model

3.3 *The Framework of Stanford CoreNLP*

To further evaluate the proposed frameworks, we compared their sentiment analysis results with the results of Stanford CoreNLP. Since Stanford CoreNLP just performs sentiment analysis at the sentence level, the sentiment sum of sentences in a document is divided by the document size (the number of sentences) to get the average sentiment for the document level. For instance, the sentence's sentiment score was defined as 0 = very negative, 1 = negative, 2 = neutral, 3 = positive, and 4 = very positive. The comment's sentiment (sentiment average) score was defined in a range as: [0; 1) = very negative, [1, 2) = negative, 2 = neutral, (2, 3] = positive, and (3, 4] = very positive. If a comment's size was nine sentences and its sentiment sum was 27 scores, the sentiment average score would be 27/9 = 3, meaning positive. In this way, we could obtain the sentiments of testing sentences for validation using a confusion matrix.

4 Experimental Results

Several experiments were carried out to evaluate and compare different machine learning methods applied to sentiment analysis in different datasets. The studied methods, naïve Bayes, decision tree, SMO, logistic regression, RNN, blending, and voting, were used to train the classifier models and are compared with the existing sentiment analysis model, i.e., Stanford CoreNLP.

Two datasets are used in the experiments. The first one was collected from the Movie Review database at http://www.cs.cornell.edu/people/pabo/movie-review-data/, including 800 comments (400 positive comments and 400 negative comments), namely Movie database. The second one was from the sentiment treebank of Stanford CoreNLP [2], including 800 comments (400 positive comments + 400 negative comments), namely CoreNLP database. Both datasets were divided into a trainset (600 comments = 300 positive ones + 300 negative ones) and a test set (200 comments = 100 positive ones + 100 negative ones) for validation, and then converted into the ARFF format by using Text Directory Loader in the Weka library. Each comment in the CoreNLP database contains only one sentence, while each comment in the Movie database has more than one sentence.

The performance evaluation results are presented in Tables 1–4. These results were made by computing the averages of precision, recall, and F1 score values for each algorithm. For the Movie database, voting method yielded the best classification results (Tables 1 and 2) while RNN yielded the best results in the CoreNLP database (Tables 3 and 4).

Table 2 shows that the voting method outperforms the other methods.

Table 4 shows that the RNN-based classification model overcomes the other models in terms of accuracy.

Table 1 The precision, recall, F1 score of various algorithms for the Movie dataset

	CoreNLP	J48	Blending	NB	SMO	Voting	RNN
Precision	0.965	0.955	0.965	0.975	0.970	**0.980**	0.875
Recall	0.890	0.955	0.965	0.975	0.970	**0.980**	0.870
F1 score	0.925	0.955	0.965	0.975	0.970	**0.980**	0.870

Table 2 The accuracy of various algorithms for the Movie dataset

	CoreNLP	J48	Blending	NB	SMO	Voting	RNN
Accuracy (%)	89	95.5	96.5	97.5	97	**98**	87

Table 3 The precision, recall, and F1 score of various algorithms for the Stanford CoreNLP dataset

	J48	NB	SMO	Blending	Voting	RNN
Precision	0.560	0.585	0.510	0.550	0.570	**0.720**
Recall	0.545	0.580	0.510	0.550	0.565	**0.715**
F1 score	0.515	0.575	0.510	0.550	0.560	**0.715**

Table 4 The accuracy of various algorithms for the Stanford CoreNLP dataset

	J48	NB	SMO	Blending	Voting	RNN
Accuracy (%)	54.5	58.0	51.0	55.0	56.5	**71.5**

5 Conclusions

This study has shown that the voting algorithm is able to give the best classification results in the Movie database (one comment has more than one sentence). It means that the voting method is more suitable in complex databases. In contrast, recurrent neural networks can provide the best analysis results for single sentences (one comment has just only one sentence). As we can see, RNN is more suitable in the simple database, but it can be adjusted to be better.

To sum up, sentiment analysis remains the best tool for gaining critical insight into various data and automating processes. It can open up the gold mines in the customer's opinion and help businesses automate processes while increasing sales.

References

1. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60

2. Song M, Chambers T (2014) Text mining with the Stanford CoreNLP. In: Ding Y, Rousseau R, Wolfram D (eds) Measuring scholarly impact. Springer, Cham
3. Ngo VM (2014) Machine learning based English sentiment analysis. J Sci Technol 52(4D):142–155
4. Weka (2020) The workbench for machine learning. https://www.cs.waikato.ac.nz/ml/weka/. Accessed 15 Nov 2020
5. Brownlee J (2014) Make better predictions with boosting, bagging and blending ensembles in Weka. In: Weka machine learning. https://machinelearningmastery.com/improve-machine-learning-results-with-boosting-bagging-and-blending-ensemble-methods-in-weka/. Accessed 28 Feb 2014
6. Lang S, Bravo-Marquez F, Beckham C, Hall M, Frank E (2019) WekaDeeplearning4j: a deep learning package for Weka based on Deeplearning4j. Knowl Based Syst 178:48–50
7. Salton G, Yang CS (1973) On the specification of term values in automatic indexing. J Doc 29:351–372
8. Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. Stanford University, Stanford, CA 94305, USA, pp 1–12
9. Han J, Kamber M, Pei J (2012) Classification: basic concepts data mining, chap. 8, 3rd edn. Morgan Kaufmann, Boston, pp 327–391
10. Flake GW, Lawrence S (2002) Efficient SVM regression training with SMO. Mach Learn 46(1):271–290

Discovering Associative Patterns in Healthcare Data

Diego de Castro Rodrigues, Vilson Siqueira, Fabiano Tavares, Márcio Lima, Frederico Oliveira, Lucas Osco, Wilmar Junior, Ronaldo Costa, and Rommel Barbosa

Abstract Health care has several knowledge discovery techniques. Among them are association rules, which provide quick access to standards. However, classic algorithms can generate many patterns or fail to identify rare cases relevant to healthcare professionals. This study identified asymmetric associative patterns in health-related data using the Health Association Rules (HAR) algorithm. We use a combined strategy of six metrics to filter, select, and eliminate contradiction steps to find patterns and identify possible rare cases. The proposed solution uses adjustment mechanisms to increase the quality of standards with knowledge of the health professional. The HAR assists health researchers and decision support systems. A survey of 597 studies identified the primary needs and problems of associative patterns in the health context. The HAR identifies characteristics with the highest cause and effect relationship. The experiments were carried out on 13 datasets, where we identified the most pertinent patterns for the datasets without losing relevant knowledge.

Keywords Medical data mining · Association rules · Health care · Asymmetric association rule · Data mining

1 Introduction

Numerous diseases have become obsolete with the development of medical technologies. Data analysis related to health care has combined medical knowledge with data mining technologies [4, 8, 11]. These studies indicate that it is vital to ensure safe analysis of patterns discovered in medical data. However, with the continu-

D. de Castro Rodrigues (✉) · V. Siqueira · F. Tavares
Federal Institute of Tocantins, Palmas, Brazil
e-mail: diego.rodrigues@ifto.edu.br
URL: http://www.ifto.edu.br

D. de Castro Rodrigues · V. Siqueira · M. Lima · F. Oliveira · R. Costa · R. Barbosa
Federal University of Goiás - UFG, Goiânia, Brazil

L. Osco · W. Junior
University of Western São Paulo - UNOESTE, São Paulo, Brazil

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_35

ous growth in the volume of information and analysis techniques (Machine learning, Data Mining, Medical Data Mining), an imbalance has emerged among data demand, analysis capacity, and pattern discovery for specific contexts in the dataset of health. Therefore, data mining techniques are becoming more generic to analyze any data, leading to dissatisfaction in finding specific patterns and higher costs of time and knowledge to identify trends.

The association rules demonstrate simplicity to understand the standards obtained in fields such as engineering [10], recommendation systems [6], and clinical diagnostics [5]. The discovery of associative patterns is one of the main tasks of data mining, with emphasis on the use of association rules based on the **Apriori model** in health care [3]. Elham Buxton [2] presents some of the limitations pointed out in the classic associative rules algorithms, initially proposed by Agrawal [1]. The limitations of traditional algorithms have insufficiency related to the amount of generated standards, redundancies, selection of better rules, and elimination of standards.

This article presents a study conducted on data from a systematic review and experiments on real datasets. We developed an associative analysis algorithm applied to the healthcare context. The algorithm identifies patterns using probabilistic metrics, pruning, filters, and custom metrics. The objective of this study is to identify asymmetric associative patterns in data related to health care, selecting the associative patterns based on a set of probabilistic asymmetric metrics. Thus, obtaining better results in identifying trends, selection, and ranking, valuing the casual relationship, and identifying possible rare patterns. Some highlights of the study: (1) Health Association Rules Algorithm (HAR); (2) Use of alternative metrics to the Support/Confidence model; (3) Identification of rare patterns, and; (4) Application in real databases.

2 Approach of Method

To identify the best strategies for the research problem and build the HAR algorithm, we conducted a systematic review of the literature with articles that applied association rules algorithms to medical data published between the period 2015 to 2019.

The articles' selection was carried out in six bases: Science Direct, PubMed, ACM, Springer Link, IEEE, and Google Scholar. Initially selecting 597 papers and after applying the inclusion and exclusion criteria, and chose 51 studies. The systematic review protocol, as well as its details, are available in their entirety in the supplementary material (https://cutt.ly/3fQQw6E).

Based on the Apriori algorithm, the Health Association Rules (HAR) presents customized processes and adds steps in the original operation of the Apriori algorithm. Figure 1 indicates the levels of the HAR and their interactions.

Process I analyzes the database in the same way as the Apriori algorithm, maintaining its original functioning in this process. Process II, the value of minimum support and minimum confidence is calculated by $\frac{1}{N}$, where N is the number of ele-

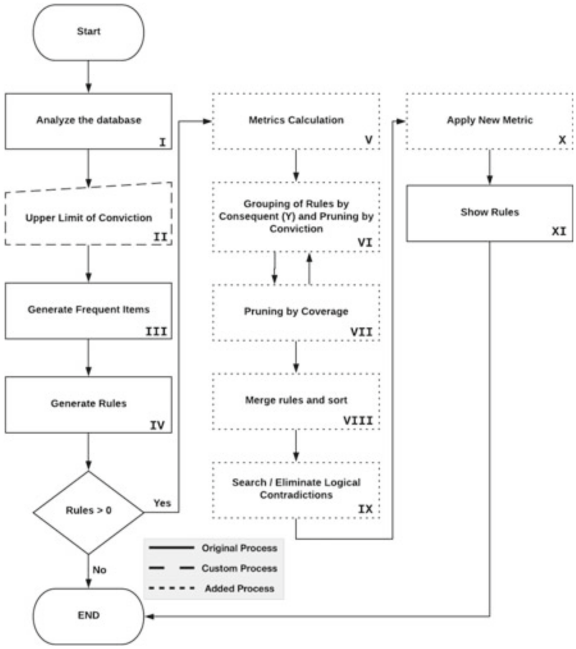


Fig. 1 HAR operation flowchart

ments in the database. Thus, the lowest possible value of Support will be assigned to the dataset, thus obtaining the most frequent items and association rules. The upper limit is defined, and these are informed through the user of the HAR algorithm. Based on the *Conviction* metric, which varies between 0 and ∞.

The generation of frequent items, Process III, occurs similarly to Apriori, the difference is in using the minimum values of *Support* obtained automatically in Process II. It is thus generating the set of frequent items.

Process IV uses the set of frequent items from the previous step together with the minimum *Confidence* value to generate the set of association rules. In the original algorithm, only *Confidence* is calculated for each of the rules. However, the V Process is added to the HAR, which calculates the metrics for all association rules.

Pruning by *Conviction* Process VI, uses the upper limit determined in Process II and a lower limit, all association rules that do not respect the established values are eliminated. The pruning process's remaining rules are grouped in Process VI, where all the rules $X \rightarrow Y$ are separated into subsets of rules. The lower limit of *Conviction* is determined by the arithmetic mean of each of the subsets of Consequents.

Each subset is organized by its consequent Y value. They were dividing the set of rules resulting from Process IV and VI into N subsets. The top rules of each of the subsets are selected and checked with the other rules, recursively looking for rules covered by another, until all the rules and subsets are checked.

Based on the study Toivonen [9], which proposed the idea of eliminating redundant rules using structural rule coverage, Process VII selects in each subset all rules with the first X antecedents. As an example, the subset [*Dengue* = *yes*] with the rules [***Pain = strong*** and PainEyes = yes → Dengue = yes], displays two antecedents the second rule with a single antecedent [***Pain = strong*** → dengue = yes].

The rules with their first identical antecedents are selected and compared for presenting similar information, regardless of the second antecedent X_2 [*PainEyes* = *yes*] the first antecedent X represented in the two rules by [***Pain = strong***] always presents a relationship with [*Dengue* = *yes*].

Thus, the rules with the lowest average between *Hyper-Confidence* and *Mutual Information* are eliminated. The first metric ensures that rules are chosen with the least chance of being generated randomly. The second one measures the information gain of the consequent Y provided through the antecedent X.

The search and elimination of logical contradictions were implemented at HAR in the IX Process, seeking contradictions of meaning. The contradiction of meaning is determined in rules [Aches = strong → Dengue = yes] and [Dengue = yes → Aches = strong]. In both situations, it is not trivial to define which could be eliminated.

The average between the *Confidence* and *Kalczynski* metrics verifies the slope patterns that take into account the relationship of $X \rightarrow Y$ and $Y \rightarrow X$. Applying these metrics to choose the rules in contradictions of sense, selecting the rule with the highest slope value (average), and eliminating the other.

The X Process displays the Difference from Sample Means (DMA), which orders the rules with the greatest asymmetric relationship. The set of final association rules is displayed in the XI process, accompanied by its metrics and ordered. following the pattern of $X \rightarrow Y$ with the measurement values *Hyper-Confidence*, *Mutual Information*, *Imbalance Ratio*, *Kalczynski*, *Gini Index*, and *DMA*.

2.1 Evaluation of the Proposed Method

The HAR method consists of four steps. In the first step, the Dataset is provided as an input; The second step (Algorithms), runs the classic Apriori algorithm for the generation of association rules. It also executes the HAR method with a configuration similar to its standard objective metrics; In the third step, (Individual Result) compares the rules generated by each step algorithm (Algorithms). The results of each algorithm are analyzed in the step (Analytics) employing objective metrics to understand the reasons for a hypothetically good rule, not being selected in the HAR or the classic algorithm.

2.1.1 Custom Measure

The ideal rule is composed of values of objective metrics (*Hyper-Confidence*, *Gini Index*, *Mutual Information*, *Imbalance Ratio*, *Kalczynski*) with the respective default values (0.95, 0.3, 1, 1, 0.6) that together define the rule (orange line) with the greatest potential for the data context. The gray lines show the behavior of the rules discovered in the HAR. When comparing the (ideal) rule with the HAR rules, it is possible to rank the best rules not in two metrics as in the classic algorithm but a set of six rules.

The customized measurement is performed employing the distance from the rules, which is calculated by the Difference of the Sample Means (DMA) introduced by [7]. We use DMA to calculate the distances from the ideal rule with the dataset rules (Eq. 1).

$$DMA = \bar{X_i} - \bar{M} \tag{1}$$

The $\bar{X_i}$ indicates the average of the Hyper-Confidence, Gini Index, Mutual Information, Imbalance Ratio, and Kulczynski metrics. The arithmetic mean of the ideal rule is defined by $\bar{M}$, the closer to 0 the DMA, the better the rule ranking in the HAR.

The standard values of the metrics for calculating the DMA are defined to value the relationship $X \rightarrow Y$ such that the asymmetric relationship X and Y is shown. The default values can be customized to meet particulars of the dataset when necessary, DMA is used to rank the rules.

2.1.2 Data Organization

The dataset is composed of data from health care, Parkinson's disease, heart disease, physiological complexity, mental health, and frequency of disorders. The databases chosen for the experiments in this study aim to diversify the tests, including related unidentified data on the health (Table 1).

3 Results and Discussions

We present the behavior of the patterns of the datasets of Table 1 individually. The datasets (01, 02, 05, 07, 08, and 12) stand out for presenting rules with a low value of *Hyper-Confidence*. This behavior is justified by low Support values found in real datasets (Fig. 2).

We believe that the results presented through the execution of the HAR are related to its harmonious functioning with different metrics to validate relations between $X \rightarrow Y$. We present a reduced number of rules compared to the classic Apriori (Fig. 3).

When a dataset presents values close to zero in most of the metrics used, it indicates that the dataset in question is not appropriate for associative patterns. The datasets (01, 02, 05, 07, 08, and 12) also indicate that these data may present rare patterns, which would be eliminated by the classic algorithm due to the low support values (Fig. 4 Rules Distribution—Datasets 01 and 13).

Featured datasets, regardless of the number of instances, can generate a large number of patterns. Figure 3 presents information about the datasets and their execution in the Classic and HAR algorithm.

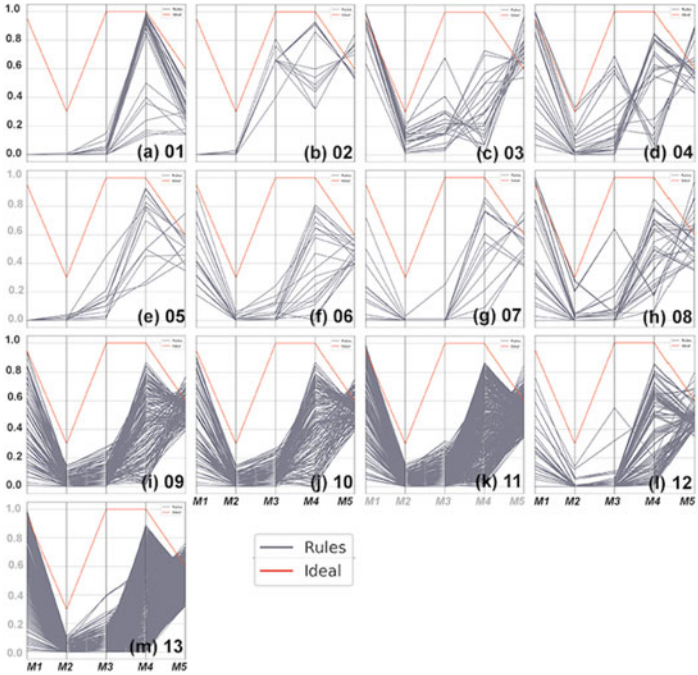


Fig. 2 HAR experiments (acronyms: M1—*Hyper-Confidence*, M2—*Gini Index*, M3—*Mutual Information*, M4—*Imbalance Ratio*, M5—*Kalczynski*)

Fig. 3 Dataset

Table 1 Datasets

Id	Domain	Instance	Attributes	Name	Source
01	Clinical data	24978	7	eICU collaborative	PhysioNet
02	Intensive therapy	1761	5	MIMIC-III clinical	PhysioNet
03	Parkinson's disease	16	52	Parkinson disease	PhysioNet
04	Pharmacology	1934	30	CiPA ECG	PhysioNet
05	Oxygen saturation	36	6	Oxygen saturation	PhysioNet
06	Physiology	196	21	Tai Chi, physiological	PhysioNet
07	ECG	4211	33	ECG effects	PhysioNet
08	Electrophysiological	5232	32	ECG Ranolazine	PhysioNet
09	Heart disease	303	14	Heart disease UCI	Kaggle
10	Heart disease	1025	14	Heart disease dataset	Kaggle
11	Disorders	1259	27	Mental health	Kaggle
12	Healthcare	43400	12	Stroke data	Kaggle
13	Healthcare	649	33	Student alcohol	Kaggle

Datasets 03, 04, and 13 showed the highest number of rules generated in the Classic algorithm and the HAR, even without being the sets with higher instances as highlighted in the upper part of Fig. 3. When comparing the classic algorithm with the HAR, we noticed that the number of generated rules had a strong relationship with the number of attributes (characteristics) present in each dataset.

We highlight datasets 01 and 13 for an individual analysis of its results in the classic algorithm and the HAR. All rules obtained in dataset 01 using the Classic algorithm Fig. 4a with a Support value of up to 10% have their knowledge to the left of the line eliminated in the classic approach. Studies with medical data presented in the supplementary material (https://cutt.ly/3fQQw6E). Use traditional algorithms and, by definition, need to determine a minimum Support value, so part of the knowledge can be eliminated.

The HAR in Fig. 4b is composed of different metrics, such as *Conviction, Hyper-Confidence, Gini Index, Mutual Information, Imbalance Ratio, Kulczynski, and DMA*, seeks to identify the standards regardless of their occurrence. Thus, the HAR can identify common and rare patterns in the dataset, always seeking the most balanced rules.

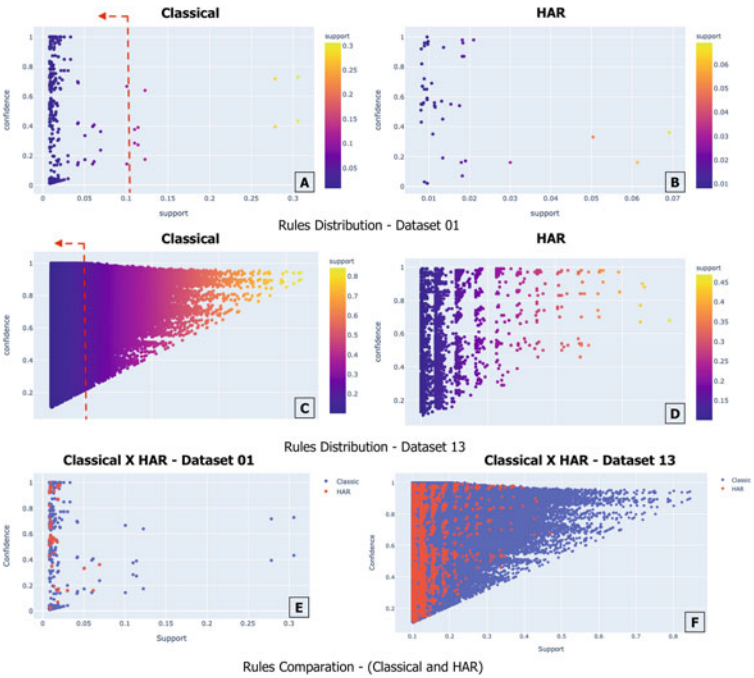


Fig. 4 Comparisons

When looking at Fig. 4a we notice that the classic selected rules with Confidence values less than 80% even the dataset showing situations with 100% Confidence, this occurrence is due to the Support being the main pattern cutting factor in the classic algorithm. In Fig. 4, it is highlighted in (e and f) the rules selected by HAR in comparison to the classic algorithm.

The behavior of dataset 13 was presented because it is more significant in the number of rules. In this way, with a support limit of 20%, all knowledge to the left of the red line would be eliminated using the classic algorithm (Fig. 4c).

When it comes to healthcare data, discarding patterns may not be an exciting solution, so HAR (d) searches for relevant patterns and uses metrics to validate the real implication among associative patterns.

In Fig. 4d, it is possible to observe the points with the best-ranked rules. The classic algorithm rules have a very high support value, which can be expected or even irrelevant standards when observed only the Support/Confidence model.

By performing filters in different stages and using a set of metrics to validate the standards found, the HAR selects the rules of the dataset closest to the ideal through customized objective metrics. It is taking into account the particularities of each dataset. Identifying and selecting patterns that could be discarded when compared to a classic approach.

4 Conclusions

The algorithm (HAR), selects associative patterns in a set of data, for this, it chooses the best rules of the group of metrics, in order to identify the most appropriate relationship of the Antecedent (X) and Consequent (Y) in associative analyzes. HAR seeks to find more balanced rules through the composition of six metrics (Hyper-Confidence, Conviction, Gini Index, Mutual Information, Imbalance Ratio, Kulczynski, and DMA). Together, select standards that value knowledge from the database, identifying rare and most common patterns, and eliminating redundancies and contradictions. Our algorithm values the meaning of the $X \rightarrow Y$ implication and eliminates potentially uninteresting rules, generating a smaller set of rules. As it is not limited to the Support/Confidence model of the classic algorithm (Apriori), HAR does not eliminate knowledge in real datasets, which may have a low Support value.

References

1. Agrawal R, Srikant R et al (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on very large data bases, VLDB, vol 1215, pp 487–499
2. Buxton EK, Vohra S, Guo Y, Fogleman A, Patel R (2019) Pediatric population health analysis of southern and central Illinois region: a cross sectional retrospective study using association rule mining and multiple logistic regression. Comput Methods Programs Biomed 178:145–153. https://doi.org/10.1016/j.cmpb.2019.06.020
3. Han J, Kamber M, Pei J (2012) Chapter 7—advanced pattern mining. In: Han J, Kamber M, Pei J (eds) Data mining, 3rd edn. The Morgan Kaufmann series in data management systems, pp 279–325. Morgan Kaufmann, Boston. https://doi.org/10.1016/B978-0-12-381479-1.00007-1
4. Jeong H, Ohno Y (2019) Cordless monitoring system for respiratory and heart rates in bed by using large-scale pressure sensor sheet. Smart Health 13:100,057. https://doi.org/10.1016/j.smhl.2018.07.025
5. Lakshmi K, Vadivu G (2017) Extracting association rules from medical health records using multi-criteria decision analysis. Procedia Comput Sci 115:290–295. https://doi.org/10.1016/j.procs.2017.09.137. 7th international conference on advances in computing & communications, ICACC-2017, Cochin, India, 22–24 Aug 2017
6. Osadchiy T, Poliakov I, Olivier P, Rowland M, Foster E (2019) Recommender system based on pairwise association rules. Expert Syst Appl 115:535–542. https://doi.org/10.1016/j.eswa.2018.07.077
7. Rothschild M, Stiglitz JE (1970) Increasing risk: I. A definition. J Econ Theory 2(3):225–243. https://doi.org/10.1016/0022-0531(70)90038-4
8. Sethia D, Gupta D, Saran H (2019) Smart health record management with secure NFC-enabled mobile devices. Smart Health 13:100,063. https://doi.org/10.1016/j.smhl.2018.11.001
9. Toivonen H, Klemettinen M, Ronkainen P, Hätönen K, Mannila H (1995) Pruning and grouping discovered association rules
10. Xu C, Bao J, Wang C, Liu P (2018) Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. J Saf Res 67:65–75. https://doi.org/10.1016/j.jsr.2018.09.013
11. Xue Q, Chuah MC (2019) Explainable deep learning based medical diagnostic system. Smart Health 13:100,068. https://doi.org/10.1016/j.smhl.2019.03.002

Systematic Literature Review of Essential Enterprise Architecture Management Dimensions

Marimuthu Trishan, Merwe Van der Alta, and Gerber Aurona

Abstract Many organisations turn to enterprise architecture (EA) to assist with the alignment of business and information technology. While some of these organisations succeed in the development and implementation of EA, many of them fail to manage EA after implementation. Because of the specific focus on the management of EA during and after the initial implementation, the enterprise architecture management (EAM) field is developed. EAM is characterised by many dimensions or elements. It is a challenge to select the dimensions that should be managed and that are vital for successful EA practice. In this study, we executed a systematic literature review (SLR) of primary EA and EAM literature with the aim of identifying dimensions regarded as key areas of EAM. The main contribution of this work is a concept map of the essential EAM dimensions with their relationships. The results of the SLR indicate that dimensions that used to be considered important or seemed to be the most essential for EA, such as frameworks, EA principles and reference models, are no longer emphasised as strongly and more focus is placed on people, skills, communication and governance when considering EAM literature and EAM maturity.

Keywords Enterprise architecture · Enterprise architecture management · Maturity · Dimensions

1 Introduction

The term architecture is a common notion in the building/construction industry. Even in this industry, the term is not unambiguous; it could mean the art or science of design or the style of design [1]. Architecture at an organisational level is known as enterprise architecture (EA) and can be described using the four main domains of

M. Trishan (✉) · M. Van der Alta · G. Aurona
Department of Informatics, , University of Pretoria, Pretoria, South Africa

G. Aurona
Centre for AI Research (CAIR), Pretoria, South Africa

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_36

architecture, namely, business architecture, application architecture data/information architecture and technology architecture [2]. Looking at the organisation holistically, this approach focuses on both the business and technical aspects [3] and assists organisations to structure how they meet strategic objectives [4]. This approach is commonly referred to as enterprise architecture planning (EAP). The goal of EAP is to "improve the efficiency, timeliness and reliability of business information" [4], while the goal of Enterprise Architecture Management (EAM) is to maintain and manage EA after implementation [5].

Huge investments are continually made to increase organisations' effectiveness and efficiency, yet many organisations constantly find it challenging to bring strategy to execution [6]. EA is generally believed to be the solution that aligns information and communication technology (ICT) strategy and business strategy, and many organisations make substantial efforts to implement EA but neglect the efforts required to follow through with aspects related to enterprise architecture management (EAM) [7, 8]. EAM is recognised as the accepted solution by organisations to assist with strategic transformation in a rapidly changing economic, technical and regulatory environment [9], yet there is no commonly accepted understanding of maturity in EAM [9, 10]. It is consequently difficult for organisations to identify and agree on the essential components of EAM and the level of maturity that is required to be successful in managing the EA practice.

Research has provided many different frameworks, methodologies and models, but all have a generalised view and often need to be customised or require the organisation to select what is appropriate to its objectives in order to derive value from them [11, 12]. While preliminary investigations suggest that there is little research in the area of EAM maturity in information systems (IS), we cannot embark on such studies without understanding where EAM dimensions are placed within the existing frameworks and the implications when considering people, processes and technologies. For the purpose of this study, an EAM dimension is described as a measurable component of EAM that could have different maturity level associations. When these EAM dimensions are identified, questions such as the following can be investigated: What dimensions increase success in EAM? Which part of existing EA frameworks can be adapted to support EAM success? How do people facilitate EAM success? In our study, we also included the identification of dimensions that add to the challenges or failures of EAM.

While an increasing number of studies on the value and the importance of EAM and the adoption of EAM are being undertaken, there is little information on the essential dimensions and how to mature them throughout their existence in an organisation. Olsen and Trelsgard [13] argued that many organisations choose to implement an EA strategy but most fail, sharing the same EA challenges. The literature provides enough information on frameworks and methodologies for EA but little or no information on what to assess and improve in EAM to increase its maturity and to support success. Failure is often due to a lack of understanding and an unclear vision of how the different components of EAM fit together, therefore understanding the different dimensions and the level at which they need to operate is imperative for success.

To identify the most essential dimensions of existing frameworks, the following question is the focus of the paper: *What are the EAM dimensions in existing EA and EAM literature that support EAM success?*

The motivation for the research question is to separate those dimensions into what contributes to the success and what provides little or no value. In this paper, the findings from a systematic literature review (SLR) on dimensions that are essential to EAM were identified. Limited comprehensive research has been published that provides a view of the dimensions in EAM that are key to the success of EA after implementation. By starting work in this area, research should assist in determining a baseline that can be critiqued and improved as further studies are conducted. This systematic approach to identifying and analysing published studies allows for the selection of reliable literature with specific themes that present recurring or conflicting findings, as well as gaps in the existing body of knowledge.

The structure of the paper includes background information in Sect. 2, followed by the method used for the SLR in Sect. 3 and the results, accompanied by a discussion, in Sect. 4. The paper is concluded in Sect. 5.

2 Foundation

"A well-defined architecture is an important asset in positioning new developments within the context of the existing processes, IT systems, and other assets of an organisation, and it helps in identifying necessary changes" [1]. Jonkers et al. [1] continue to explain that having an integrated practice of business and information technology (IT) is necessary for understanding the impact of introducing a new product into the organisation. It may require new people, processes and technologies to support the change; it may even require a change in organisational structure.

When EA started in 1975, it was known as business system planning (BSP), developed by IBM. Zachman's framework [14] was the first evolution of EA and introduced a framework that conceptualised architecture from multiple perspectives to information system professionals. During the 1990s, EA went through its second evolution. As a result of technological advancements and an increase in organisations wanting to digitise their business processes, it introduced a new era of complexities expanding to the different stakeholders of organisations. Today we see the next evolution of how EAs have evolved with the relevance of EAM.

Most industries find that it is a challenging task to manage EA in their organisations once implemented [10]. Every organisation experiences a set of unique challenges together with some common challenges [9, 13]. The common challenges that impede the process of moving towards a common understanding of EAM include [9, 13]:

- Executive management's understanding of EA
- Unclear EAM roles
- Ineffective communication and adoption
- Low EA maturity and governance.

Further to the common challenges are some unique ones that include socio-technological literacy and lack of process integration within the organisation or a low level of this. To gain a better understanding, these challenges need to be examined further to determine what makes them essential components of EAM. The literature provides enough information on frameworks and methodologies for EA, but limited information on what to assess and improve in EAM to increase its maturity to ensure success.

The concept of EAM is far more than just modelling. A big part of it includes the management of tasks related to planning and executing business changes from an architectural point of view [15]. EAM has many dimensions and not all may be relevant to all organisations. Understanding which dimensions are pertinent to meeting one's objectives is the key to EAM success.

There is a strong association with EAM as part of the EA project, which suggests that there is a beginning and an end to EAM, but there is another aspect of EAM, which relates to the continuous management of the functions and other relationships of EA at an enterprise level after implementation. For the purpose of this paper, we are interested in this definition of EAM. While this section aims to identify and relate to the key areas within EAM, these have not reached the maturity level required, considering their contribution to the competitiveness of enterprises [16].

3 Method

The steps detailed in this section ensured that most of the important content was included and that all steps could be repeated easily, maintaining the integrity of the research. There are many SLR approaches that can be followed [17–19]. For the purpose of this research, the phased approach to conducting an SLR was described by Okoli and Schabram [19].

This protocol was first piloted by performing searches based on rules within the protocol. Through this exercise, the pilot study uncovered a few concerns, which contributed to the quality of the results. Changes were accordingly implemented to correct these concerns and strengthen the quality, using the guidelines recommended by Kitchenham [18], which focus on three viewpoints, namely, the population, intervention and outcomes.

Keywords were used in the SLRs to search databases in a specific academic realm. These keywords were expanded to include synonyms referring to the constructs identified for each research question.

The source selection criteria included the evaluation of all relevant electronically available scientific databases. It also included features and functionality that allowed for the use of keyword searches for abstracts, titles and citations with an advanced searching capability that allowed for the use of concatenation and truncation functions. All databases used for the searches were required to be in the English language and of an academic nature and include: IEEE Explore, ACM Digital Library, Google Scholar, ScienceDirect, ISI Web of Science and Springer Link.

Executing the search string in a database yielded a large number of articles in the resulting dataset. It was not practical to analyse each and every one of these articles, and many were deemed inapplicable to the research question of this review. This step was required to conduct an early-stage separation of articles that might have potential and those that would not contribute to the value of this SLR. This was achieved by reading through the abstract of the article and applying criteria that were either in favour of proceeding with the article or discarding it.

The quality appraisal assessment was completed for all research articles using assessment criteria and was applied to all research papers that satisfied the inclusion and exclusion criteria assessments. The inclusion and exclusion criteria, together with the assessment scoring, provided aggregated quality criteria as a guide to select quality research papers. The score alone had no bearing on the quality of the article itself. Quality was based on a combination of the criteria, the context of the paper and the score.

To ensure the reliability and replicability of this research, a set of steps were followed with prescribed data input values and templates. Following this, process and procedure yielded 45 primary studies that were used in answering the research question of the SLR, which is to identify dimensions that are vital to the success of EAM. A pre-defined search string was used within each qualifying search database. In cases where some database engines had limitations in the manner that the search string could be executed, the search string was altered to adhere to the manner in which that database required to provide a successful result. This was done in such a way that the altered search string still provided correct and valid results. The results from each database were captured individually against the respective database and then consolidated to provide a combined list of searched results.

The search results retrieved by applying the search criteria across the identified databases equated to 618 studies. Working through the exclusion criteria specified in the previous section, all duplicates were then removed. This was followed by removing any study that was listed as a book, non-English studies and non-full-text studies, based on the review of the studies' titles, abstracts or conclusions. This brought down the number of primary studies for a full review.

The primary studies that remained were then evaluated against the inclusion criteria. All studies that did not meet the criteria were moved to the exclusion list as non-related studies. By applying the exclusion and inclusion criteria, the number of studies was reduced by 473, leaving the researchers with 145 studies. Using the scoring appraisal mentioned in the previous section, each study was further evaluated and scored according to the criteria. Any study that satisfied the criteria was regarded as a selected primary study to be used to answer this SLR's research question through a critical review. After applying the quality appraisal and scoring, the total number of primary studies was reduced to 45. Once the primary studies had been identified, the data extraction process could begin. The purpose of the data extraction process was to answer the research question outlined in the section above.

4 Discussion

In-depth understanding of the EAM dimensions was acquired through the execution of this SLR. Many conclusions were extracted through synthesis of the results. For example, there is no consensus on what the exhaustive list of essential dimensions should be, but some agreement on a few dimensions. One of the key contributions from this study is that many of these dimensions have direct and indirect implications for one another. The key dimensions should be directly linked to the objectives of one's EAM mandate and a strong focus on people rather than technology for EAM to succeed. This can be viewed at a very abstract level in Fig. 1. Integration was identified as a vital dimension with the focus on people integration and process integration, tightly coupling governance and policy enforcement to these areas. EAM frameworks have lost focus or at least it is assumed that these frameworks are in place and are mature. Repositories and methodologies are becoming more important and are based on the frameworks. Communication has become more important as well. It is not enough for IT departments to understand and adopt EAM; business departments should be involved as well. Visibility is important for selling the EAM service so that value can be derived at all levels. The above analysis highlights areas for future research with a more detailed focus. Some other equally important dimensions that

Fig. 1 EAM dimensions and their relationships

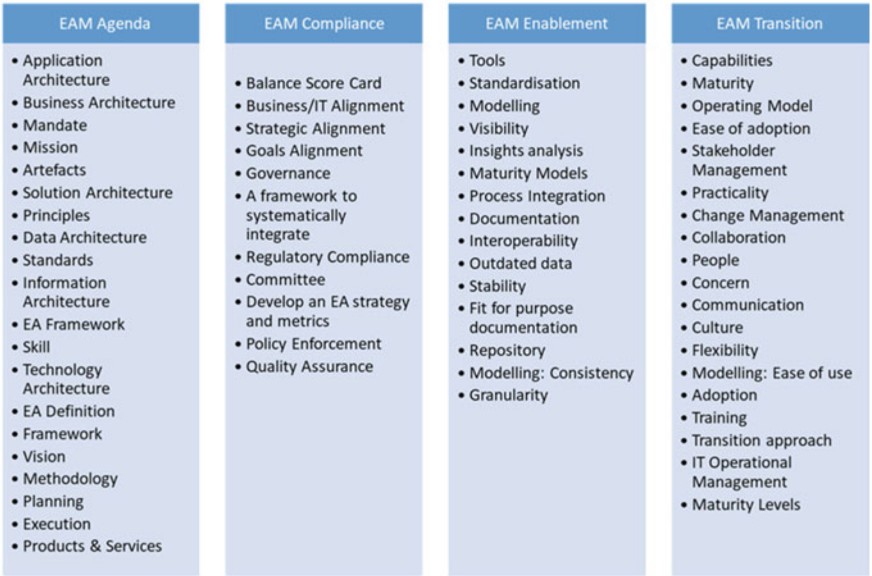


Fig. 2 List of EAM dimensions

did not feature very often in this SLR was EAM adoption, the EAM mandate, EAM change management and the operating model, to mention a few needs to be explored further. Given that academic studies in this area are still growing, more practical case studies should be developed to test and verify the dimensions.

The paper concludes with the identification of a list of vital EAM dimensions extracted from 45 studies. The list of key dimensions is depicted in Fig. 2. Based on the dimensions identified, four categories emerged, namely, the *EAM agenda*, *compliance, enablement* and *transition.*

4.1 EAM Agenda

The literature provides a huge spectrum of dimensions for the EAM agenda as depicted in Fig. 2. It is fair to assume that some of the common dimensions, such as frameworks, methodologies, principles and domain architecture deliverables, are strongly evident. Skills, on the other hand, are generally seen as something that will be developed over time.

The skills of resources in an EA team are cross-cutting and there are often small overlaps among the different domains. The expectation is then that over and above the specialist skills of a particular domain, practitioners must have good all-round experience in other domains as well. Second to skills as an essential dimension is the EAM mandate. Many organisations are still very unclear when it comes to how decisions are to be made and the processes to be followed [20]. Low integration into the organisation leads to the EAM team building ivory towers. Ultimately, this will result in the organisation not making use of EAM services, leading to the failure of the practice.

While the relevance of an EAM methodology is obvious, many problems have been identified related to this dimension including the image issue affecting EAM as a result of poor implementations of a methodology. Organisations adopt complete methodologies without aligning and adapting these to the existing methodologies in the organisation. The existing EAM methodologies are guiding principles and not prescriptions for implementation.

4.2 EAM Compliance

While it is not surprising that governance was identified as the most essential dimension under compliance, organisations do place enough focus on this area in general. However, the analysis provided a little more insight into some of the less prominent detail of governance implications. Some of these lower-level implications were EAM goals alignment, policy enforcement and process integration. There is a stronger focus on the integration between the different areas of the organisation and EAM. It is not enough to create principles, standards and policies; adoption has to be increased through awareness and socialisation. Figure 1 provides a view of the relationships with other key dimensions that are prominent in current EAM practice. Governance assists with the communication of EAM artefacts among different stakeholders. This form of communication ensures compliance and alignment between business and IT, which fall outside other existing forms of governance, such as IT and corporate governance.

Stakeholder management was the second-highest dimension identified within compliance, as well as throughout the 45 primary studies. In many of the studies, it was highlighted as an area of concern and one of the biggest contributors to EAM failure [13, 20]. Understanding who the key stakeholders from both business and IT becomes crucial for EAM, including having them participating in decision-making committees. It is clear that top management buy-in is essential for EAM, but middle management involvement is becoming equally important to drive the EA agenda. High stakeholder participation is required for improved decision-making. To ensure that stakeholder participation is engaged at the correct level, understanding the stakeholder hierarchy is crucial.

4.3 EAM Enablement

Tools and technology were always important; however, the mindset has shifted from entry-level platforms/applications such as Excel sheets and modelling tools as standalone solutions to more insight-driven, relationship-mapping software that can automatically model architectures or dynamically produce dashboards and views for quicker decision-making. These capabilities tie into the second and third most essential dimensions in EAM enablement, namely, collaboration and fit-for-purpose documentation.

Making the information that one already has available to all those who require it is fundamental to EAM success. There is no benefit in putting full effort into recording and documenting architectural information if it is not available for use. Not being able to use information was noted as one of the most common problems experienced, as no one knows where or how to find the relevant information. On the other hand, one might be able to find what one wants, but it is not at the right level of detail. Finding the balance of relevant information for each stakeholder is key to understanding that tools should not be adopted to provide something new, but rather enhance the existing process to satisfy information needs, which will lead to long-term success [21].

4.4 EAM Transition

As with EAM compliance, stakeholder management is an important dimension of EAM transition. Understanding the stakeholder hierarchy and its needs is a concern identified through the SLR. It is necessary to satisfy stakeholders' needs while providing the value that is required by EAM. To affect culture change, a combination of EAM agenda requirements comes into play. These can be described as policy enforcement, rewards and recognition, support, training, awareness programs, roadshows, etc. Communication and marketing become essential for driving buy-in across the organisation. Educating the entire organisation assists in working towards the same goal; this ties in with goal alignment dimensions in the EAM agenda.

Communication was identified as one of the key challenges in EAM. This stems from the lack of a shared dictionary [22]. Every area of the organisation develops key/common terms in its respective realm. Often stakeholders, EA practitioners and IT personnel assume that these terms are also common outside their area. This leads to misunderstanding and misalignment of outcomes and objectives. Communication extends further than this, as it touches on everything EA practitioners do, including communicating the services they provide so that the rest of the organisation is aware of them and more importantly, makes use of them. If the EAM team is not providing a service to satisfy a particular stakeholder's needs, it is seen as ineffective in the organisation, which ultimately leads to its demise.

5 Conclusion

In this report, the findings presented were based on the SLR identifying the key dimensions considered essential for EAM success. These dimensions with their relationships are depicted in a concept map, which is the main contribution of the study. Since the inception of EA and EAM with their associated frameworks and methodologies, no research results have been published that provide a view of what the key dimension for EAM success is. By initiating work in this area, this research should assist enterprises in understanding their current state of EAM and focus on key dimensions, which could accelerate the maturity of EAM.

References

1. Jonkers H, Lankhorst M, ter Doest H, Arbab F, Bosma H, Wieringa RJ (2006) Enterprise architecture: management tool and blueprint for the organisation. Inf Syst Front 8:63–6
2. Kotusev S (2018) TOGAF-based enterprise architecture practice: an exploratory case study. CAIS 321–59
3. Lindström Å, Johnson P, Johansson E, Ekstedt M, Simonsson M (2006) A survey on CIO concerns-do enterprise architecture frameworks support them? Inf Syst Front N Y 8:81–90
4. White SK (2018) What is enterprise architecture? A framework for transformation. CIO. https://www.cio.com/article/3313657/enterprise-architecture/what-is-enterprise-architecture-a-framework-for-transformation.html. Accessed 17 Feb 2019
5. Ahlemann F (ed) (2012) Strategic enterprise architecture management: challenges, best practices, and future developments. Springer, Berlin, New York
6. Simon D, Fischbach K, Schoder D (2014) Enterprise architecture management and its role in corporate strategic management. Inf Syst E Bus Manag 12:5–42
7. Aier S, Schelp J (2010) A reassessment of enterprise architecture implementation. In: Service-oriented computing—ICSOC 2007, vol 4749. pp 35–47
8. Parks R, Xu H, Chu C-H, Lowry PB (2017) Examining the intended and unintended consequences of organisational privacy safeguards. Eur J Inf Syst 26:37–65
9. Buckl S, Dierl T, Matthes F, Schweda CM (2010) Building blocks for enterprise architecture management solutions. In: Harmsen F, Proper E, Schalkwijk F, Barjis J, Overbeek S (eds) Practice-driven research on enterprise transformation. Springer, Berlin, Heidelberg, pp 17–46. https://doi.org/10.1007/978-3-642-16770-6_2. Accessed 7 Mar 2019
10. Simon D, Fischbach K, Schoder D (2013) An exploration of enterprise architecture research. Commun Assoc Inf Syst 32. https://aisel.aisnet.org/cais/vol32/iss1/1. Accessed 21 Feb 2019
11. Bandar MA (2017) Enterprise architecture frameworks: a critique review from a security perspective. Int J Comp Appl 174:9–15
12. Buckl S, Ernst AM, Matthes F, Ramacher R, Schweda CM (2009) Using enterprise architecture management patterns to complement TOGAF. In: 2009 IEEE international enterprise distributed object computing conference. IEEE, Auckland, New Zealand, pp 34–41. http://ieeexplore.ieee.org/document/5277730/. Accessed 27 Sept 2019
13. Olsen DH, Trelsgard K (2016) Enterprise architecture adoption challenges: an exploratory case study of the Norwegian higher education sector. Procedia Comput Sci 100:804–811
14. Zachman JA (1999) A framework for information systems architecture. IBM Syst J Armonk 38:454–70
15. Aier S, Gleichauf B, Winter R (2011) Understanding enterprise architecture management design—an empirical analysis. In: Bernstein A, Schwabe G (eds) Proceedings of the 10th

international conference on wirtschaftsinformatik. Zurich, pp 645–54. http://www.zora.uzh.ch/60516/1/20120227114121_merlin-id_2354.pdf. Accessed 23 June 2019
16. Winter K, Buckl S, Matthes F, Schweda CM (2010) Investigating the state-of-the-art in enterprise architecture management methods in literature and practice. In: Mediterranean conference on information systems, vol 13
17. de Almeida Biolchini JC, Mian PG, Natali ACC, Conte TU, Travassos GH (2007) Scientific research ontology to support systematic review in software engineering. Adv Eng Inform 21:133–151
18. Kitchenham B (2004) Procedures for performing systematic reviews, p 33
19. Okoli C, Schabram K (2010) A guide to conducting a systematic literature review of information systems research. SSRN J. http://www.ssrn.com/abstract=1954824. Accessed 19 June 2019
20. Hacks S, Brosius M, Aier S (2017) A case study of stakeholder concerns on EAM. In: Halle S, Dijkman R, Lapalme J (eds) Proceedings of the 2017 IEEE 21st international enterprise distributed object computing conference workshops and demonstrations (EDOCW 2017). IEEE, New York, pp 50–56
21. Seppänen V (2014) From problems to critical success factors of enterprise architecture adoption. In: Jyväskylä studies in computing, University of Jyväskylä
22. Kotusev S, Singh M, Storey I (2015) Investigating the usage of enterprise architecture artifacts. In: Association for information systems AIS electronic library, vol 15

Role of Gamification in Cultural Heritage Dissemination: A Systematic Review

Imran khan, Ana Melro, Ana Carla Amaro, and Lídia Oliveira

Abstract In the current modern world, gamification and its innovative methods become one of the primary tools for advanced communication and socialization with users in a variety of sectors—finance, marketing, business training and entertainment (Kapoor et al (2018) Inf Syst Front 20:531–558). Studying the skills of the game link shows that game plays a powerful and productive role in improving the key skills and competencies required for success in education, real life and professional realization (Pellegrino JW, Hilton ML (2012) Education for life and work : developing transferable knowledge and skills in the 21st century. The National Academies Press. http://www.nap.edu/catalog.php?record_id=13398). Comprising the historical-cultural heritage in the field of gaming cultural heritage training might be one of the most promising learning approaches for visitors of all ages. This article presents the results of a methodical review, aimed to understand how gamification environment used, for the past 5 years, for promoting cultural heritage sites, activities and enhancing knowledge. Results indicated that a total of 72 studies have been published between 2015 and 2020, and 45 were selected for a systematic review.

Keywords Gamification · Cultural heritage · IoT technology · Playful · Engagement

1 Introduction

Gamification is the process of introducing game strategies and components into some scenarios and situations that are not a game [3]. Gamification can be considered the

I. khan (✉) · A. Melro · A. C. Amaro · L. Oliveira
Department of Communication and Art, University of Aveiro, Aveiro, Portugal
e-mail: Imran10111@ua.pt

A. Melro
e-mail: anamelro@ua.pt

A. C. Amaro
e-mail: aamaro@ua.pt

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_37

meaning to engage people in tasks [4], promote relationship [5] or improve motivation [6]. Khaled, Deterding, Dixon and Nack [7] suggested a definition of gamification as “the use of game design elements in a non-game context”.

Due to gamification elements, the gamification can be useful for teaching and learning activities because it can promote student engagement. The difficulties educationalists face related to student interest and engagement in their classroom are not new. In the past, educationalists have attempted to use a variety of inventions, including the use of motivational tactics. However, the impact of intervention lasted for only short period. Due to its playful and natural characteristics, gamification can be a great solution to help solve student engagement and participation in learning activities [8]. The primary function of a serious game is its aim of helping the player to achieve a learning target via an enjoyable experience [9].

This paper reviews in recent times the role of gamification environment for the promotion of cultural heritage sites and activities. In this way, it seems essential to understand how the gamification environment can be developed, how it promotes cultural heritage sites and activities in a way that the visitors can gain knowledge and entrainment. Therefore, an extensive systematic review has been conducted, from 2015 to 2020, to understand the gamification environment, which could promote cultural heritage sites and effectively convey knowledge of cultural heritage sites to visitors.

2 Background

The growth of games in non-leisure contexts is transforming daily lives and most importantly introducing more fun in everyday life [10], it is the power of games to motivate, engage and immerse [11]. Over the past few years, there has been a massive development in entertainment technology, i.e. today’s games are exponentially more sophisticated, smart and responsive than those of just a few years ago [12], which, in turn, intake the lead to remarkably high customers’ expectation. Modern real-time computer graphics can accomplish near virtual games, and photorealism, worlds are generally populated with a massive amount of high-value content producing a richer experience. The project “Discovering the territory of Old Peucetia” [13] describes the usage of a tangible consumer interface game and how he/she understands cultural heritage, particularly considering the castle Caracciolo Sammichle at Bari in Apulia, Italy. The main objective of this project is to promote the local heritage of old Pecucetia in Apulia among primary school students. This project provides a network of activities and educational workshops, which have been arranged in the historical buildings of towns of old Pecucetia to allow students of primary school to learn the history and the environment of the area by gamification method. Offered workshops rely on interactive tools and multimedia installations, which were designed, developed and installed. Via events and simulation, familiar visitors and students of the exhibitions can engage themselves in ancient times and uncover the history of their area. Another study offers the possibility to explore the cultural heritage sites under the water through AR gamification method. The title of this study is "A hybrid augmented reality guide for underwater cultural heritage sites", conducted by Čejka, Zsíros and Liarokapis [14]. This study offers a very innovative augmented reality model for the underwaters' divers to introduce ancient, abandoned buildings at underwaters' archaeological locations. The study offers a prototype that runs on smart devices (smartphones), which are sealed in a waterproof casing and uses a hybrid method to pinpoint the location of divers—the prototype of this study experiment conducted at Baiae, Italy. A total of 10 professional divers took part in this study, and their gamification experience level showed in the result that underwater AR substantially improves the user experience and their entertainment level.

3 Methodology

3.1 Research Question

The presented systematic review seeks an answer to the following question:

How gamification contributes to the dissemination of cultural heritage?

3.2 Research Procedure

For the systematic review, this research was carried out using the Scopus database, using one main query with two main keywords. First, "Gamification" and the second one is "Cultural Heritage (CH)". The reason behind the use of one query with two keywords was to focus only on gamification and cultural heritage studies and avoid the dispersion of results and consequently avoid the loss of focus from the main topic.

3.3 Inclusion Method

The selected studies for systematic review cumulatively met the following inclusion procedures:

(i) Studies have been published between 2015 and 2020. (ii) Address gamification from the perspective of cultural dissemination. (iii) Sufficiently described the design and usage of gamification application for the promotion of cultural heritage. (iv) Selected studies present sufficient knowledge and highlight the importance of gamification applications in a cultural heritage context.

4 Research Results

There were n = 72 studies located in references, including n = 4 duplicates (Figs. 1 and 2).

In the first step, 68 studies have been reviewed. However, 68 studies have been examined in which n = 48 studies are starting with an appraisal of their titles and abstracts. These n = 48 studies are selected because their titles and abstracts met

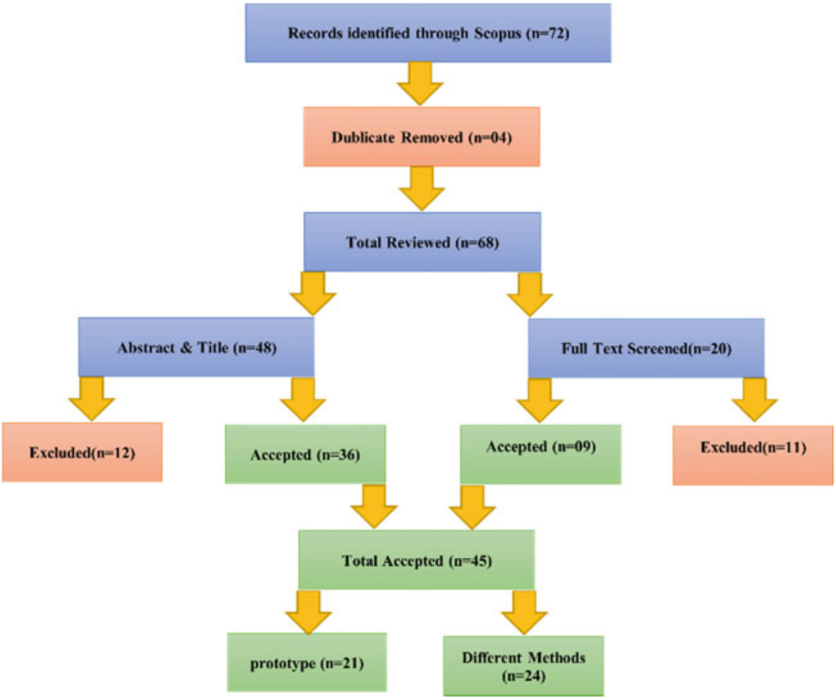


Fig. 1 Graphical representation of results

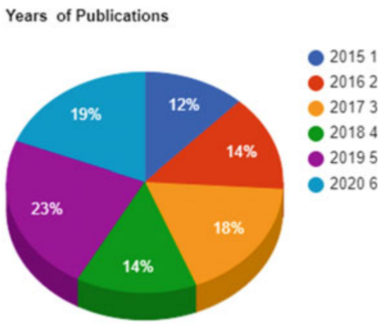


Fig. 2 Percentage of publication in different years

inclusion criteria, but when we reviewed in-depth n = 36, these studies met the inclusion criteria and accepted, and remaining 12 studies have been excluded. Furthermore, remaining n = 20 studies with full text have been screened because there were some doubts in their titles and abstracts so they decided to screen and analyze the full text. Consequently, n = 11 studies have been accepted but remaining n = 9 studies were excluded. Collectively, n = 45 studies have been accepted in which n = 21 are prototype and 24 used other different methods to offer gamification.

4.1 Quality Appraisal

All studies' titles, references and abstract were examined. However, n = 20 full text of articles was analyzed and read, and n = 11 articles did not meet the inclusion criteria of the study, which lead to the exclusion of the systematic review. Then an assessment of quality was done. Higgins and Altman [15] recommended against rating using statistical scales, so emphases were placed on an assessment of bias and in the assessment of the risk of bias from trial attrition, selection and selective conclusion reporting.

4.2 Data Extraction

Data were extracted for the research setting, research publishing date, design of the presented study, objective of the study, novel approaches used for presented studies and outcomes. The information abstracted was checked against the full-text articles independently.

Dates of publication and data collection: The included studies were conducted, data collected and published between 2015 and 2020.

Countries: Most studies were published and conducted in European countries like Italy 16 and Greece 9 studies published between 2015 and 2020, and then Asia in which China, Saudi Arabia and India stand out.

4.3 Reviews Studies' Designs and Methods

All selected 45 studies are based on the gamification approach for cultural heritage promotion. Twenty-one studies are based on the development of prototype applications and system, remaining 24 studies are based on literature review and survey. Most of the selected studies used AR and VR approaches to provide the gamification environment for cultural heritage exploration and to learn about cultural heritage. These results can also be analyzed from the VR and AR analyses, which are already being prepared for future publication.

4.4 Gamification with Various Technological Methods

In the present technological era, researchers are promoting cultural heritage via VR and AR gamification. However, many researchers are using different techniques and approaches to develop a gamification environment for cultural heritage dissemination. Therefore, the authors review 24 studies out of 45 that used 24 approaches to develop a gamification platform or to understand the gamification application for cultural heritage enhancement. The following 24 studies [16–38] use different approaches and methods to create a gamification environment for cultural heritage promotion.

5 Discussion and Conclusion

Overall assessment of quality across the studies reviewed, the design and methods were explained clearly and, in most cases, would be replicable. In all reviewed studies, convenient nonrandomized approaches and procedures were employed for the development of gamification atmosphere or to understand the gamification impact on cultural heritage.

In this systematic review, a comprehensive understanding of gamification for the dissemination of cultural heritage applications and consideration of open issues was achieved. In this literature, the study uses Scopus data platforms for exploration query on 72 papers that were published from 2015 to 2020. Finally, the study reviewed 45 papers that emphasized gamification and cultural heritage. For gamification, AR and VR approaches have the highest percentage of the application approaches by 46.66% (21 out of 45 studies) of quotas in the literature that enhance the dissemination of cultural heritage knowledge. The high percentage of AR and VR shows that this technology has significant influences in the tourism sector and this influence pushes the researchers to develop AR and VR applications in which tourists can explore the cultural heritage through augmented reality and virtual reality.

Furthermore, in a systematic review, 53.33% of studies (24 out of 45 studies) are based on different technological methods that develop a gamification environment for the promotion of cultural heritage among the tourists. The percentage of different methods is very high as compared with VR and AR technologies, which indicated that researchers are not only to rely on augmented reality and virtual reality but also they are using diverse and new innovative techniques to develop a gamification application for cultural heritage dissemination.

The review studies described that gamification does not only offer entertainment platforms but also offers tourists to enhance their knowledge level by planning serious games. Conclusively, this systematic review provides extensive evidence that indicated gamification is playing an essential role to promote cultural heritage and its knowledge effectively.

References

1. Kapoor KK, Tamilmani K, Rana NP, Patil P, Dwivedi YK, Nerur S (2018) Advances in social media research: past, present and future. Inf Syst Front 20(3):531–558
2. Pellegrino JW, Hilton ML (2012) Education for life and work : developing transferable knowledge and skills in the 21st century. The National Academies Press. http://www.nap.edu/catalog.php?record_id=13398
3. Robson K, Plangger K, Kietzmann JH, McCarthy I, Pitt L (2015) Is it all a game? Understanding the principles of gamification. Bus Horiz 58(4):411–420
4. Nicholson S (2015) A recipe for meaningful gamification. Gamification Educ Bus 2015:1–20
5. Caporarello L, Magni M, Pennarola F (2017) Learning and gamification: a possible relationship? EAI Endorsed Trans E Learn 4(16):153488
6. Sailer M, Hense JU, Mayr SK, Mandl H (2017) How gamification motivates: an experimental study of the effects of specific game design elements on psychological need satisfaction. Comput Human Behav 69:371–380
7. Deterding S, Dixon D, Khaled R, Nacke L (2011) From game design elements to gamefulness: defining 'gamification'. In: Proceedings of the 15th international academic MindTrek conference: envisioning future media environments. MindTrek 2011, pp 9–15
8. Kim S, Song K, Lockee B, Burton J (2018) Gamification in learning and education
9. Anastasiadis T, Lampropoulos G, Siakas K (2018) Digital game-based learning and serious games in education. Int J Adv Sci Res Eng 4(12):139–144
10. Arnab S, Clarke S (2017) Towards a trans-disciplinary methodology for a game-based intervention development process Sylvester. Br J Educ Technol 48(2):279–312
11. Alsawaier R (2017) The effect of gamification on motivation and engagement. Int J Inf Learn Technol 45:49
12. Boomsma C, Hafner R (2018) Should we play games where energy is concerned ? Perceptions of serious gaming as a technology to motivate energy behaviour change among social housing residents. Sustainability 18
13. Ferdinando C et al (2019) Gamification in cultural heritage: a tangible user interface game for learning about local heritage. In: Digital cultural heritage, vol 9. Springer, Cham, pp 411–422
14. Čejka J, Zsíros A, Liarokapis F (2020) A hybrid augmented reality guide for underwater cultural heritage sites. Pers Ubiquitous Comput 14
15. Higgins JP, Altman DG (2017) Assessing risk of bias in included studies
16. Markovic L, Sofronijevic A (2015) Building a gamified system for capturing MOOC related data-smart city learning community as its most precious source of intangible cultural heritage. In: 2015 international conference on culture and computing building, pp 175–182
17. Smaniotto C et al (2019) CyberParks—the interface between people, places and technology. Springer open
18. Jia X, Tan Y, Xiong Q, Li Y, Hai W (2019) Designing edutainment games for intangible cultural heritage: the BNREG Project, vol 9, pp 1–6
19. Vayanou M, Lumos G (2019) Designing performative, gamified cultural experiences for groups, pp 1–6
20. Kumar A (2016) Introducing new age to ICH : revering matters of significance. In: HCI'16: proceedings of the 8th Indian conference on human-computer interaction, pp 132–138
21. Corallo A, Fortunato L, Renna C, Sarcinella ML, Spennato A, De Blasi C (2017) Mobile app for promoting cultural heritage: geostatistic and textual analysis, April 2018
22. Lee BC (2019) sustainability the effect of gamification on psychological and behavioral outcomes: implications for cruise tourism destinations. Sustainability 11(11):3002
23. Aydin S, Schnabel MA (2016) The museum of gamers: unmediated cultural heritage through gaming. In: Borowiecki KJB, Forbes N, Fresa A (eds) Cultural heritage in a changing world. Springer, Berlin, pp 125–144
24. Chortaras A et al (2018) WITH: human-computer collaboration for data annotation and enrichment, pp 1117–1125

25. Fakhour M, Azough A, Kaghat FZ, Meknassi M (2020) A cultural scavenger hunt serious game based on audio augmented reality. In: Advanced intelligent systems for sustainable development (AI2SD'2018), vol 1102. Springer, Cham
26. Robles-Bykbaev V, López-Nores M, Ochoa-Zambrano J, García-Duque J, Pazos-Arias JJ (2015) A game-like application for dance learning using a natural human computer interface Alexandros. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 9177, pp 525–534
27. Grammatikopoulou A, Laraba S, Sahbenderoglu O, Dimitropoulos K, Douka S, Grammalidis N (2019) An adaptive framework for the creation of exergames for intangible cultural heritage (ICH) education, vol 6. Springer, Berlin, Heidelberg, pp 3
28. Olivieri AC, Schegg R, Sokhn M (2016) Cityzen: a social platform for cultural heritage focused tourism. In: 8th International conference on management of digital ecosystems. MEDES, pp 129–136
29. Basaraba (2018) Co-constructing cultural heritage through a web-based interactive digital narrative, vol 1. Springer International Publishing, New York
30. Lee J, Yi JH, Kim S (2020) Cultural heritage design element labeling system with gamification. IEEE Access 8:127700–127708
31. López-Martínez A, Carrera Á, and Iglesias CA (2020) Empowering museum experiences applying gamification techniques based on linked data and smart Objects. Appl Sci 10(16)
32. He H, Li Z, Cheng X, Wu J (2019) Gamified participatory museum experience for future museums. In: Veale G, Dogan H, Murphy J (eds) Design, user experience, and usability, vol 11585 LNCS, pp 195–208
33. Prandi C et al (2019) Gamifying cultural experiences across the urban environment. Multimed. Tools Appl. 78(3):3341–3364
34. López-Gonzalez L Gaming as a gateway ensuring quality control for crowdsourced data. In: Cooperative design, visualization, and engineering, pp 252–260
35. Ioannides M et al (2016) Imaging Novecento. A mobile app for automatic recognition of artworks and transfer of artistic styles. In: Digital heritage progress in cultural heritage: documentation, preservation, and protection, pt 1, pp 781–791
36. Quattrini R, Pierdicca R, Lucidi A, Di Stefano F, Malinverni ES (2019) The European research night: new ways for communicating science with ICT and video mapping. In: ISPRS annals of the photogrammetry, remote sensing and spatial information sciences, vol 42, no 2/W9, pp 647–654
37. Papathanasiou-Zuhrt D, Weiss-Ibanez DF, Di Russo A (2017) The gamification of heritage in the UNESCO enlisted medieval town of Rhodes. CEUR workshop proceeding, vol 1857, pp 60–70
38. Moreno Sánchez I, Navarro Newball AA (2015) La ciudad escondida. Toledo, laboratorio de comunicación transmedia. Opcion, vol 31, pp 806–827

Common Pitfalls When Explaining AI and Why Mechanistic Explanation Is a Hard Problem

Daniel C. Elton

Abstract Recently researchers have started using explainability techniques for several different applications—to help foresee how a model might operate in the field, to persuade others to trust a model, and to assist with debugging errors. A large number of explainability techniques have been published with very little empirical testing to see how useful they actually are for each of these use cases. We discuss several pitfalls one can encounter when trying to utilize explainability techniques. We then discuss how recent work on the double descent phenomena and non-robust features indicate that mechanistic explanation of deep neural networks will be very challenging for most real-world applications. In some cases, one may be able to use an easily interpretable model, but for many applications deep neural networks will be more accurate. In light of this, we suggest more focus should be given to implementing out-of-distribution detection methods to detect when a model is extrapolating and thus is likely to fail. These methods can be used in lieu of explainability techniques for increasing trust and debugging errors.

Keywords Interpretability · Explainability · Artificial intelligence · XAI · Out-of-distribution detection · Deep learning · Machine learning

1 Introduction

There is growing interest in developing methods to explain the inner functioning of deep neural networks. In this paper, we survey some of the pitfalls that are easily encountered when trying to "explain an explanation", some of which are not well appreciated in our experience in medical AI and other applied areas of applied AI. We first distinguish several motivations for interpretation in medical imaging. We argue that mechanistic interpretation (i.e., elucidating the underlying mechanism,

D. C. Elton (✉)
Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892, USA
e-mail: daniel.elton@nih.gov

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_38

similar to how we explain human decision-making) is potentially the most useful for increasing trust in AI. However, the recent discovery of double descent indicates that deep neural networks such as the convolutional neural networks work through the brute force local interpolation of data points. This is in contrast to popular narratives that deep neural networks work by extracting a few high level rules that we might explain in a few sentences [1]. As we will discuss, it follows that such networks are intrinsically hard to interpret and also have very little chance at being able to extrapolate outside their training distribution.

2 Motivations for Explanation

Different practitioners understand the term "intepretability" in different ways, leading to a lack of clarity on the matter [2]. Here we take the terms "explanation" and "interpretation" to be synonymous. We note several reasons we might be interested in explanations:

- **Prediction**—it can be useful to be able to predict whether a model can generalize or extrapolate to different conditions. For instance, will the model still be able to function if there is increased scanner noise or cropping?
- **Persuasion**—often we want to convince clinicians or other stakeholders that they can trust an AI system. Providing explanations can increase user's trust in a model, even if the explanations are not correct [3].
- **Debugging**—often we'd like to "open the black box" and understand why a model fails in particular cases. This can allow for iterative refinement.

While many fine distinctions can be made between different types of explanations, we find two high level definitions to be useful. We define a **descriptive explanations** as an account of model function which is descriptively accurate and relevant to the end user [2]. By "descriptively accurate", we mean that the explanation can reproduce the input–output mappings of the model to some degree. This type of explanation is typically boiled into natural language statement relevant to the domain of application. This type of explanation can be distinguished from a **mechanistic explanation**, which captures, at least approximately, the actual data manipulations occurring in the model. Only mechanistic explanation accurately predicts what the network will do when new data is presented.

Whether an explanation is descriptive or mechanistic can be distinguished by seeing if the explanation allows the user to predict the model's behavior when new inputs are presented from outside the model's training distribution or under counterfactual testing where parts of the image are removed. Most present day methods do not work very well in this regard, and thus fall under the category of descriptive explanation. Recently, Hase et al. have performed tests with a large pool of human subjects to see if different explainability techniques help users predict a neural network's behavior [4]. Out of the methods they compared, the only one that helped users for image data was the "This looks like That" approach [5], a method which was designed with explainability in mind [4]. This is not surprising because the "Rashomon Effect", [6] which says that for any set of noisy data, there are a multitude of models of equivalent accuracy but which differ significantly in their internal mechanism. To give an example of why mechanistic explanation may be useful, consider task of pancreas segmentation in CT images, which is challenging due to the variable shape and low contrast of the pancreas relative to background tissues. A robust way of finding it would be to locate a higher-contrast and easier to locate organ first, such as the liver. The pancreas has a relative position to the liver which is fairly consistent. A somewhat more brittle method would be trace the lower intestine to the duedenum, to which the pancreas is attached. A very brittle method would be to look for the pancreas in the center of the image.

3 Pitfalls

Recently, there has been a "cottage industry" of research showing problems with saliency-based methods [7, 8] and related "heatmapping" methods such as layer-wise relevance propagation [9]. At a high level, saliency maps may show where a model is not looking, but not what is doing. Saliency maps for different classification outputs may look similar, making it hard to distinguish why the network chose the output it did [6]. Many outputs of saliency-based methods are very similar to the output of an edge detector. Confirming this, it was found that, even if most of the later layers of a neural network are randomized, saliency maps do not change much [8]. Additionally, if labels or features are scrambled and the model is retrained, the outputs do not change [8] Indeed, it has been found that features "important" by the explanations are actually no more important than randomly specified features [10].

Recently, Olah et al. hypothesized that deep neural networks are explainable in the mechanistic sense, given enough careful study of what features each node represents and the connections between them [11]. Olah's et al.'s techniques are based on activation maximization, where a neuron or group of neurons is related to an interpretable feature (or more rarely, a combination of features). We see two issues with this type of approach. The first is that pure activation maximization leads to images which look like noise to the end user, so many "regularization" constraints have to be applied (in particular, Olah uses constraints developed for the highly publicized "deep dream" visualization). We are skeptical of this procedure due to its artificial nature. The method is tailored to provide the end user a pretty picture rather than remaining true to visualization the mechanism of the network. The "noise" which naive activation maximization shows may actually be "non-robust" features (see [12]). The second problem is that, if you take a linear combination of units from a given layer instead of a single unit (or more precisely perform a random rotation / change in basis), and maximize that instead, you end up with similar types of explanations for what each unit is sensitive to [13]. While Olah has discussed this issue in one of his previous works from 2017, in our view, it has not been adequately addressed in his most recent work [11].

There are many sources of variance which are often not taken into account when performing explanations. In several different contexts, it has been shown that the output of explainability methods often varies between test cases [14, 15]. When interpretations are shown, many should be given and they should be randomly chosen rather than "cherry-picked". If possible, a holistic analysis should be done, averaging interpretability results from many cases. Non-rigid registration to a reference case may be useful here in certain contexts such as medical imaging. Another source of variance is that in some cases visualizations for different train-test folds appears different, even when the resulting models are of equivalent accuracy [14, 16]. One way of mitigating this issues is to average results over a few different visualization methods [17]. Another source of variation that effects interpretation is the hyperparameter settings used. For instance, changing the LIME hyperparameters slightly has been shown to significantly change the output of the visualization in some cases [18].

An alternative method of explanation is to train a "post-hoc" model which is simpler and more interpretable to reproduce the output of the hard to interpret model. For instance, a decision tree can be trained to reproduce the output of a CNN. This procedure is the same as model distillation, where a large model is distilled on a smaller one by running the large one on a large unlabeled dataset and training the smaller model to reproduce the output of the larger one. Lillicrap & Kording show that this technique has limits however, and distilled models for image classification with equivalent accuracy are still quite large, with millions of parameters [19]. Thus, distilling further to a small interpretable model will incur a large decrease in accuracy, and therefore, won't be properly reproducing the input–output behavior of the original model.

Several recent works add an "explanation branch" to "explain" the output of a different branch of the network (the "prediction branch") [20, 21]. For the case of diagnosing lung nodules in chest CT, an example is illustrated in Fig. 1. The explanation branch in this case was trained to predict several attribute scores which clinicians consider important for lung nodule diagnosis. By seeing which attributes were predicted, the idea is that this constitutes an "explanation" of how the network arrived at its prediction. The issue with this sort of approach is that, it is not clear how the two branches are related—in principle they could be computed independently. In a recent work, we described a possible solution, which is to use a measure of mutual information overlap to make sure the outputs of the explanation branch and the prediction branch are related [22].

4 Why Mechanistic Interpretation Is Difficult

It has been noted for several years that the most successful deep learning models have millions of parameters and appear to be vastly underdetermined, yet they still generalize. More recently, it has been shown that the bias-variance trade-off breaks down in large enough networks [23]. Belkin et al. call this the "double descent phenomena" [23]. In the regime where deep neural networks operate, they are able

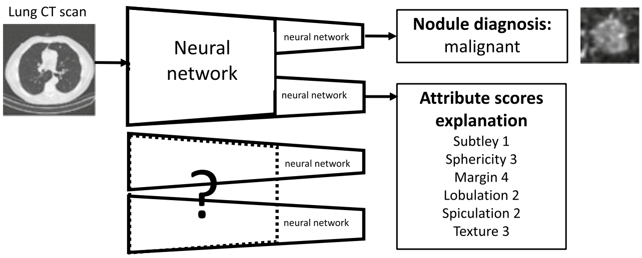


Fig. 1 A network for diagnosing lung nodules on a CT scan is based on a previously published work [20, 21]. The model contains an explanation branch in addition to the diagnosis branch, but it is not clear how the computations underlying each branch's output are related

to interpolate each data point in a "direct" way which does not exhibit the undershoot or overshoot which is typical of overfitting [1]. An illuminating example of direct fitting is given by Hasson et al. showing direct fitting of a parabolic function with noise [1]. The computations involved are clearly local—similar to nearest neighbors type computations—and the global trend ($y \propto x^2$) is not extracted. Because of this, there is clearly no hope for extrapolation. One the other hand, the model is flexible enough to fit any data. These observations call into question the popular idea that deep neural networks work by extracting high level features that are of particular interest, such as the whiskers of a cat. In actuality, it seems they are interpolating between a very large number of small "non-robust" features, some of which are particular to the training data [12]. It is tempting to tell "just-so" stories on how a deep neural network is functioning using explainability methods. These stories can mislead from what models are actually doing internally.

5 What Can Be Done?

The Rashomon effect (discussed above) suggests that in many cases, interpretable models may exist which have equivalent accuracy [6]. This seems to be especially the case with tabular data, where variants of linear regression often perform just as good as deep neural networks. However, in many other domains, deep neural networks are currently dominant. If the world is messy and complex, then neural networks trained on real-world data will also be messy and complex. Still, there are some models developed for images with interpretability in mind—one example is the "This Looks Like That" approach developed by Rudin et al. [4], which references parts of training examples.

Deep learning systems are notoriously bad at extrapolation, often failing spectacularly when small distributional shifts occur. For instance, a deep learning system for diagnosing retinopathy developed by Google's Verily Life Sciences which reached human-level diagnostic ability in the lab performed very poorly in field trials in Thai-

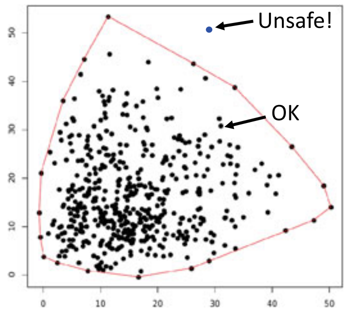


Fig. 2 Illustration of a simple approach to out-of-distribution detection

land due to poor lighting conditions, lower resolution images, and images which had been stitched together [24]. As another example, a deep learning system developed by *DeepMind* to play *Atari* games was shown to fail if minor changes are made, such as moving the paddle 3% higher in the game of Breakout [25]. In light of the fact that deep neural nets work through a sort of brute force fitting, this brittleness is not very surprising. If deep neural nets work by local interpolations over a massive number of data points, this implies an inability to extrapolate [1]. Explainability techniques can help understand how a model works within the dataset but they cannot help understand what happens when the network is asked to perform outside the training distribution. Fortunately, many techniques have been developed which can provide a “warning light" if a network is being asked to extrapolate. These techniques go under several different names—“applicability domain analysis”, “out-of-distribution detection”, “change point detection”, and “outlier detection” [26]. A full analysis and comparison of the many different techniques that have been developed is outside the scope of this paper. A simple illustration of how many such methods work is shown in Fig. 2. In Fig. 2, the “domain of interpolation” is delineated by projecting the data into 2 dimensions and then looking at the convex hull of the data points. If the input/latent vector of a test data point (projected into 2D) falls outside the convex hull, then the model is extrapolating and a warning should be given. Typically a dataset will form one or more clusters when projected into a low dimensional space. It has been observed empirically that the average error for test data samples depends on how close to the center of the training data clusters the test point lies [27].

6 Conclusion

In this work, we discussed some of the motivations for explanation in deep learning systems and distinguished descriptive explanations from mechanistic ones. We believe mechanistic explanations to be the most important for increasing trust and ensuring robustness to distributional shift. However, recent work on double descent [22] and adversarial examples [12] indicate that mechanistic explanation is difficult, since deep neural networks operate by brute force interpolation over large datasets rather than by simple heuristics with high level features. We discussed a few possible solutions to the issues raised—using explicitly interpretable models, adding an explanation branch, and implementing out-of-distribution detection methods.

Funding and Disclaimer No funding sources were used in the creation of this work. The author wrote this article in his own personal capacity. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

References

1. Hasson U, Nastase SA, Goldstein A (2020) Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. Neuron 105(3):416–434
2. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci 116(44):22071–22080
3. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld DS (2020) Does the whole exceed its parts? the effect of AI explanations on complementary team performance. arXiv:2006.14779
4. Hase P, Bansal M (2010) Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? arXiv:2005.01831
5. Chen C, Li O, Tao D, Barnett A, Rudin C, Su J (2019) This looks like that: Deep learning for interpretable image recognition. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, 8–14 Dec 2019. Canada, Vancouver, BC, pp 8928–8939
6. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
7. Yeh CK, Hsieh CY, Suggala AS, Inouye DI, Ravikumar P (2019) On the (in)fidelity and sensitivity for explanations. arXiv:1901.09392
8. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. In: Proceedings of the 32nd international conference on neural information processing systems, NIPS 18. Curran Associates Inc., Red Hook, NY, USA, 95259536p
9. Lie C (2019) Relevance in the eye of the beholder: diagnosing classifications based on visualised layerwise relevance propagation. Master's thesis, Lund Unversity, Sweden
10. Hooker S, Erhan D, Kindermans P, Kim B (2019) A benchmark for interpretability methods in deep neural networks. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: NeurIPS 2019, 8–14 Dec 2019. Canada, Vancouver, BC, pp 9734–9745
11. Olah C, Cammarata N, Schubert L, Goh G, Petrov M, Carter S (2020) Zoom in: an introduction to circuits. Distill 5(3)
12. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A (2019) Adversarial examples are not bugs, they are features. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, 8–14 Dec 2019. Canada, Vancouver, BC, pp 125–136
13. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R (2014) Intriguing properties of neural networks. In: Bengio Y, LeCun Y (eds) 2nd International conference on learning representations, ICLR 2014, Banff, AB, Canada, 14–16 Apr 2014
14. Eitel F, Ritter K (2019) Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In: Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support. Springer International Publishing, pp 3–11

15. Barnes BC, Elton DC, Boukouvalas Z, Taylor DE, Mattson WD, Fuge MD, Chung PW (2018) Machine learning of energetic material properties. arXiv:1807.06156
16. Sutre ET, Colliot O, Dormont D, Burgos N (2020) Visualization approach to assess the robustness of neural networks for medical image classification. In: Proceedings of the SPIE: medical imaging
17. Rieke J, Eitel F, Weygandt M, Haynes JD, Ritter K (2018) Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. In: Understanding and interpreting machine learning in medical image computing applications. Springer International Publishing, pp 24–31
18. Alvarez-Melis D, Jaakkola TS (2018) Towards robust interpretability with self-explaining neural networks. In: Proceedings of the 32nd international conference on neural information processing systems NIPS 18. Curran Associates Inc., Red Hook, NY, USA, 77867795p
19. Lillicrap TP, Kording KP (2019) What does it mean to understand a neural network? arXiv:1907.06374
20. Shen S, Han SX, Aberle DR, Bui AA, Hsu W (2019) An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Syst Appl 128:84–95
21. LaLonde R, Torigian D, Bagci U (2020) Encoding visual attributes in capsules for explainable medical diagnoses. In: Medical image computing and computer assisted intervention—MICCAI 2020. Springer International Publishing, pp 294–304
22. Elton DC (2020) Self-explaining AI as an alternative to interpretable AI. In: Artificial general intelligence. Springer International Publishing, pp 95–106
23. Belkin M, Hsu D, Ma S, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc Natl Acad Sci 116(32):15849–15854
24. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, Vardoulakis LM (2020) A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI conference on human factors in computing systems. ACM
25. Kansky K, Silver T, Mély DA, Eldawy M, Lázaro-Gredilla M, Lou X, Dorfman N, Sidor S, Phoenix DS, George D (2017) Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, ICML 2017, Sydney, NSW, Australia, 6–11 Aug 2017 (Proceedings of machine learning research, vol 70, pp 1809–1818). PMLR (2017)
26. Hendrycks D, Mazeika M, Dietterich TG (2019) Deep anomaly detection with outlier exposure. In: 7th International conference on learning representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019
27. Philipsen MP, Moeslund TB (2020) Prediction confidence from neighbors. arXiv:2003.14047

Joint Position Estimation and Synchronization of Clocks in WSN

Nikhath Tabassum, D. D. Geetha, and Rajashekhar C. Biradar

Abstract The data collected by the sensor nodes in a Wireless Sensor Network (WSN) is relevant only if the location and the time of data generation are known. Therefore, the position and time synchronization of the sensor nodes is important. In this paper, we treat these as one problem, and therefore, we proposed JLCS, a method that jointly localizes and synchronizes the nodes in the network. We utilize the Time of Arrival (TOA) method to determine the distance between the sensor nodes. During this time, the grading factor is exchanged between nodes to synchronize. Clock synchronization is achieved by taking square root of the clock time as many times as the grading factor. By simulation, we have compared the proposed method (JLCS) with the Unconstrained Weighting Least Square (UWLS) method for synchronization and localization errors. The proposed method has better synchronization and localization accuracy and suitable for larger networks.

Keywords Localization · Clock synchronization · Wireless sensor network · Time of arrival · Sensor nodes · Position estimation · Time synchronization · Convergence time

1 Introduction

Wireless sensor network (WSNs) consists of numerous sensor nodes which are distributed and connected. They are used for applications, such as surveillance [1], data collection [2], event detection [3], smart farm application [4] and monitoring gas pipelines [5]. In order for these applications to be useful, the geographical location of the origin of the data is necessary and also the time at which these samples were taken are also important. For example, if sensor nodes are used for forest fire detection. It is important to know the exact location of the sensor node which had detected the fire. It is also necessary to note the moment at which this event was recorded by the sensor node to ascertain the exact time of occurrence of the fire. The former

N. Tabassum (✉) · D. D. Geetha · R. C. Biradar
REVA University, Rukmini Knowledge Park, Yelahanka, Bengaluru, Karnataka, India

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_39

problem of locating the position of sensor node is known as localization and the latter is a problem related to time synchronization. These two problems were treated separately and the solution was derived separately. But the problem of localization and synchronization can be solved as a single problem.

Looking at the solutions for localization problem, Global Positioning System (GPS) looks like a good and easy choice. But GPS module is bulky, consumes more power, it is costly and does not give accurate result in indoor conditions. The next method for localization is through the Received Signal Strength (RSSI). The unknown sensor nodes estimate their distance to any 3 anchor nodes (nodes whose position are known) by the signal strength received from the anchor nodes. There are various algorithms [6, 7] that are based on RSSI. The implementation is simple but its accuracy depends on the Radio Frequency (RF) model that is chosen and the accuracy decreases if there are obstacles and multipath fading.

Coming to time synchronization, every node in the network has a clock time (CT) period. CT is not same for the nodes in an asynchronous network. The variation in CT is due to variation in operating temperature, humidity and ageing of crystal oscillator. Due to this, the CT can have two issues namely clock offset and clock skew. The clock offset is the constant difference between clock time. Clock skew is the persistent difference in the clock time even after the clock offset is set to zero. The time synchronization, is achieved when both clock offset and clock skew are eliminated. It is done, either by recursively modifying the clock parameters to reach a predetermined reference clock value [8] or by having some consensus [9]. In consensus-based method, all the sensor nodes in the network agree to reach a common time. The common time could be the maximum time among the participating nodes [10] or the average time [11] of all the sensor nodes clocks or Least common multiple (LCM) [12] of all the sensor clocks in the network. The latter consensus-based algorithms are easy, but they are prone to changes in the common time with the additions of new nodes to the network. The former recursive synchronization methods have high dependency on the reference clocks and have accumulation error as the network grows bigger. In order to overcome the clock dependency, to reduce accumulation error and to have higher accuracy for time synchronization in sensor nodes, we have proposed a novel clock synchronization method. In this proposed method, the Clock Time (CT) of the sensor nodes is synchronized to unity by taking square root of CT multiple times. So, if the nodes are time synchronized, then utilizing Time of arrival (TOA) [13] method becomes very easy. In the proposed TOA method, the sensor nodes calculate the propagation time to their neighbouring sensor nodes by a round trip message. By knowing the propagation time and the transmission speed of the signal, the distance between the nodes can be calculated. The main contributions of this work are as follows:

(1) The proposed time synchronization method eliminates the clock skew and clock offset effectively.
(2) The accumulation error is nil for time synchronization as there is no sharing of clock parameters between the nodes.

(3) The clock time can be scaled after synchronization as the resultant time synchronization achieved is unity.
(4) The localization of sensor nodes is implemented using TOA technique.
(5) The localization accuracy is higher and it does not deteriorate as the network size increases.

This paper is organized as follows. Section 1 consists of Introduction, Sect. 2 has Research method, Sect. 3 contains Result and Discussion, and Sect. 4 has Conclusion.

1.1 Related Works

The most recent related works of joint localization and clock synchronization are discussed in this section. We focus on only the TOA and TDOA techniques which are considered for the joint localization and synchronization of the nodes. The algorithms that are based on TOA technique with stationary anchor nodes are listed below. In a TOA-based technique [14], for synchronization and localization, it utilizes the Maximum Likelihood (ML) estimator by exchanging time stamped messages. The initial values are set for synchronization and localization. Then they are alternately solved, minimizing the cost function. But, it is good only in low noise level conditions. Another method [15], is efficient for high SNR as it utilizes weighted least square (WLS) technique by introducing an extra variable. Though it yields good result, it does not consider the effect of clock skew on time synchronization. The clock skew and position biased error are considered in [16]. This employs the Mean Square (MS) method to refine the positions by exploiting the relationship between nuisance parameter and sensor node. This method gives good results only if the initial value of position and clock time are close to minima. The initial value problem is overcome in [17]. The network is divided into L groups with P nodes. Each group is assumed to have common clock offset and linear equations are derived for each group separately. The clock offset is eliminated by subtracting two-time difference measurement from the same receiver group. However, the convergence time to reach localization is high is in this method. The next set of methods utilize the TDOA method with stationary anchor nodes. The TDOA technique [18] does not depend on clock parameters of reference nodes. This is the UWLS state method that we use to compare our proposed method with. The difference between the proposed method and UWLS is that, UWLS is a recursive algorithm and uses an initial estimate for estimating the location and synchronization. There are two approaches in UWLS method, first approach is to use linear estimator while taking distance measurements between anchor and sensor nodes. The second approach is to have Semi Definite Program (SDP) for the distance measurements between source nodes. The initial estimate of clock skew is taken as 1 and then arrive at the actual minima for synchronization. However, the accuracy is good, but there is inherent accumulation error as the network grows. But if the sensor nodes are distributed in form of an array [19], initially location estimate is obtained from ML estimator by the relative placement of nodes in the network based on direction of arrival (DOA) observed at each array. Time synchronization is achieved by linear LS by TDOA method. But the accuracy depends on the alignment of the array elements which have to be carefully placed for higher accuracy.

2 Research Method

Consider a network model consisting of N nodes. This network can be expressed as a graph of $G = (Q, E)$, where every node is represented as a Vertex $Q_i, i = 1, 2, 3 \ldots N$ and the edges are represented by E. An edge is present between two, if, they are within the radio range of each other.

We have the following assumptions in our model:

(1) The model considered is a dense network.
(2) There are no lonely nodes, all the nodes are connected in the network.
(3) The nodes know their clock time period (CT).

2.1 *Proposed Algorithm for Time Synchronization and Localization of Nodes*

The proposed algorithm is implemented by pairwise message exchanges among sensor nodes. Consider two nodes Q_i and Q_j communicating with each other as in Fig. 1. The following steps are employed to synchronize and localize the nodes.

- **Step 1** The first step of the proposed method is to remove clock skew. A packet of data containing the grading factor Ω and C_0 is transmitted from node Q_i at T_{i0} and received by Q_j at T_{j0}. Let C_0 be a time between the receiving and transmission of a packet from a node. The purpose of Ω is to set a bound for calculating the square root of CT for synchronization, as we discuss this later in this step. Consider the the clock equation

$$T_i = \alpha_i t + \beta_i \tag{1}$$

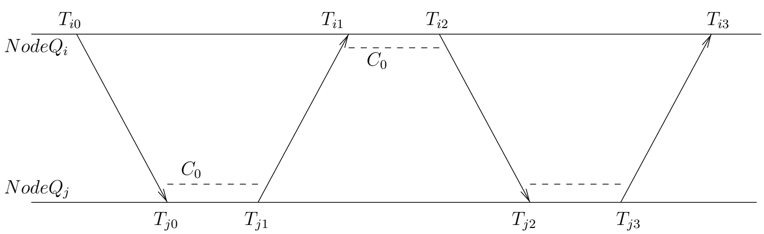


Fig. 1 Communication between nodes Q_j and Q_k with propagation time T_p

Where T_i is the CT of a node Q_i, α_i is the clock skew, β_i is the clock offset and t is the actual synchronized time. The square root of the CT is taken multiple times till it reaches unity. We take the square root of (1) to get (2)

$$T_i^{\frac{1}{2}} = (\alpha_i t + \beta_i)^{\frac{1}{2}}, \tag{2}$$

Taking square root Ω times of (2), we get

$$T_i^{1/2^{\Omega}} = (\alpha_i t + \beta_i)^{1/2^{\Omega}}, \tag{3}$$

If the grading factor Ω, is very large, the CT will reduce to unity. This is proved by taking Ω tending to infinity as in (4)

$$\lim_{\Omega \to \infty} (T_i)^{1/2^{\Omega}} = \lim_{\Omega \to \infty} (\alpha_i t + \beta_i)^{1/2^{\Omega}} = 1 \tag{4}$$

From (4), it can be seen that by employing this method both the nodes Q_i and Q_j Clock Times (CTs) reach unity. In order for their CTs to be unity, α should also also be unity, i.e. clock skew should also be unity. Therefore, the clock skew is overcome, without sharing any clock parameters among the two nodes. The sharing is only by means of an externally determined grading factor Ω which is common to all the nodes of the network.

– **Step 2** The proposed localization method utilizes the Time of Arrival (TOA) to estimate the distance between the neighbouring sensor nodes. The TOA is better than the widely used Received Signal Strength Indicator (RSSI). The distance measurement using RSSI method suffers from multipath and reflection [20]. The RF model is not very accurate. Hence, we choose TOA method which is efficient in estimating distance between nodes, even for larger distances. At this step, the nodes Q_i and Q_j have reached unity time with clock skew as unity. The node Q_j calculates the propagation time T_p from the two-time stamps T_{j1} and T_{j2} as in (5)

$$T_p = [T_{j1} - T_{j2} - C_0]/2 \tag{5}$$

Assuming the signal from one sensor to another sensor node, travels with the speed of light, with $C = 3 \times 10^8$ m/s, then the distance between node Q^i and Q^j is given by (6)

$$D_{(ij)} = T_p \times C \tag{6}$$

– **Step 3** In this step Q_i receives T_p along with the time stamp T_{j3} from Q_j and it calculates the distance to Q_j from (6). Though the clocks of the two nodes have reached unity and overcome the clock skew. They have to start at the same time to eliminate clock offset β, which is a constant difference that exists between the clocks. In order for the clocks to start at the same time, the time stamp T_{j3} is important as discussed below. β_j is the clock offset of node Q_j and β_i is the clock offset of node Q_i as in the (7) and (8), respectively, and t is the synchronized unity time.

$$T_{j3} = t + \beta_j, \tag{7}$$

$$T_{i3} = t + \beta_i, \tag{8}$$

When node Q_j transmits a packet at T_{j3}, this time stamp becomes the reference for Q_i to eliminate the clock offset between the nodes as in (10)

$$T_{j3} - T_{i3} = (t + \beta_j) - (t + \beta_i), \tag{9}$$

$$T_{j3} - T_{i3} = \beta_j - \beta_i \tag{10}$$

Therefore the clock synchronization and distance calculation between the nodes is achieved simultaneously.

2.2 Error Analysis

The time synchronization error, E_i for a node Q_i is measured by the difference in the synchronization achieved and unity time as in (11).

$$E_i = |T_i - 1| \tag{11}$$

For N nodes in the network, the root mean square (rms) error is E_{rms}, is given by (12)

$$E_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |T_i - 1|^2} \tag{12}$$

The rms error for the localization is D_{rms}.

$$D_{rms} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} [D_{i(act)} - D_{i(est)}]^2} \tag{13}$$

where $D_{i(act)}$ is the actual distance and $D_{i(est)}$ is the estimated distance from the proposed method.

3 Result and Discussion

The simulation was done in MATLAB. The nodes were distributed randomly in an area of 200 m $\times$ 200 m. The CT intervals of [0.1 ms, 0.9 ms] and [1 ms, 10 ms] were considered to see the results of the proposed method. In each trial, 100 random values were chosen from the above-mentioned time intervals with varying standard deviation. The error values were averaged over 50 such independent iterations.

3.1 Clock Synchronization Error and Localization Error Comparison of the Proposed Method JLCS with UWLS

For comparison with UWLS, the data set considered is in the interval $[0.1ms, 0.9ms]$ with a standard deviation of $\sigma = 0.1$ as in Fig. 2. The curve for the proposed method is almost constant. The reason for this behaviour is that, in the proposed method, the accumulation error is nil as there is no propagation of clock parameters between sensor nodes for synchronization and also because the clock skew is unity. Hence, the network size does not affect the synchronization error. It is observed in Fig. 2, that the clock synchronization error E_{rms}, for UWLS rises as the number of nodes increase, due to accumulation effect. Therefore, as the network grows bigger, the error also increases.

Since the clock synchronization is affected by accumulation error. The position estimation also suffers an error as the number of nodes increase in UWLS. The

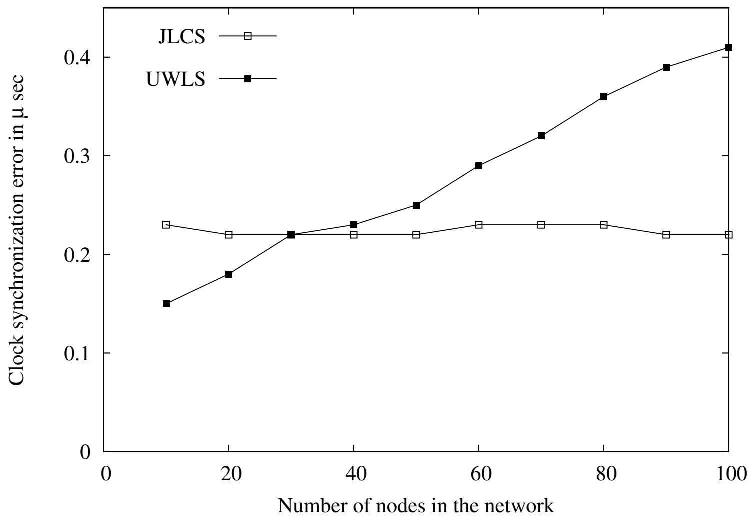


Fig. 2 Comparison of Clock synchronization error of UWLS and JLCS in the range [0.1 ms, 0.9 ms] with $\sigma = 0.1$

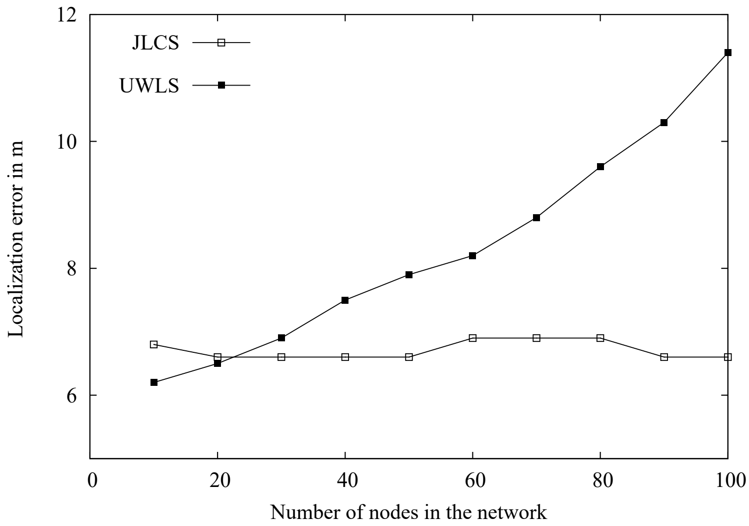


Fig. 3 Comparison of Localization error of UWLS and JLCS in the range [0.1 ms, 0.9 ms] with σ = 0.1

proposed method JLCS has almost a constant localization error, irrespective of the size of the network as in Fig. 3.

3.2 Convergence of CT of Sensor Nodes

To demonstrate the convergence of clock time period to unity, the range was taken from [0.1, 10] and the grading factor Ω, was fixed to 16 by trial and error. For a grading factor of 16, the CT reached to 1 and corrected up to 4 decimal places, i.e. all the values in the above-mentioned interval reached a value of 1.000 irrespective of the starting value as in Fig. 4. This value of unity can also be scaled up to a different value.

4 Conclusion

The proposed method jointly synchronizes and localizes the sensor nodes. It is easy to implement. The simulation results indicate that the clock synchronization and localization errors are almost constant as the network size increases. It indicates that this method is scalable for larger networks too. It overcomes clock skew and clock offset effectively. The clock value, once it is synchronized to unity, can be scaled to a desirable value also. This is a flexibility offered by the proposed method

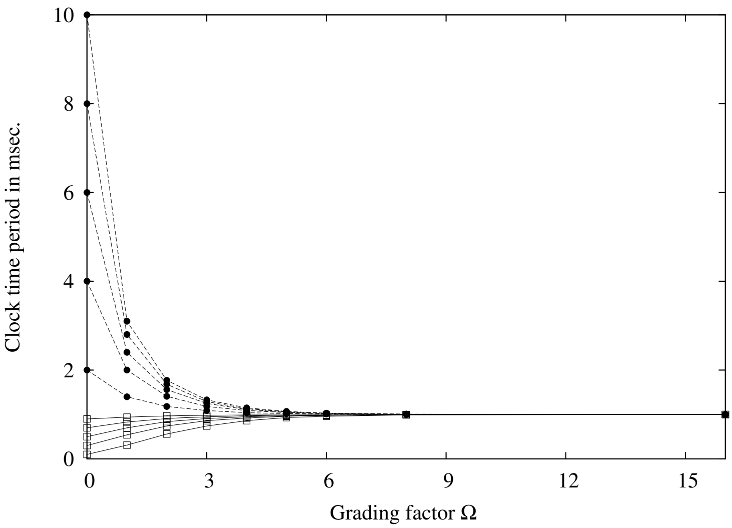


Fig. 4 Convergence of different clock time to unity

for synchronization. The localization accuracy is dependent on the synchronization accuracy of sensor nodes. The simulation results of the proposed method, match up with the state of the art UWLS method and performs better as the number of nodes increase.

References

1. Arjun D, Indukala PK, Unnikrishna Menon KA (2019) PANCHENDRIYA: a multi sensing framework through wireless sensor networks for advanced border surveillance and human intruder detection. In: International conference on communication and electronic systems, pp 295–298
2. Gao C, Wang Z, Chen Y (2019) On the connectivity of highly dynamic wireless sensor networks in smart factory. In: International conference on networking and network applications, pp 208–212
3. Kundu S, Das N, Das A (2020) Time series snapshot of event boundary detection and area estimation in wireless sensor networks. In: International conference on communication systems and networks, pp 563–566
4. Baldovino RG, Valenzuela IC, Dadios EP (2018) Implementation of low power wireless sensor network for smart farm applications. In: IEEE International conference on humanoid, nanotechnology, information technology, communication and control, environment and management, pp 1–5
5. Ivanov O, Avdeuk O, Bushmeleva K, Ivanov I, Uvaysov (2018) Model of calculating the reliability of a wireless sensor telecommunication system for monitoring the gas transmission network state. In: Moscow workshop on electronc and networking technologies, pp 1–5
6. Xu H, Sun H, Chang Y, Liu H (2016) Wireless sensor networks localization based on graph embedding with polynomial mapping. Comput Netw 106:151–160

7. Ruiz JC , Rivas-Perea R, Sfuentes E, Landaeta RG (2016) A recursive shortest path routing algorithm with application for wireless sensor network localization. IEEE sensors J 16:4631–4637
8. Perez-Solano Juan J, Felici-Castell S (2017) Improving time synchronization in wireless sensor networks using Bayesian inference. Netw Comput Appl 82:47–55
9. Panigrahi N, Khilar PM (2017) Multi-hop consensus time synchronization algorithm for sparse wireless sensor network; a distributed constraint-based dynamic programming approach. Ad-Hoc Netw 61:124–38
10. Zhang A, Bi C, Wan Y (2015) Time synchronization in wireless sensor networks under maximum value based time synchronization protocol with switching topology. In: IEEE 27th Chinese control decision conference, 16974p
11. Wu J, Jiao L, Ding R (2012) Average time synchronization in wireless sensor networks by pairwise messages. Comput Commun 35(2):221–33
12. Tabassum N, Devanagavi GD, Biradar RC (2017) Clock synchronization in wireless sensor networks using least common multiple. AEU-Int J Electron Commun 82:446457
13. Spano D, Ricciato F (2017) Opportunistic time of arrival localization in fully asynchronous wireless networks. Pervasive Mobile Comput 37:139153
14. Tian Q, Feng DZ, Hu HS, Yang F, Wei L (2019) Bi-Iterative algorithm for joint localization and time synchronization in wireless sensor networks. Signal Process 154:304–313
15. Huang J, Xue Y, Yang L (2013) An efficient closed form solution for joint synchronization and localization using TOA. Future Gener Comput Syst 29:776–781
16. Weng Y, Huang J, Yang L, Xue Y (2015) TOA based joint synchronization and source localization with random errors in sensor positions and sensor clock biases. Ad-Hoc Netw 27:99–111
17. Wang D, Yin J, Tang T, Chan X, Wu Z (2018) Quadratic constrained weighted least square method for TDOA source localization in the presence of clock synchronization bias: analysis and Solution. Digital Signal Proc 82:237–257
18. Wu X, Gu Z (2017) A joint time synchronization and localization method without known clock parameters. Pervasive Mobile Comput 37:154–170
19. Wozniak S, Kowalczyk K (2019) Passive joint localization and synchronization of distributed microphone arrays. IEEE Signal Process Lett 26(2):292–296
20. Yim J (2012) Comparison between RSSI-based and TOF-based indoor positioning methods. Int J Multimed Ubiquit Eng 7(2):221–234

Minimization of Cyber Security Threats Caused by COVID-19 Pandemic

Liqaa F. Nawaf, Chaminda Hewage, and Fiona Carroll

Abstract Whilst the world is preoccupied in its struggle with the Coronavirus pandemic, cyber-criminals are busy every day, spreading their own viruses, by phishing emails, data breaches, frauds, denials of service, and taking advantage of the vulnerabilities created by this crisis. In many ways, we, as a nation, are handing over our data without realizing it, without fully thinking it through or even being aware of cyber threats, which will ultimately have a tremendous impact on the governments and citizens both personally and at work. The goal of this paper is to investigate the correlation between the cyberattacks before the coronavirus and during the coronavirus in order to build an understanding of what is happening. To optimize cyber security and provide effective ways to tackle cyber security attacks during COVID-19 or something similar, we need to consider extra precautions and take a more secure approach to protection. To minimize the universal risks of data breaches and other cyber incidents, we need to enforce practical steps to deal with and if possible limit those risks. This requires not only thoughtful consideration, but also a good understanding of the opportunities that COVID-19 provides to cyber-criminals. The aim of this research paper is to investigate the growth of and reasons for the increase of cyberattacks during the COVID-19 pandemic. In order to make better cyber security decisions, we need to address and maximize the level of cyber security awareness and precaution taken during COVID-19. A set of practical steps to minimize the risk of cyber-attack is provided to compensate for the vulnerabilities associated with COVID-19.

Keywords Cyber-attack · COVID-19 pandemic · Cyber-criminal · Risk · Phasing · Data breach

L. F. Nawaf (✉) · C. Hewage · F. Carroll
Cardiff School of Technologies, Cardiff Metropolitan University, Cardiff, UK
e-mail: LLLNawaf@Cardiffmet.ac.uk

C. Hewage
e-mail: CHewage@Cardiffmet.ac.uk

F. Carroll
e-mail: FCarroll@Cardiffmet.ac.uk

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_40

1 Introduction

All over the world millions of network engineers, technicians, and system administrators are working determinedly to maintain safe and effective cyber systems for everyone. In a scenario where millions more people are working from home, they are faced with a high volume of people online, high network demands, and consequently a rising number of cyber threats. The COVID-19 pandemic is a worldwide threat to health and society that requires effective and immediate action by governments, businesses, and individuals. From the data collected, we can see that a number of threats have come from people using COVID-19 related lures to deploy malware.

Clearly, we feel that certain businesses have a significant role to play in reducing the possibility of cyber threats on society. It is also believed that prompt and effective action will reduce the risks to employees and the long-term costs to businesses. For example, the restrictions on face to face communication and social distancing requirements have made it vital for many households and organizations to go online to connect with work colleagues and family. This opens up opportunities to cybercriminals and hackers to exploit new weak points in the access to data and networks; the sheer volume of people socializing and working online gives these people more ways to install a malicious code or software on poorly protected computers.

Many organizations have implemented an information security management system to protect their data and networks, but it is the human element that poses the greatest threat. All organizations, which depend on business continuity, have found that the COVID-19 pandemic has created such a high risk for their day-to-day activities that they have had to reassess their security strategies carefully. For example, phishing emails went from 25,000 a day to 125,000—that equates to 500 per cent increase [1]. This indicates that the risk exists and is serious, whereas the firewalls included within the local broadband routers are adequate for certain attacks, but not perhaps for the criminal attacks that are on the rise in response to the increased personal use of computers and the unprecedented amount of homeworking. Moreover, while the NHS is working hard to mitigate the health risks from Covid-19, the cyber-criminals are in full control of malicious code which they install to increase the risks of a cyber-attack. This paper will investigate a sensible and practical set of steps to limit those risks and help secure transactions online in these very difficult times. The novelty of the paper lies in the layered approach to cyber security that we proposed and particularly the practical steps required to minimize and limit cyber security risks. This paper will firstly look at related research, it will then propose and discuss applying a layered approach to cyber security and finally provide a discussion centred on data from a Covid-19 cyber security perspective. The paper is organized as follows. Section 2 provides a literature review of related approaches. Section 3 describing and applying a layered approach to cyber security: From multi-factor authentication to cyber security awareness. Followed by Sect. 4, A COVID-19 'Data' perspective on cyber security that discuss cyber threats. Having evaluated the results, we then proceed the final section to draws some conclusions.

2 Related Work

Cyber-attackers, including those from aggressive nation-states, are more and more focused on targeting their actions on critical infrastructures [2]. Cyber criminals come in many different forms, each has an individual purpose, such as theft or interrupting communication and cause a significant loss of time and money. For example, the International Criminal Police Organization has recently exposed a €1.5 m face masks scam in Germany, Ireland, and the Netherlands [3]. Preying on the world's desperate need of face masks at present, some money was swiftly wired to Nigeria, but luckily the fraud was prevented by action from Interpol. In fact, successful attacks are not always sophisticated; even an unskilled attacker can sometimes achieve his/her illegal aims.

The issues, as realized in many of the recent news accounts of data breaches that few people know enough about these cyber security threats to protect private and powerful information [4]. Therefore, it is crucial for users to understand the cyber security issues that might affect them and to be equipped to combat cyber-attacks. To be specific, a recent survey stated that Denial of Service (DoS) attacks are considered to be the most widespread. They interrupt online services for one and sometimes many end-users [5]. Clearly, knowing about these types of attack may reduce the danger to their privacy and security.

As recent news reports highlight, the vulnerabilities and threats generated by the COVID-19 pandemic have demonstrated the high risk to cyber security. Cloudflare [6] which is the foundation supporting infrastructure, applications and teams, states that over the past few weeks online threats have risen to almost six times their usual level. The current pandemic provides a new impetus for cyber-attacks. One UK news report [7] claims that hacking and phishing attempts have risen by 37% month-on-month throughout the pandemic, whereas on some days, security was obstructed between four and six times as often as usual. Research from DARK Reading [8] also indicates the increase in threats and highlights. The top threat as that of phishing (55% of respondents), followed by wrong and fake information about COVID-19 on websites, malware, and ransomware attacks. The BBC News [9] stated that social networks such as Twitter, YouTube, TikTok, and Facebook have nurtured COVID-19 misinformation, "fake news", including viral messages asserting that 5G is related to Covid-19. YouTube has since banned these videos and other incorrect/fake statements.

It is clear that whilst risk has risen and the effects of threat are continuous, people worry that there will be a long-term effect on society and the cyber security industry. Thus, we believe that further awareness and precautions are essential, we need more sophisticated approaches (set of steps/procedures) to equip people with the knowledge to avoid a malicious attack (Fig. 1).

The graph shown in Fig. 1 demonstrates all the coronavirus-related cyber-attacks that have been identified by Check Point [10]. As can be seen, different types of threat technology extend across networks, endpoints and mobile devices. The graph highlights a continuous and rapid increase in cyber-attacks. However, it is not astonishing

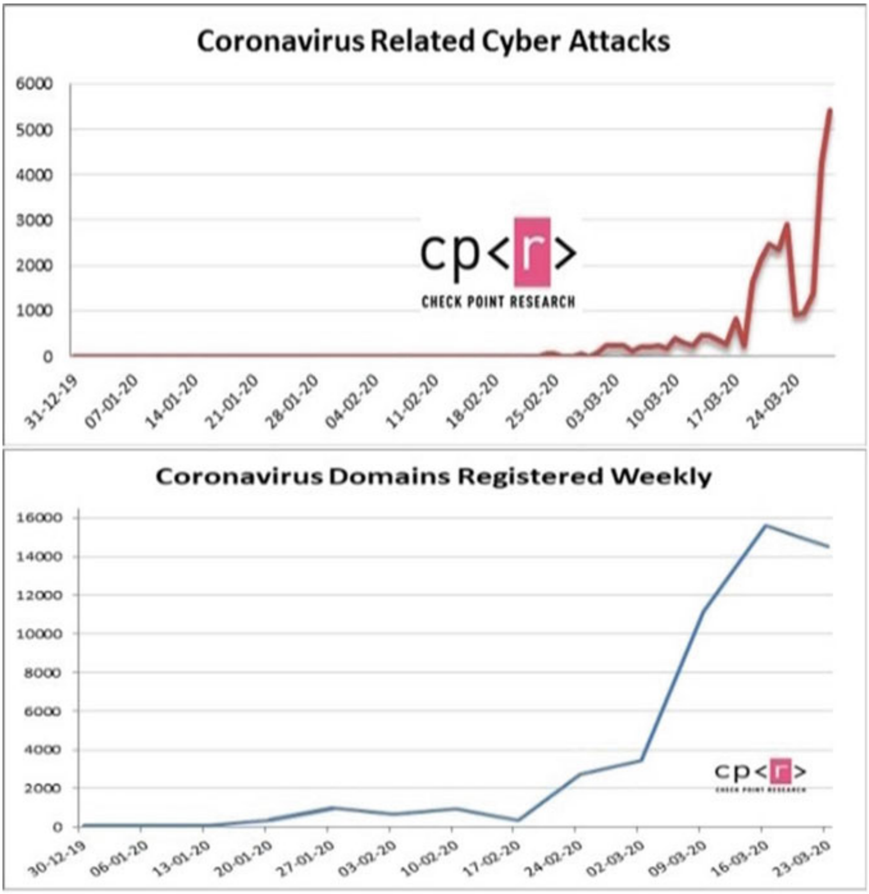


Fig. 1 Coronavirus-related cyberattacks and domains registered [10]

that the pandemic has caused a sharp increase. In recent weeks, a significant growth in the number of phishing attacks has been perceived by websites posing as Netflix sites [11]. This indicates the correlation between cyber-attacks and the increase in vulnerabilities caused by the pandemic.

Dark web-based drug (medical) dealing is on the rise. COVID-19 has caused a scarcity of essential medical supplies (e.g. Paracetamol, PPE, hand sanitizers) due to problems of sourcing and supply. Cyber-criminals have used this opportunity to advertise these goods on the dark web at significantly reduced prices to the large population of users who have tended for some time to buy certain items from this source. The items on the dark web market range from N95 masks, hand sanitizers, COVID-19 related drugs such as Chloroquine and even sell home testing kits for COVID-19 from some vendors. Since most of these products are new to the dark web marketplace, they carry no reviews from buyers. Hence, buyers inherently risk

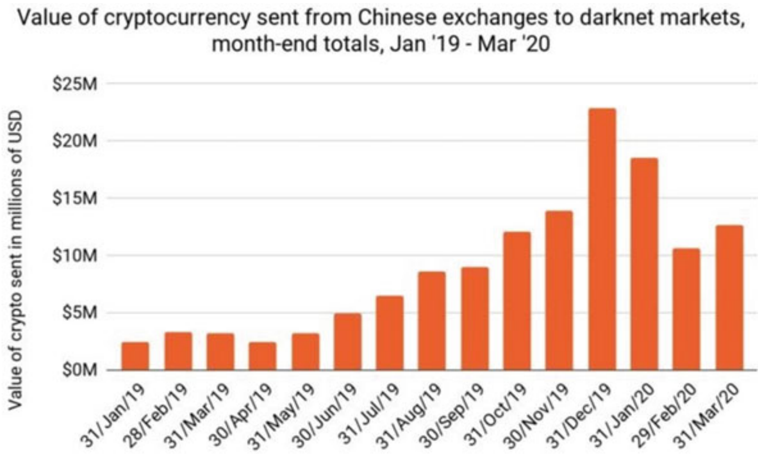


Fig. 2 The value of Bitcoin sent from Chinese exchanges to dark net markets [12]

not getting what they have paid for. It seems that the dark web is also facing supply chain problems due to the pandemic [12]. Hence, the amount of Bitcoins spent on the dark web has over the last two months has gone down. Figure 2 shows, according to the blockchain analysis company, Chainalysis, that the value of Bitcoins sent from Chinese exchanges to dark net markets has fallen significantly.

3 Applying a Layered Approach to Cyber Security: From Multi-Factor Authentication to Cyber Security Awareness

The previous section has indicated from the cyber security perspective the great increase of cyber-attacks and the impact of COVID-19 in the form of cyber-criminals exploiting the crisis. These cyber-attacks need to be known by the general public to prevent these and other cybercrimes from growing.

Currently, the healthcare specialists and hundreds of thousands of allied professionals are using the full capacity of their knowledge and resources to fight the current pandemic, though no one can as yet predict which approach will achieve the most successful cure. Regarding viruses of the other type, we feel that everyone can play a part by finding and sharing cyber security knowledge to create a solution that can minimize cyber-attacks. Some companies are working to identify measures used in banking and other services where sensitive data is held to better defend systems from cyber-attack. The technique is commonly known as two-factor or multi-factor authentication, which uses an application installed on a smartphone, or sends out an SMS or email to verify who is logging in. In more detail, this provides an extra layer of security, in the form of requirements, the user must fulfil in order to be granted access. This technique is essential for protecting data from a cyber-attack and depends on more than simply relying on knowing the username and password to access a secure system. To reduce the current risk level, a whole series of procedures need to be accepted and end-users need to be aware of the threats to their security and safety. Moreover, in light of COVID-19, the joint advisory staff from the UK's National Cyber Security Centre and the USA's Cyber security and Infrastructure Security Agency are working with law enforcement and industry partners to find new ways to detect and prevent malicious COVID-19 cyber activities [13]. We feel that to improve the resilience against the phishing attack and/or other criminal attacks, we should extend the defences to combine the more technical measures with a heightened cyber security awareness.

4 A Covid-19 'Data' Perspective on Cyber security

Having demonstrated the possibility of minimizing the risk level, we now report further work that was done to expand the range of measures and investigation of cyber-attack. A simple algorithm was used to generate an appropriate data set taken from MISP Project-Threat Sharing, which is an open source software program and set of standards to share, create, and confirm threats to intel and intelligence. The data set listed places at random that are exploiting the impact of the COVID-19 pandemic. The data set of the intrusion detection systems for a given attribute was generated. This data set contains various categories, types, values, object names, and countries/regions. We analyzed the data set as follows to evaluate the solutions in this paper:

The data set included different types of cyber threat during the current pandemic which have to be minimized or if possible stopped.

The data set contains various types of attachment for different categories of malware attack, (see Fig. 3a and b) showing that the most often used attachment to deploy malware on the victim's computer is a file. Different categories of malware

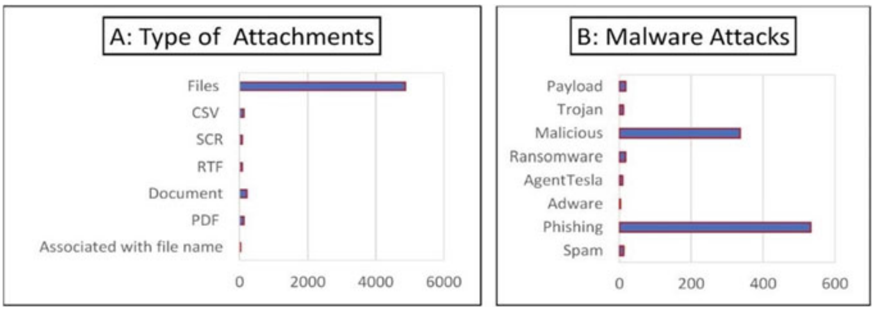


Fig. 3 **(a)** Type of attachments and **(b)** value of malware attacks

attack reveal that a great number of phishing and malicious attacks were used against victims. In addition, a small number of payload, Trojan, ransomware, adware, agent tesla, and spam, was applied. This indicates that phishing awareness must be a highly eligible way to help victims mitigate the cyber threat.

According to the indicators of compromise (IOCs), it is evident that there is an increased cyber activity related to COVID-19. Cyber-criminals use the panic situation or the uncertainty created by the pandemic to lure general users to click on a malicious link or use some other path to hack into a victim's computer and steal data or to direct further attacks. This has been facilitated by the nationwide lockdowns currently imposed in many countries. As a result, many more people are using the Internet to learn, socialize, and work from home. This has paved the way for the would-be attackers to invade users' privacy and security controls. According to Fig. 4, it is clear that these have increased their domain name registration with some form of reference to COVID-19. These domain names can be used by cyber-criminals for disinformation, malvertising and as malicious domains which closely align with genuine websites (Cyber-squatting). Hence, domain names need to be closely monitored for their use in order to tackle all forms of malicious activity not only during the present pandemic, but also afterwards. A significant increase of hostnames which refer to COVID-19 are also apparent. These staged usernames can be used to lure victims in phishing campaigns. Furthermore, a number of malware payloads have been discovered. Some of them are being currently investigated by organizations such as Virustotal. It is highly likely that these malwares were not generated during this pandemic, but have lately been reshaped for activation during the pandemic.

The way that cyber-criminals work may not have changed since before COVID-19. They have simply seen an added opportunity thanks to the infodemic and increased use of the Internet. Whilst the main attack weapons used by the criminals remain the same (e.g. phishing attacks), they may change in format and appearance to attack via the disease. The key solution is user awareness. Whilst more than a quarter of the global population is in lockdown, the hackers roam freely. They will use every opportunity available for them to steal important data or harm individuals or organizations. Hence, it is important to increase the awareness of the public and ask them to always use standard security measures for and preserving privacy.

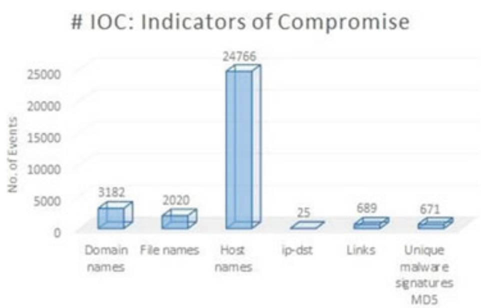


Fig. 4 IOC for cyber related incidents with COVID-19

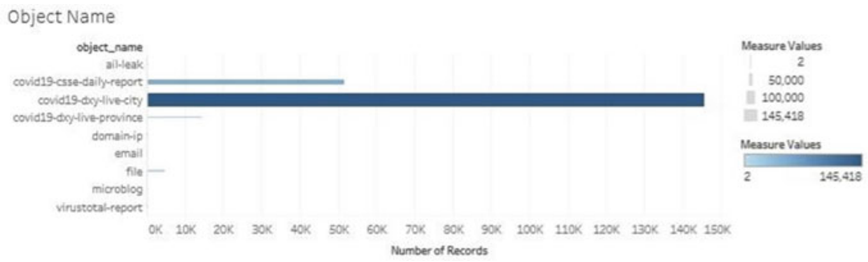


Fig. 5 The main source of cyber threat and number of records/websites carrying threats

Endsley talks about the importance of situational awareness for humans to make decisions within a wide range of environments [14]. She defines situational awareness as the perception of the elements in the environment, comprehension of the situation, and projection of future status. Aligning with this ideology, we envision a set COVID-19 cyber security awareness guidelines as, first, supporting the user in becoming aware of the elements in their online environment. As we have found in this research, 'hostnames' linked to COVID-19 are proving to be frequent sources of threat, and as end-users we need to be cautious of these COVID-19-related sites. Second, on these websites (see Fig. 5, which highlights daily reports and city reports of COVID-19 as types of websites carrying threats), and/or receiving emails about COVID-19-related subjects, we need to know how to identify malicious characteristics and elements such as files, or inferences drawn from these files or, in sum, phishing attacks. Third, particularly during the present crisis, we need to have the foresight to predict from these inferences so as to protect everyone's security. Although the COVID-19 is affecting over 210 countries, and people across the world are facing the same devastating impact from its spread, they should be aware that cyber-criminals feel no sympathy and that they will take advantage of people's fear to attract them and lure them to the trap. Therefore, we feel that this three-pronged approach (supported by reliable data) will help support end-users to identify the visible threats, reduce their areas of weakness, and enhance their cyber security awareness, especially these days.

5 Conclusion

The vulnerabilities and threats associated with COVID-19 pandemic have been highlighted and discussed in the previous sections. To mitigate those risks, security protection such as providing an employer's internal computer system with the following is needed: imposing a working time restriction even on staff who work from home; and employing the Multifactor Authentication of the state of the pandemic. We advocate for a layered approach to cyber security and the practical steps to minimize and limit those risks must be taken. In addition to the technical, people must be made aware of the sophisticated human threats in order to enable them to avoid COVID-19 fake news and deception completely. The social networks claim to take significant steps in fighting awkward coronavirus posts and suggesting or calling for apps to catch fake COVID-19 news that could cause harm [9]. As we have seen from the data, the threats are extensive and it is very challenging to eliminate all the risks of a data breach and other cyber-attacks in the face of the fast changing and increasing uncertainty of coronavirus in the population. Saying that, we feel that a layered approach that will provide end-users with the required knowledge to be able to identify malicious characteristics and other adverse cyber signals is essential. Furthermore, the government and organizations must be capable of reacting quickly because of the possibility of non-stop cyber-attacks if action is not taken.

References

1. Tyler N (2020) Covid-19 cyber security threat to impact businesses. Newelectronics, 1 Apr 2010. https://www.newelectronics.co.uk/electronics-blogs/covid-19-cyber-security-threat-to-impact-businesses/225701
2. Thatte R (2018) Is cyber security the biggest technological challenge that we are facing today? 6 July 2018. https://scet.berkeley.edu/reports/is-cyber-security-the-big-gest-technological-challenge-that-we-are-facing-today/. Accessed 4 Apr 2020
3. Admin (2020) Interpol foils €1.5m face masks fraud traced to Nigeria, 16 Apr 2020. https://observerstimes.com/2020/04/16/interpol-foils-e1-5m-face-masks-fraud-traced-to-nigeria/. Accessed 16 Apr 2020
4. Weiss GL (2018) Network effects are becoming even more important on emerging plat-forms, 18 Mar 2018. https://www.forbes.com/sites/startupnation-central/2018/03/18/why-a-network-effect-is-the-only-way-your-startup-can-win/#1ef866de7527. Accessed 4 Apr 2020
5. Kolias C, Kambourakis G, Stavrou A, Gritzalis S (2016) Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset. IEEE Commun Surv Tutor 18(1):184–208
6. CloudFlare (2020) Cloudflare was built for this, 28 March 2020. https://www.cloudflare.com/en-gb/. Accessed 5 Apr 2020
7. Muncaster P (2020) Cyber-attacks up 37% over past month as #COVID19 bites. UK/EMEA news report, 1 Apr 2020. https://www.infosecurity-maga-zine.com/news/cyberattacks-up-37-over-past-month/#at_pco=smlrebh-1.0&at_si=5e89dbda83e6cab5&at_ab=per-2&at_pos=3&at_tot=4. Accessed 4 Apr 2020
8. Staff DR (2020) Vulnerabilities-threats 71% of security pros see threats jump since COVID-19 outbreak, 7 Apr 2020. https://www.darkreading.com/vulnerabil-ities-threats/71-of-security-pros-see-threats-jump-since-covid-19-outbreak/d/d-id/1337498?fbclid=IwAR3N8EF83CPJW89Z1u2RtIwpIjmAEVsJSKGwqAJvMdjGWjvi-DJyJLSekIc. Accessed 9 Apr 2020
9. DM, Kelion L (2020) Coronavirus: call for apps to get fake Covid-19 news button. BBC News, 9 Apr 2020. https://www.bbc.co.uk/news/technology-52157202. Accessed 9 Apr 2020
10. Fauci DA (2020) Coronavirus update: in the cyber world, the graph has yet to flatten, 28 Mar 2020. https://blog.checkpoint.com/2020/04/02/coronavirus-update-in-the-cyber-world-the-graph-has-yet-to-flatten/. Accessed 5 Apr 2020
11. Chadwick J (2020) Explosion in cybercrime targeting Netflix users detected since the coronavirus outbreak with hundreds of phishing schemes and fraudulent websites set up to trick viewers. Associated Newspaper Ltd, 15 Apr 2020. https://www.dai-lymail.co.uk/sciencetech/article-8221407/Netflix-users-targeted-cybercriminals-COVID-outbreak.html. Accessed 16 Apr 2020

12. Hamacher A (2020) Coronavirus is driving drug users to the dark web and Bitcoin. Chainalysis, 1 Apr 2020. https://decrypt.co/24193/coronavirus-is-driving-drug-us-ers-to-the-dark-web-and-bitcoin. Accessed 20 Apr 2020
13. National Cyber Security Centre & Cybersecurity Infrastructure Security Agency (2020) Advisory: COVID-19 exploited by malicious cyber actors, 8 Apr 2020
14. Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. Human Factors J 37(1):32–64. https://doi.org/10.1518/001872095779049543

SMEs in South Africa: The Era of Adopting Mobile Payment Solutions

Alick Chingapi and Adriana A. Steyn

Abstract This research paper is the result of analysis into critical factors impacting the adoption and wide spread usage of mobile payments by South African small medium enterprises (SMEs). The research involved a qualitative research approach utilising interviews as a data collection method. The findings from the interviews with the SMEs identified critical factors and themes to be considered regarding mobile payment usage and adoption by SMEs. The findings revealed that the factors impacting mobile payment adoption by South African SMEs were risk, convenience, ease of use, trust in service providers, system features, device features and issues, cost of fees, company image and credibility, Bluetooth connection, customer service and integrated systems. The research gives key recommendations in the form of a framework for adoption of mobile payments by South African SMEs.

Keywords Acceptance · Adoption · Framework · Merchants · Mobile payments · South African SMEs

1 Introduction

SMEs can be seen as the building blocks of a society and its economy due to the impact that they have on job and employment creation [11, 23, 47, 63, 64]. They are critical in the building of an economy through economic and social development [24]. [64] as well as [55] discusses how SMEs are important in many economies in the world and how the adoption of information communication and technology (ICT) is seen as a factor in the growth of SMEs.

The mobile device has become an extension of the human body. It is the gateway to our social lives through social network sites such as Facebook, Instagram and Twitter, our work lives through email, work applications, our personal lives (personal contacts, SMS, instant messaging services) and our financial lives [17, 56, 64]. SME's

A. Chingapi (✉) · A. A. Steyn
University of Pretoria, Pretoria, South Africa

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_41

acceptance of payment methods is understudied yet they are the providers of the goods and services that consumers will spend their money on.

South Africa is one of the most competitive economies in sub-Saharan Africa according to the World Economic Forum, Competitiveness 2019 Report [58]. South Africa is also one of the largest economies in Africa and is exposed to the international community through mineral export and oil import, technology, foreign direct investment and economic activity [67]. Firms operating in such an environment need to be able to adapt dynamic capabilities to survive these environments [65].

From 2007 to 2015, mobile payment research has spread to Asian countries such as South Korea and Japan with limited research in developing countries such as Africa. The following mobile payments research has focused on merchants in a South African context: merchants in South Africa using Snapscan [51], merchants in greater Cape Town area using mobile payments [31] and mobile payments in South African townships [40, 71]'s study on South Africa consumers. Research of [36] was focused on the European market while [45] focused on mobile payments in developed countries, highlighting the failures and the role of banks in the ecosystem [45]. [44] carried out a review of the efforts of mobile payment platforms between 2005 and 2015. This was based on the repeated failures of launches within the European markets, in countries such as Holland, Spain, Germany and Norway. [44, 28] viewed the mobile payment systems as multi-sided platforms as they bring together more than one set of users: consumers and merchants. [45] argue that mobile payment solutions are part of a platform with multiple parties that need to interact with each other before delivering value to the consumer and merchant. These parties include mobile network operators, financial institutions, content providers, credit card companies and technology firms [36].

This research proposes a conceptual dynamic framework for the mobile payment adoption by SMEs with a specific focus on the relationship to dynamic capabilities. This study contributes by developing a framework for the adoption of mobile payments by SMEs in South Africa. This framework allows for better understanding of drivers and barriers to mobile payments in South Africa by SMEs within a developing country. The outcomes of this paper would enable mobile payment service providers to have a holistic view of what they need to do for mobile payment services to be more readily accepted by the public (merchants and consumers) in South Africa. Given the multi-sided nature of the market under research, there is a need to understand what value mobile payments provide to the different actors in the mobile payments' ecosystem, which is currently under-researched [14]. Reference [70] argues that merchant and customer acceptance should be studied separately despite the high interdependence between the two.

The research question, this paper will address, is thus: What factors impact the SMEs adoption and usage of mobile payments in South Africa?

The rest of the paper is structured as follows. Section 2 details a literature review into SME merchants and mobile payments followed by the theoretical discussion which forms the basis of the framework. This is followed by a discussion on methodology as well as the data collection and analysis process which was followed. The results are presented along with a discussion and conclusion.

2 Literature Review

Small businesses are the key enabler to employment creation to counter the increasing unemployment rates of 27.2% [2, 7, 8, 20, 39, 47, 49]. SMEs contributed 36% to the GDP of South Africa in 2015 [19], however ranks very low in terms of ease of starting a business, in fact, ranks 74/190 compared to other Africa countries such as Mauritius (49th), Rwanda (56th) and Morocco (68th) [26].

The importance of SMEs especially in the African context cannot be overstated [29]. SMEs can be seen as the building blocks of a society and its economy due to the impact that it has in terms of job and employment creation [8, 20, 39]. They are critical factors in the building of an economy through economic and social development [18, 19, 24]. [55] acknowledges the importance of SMEs in many economies across the world and how the adoption of ICT is seen as a factor in the growth of SMEs.

According to [55], micro-businesses employ 1 to 9 people, small businesses employ 10 to 49 employees and medium businesses employ 50 to 500 employees, however this definition differs from country to country [42]. The classification of SMEs can be further extended by market sector, location, innovation rate, asset value and organisation [3]. For the purposes of the research, the researcher applied the definition as per the [42] as defined above.

At the end of 2015, there were more than 2.2 million small medium and micro-enterprises (SMMEs) in South Africa with more than two-thirds (1.49 million) of these SMMEs being informal [7]. A year later, there were 2.34 million SMMEs but this number decreased to 2.25 million by end of 2017 and was attributed to people being employed into formal sector jobs [59]. According to the [7], the more formal SMMEs are residing in Gauteng and the Western Cape, whereas the informal SMMEs operates in more rural areas. A worrying trend is highlighted in the [59] report indicating an increased number of survivalists' SMMEs.

2.1 Mobile Payments

There is limited literature on mobile payment adoption in developing countries, especially in Africa, and this study aims to add to the body of knowledge. [44] call for more research into the field of mobile payment systems as there is still not enough reasoning or agreement as to why mobile payments have not been more widely accepted as initially expected. [16] confirmed the lack of studies into merchant the mobile payments ecosystem. Literature review of [14] found 188 papers in mobile payments that focused on the technological and consumer aspects between the period 2007 and 2014, supporting the findings of Dahlberg et al. (2008) on the limited research into merchant's role in the ecosystem. The study conducted a systematic literature review and found there is limited knowledge and research to understand the merchants' role and adoption of mobile payments especially in the African context. This supports the findings by [21] of the limited attention paid to the role of merchants in the mobile payment's ecosystem. The gap in knowledge is crucial because the growth of mobile payments usage in South Africa presents opportunities for vendors, merchants and consumers alike. Using previous research in mobile payments in Europe, Asia and South America, this research aims to understand the reasons, the choice of payment technologies in the stated market and produce a dynamic conceptual framework that can lead to the increased understanding and adoption of mobile payment technologies by SMEs in South Africa. This study combines the TOE theory with dynamic capabilities to provide insights into business' capabilities to be considered for merchants in the context of mobile payment ecosystem.

A mobile payment is defined as the process in which the payer employs a mobile device to initiate an electronic procedure that initiates a financial transaction [37, 52]. [33] viewed mobile payments as a subset of mobile commerce whereas [9] define mobile payments as a subset of electronic commerce where at least one side is processed through a mobile device. [46] state that mobile payments are carried out wirelessly through a mobile device. [16] expand on this stating that mobile payments are a form of payment for goods and services through a mobile device using wireless and communication technologies. Guo and Bouwan (2016b) update the definition to include the initiation, authorisation and confirmation of payment processes using wireless or other communication technologies. Based on the definitions, it can be summarised that a mobile payment is any form of payment that is initiated, processed and completed using a mobile device through the wireless application protocol, SMS service, mobile wallet or some form of application.

Payments are the critical process in the provision of good and services such as purchasing electricity, food, water, paying taxes, healthcare, salaries and education as all these services help an economy function. [46] noted that the digitisation of the payment process would become essential due to the success of e-commerce. This can be argued to be part of the evolution of the shopping process, with mobile payments being a part of the process [46]. The merchant adoption of mobile payments was seen as a challenging process as merchants are the ones who have to pay the mobile payment provider for the service similar to credit card fees [70].

In light of the framework of [66], [16] refined the model and put together the framework in Fig. 1. This was further refined by Guo and Bowman (2016) based on their studies of merchants operating in China. The framework indicated that consumers and merchants have influence on the mobile payment service provision which are all influenced by external factors such as culture, technology, legal, regulatory and commercial factors. This framework takes into account the factors that form the basis of the Technology-Organisational and Environmental (TOE) theory and hence from a research perspective makes the TOE theory a suitable choice for this paper.

The four corners in Fig. 1 indicate the external factors that have an impact on the mobile payment ecosystem. Examples are

- Changes in social/cultural environment: Covid-19 impact on social lives [69].
- Changes in technological environment: The introduction of a smartphone by Vodacom that cost less than 1000 Rand [53].

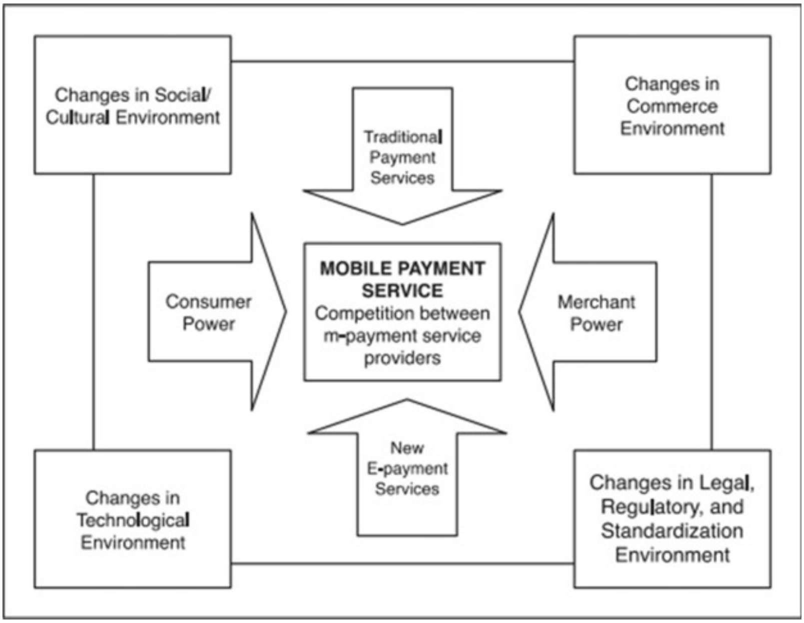


Fig. 1 Framework: mobile payment ecosystem [32] [16]

- Changes in legal, regulatory and standardisation environment: commencement of the POPI Act in South Africa in June 2020 [61], Commission and Vodacom (South African service provider) settlement agreement on data prices [12].
- Changes in commerce environment: Impact of the Steinhoff collapse [41]; lockdown impact on South African economy [62].

All these external factors have an impact on economic and business within the mobile payment ecosystem. Despite the alternative payment forms in developed countries, such as mobile payment technologies and card payments, cash is still the pre-dominant form of payment worldwide with 85% of the value of transactions being in cash [38]. In selected developed countries (Germany, France, USA, Austria, Canada and Netherlands), cash was still in dominant use and not just for small payments [5]. The European Central Bank noted that 75% of payments in the euro-zone are made in cash despite the digital age [54]. [5] identified through their model that consumers will prefer to use cash for payment if they have enough cash on hand despite alternative payment method options. [25] proposes that due to the anonymity that cash provides, it is the easiest form of payment for consumers to use as there are no pin numbers, account numbers to memorise or limits in most instances.

Mobile payments can be categorised into two sectors based on their transaction value [52]. Firstly, macro-payments: these are typical normal transactions, e.g. purchasing plane tickets, bill payments and remittances and are defined as payments of a higher value 50 Rands and higher [33]. Methods of payments are carried out through short message service (SMS), near-field communications (NFC), quick response (QR) codes, mobile applications and web browser for mobile phones [32]. These methods allow for the transmission of financial messages.

Micro-payments: These are payment values up to 50 Rands [33]. [14] and [46] define micro-payments as peer-to-peer (P2P) and are small value transactions which are usually paid for by cash or debit card to the value of 205 Rands.

Reference [43] defined possible handset designs, multi-application chip card, dual sim phone, external card reader, dual-slot phone and payment software built in the phone. The multi-application chip card has a SIM card and a wireless identification module combined in a single card [43].

2.2 SME Mobile Payment Adoption Factors

Transactions Fee

The transaction fee is regarded as a major factor impacting adoption of mobile payments by merchants [66, 70]. Large retailers can afford the cost of investing in mobile payment technologies and rolling out devices at their own costs; however, SMEs are already under significant pressure financially and hence this would be an extra overhead they cannot afford. [70] details how most merchants are small and medium enterprises and would prefer a cheaper alternative such as cash payments.

Ease of Use

Positive adoption factors were found to increase the prospect of increased sales and reduction in transaction costs for payments [37]. Ease of use on the merchant's side was identified as a cause for concern, i.e. in the payment process when an employee did not understand how the mobile payment system worked and how to use it and resolve issues when faced with a customer who wants to pay using a mobile device [70].

Perceived Trust and Security

[37] cite a lack of trust by merchants in any of the multiple parties involved in the payment process being a significant barrier to adoption; however, [37] further state that few studies have examined security in mobile payments from the view of merchants. Trust and security are deemed inhibitors to mobile payments if not addressed [70]. A factor such as security was identified from the initial research and is still a pertinent factor for adoption and mass acceptance from both consumers and merchant perspective [43].

More trust is placed in larger telecommunication companies and financial institutions [37]. This was raised as an inhibitor to adoption by [66] with more than 50% of the responses citing secuity as an issue. [43] identifies authentication, availability, data integrity, confidentiality as some of the criteria relating to security that need to

Table 1 Identified constructs

Incompatibility with existing business	[35, 37]
Trust and security	[35, 37, 40, 51, 52]
Perceived lack of standardisation	[37, 46, 51, 70]
Cost (relative to substitutes)	[1, 35, 37, 40, 52, 66, 70, 46]
Ease of use (relative to substitutes)	[1, 31, 51, 70, 46]
Perceived risk	[70]
Network effects	[51, 52]
Network externalities	[51]
Business model	[35, 37, 46, 51]
Technology compatibility	[35, 37, 46, 51]

be addressed as part of ensuring that security and trust are embedded in the system and process.

Network Externalities

Network externalities refers to the benefits that consumers will enjoy due to the presence of a network [15]. Universal usage and wide-scale acceptance were identified as important prerequisites for mobile payment penetration. [37] discuss critical mass as a part of network externalities. Favourable network effects can result in critical mass adoption which is a component for success in the mass adoption of mobile payments [44]. A large number of customers using or asking about a certain payment option such as mobile payments would convince merchants to adopt and offer mobile payments [66]. High usage rates in the long term are identified as consumer patterns that would likely convince merchants to adopt mobile payments [35]. Another reason for not adopting is the lack of merchant involvement in the development process. Merchants seek a process that is easy to use and most importantly a strong element of trust [13]. Table 1 summarises the constructs that will be used in this research, based on the literature.

A literature review and assessment of the T-O-E provided a basis for categorising these constructs within the dynamic framework to provide a theoretical basis paper, see Fig. 2.

3 Theoretical Foundation and Research Framework

More than 91% of the research conducted in the mobile commerce field between 2008 and 2016 was carried out through quantitative methods indicative of a strong positivist paradigm by the researchers [10]. Despite calls by [14] in the field to diversify the research methods and to enrich the data of mobile-based research, the quantitative

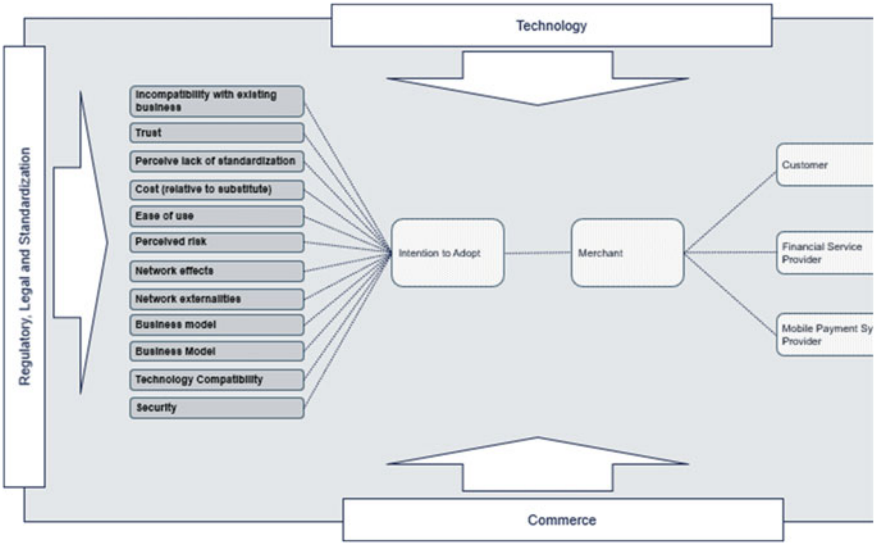


Fig. 2 SME mobile payment conceptual framework

trend still continues. The following sections and sub-sections provide the two theories applied in this study. Given the setting of the SME Mobile Payment Conceptual Framework (SME MPCF) and exposure to various external and internal forces, the TOE theory provides a complimentary lens into the ecosystem for this research. Given the early stages of mobile payments and maturing business models and two-sided market views, the dynamic capabilities framework enables the deeper understanding of the business models that are in effect in the mobile payments ecosystem.

Reference [65] indicates that a framework is derived from reality and attempts to classify constructs and depict their relationships in a meaningful way. The framework that has been proposed here is underpinned by the theoretical lens of the T-O-E. A conceptual framework is a useful way for a researcher to express ideas, concepts and to represent their findings [60]. T-O-E considers environmental, internal and external aspects. Another reason is that [50] argues that there are a limited number of firm-level theories that can be used to study the adoption of IT in firms. Reference [4] states that a combination of T-O-E and Diffusion of Innovation (DOI) has been shown to be most effective given the consistency and applicability of DOI. Reference [6] further state that due to the highly differentiated nature of technological innovations, no one theory or model can be all encompassing to explain the adoption of innovations by firms. Reference [34] add that T-O-E is more ideally suited for this study than the DOI as it takes into account the environment context.

Due to the nature of the innovation of mobile payments focusing on business-level research, the T-O-E is the most appropriate framework to use [4]. Environmental and organisational factors have been shown to influence e-commerce adoption by

SMEs in Tanzania [30]. This theory examines how firms adopt technology innovation through the different contexts of environmental, technologically and organisationally [4, 50]. The T-O-E framework has been used extensively in explaining organisational adoption of information systems and new technologies such as cloud computing adoption by Saudi SMEs [4], adoption of green practices in SMEs [50] and e-Government in Jordanian companies [68]. The T-O-E framework has been used in different research strategy approaches with both qualitative and quantitative data collection methods [27].

4 Methodology

4.1 Research Design

To gain a deeper understanding of the phenomena at the heart of the study, this research used semi-structured interviews as the main data collection method due to the ability to enable an in-depth analysis of the participant's views, thoughts, actions and behaviours. The questions were based on the constructs within the context of the TOE framework. In the research design process, the interview questions were designed in accordance with the objectives of the research and based on literature. The questions were divided into sections, starting with a brief overview of the SME followed by questions relating to technological, organisational and environmental aspects.

The interviews were conducted with 18 SMEs, predominantly the owners of the SME businesses that operated in different locations around South Africa. This was to ensure a broad perspective given the different conditions faced by SMEs in different parts of South Africa. In person, interviews were conducted with the sessions lasting between 20 and 60 min each.

The SME interviewees were selected based on having already engaged in e-commerce, m-commerce activities and who possessed at the least a point-of-sale device. Contact was made with SMEs or SME owners in the urban and semi-urban areas of the nine provinces of South Africa (Gauteng, North West, Northern Cape, Western Cape, Eastern Cape, Free State, KwaZulu Natal, Mpumalanga and Limpopo). The list and details of the interviewees are detailed in Table 2.

The list of SMEs was obtained from each province's business directory sites as well as by referencing mobile payment provider websites. The interviews were conducted from May 2019 to July 2019. In the interview process, the interviews were recorded on a digital device and the researcher made notes to supplement the audio recordings. Upon completion of the interviews, these were transcribed and each response to the questions was analysed and discussed in the section below.

Table 2 SMEs industry and years in operation

SME	Positions	Number of employees	Payment technology	Industry and business focus	Number of years in operation
SME A	Owner	None	Yoco and Snapscan	Healthcare—online health store	5 years
SME B	Owner	None	Yoco and Cash	Publishing—comic books and novels	3 years
SME C	Owner	None	EFT and Cash	Retail—natural hair products	3 years
SME D	Owner	1	Nedbank Pocket POS	Retail—shoe accessories	2 years
SME E	Owner	9	Bank POS and Cash	Hospitality—fast food outlet	2.5 years
SME F	Owner	9	Zapper and Bank POS	Hospitality—restaurant	2 years
SME G	Owner	10	Yoco, iKhokha and Cash	Hospitality—restaurant and catering	5 years
SME H	Owner	1	Yoco, Snapscan, Bank POS and Absa Pebble	Retail—clothing	12 years
SME I	Owner	100	Yoco, Snapscan, Bank POS and Zapper	Hospitality—restaurants	6 years
SME J	Owner	None	Yoco and Snapscan	Retail—women's clothing	2 years
SME K	Manager	20	Yoco and Snapscan	Retail—women's clothing	7 years
SME L	Owner	2	Bank POS and Nedbank Pocket POS	Retail—women's clothing	2 years
SME M	Owner	2	iKhokha, Snapscan and Bank POS	Education—second-hand books store	7 years
SME N	Owner	6	Yoco, Snapscan, Zapper and Bank POS	Hospitality—street food	6 years
SME O	Owner	7	Yoco	Retail—clothing	6 years

(continued)

Table 2 (continued)

SME	Positions	Number of employees	Payment technology	Industry and business focus	Number of years in operation
SME P	Owner	1	Yoco and Cash	Education—book store	2 years
SME Q	Owner	6	Yoco	Hospitality—restaurant	2 years
SME R	Owner	None	Zipzap and EFT	Manufacturing—handbags	2 years

4.2 Data Analysis

Using the TOE framework, the researcher analysed and coded the interviews as well as the notes taken during the interviews. The in vivo coding process was carried out on the transcripts before axial coding was carried out. [57] recommended axial coding to complement a process such as in vivo to enable a deeper understanding and bring out more connnections, causalities and relationships between codes. The axial coding process entailed looking for similarities among codes, repetitions, relationships, causalities and looking for different perspective to view the codes and to ask if some were really codes. A refinement in the process occurred as some of the codes were too narrow, i.e. only one instance or too broad, and on further examination there were two or three codes. The axial coding refined the codes to just under 150 unique codes. Table 2 is an overview of all the SMEs interviewed.

This research carried out the process at a code, category and theme level.

5 Results

Technology adoption decisions are impacted by technology, organisation and environmental impacts.

System Features

System features play an important role in the themes described by the interviewees as they enable the business operations, payment processes as well as aspects such as customer experience. System features have a link to the customer experience and hence ease of use factor.

"Reporting is an issue as we only receive SMS notifications, so we go through those one by one" (SME D).

Some of the mobile payment systems facilitated their processes manually before the introduction of the mobile payment systems and hence the positive mentions as per below in terms of the system features.

"It allows us to track stock and orders" (SME K).

However, the evidence is not consistent as different SME merchants had alternative experience. This was not a consistent feature available on all systems and hence resulted in some manual processes in the stock taking process remaining. In the instance of the feedback shared by SME K, this was noted as they had two mobile payment solutions (Yoco and Snapscan) and hence the feel below feedback referred to Yoco.

"Yoco does not manage stock so right now we use excel spreadsheets" (SME K).

As a construct, system features had both positive and negative mentions and hence features as a construct in the updated model.

Trust in service providers

Trust is a strong factor in mobile payment acceptance literature and also featured strongly in the feedback from the interviewees, SME D stated:

"It is a bank approved product"

This is indicative of the strength of the company brand and the inherent trust based on the fact that the provider is a bank, an established financial entity. However, SME H said "The trust is inherent until there is an issue".

This contrasts other feedback that indicated that the type of a service provider, a bank or an independent/private mobile payment system provider played a role when looking at trust.

Device features/issues

Device issues had a very high frequency in the codes and categories during the analysis. Interviewees, who had mobile payment systems involving interaction between the mobile phone and the mobile payment device had this construct as a common code which emerged.

"It does go down, I had issues when it went down" SME G.

Instances of mobile payment applications crashing were noted strongly. Given that entrepreneurs exist within an ecosystem, this feedback is shared by other entrepreneurs when inquiring about payment systems to use. As this extended to physical devices, it was noted that there was sometimes a syncing issue between the device and the mobile phone:

"Well it crashed the one time when I was at a pop-up market, it was not syncing" (SME)

In this particular instance, this resulted in the business owner cancelling that particular mobile device, deferring back to the traditional point-of-sale units from a bank. The device features and frequency of issues with the system are real concerns; however, this can also be viewed as poor customer service from the mobile payment system provider as well.

Cost of fees

Cost emerged in the literature as an important construct which was confirmed during the interviews. In the analysis, cost of fees was noted as a code and as a category with the highest number of codes. The evidence shared by the interviewees ranged from positive to negative opinions based on mobile payment systems that was based on a need to reduce costs:

"The cost of using these devices is less than the cost of depositing cash" (SME F).

"The most important thing that I look at these mobile things is really the fees" (SME).

This is mirrored by similar feedback that was consistent throughout the interviews. One very passionate and incensed interviewee operating four restaurants stated:

"Yes, the ****** fees, the fees are too high, too high" (SME A).

The evidence suggest that cost is a serious consideration to take into account before making a decision on the type of payment system to use in the business. These costs are both the once off costs of the device and system as well as the monthly/transactional cost associated with using the system.

Simple to use/Difficult to understand

The evidence from the interviewees indicated that the simplicity to use the mobile payment devices was not as straightforward. The most dominant feedback was that the systems were a bit more complex to use and were not as intuitive or easily understood when in operation. One SME stated that the type of mobile payment system that they had taken up from a financial institution required that they use a certain version of smartphone and this forced them to incur more costs by buying new smartphones for the business:

"So, it meant you had to have proper phones and everything, which is a challenge" (SME G).

"That was the most challenging part, trying to figure it out from the start, it was not straight forward" (SME D)

This indicated the lack of intuitiveness of the devices and processes once the mobile payment devices were in operation and hence posed a challenge to the SMEs.

Risks

Risk emerged as a consistent factor in the feedback from the interviewees, including things like data breaches, security, loss of cash and physical security. These factors had a positive impact on the decision to use mobile payment systems:

"It was to reduce cost and to give the customer the option of not carrying cash" (SME G)

It was also noted that the use of electronic devices came with a risk and hence this was viewed as a two-sided factor. There were positive reasons to consider in the use of mobile payment devices, such as mitigating risks including theft, physical harm in cases of robbery. The risks noted with the use of mobile payment devices were encapsulated with: "There is always a risk when you give credit card information to a third party like Zapper" (SME D).

Statements such as this also speaks to an earlier mentioned construct of trust. There seems to be evidence that constructs such as risk and trust go hand in hand in influencing the decision to acquire a mobile payment solution.

Company image and credibility

The use of electronic payment systems seems to be an indicator of a better company image and some level of credibility compared to the use of cash in the business transactions. SME P stated that the use of mobile payment technologies does improve the image that their business portrays:

"Yes, it does improve my company image, that is the first thing people ask, when they walk into the shop."

This is supported by SME Q, an owner of an African cuisine restaurant stating that

"Yes. A lot of African Restaurants operate on a cash only basis. It makes us look more professional that people can pay via their cards using Yoco."

The evidence was not as consistent, it indicated that some SME merchants were thinking of customer's perceptions based on the technology in store.

Bluetooth connection

A large majority of the more well-known mobile payment systems that involved the pairing of a smartphone to an additional payment device involved the use of Bluetooth. The Bluetooth connection factor raised issues and resulted in instances where there was failed connections at vital points at markets. This resulted in lost sales, as customers would not wait for the devices to pair up. This was as a sticky issue in a process that already involves friction in the payment instance at check-out/sale process:

"I wish I didn't have to pair it with my phone to work" (SME D)

There is evidence of issues linked to the Bluetooth connection factor leading to misunderstandings in the process in instances where the business owners thought the device had failed in the transaction process:

"There is always like a delay between the phone and the machine, so you don't ehm, you don't take that into consideration, so you think the transaction has failed, in the meantime it hasn't, it is just a delay" (SME N).

Business processes/Customer service/Payment options strategy

This is a new construct and speaks to the impact that the mobile payment system has on the business process and customer service. The evidence shows linkages in the three constructs, payment strategy, business processes and customer service. SME D shared that: "It hasn't improved business processes (laughs)".

This indicates that the interviewee had expectations of an improvement in business processes that could have resulted in quicker operations times, improved customer service, improved sales and this was not materialising. SME I saw mobile payments as part of their strategy as the economy toughened and they were affected by electricity shortages.

"I think it became part of our adaptive strategy, so when we first opened, no one thought we would have rolling black outs so often."

The emergence of this construct is partly based on the environmental factors that have impacted businesses in South Africa such as increased power outages. To increase or continue offering the same customer service, there is evidence of the use of payment solutions that can be used during power outages.

Integrated systems

The implementation of another system had cost and complexity implications with the addition of another device and system impacting the existing business processes. This resulted in additional processes or amendments to an existing process:

"Our systems are not merged, they are separate" (SME A).

The interviewee noted that there was a need to integrate the systems so that reporting could be carried out from one place.

"If there was some-way to integrate everything holistically so that you get all of your statements through one portal per se" (SME A).

SME D noted that she had to use a book and write down her sales while scanning the phone for the SMS's that confirmed payment because the system had no reporting function. This summarised that integrated system was a note of concern in the payment process.

Convenience

Given that most of the mobile payment systems involve the use an additional device, this construct was driven by the need to ensure that customers did not suffer in the process of making a payment. SME L shared that:

"So, it was more the convenience that you were offering to your customer that as a form of payment".

SME N stated that "I think that from a convenience view you are able to accept credit cards, debit cards for any event, any market, any time so that's it".

So, the convenience construct is also a two-sided factor as it impacts both the customer and the SME merchant in the process.

6 Discussion and Conclusion

This section discusses the research impacting mobile payment adoption by SMEs in South Africa. The insights based on the analysis indicate that there are more factors to take into account than the literature suggested. The identified constructs within the framework are grouped into the following themes: business decision-making; customer access and marketability; the impact of payment systems and payment process on the business, infrastructure setup, support and connectivity and operating a business and its processes. The identified constructs all play important roles in the adoption decision by merchants; however, the themes that they are grouped into also play an important role in influencing the decision to adopt mobile payments. The TOE describes an all-encompassing view of both internal and external lenses.

The study has shown the need for an all-encompassing understanding of technological, organisational and environmental factors in the SMEs usage of payment technology. Factors such as cost, ease of use and risk have been highlighted consistently in findings from consumer research as significant factors in adoption and this is mirrored by SMEs who offer the payment options as a factor that impacts the adoption and use of the technologies. The updated constructs from the study are detailed on the right-hand side. There are common constructs such as trust, costs, ease of use and risk which are supported by the literature reviews. The new factors identified, such as customer service, company image and credibility and convenience, show that non-technology factors also play a significant role in influencing the adoption of mobile payments by SMEs.

As per the code, category and theme analysis, these factors had a more significant impact on an SME's use, decision-making process than the initial set of constructs in the framework.

The research adds value to the academic body of knowledge through a comprehensive review of existing literature on mobile payments and SMEs in South Africa. The findings have revealed that for mobile payment adoption to take effect and be impactful in SMEs in South Africa, a multi-disciplinary approach must be considered. The research extends the current understanding of mobile payments especially in developing countries, specifically an African country. It highlights critical themes that must be taken into account by financial institutions, information technology ministries and private firms offering mobile payment technologies.

There is limited literature that seeks to establish a framework depicting adoption of mobile payments by SMEs. By answering the research calls to provide more research into the SMEs role in mobile payments, an extension of knowledge about mobile payments has been bridged. The focus on South Africa adds further value to the research and body of knowledge of ICT adoption by SMEs in Africa. This is an area that has received less attention compared to developed countries as shown in the literature review. The research was not exhaustive and there is a need to validate and authenticate the results by applying the study to a wider range of SMEs across South Africa. This can be complemented by a multi-disciplinary data collection approach. A comparative study of other African countries that have similar economic, socio-economic, political climates to ascertain if there would be noticeable differences would be helpful in establishing and broadening the legitimacy of the study and results.

The research has contributed to the body of knowledge by identifying key factors that significantly impact the adoption and use of mobile payment technologies by SMEs in South Africa. The study highlighted this through the use of the TOE model as a baseline. The study shows that technology plays a key role in the adoption but organisational and environmental factors play a pivotal role in the decision process and hence use of mobile payment technologies by SMEs in South Africa. By incorporating this framework, SMEs in South Africa, and hopefully other developing countries, will be able to adopt and gain a competitive advantage through the successful implementation of mobile payment devices.

References

1. Abebe F, Lessa L (2020) Factors affecting mobile payment adoption by merchants in Ethiopia. In: The 6th Africa conferecne on information systems and technology (ACIST2020), Addis Ababa
2. Adeniran VT, Johnston AK (2016) The Impacts of ICT utilisation and dynamic capabilities on the competitive South Africa SMEs. Int J Inf Technol Manag 15(1):59–89
3. Akhavan P, Fathian M, Hoorali M (2008) E-readiness assessment of non-profit ICT SMEs in a developing country: the case of Iran. Technovation 28(9):578–590
4. Alharbi F, Atkins A, Stanier C (2016) Understanding the determinants of cloud computing adoption in Saudi healthcare organisations. Complex Intell Syst 2:155–171
5. Arango-Arango CA et al (2018) Cash remains top-of-wallet! International evidence from payment diaries. Econ Model 69:38–48
6. Awa HO, Ukoha O, Igwe SR (2017) Revisiting technology-organisation-environment (T-O-E) theory for enriched applicability. Bottom Line 30(1):2–22
7. Bureau for Economic Research (2016) The small medium and micro enterprise sector of South Africa. University of Stellenbosch, Stellenbosch
8. Cant CM, Wiid AJ, Hung Y (2016) Internet-based ICT usage by South African SMEs: the barriers faced by SMEs. J Appl Bus Res 32(6):1877–1888
9. Carat G (2002). ePayment systems database: trends and analysis. http://epso.jrc.es/Docs/Background-9.pdf. Accessed 29 Apr 2018
10. Chhonker MS, Verma D, Kar AK (2017) Review of technology adoption frameworks in mobile commerce. Procedia Comput Sci 122:888–895
11. Chiware ER, Dick AL (2008) Information needs and information seeking patterns of small, medium and micro enterprises in Namibia. Electron Libr 24(1):24–36
12. Competition Commission South Africa (2020) Commission and vodacom settlement agreement on data prices. Competition Commission South Africa, Pretoria
13. Dahlberg T, Anssi O (2007) In: Proceedings of the 40th Hawaii international conference on system sciences. Hawaii
14. Dahlberg T, Guo J, Ondrus J (2015) A critical review of mobile payment research. Electron Commer Res Appl 14:265–284
15. Dahlberg T, Mallat N (2002) Mobile payment service development—managerial implications of consumer value perceptions. s.l., ECIS
16. Dahlberg T, Mallat N, Ondrus J, Zmijewska A (2007) Past, present and future of mobile payments research: a literature review. Electron Commer Res Appl 2(1):1–117
17. Davel C (2017) The mobile phone as an extension of the self: a study among adolescents in secondary school. University of South Africa, Johannesburg
18. Esselar S, Christoph S, Ndilwana A, Deen-Swarray D (2007) ICT usage and its impact on profitability of SMEs in 13 African countries. Inf Technol Int Dev 4(1):87–100
19. GEM (2016) Global report 2016/17, Global Entrepreneurship Research Association (GERA), New York
20. Gono S, Harindranath G, Ozcan GB (2016) The adoption and impact of ICT in South African SMEs. Strat Chang 25(6)
21. Guo J, Bouwan H (2016) An ecosystem view on third party mobile payment providers: a case study of Alipay wallet. Info 18(5):56–78
22. Guo J, Bouwman H (2016) An analytical framework for an m-payment ecosystem: a merchants' perspective. Telecommun Policy 40:147–167
23. Habberton G, Notuctt K (2013) Unlocking the growth of SMEs and social businesses in South Africa. https://www.coursehero.com/file/29160874/Unlocking-the-growth-of-SMEs-and-social-businesses-in-South-Africadocx/. Accessed 11 Sept 2020
24. Harris MA, Patten KP (2014) Mobile device security considerations for small-medium-sized enterprise business mobility. Inf Manag Comput Secur 22(1):97–114
25. Hazra D (2017) Monetary policy and alternative means of payment. Q Rev Econ Financ 65:378–387

26. Herrington M, Kew P, Mwanga A (2016) Can small businesses survive in South Africa. Global Entrepreneurship Monitor, New York
27. Hoti E (2015) The technological-organisaitonal-environmental framework of IS innovation adaption in small and medium enterprises. Evidence from research over the last 10 years. Int J Bus Manag 3(4):1–14
28. Huynh K, Nicholls G, Shcherbakov O (2019) Explaining the interplay between merchant acceptance and consumer adoption in two-sided markets for payment methods. Bank of Canada, Ottawa
29. Jones P (2011) ICT impact within the SME sector. J Syst Inf Technol 13(2):1–4
30. Kabanda S, Brown I (2017) A structuration analysis of small and medium entreprise (SME) adoption of E-commerce: the case of Tanzania. Telemat Inform 34:118–132
31. Kalan A (2016) Understanding merchant adoption of m-payments in South Africa. University of Cape Town, Cape Town
32. Khiaonnarong T (2014) Oversight issues in mobile payments. International Monetray Fund, Washington
33. Kreyer N, Pousttchi K, Turowski K (2002) Characteristics of mobile payment procedures. Lyon, ISMIS 2002 workshop on M-services
34. Maduku DK, Mpinganjira M, Duh H (2016) Understanding mobile marketing adoption intention by South African SMEs: a multi-perspective framework. Int J Inf Manag 36:711–723
35. Mallat N, Dahlberg T (2005) Consumer and merchant adoption of mobile payment solutions. In: Tseng A, Saarinen T, Tinnila M (eds) Managing business in a multi-channel World. IDEA Group Publishing, Hershey, p 347
36. Mallat N, Rossi M, Tuunainen VK (2004) Mobile banking services. Commun ACM 47(5):42–48
37. Mallat N, Tuunainen VK (2005) Exploring merchant adoption of mobile payment systems: an empirical study. In: International conference on mobile business (ICMB 2005), Sydney, pp 347–353
38. Mastercard (2017) World economic forum on Africa. Mastercard, Durban
39. Mavimbela R, Dube E (2016) Can an internet adoption framework be developed for SMEs in South Africa. J Entrep Innov Emerg Econ 2(2):120–135
40. Mhlongo KM (2016) Assessing the diffusion and use of mobile payment solutions: a case of the South African townships. University of South Africa, Pretoria
41. Mittner M (2017) JSE closes weaker on Steinhoff collapse as Wiese empire takes a huge hit. https://www.businesslive.co.za/bd/markets/2017-12-06-jse-closes-weaker-on-steinhoff-collapse-as-wiese-empire-takes-a-huge-hit/. Accessed 15 June 2018
42. National Small Business Act (1996) National small business act. Government Gazetter, Pretoria
43. Ondrus J (2003) Mobile payments: a tool kit for a better understanding of the market. University of Lausanne (HEC), Lausanne
44. Ondrus J, Gannamaneni A, Lyytinen K (2015) The impact of openness on the market potential of multi-sided platforms: a case study of mobile payment platforms. J Inf Technol 30:260–275
45. Ondrus J, Gaur A (2012) The role of banks in the mobile payment ecosystem: a strategic asset perspective. In: International conference on electronic commerce, Singapore
46. Ondrus J, Pigneur Y (2005) A disruption analysis in the mobile payment market. IEEE, Hawaii
47. Packham TB, Miller C, Brooksbank D (2004) The use of websites for SME innovation and technology support services in Wales. J Small Bus Enterp Dev 11:400–407
48. Patton MQ (2002) Qualitative research and evaluation methods, 3rd edn. Sage Publications, Thousand Oaks
49. Peris M, Nuttgens M, Lindermann N, von Kortzfleisch H (2013) Acceptance of professioanl Web 2.0 platforms in regional SME networks: an evaluation based on the unified theory of acceptance and use of technology. In: 46th Hawaii international conference on system sciences
50. Piaralal SK, Nair SR, Yahya N, Karim JA (2015) An integrated model of the likelihood and extent of adoption of green practices in small and medium sized logistics firms. Am J Econ 5(2):251–258

51. Pidugu K (2015) Mobile payment adoption in South Africa: merchant's perspective. Gordon Institute of Business Science, Pretoria
52. Pousttchi K (2004) An analysis of the mobile payment problem in Europe. In: Multikonferenz wirtschaftinformatick (MKWI), mobile business systems, mobile and collaborative business, techniques and applications for mobile commerce, pp 260–268
53. Reporter S (2018) Vodacom launches Android Go Smartphones. https://www.iol.co.za/business-report/technology/vodacom-launches-android-go-smartphones-15380355. Accessed 20 June 2018
54. Reuters (2017). Reuters. https://www.reuters.com/article/us-ecb-payment-cash-idUSKBN1760U6. Accessed 6 May 2018
55. Rhodes C (2017) Business statistics. House of Commons Library, London
56. Sagl G, Resch B (2015) Mobile phones as ubiquitous social and environmental geo-sensors, 1st edn. IGI Global, New York
57. Saldana J (2016) The coding manual for qualitative researchers, 3rd edn. SAGE, London
58. Schwab K (2019) The global competitiveness report 2019. World Economic Forum, Geneva
59. Small Enterprise Development Agency (2018) SMME quarterly update: 3rd quarter 2017, Small Enterprise Development Agency, Johannesburg
60. Smyth R (2004) Exploring the usefulness of a conceptual framework as a research tool: a researcher's reflections. Issues Educ Res 14:167–180
61. South African Government (2020) Protection of personal information act 4 of 2013. https://www.gov.za/documents/protection-personal-information-act. Accessed 14 Aug 2020
62. South African Reserve Bank (2020) Full quarterly bulletin—no 296—June 2020, South African Reserve Bank, Pretoria
63. Steyn AA (2018) Changing thoughts towards digital literacy interventions for South Africa entrepreneurs. Read Writ 9(1)
64. Steyn AA, Leonard CA (2012) Guidance for SMEs with the adoption of technology: a conceptual framework. S Afr J Entrep and Small Bus Manag 5:24–33
65. Teece D (2007) Explicating dynamic capabilities: the nature and microfoundation of (sustainable) entreprise perfomance. Strat Manag J 28:1319–1350
66. Teo E, Fraunholz B, Unnithan C (2005) Inhibitors and facilitators for mobile payment adoption in Australia: a preliminary study. In: International conference on mobile business, ICMB 2005, IEEE, Los Alamitos
67. The World Bank (2019) The World Bank in South Africa. https://www.worldbank.org/en/country/southafrica/overview. Accessed 31 Aug 2020
68. Thi LS, Eam LH, Issa al zoubi M (2014) Estimating influence of TOE factors on E-government usage: evidence from Jordanian companies. Int J Bus Soc 15(3):413–436
69. United Nations (2020) Everyone included: social impact of COVID-19. United Nations, Department of Economic and Social Affairs—Social Inclusion, New York
70. van der Heijden H (2002) Factors affecting the successful introduction of mobile payment systems. In: 15th bled electronic commerce conference eReality: constructing the e-economy, Bled
71. Verkijika SF (2020) An affective response model for understanding the acceptance of mobile payment systems. Electron Commer Res Appl 39

Harmonic Compensation Strategy for Increasing the Operating Range of Solar-Fed Single-Phase Cascaded H-Bridge Multilevel Inverter

Ravi Ranjan Kumar and Jayanti Choudhary

Abstract In this paper, we have presented a single-phase cascaded H-Bridge (CHB) grid-connected solar inverter, Due to the unmatched solar radiations and the atmospheric temperature variation of solar panel in a single-phase CHB solar inverter, the unmatched output power among solar cell may cause the H-bridges overmodulation with high power, arise in a distorted grid current. overcome this problem, we have developed a harmonic compensation strategy (HCS) that injects a certain quantity of harmonic into the higher power H-Bridge cell to maintain the magnitudes of their modulation waveforms turns out to be near unity, then we injected the same quantity of reverse harmonic for compensation and appropriately distributes among the remainder H-Bridge cell. The developed methodincreases the operating range of grid current and also keeps the THD of inverter grid current in an allowable range under power unmatched cases, the effectiveness and validation of the developed method are confirmed by simulation results.

Keywords Cascaded multilevel inverter · Pulse width modulation · Solar cell · Unequal power · Five-level inverter · Harmonic compensation strategy

1 Introduction

WORLDWIDE demand for renewable energy resources, especially solar energy, wind energy, are tremendously growing with reducing the cost of solar panel and inverters in terms of energy shortage in the last 12 years and environmental concerns [1–3]. These methods inject a high-quality current into the grid solar inverter with lower harmonic distortion, achieve maximum efficiencies with improve the system operation stability and reliability, and allow a transformerless connection to the grid [4, 5]. There are many multilevel develop topologies, such as CHB, flying capacitor, and NPC [6], Among them, the CHB topology is a more desirable inverter because of its more reliable, simple, and outstanding efficiency and it makes this topology

R. R. Kumar (✉) · J. Choudhary
National Institute of Technology Patna, Patna 800005, India
e-mail: ravik.pg18.ee@nitp.ac.in

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_42

we have used CHB inverter that better choice for solar applications [7]. Each H-Bridge cell in a CHB solar panel grid multilevel inverter connects only one solar module. Each solar module in the inverter can operate at maximum power (MPPT) by regulating independently DC-link voltage to improve operating range of grid current. However, the unequal ambient temperature and solar radiation of solar panels which will cause unequal DC side powers among solar panels may create the H-Bridge cell overmodulation with higher power cell, result in distorted grid current [8, 9]. Several technical papers have presented for inverter solar systems to fulfill the demand of MPPT and current control. MPPT algorithm is proposed to enhance the inverter system performance under over modulation case due to unbalance power. It is able to achieve the balance power among the H-Bridge solar module but reduce the energy harvesting. An RPCS method is described in [10–12], which can enhance the CHB solar inverter performance by decreasing the power factor of the inverter network, although a lesser power factor may confine its utilization. In [12, 13], a hybrid modulation method is described to enhance the performance of the converter. Which employs the SPWM technique for shaping grid AC current and for DC-link capacitor voltage stabilized used low-frequency switching. However, the hybrid modulation strategy increases the modulation span from 1 to a maximum around $5/\pi$, which could inflame DC-link voltages variation. In [14, 15], a THCS topology is presented to prevent the solar cell bridge from devoid of higher modulation during power unmatched cases. Although the modulation indexes of a few H-Bridge cells reach up to 1.255, the magnitude of their modulation or reference waveforms will not be more than unity and these H-Bridge cell will still be over modulation.

For this concern, we have proposed in this paper harmonic compensation topology to increase the operating range of CHB solar cell inverter. To keep the magnitude of modulation waveform of H-Bridge solar cell greater than one, the proposed topology injects the harmonic into the overmodulation cell (OMC). Therefore, the method can ensure to keep their magnitude of reference waveform being unity. Meanwhile, the equal amount of reverse harmonic for compensation injected to the H-Bridges and distributed properly to NC with lesser power, so the total output AC voltage of CHB multilevel inverter does not contain the injected harmonic sequences components. The designed strategy can make sure all H-Bridge power cell free from overmodulation during power unbalance cases.

2 Cascaded H-Bridge Inverter

The CHB inverter is a famous method and attracts widespread interest due to the higher demand for high-power medium-voltage operation. Here, we have designed a CHB inverter consisting of two units of HB solar cells. The H-Bridge (HB) power modules are generally connected in series on their output AC side to get high voltage and lower THD. The CHB inverter needs multiple units of independent DC source from solar cell, all these feed an individual to HB inverter (Fig. 1).

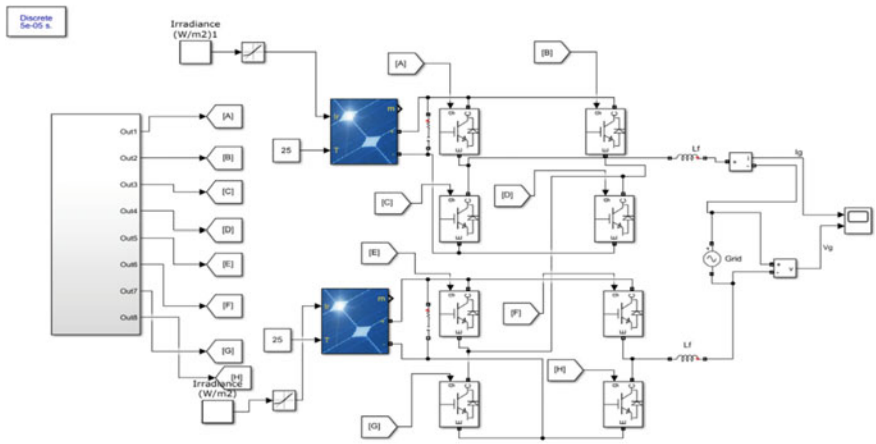


Fig. 1 Single-phase CHB multilevel inverter connected to the grid

SPWM-Sinusoidal Pulse with Modulation

In this method, to generate the PWM pulses for IGBT switches, we have compared a reference signal with a high-frequency carrier signal.

The reference signal is considered as sinusoidal where the carrier signal is considered as a triangular wave (Tables 1 and 2).

Here, I am going to discuss five-level cascade H-Bridge SPWM Inverter by using in-phase disposition pulse width modulation method the number of carrier signals required = p − 1,

where “p” is the number of levels

For five-level inverter

The number of carrier signals = 4, and reference signal = 1

Table 1 Component required for designing IPDPWM five-level cascaded H-Bridge inverter

1	Insulated gate bipolar transistor switches
2	MUX block for combining input signals into a single vector output
3	Grid and RL branch load
4	Voltage measurement blocks for measuring voltages
5	Logical operator blocks for inverting the pulses
6	AC voltage source for sinewave
7	Scope blocks for observing waveforms
8	Repeating sequence blocks for triangular waves
9	Power GUI
10	Relational operator blocks for comparing signal
11	Solar cell for DC voltage

Table 2 Switching table of five-level cascaded H-Bridge IPDPWM inverter

Pulse no	VDC	2VDC	2VDC	2VDC	VDC	−VDC	−2VDC	−2VDC	−2VDC	−VDC
Angle	30	60	120	150	180	210	240	300	330	360
Time (s)	0.00166	0.0033	0.0066	0.0083	0.01	0.01166	0.0133	0.0166	0.0183	0.02
S1, S2, S6	ON	ON	ON	ON	ON	OFF	OFF	OFF	OFF	OFF
S5	OFF	ON	ON	ON	OFF	OFF	OFF	OFF	OFF	OFF
S7, S8, S4	OFF	OFF	OFF	OFF	OFF	ON	ON	ON	ON	ON
S3	OFF	OFF	OFF	OFF	OFF	OFF	ON	ON	ON	OFF

We have used (IPDPWM) method, here above and below zero reference line, all the career signals are in the same phase.

The frequency modulation index $M1 = \frac{fc}{fp}$. The amplitude modulation index $M2 = \frac{Bm}{(p-1)Bc}$.

Here, f_c—carrier signal frequency, f_p—reference signal frequency, $\boldsymbol{B}_m$—reference signal amplitude.

B_c—carrier signal amplitude, p—Number of levels.

3 Topology Description and Analysis

A cascaded H-Bridge solar cell grid inverter diagram is exhibited in Fig. 1; the grid is connected with filter inductors. The inverter is consisting of cascaded connected two HB cells which input DC is obtaining from a solar module. Each H-Bridge is consisting of four IGBT switches and a DC-link capacitor. V_{dcj} And I_{pvj} $(j = 1, 2)$ represents the output DC voltage and current of j_{th} solar cell. v_g And i_g represent grid voltage and current, respectively. L_f Represent the grid side filter inductors.

Assuming stable states performance of the solar HB, the reference waveform of j_{th} H-Bridge can be calculated as

$$M_j = \frac{V_{Hj}}{V_{dcj}} \tag{1}$$

where V_{Hj} represents the fundamental component of jth cell output voltage $\left(V_{Hj}\right)$. The magnitude of the reference waveform is known as modulation index "MI."

From Fig. 2, P_j represents the output power of the solar panel in j_{th} H-Bridge cell and P_T denotes the total active power transferred from DC solar cell output to AC grid. Due to the small amount of total power, the power losses that occur are

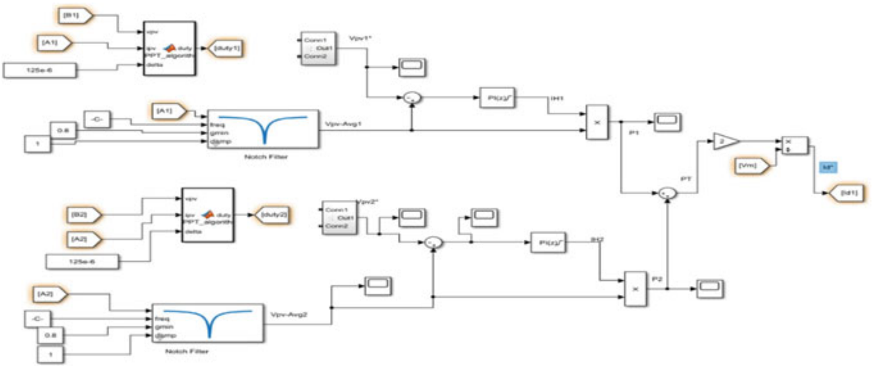


Fig. 2 H-Bridge controller or slave controller

neglected in switches and resistances of inductors. For the operation of CHB solar inverter at unity power factor, the below-given equation are getting from the power balance:

$$\sum_{j=0}^{2} P_j = P_T = V_g I_g \tag{2}$$

where V_g *and* I_g denote the grid voltage and the grid current, respectively.

3.1 Control System Design

The control block of CHB solar multilevel inverter is shown in the above Figs. 2 and 3, which is composed of two HB controllers (slave controller) and one master controller. The data transfer between controllers through communication.

Here, HB controller is liable for extract the maximum power point by using MPPT algorithm of corresponding solar panel and the switching devices of H-Bridge cell perform the operation on the basis of the reference waveform from the master controller. The master controller is totally liable for control the grid current, finding wherever to inject harmonic, and calculate the modulation or reference waveform of each HB panel based on harmonic compensation strategy. Here, the reference DC voltage $\left(V^*_{pvj}\right)$ is getting by the perturb and observe MPPT. The average voltage $\left(V_{dcj}\right)$ is obtained with the help of a 100-Hz notch filter. We have used the proportional integral controller to control the average DC voltage of $\left(V_{dcj}\right)$ to the reference DC voltage $\left(V^*_{pvj}\right)$ and obtained the average reference current $\left(I_j\right)$ as an output of the PI controller. Lastly, the power $\left(P_j\right)$ can be calculated by the multiplication of I_{Hj} and V_{dcj}. The in-phase disposition PWM method [14, 16] is used to create a switching

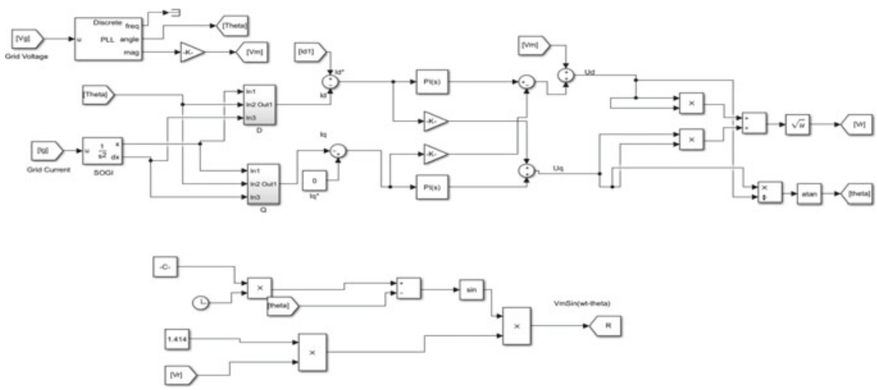


Fig. 3 Master controller

signal. In the master (main) controller, a digital PLL [17] has been taken to calculate the magnitude (V_M) and phase (θ) of v_g. The addition of power $\left(P_j\right)$ obtain from HB controller is the total power (P_T) transferred to AC-side. P_T Is divided by V_M and then multiplied by 2 to get the active power reference current $\left(I_d^*\right)$ according to Fig. 2. Generally, the reactive power reference current $\left(I_q^*\right)$ is set as zero to keep the inverter in unity power factor. Then, we have used SOGI to divided the grid current into orthogonal components i_D *and* i_Q. The active power feedback current (I_d) and reactive power feedback current $\left(I_q\right)$ can be calculated by $\alpha\beta$/dq (Park transformation). Finally, the amplitude of active power modulation voltage (U_d) and reactive power modulation voltage (U_q) can be achieved by PI controllers.

Here, in the equation obtained below, V_r represents the modulation voltage magnitude and θ_r represents the angle between the modulation voltage and the grid voltage.

$$V_r = \sqrt{{U_d}^2 + {U_q}^2} \tag{3}$$

$$\theta_r = \mathrm{arc}\, tan\left(U_d / U_q\right) \tag{4}$$

Since the current flowing through all HBs are the same, P_j of the j_{th} cell is proportional to V_j

$$\frac{V_1}{P_1} = \frac{V_2}{P_2} = \frac{\sum_{j=1}^{2} V_j}{\sum_{j=1}^{2} P_j} = \frac{V_r}{P_T} \tag{5}$$

According to (1) and (5), MI of the j_{th} cell is obtained as follows:

$$S_j = \frac{V_j}{V_{dcj}} = \frac{V_j}{V_r}\,\frac{V_r}{V_{dci}} = \frac{P_j}{P_T}\,\frac{V_r}{V_{dci}} \tag{6}$$

If the modulation index of all H-Bridge cell is not more than one, the modulation waveform of j_{th} H-Bridge is obtained by (8). Other than, the harmonic compensation strategy wasapplied to find modulation waveform.

$$m_j = S_j \sin(\theta + \theta_r), \quad j = 1, 2 \tag{7}$$

3.2 Harmonic Compensation Principle

The principle of harmonic compensation strategy is to inject some harmonic into the higher power cell and to make the grid current not contain the harmonics, the normal cell needs to compensate harmonic whose phase is reverse to injected harmonic in higher power cell or overmodulation cell. The harmonic components get canceled and only the fundamental in-phase component of sinusoidal remains in the grid current.

For the single-phase CHB solar grid-tied multilevel inverter with n number of HBS, suppose that the number of higher power cell or overmodulation cells is y, and the remaining are $\left(S_1 \sim S_y \text{ and } S_{y+1} \sim S_n \leq 1\right)$, that is normal cells, the principle of HCS can be described as follows:

To make the amplitude of modulation waveforms of higher power cells (OMC) are not a greater one, we inject some harmonic h_{Aj} into the higher power cell (OMC). HCS was used to calculate the modulation waveforms of j_{th} OMC.

$$m_j = S_j \sin(\theta + \theta_r) + h_{Aj}, \quad j = 1, 2, \ldots y \tag{8}$$

The total harmonic voltage injected into the modulation or reference waveforms of OMC is obtained as

$$v_{AT} = \sum_{j=1}^{y} h_{Aj} V_{dcj} \tag{9}$$

To make the CHB solar inverter grid voltage free from the injected harmonics, we need to compensate harmonic h_{Bj} into the reference waveform of normal cell (NC) whose phase is opposite to h_{Aj}.

$$m_j = S_j \sin(\theta + \theta_r) + h_{Bj}, \quad j = y + 1, \ldots n \tag{10}$$

So, the total harmonic voltage of CHB solar grid inverter compensated to the NC is determined as

$$v_{BT} = \sum_{j=y+1}^{n} h_{Bj} V_{dcj} \tag{11}$$

Hence, injected and compensated harmonic need to be appropriately chosen to make the work [18].

$$v_{AT} + v_{BT} = 0 \tag{12}$$

Here, the h_{Aj} injected harmonic into the overmodulation cell that reduces the modulation waveforms peak values but h_{Bj} compensated the harmonic by phase opposition to the normal cell will extend the modulation waveforms peak values correspondingly. So, the key of HCS method is to choose the suitable h_{Aj} and h_{Bj}.

4 Simulation Results

For the uncontrolled system with grid connected, the simulation is done and the waveforms analysis of the grid current and grid voltage are scoped.

For the controlled system with grid connected, the simulation is done and waveforms analysis of the grid current and Grid voltage are scoped. It is extending the range of grid current from 5 to 6.5 A (Figs. 4, 5, 6, 7, 8 and 9).

For the uncontrolled system with R-L load connected, the simulation is done and the analysis of the waveform of the load current and load voltage are scoped.

For the controlled system with R-L load connected, simulation is done and waveforms analysis of the load current and load voltage are scoped. Here, also increase the range of load current from 5.2 to 6.9 amperes.

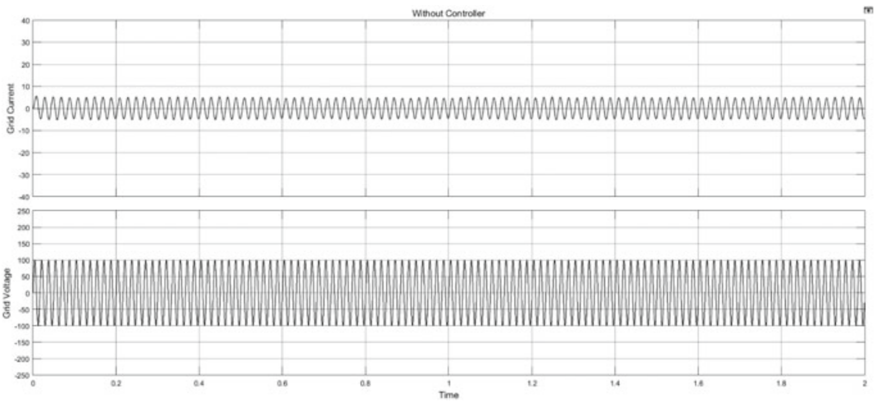


Fig. 4 Simulation results of uncontrolled system with grid connected

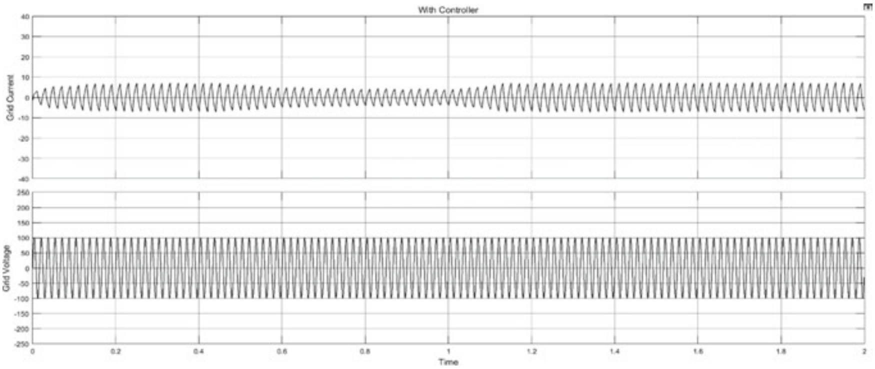


Fig. 5 Simulation results of controlled system with grid connected

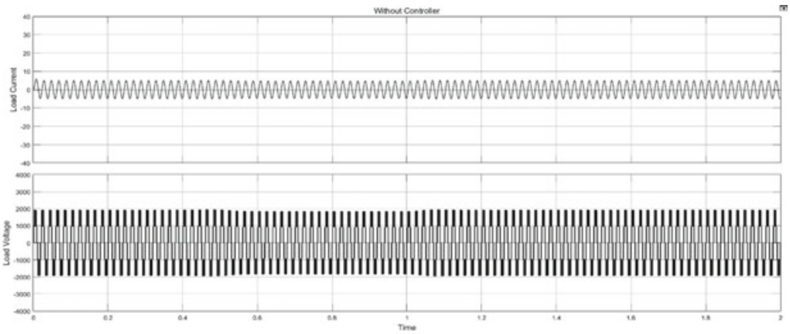


Fig. 6 Simulation results of uncontrolled system with load connected

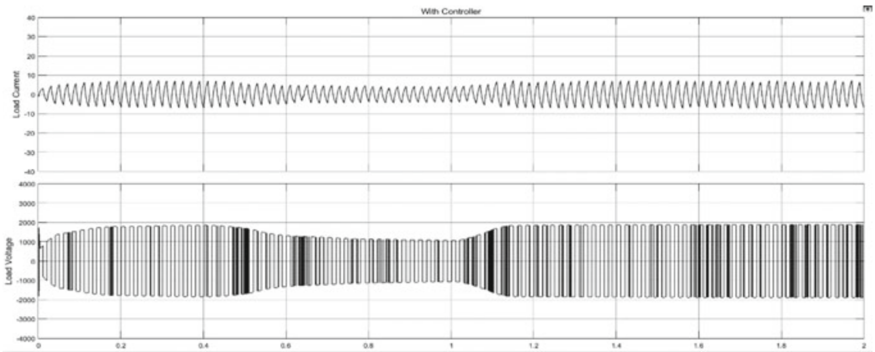


Fig. 7 Simulation results of controlled system with load connected

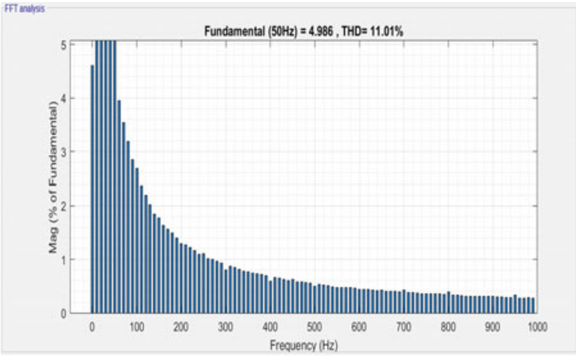


Fig. 8 FFT analysis of grid current without control system

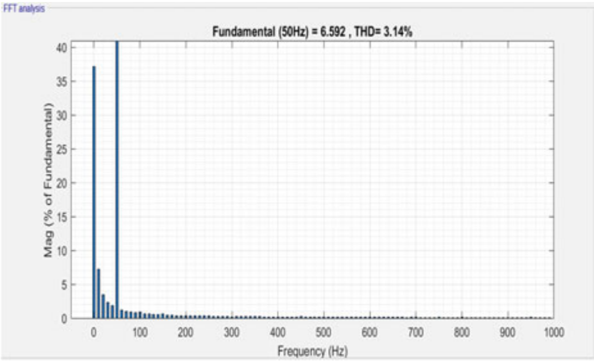


Fig. 9 FFT analysis of grid current with control system

5 Conclusions

Solar-based applications are tremendously growing day by day. Because most of the electrical device applications are in AC, we require an efficient power electronics converter for the conversion of solar cell output DC into AC. In this paper, we presented an analysis and simulation of a single-phase cascaded H-Bridge Solar cell multilevel inverter. This system is used to eliminate harmonics present in the grid current and extend the operating range with improve the modulation range of solar-fed multilevel grid inverter. Simulation results show that the range of grid current is increasing from 5 to 6.5 amperes after harmonic compensation strategy-based control systems. The method is also reducing that The THD of grid current from 11.01 to 3.14% under power unbalance cases, so it has good overall performance.

References

1. Bo Y, Wuhu L, Yi Z, Xiangning H (2010) Design and analysis of a grid connected photovoltaic power system. IEEE Trans Power Electron 25(4):992–1000
2. Ebrahimi J, Babaei E, Gharehpetian GB (2011) A new topology of cascaded multilevel converters with reduced number of components for high-voltage applications. IEEE Trans Power Electron 26(11):3109–3118
3. Nousiainen L, Puukko J (2013) Photovoltaic generator as an input source for power electronic converters. IEEE Trans Power Electron 28(6):3028–3037
4. Rodriguez J, Lai JS, Peng FZ (2002) Multilevel inverters: a survey of topologies, controls and applications. IEEE Trans Ind Electron 49(4):724–738
5. Daher S, Schmid J, Antunes FLM (2008) Multilevel inverter topologies for stand-alone PV systems. IEEE Trans Ind Electron 55(7):2703–2712
6. Rodriguez J, Bernet S, Bin W, Pontt JO, Kouro S (2007) Multilevel voltage-source-converter topologies for industrial medium-voltage drives. IEEE Trans Ind Electron 54(6):2930–2945
7. Yifan Y, Konstantinou G, Hredzak B, Agelidis VG (2015) Operation of cascaded H-bridge multilevel converters for large-scale photovoltaic power plants under bridge failure. IEEE Trans Ind Electron 62(11):7228–7236
8. Rezaei MA, Farhangi S, Iman-Eini H (2010) Extending the operating range of cascaded H-bridge based multilevel rectifier under unbalanced load conditions. In: Proceeding of PECON2010, pp 780–785
9. Eskandari VJ, Iman-Eini H, Yadollahi M (2013) Stable operation of grid connected cascaded H-bridge inverter under unbalanced insolation conditions. In: Electric power and energy conversion systems, Istanbul, Turkey, Oct 2–4 2013
10. Liu L, Li H, Xue Y, Liu W (2015) Reactive power compensation and optimization strategy for grid-interactive cascaded photovoltaic systems. IEEE Trans Power Electron 30(1):188–202
11. Zhao T, Zhang X, Mao W, Xu J, Gu Y, Zhao D, Jiang C (2017) Control strategy for cascaded H-bridge photovoltaic inverter under unbalanced power conditions based on reactive compensation. Proceeding of the CSEE 37(17):5076–5085
12. Moosavi M, Farivar G, Iman-Eini H, Shekarabi SM (2014) A voltage balancing strategy with extended operating region for cascaded H-bridge converters. IEEE Trans Power Electron 29(9):5044–5053
13. Miranbeigi M, Iman-Eini H (2016) Hybrid modulation technique for grid-connected cascaded photovoltaic systems. IEEE Trans Ind Electron 63(12):7843–7853
14. Ko Y, Andresen M, Buticchi G, Liserre M (2017) Power routing for cascaded H-bridge converters. IEEE Trans Power Electron 32(12):9435–9446
15. Zhao T, Zhang X, Mao W, Wang F, Xu J, Gu Y, Wang X (2018) An optimized third harmonic compensation strategy for single-phase cascaded H-bridge photovoltaic inverter. IEEE Trans Ind Electron 65(11):8635–8645
16. Tsoumada YH, Kiozumi H (2014) Level- and phase-shifted PWM for seven-level switched-capacitor inverter using series/parallel conversion. IEEE Trans Ind Electron 61(8):4011–4021
17. Ciobotaru M, Teodorescu R, Blaabjerg F (2006) A new single-phase PLL structure based on second order generalized integrator. In: IEEE power electronics specialists conference, pp 1–7
18. Zhao T, Zhang X, Mao W, Wang M, Wang Fs, Wang X, Xu J (2019) Harmonic compensation strategy for extending the operating range of cascaded H bridge PV inverter. IEEE J Emerg Sel Top Power Electron

Presence in VR Experiences—An Empirical Cost–Benefit Analysis

René Peinl and Tobias Wirth

Abstract Virtual reality (VR) is on the edge of getting a mainstream platform for gaming, education, and product design. The feeling of being present in the virtual world is influenced by many factors and even more intriguing a single negative influence can destroy the illusion that was created with a lot of effort by other measures. Therefore, it is crucial to have a balance between the influencing factors, know the importance of the factors and have a good estimation of how much effort it takes to bring each factor to a certain level of fidelity. This paper collects influencing factors discussed in literature, analyses the immersion of current off-the-shelf VR-solutions, and presents results from an empirical study on efforts and benefits from certain aspects influencing presence in VR experiences. It turns out, that sometimes delivering high fidelity is easier to achieve than medium fidelity and for other aspects it is worthwhile investing more effort to achieve higher fidelity to improve presence a lot.

Keywords Virtual reality · Presence · Cost–benefit ratio · Off-the-shelf headset · Interactivity

1 Introduction

The hype around virtual reality (VR), that already arose in the 1990s, was revived with the availability of Oculus Rift, and now seems to be here to stay. The quality of head-mounted displays (HMDs) together with the low latency of sensors and enough processing power in CPUs and GPUs allow for immersive experiences at comparably low cost that were not possible 25 years ago. Since then, an abundance of VR companies has emerged and a number of HMDs are available for professionals and

R. Peinl · T. Wirth (✉)
Institute of Information Systems, Hof University of Applied Sciences, Alfons-Goppel-Platz 1, 95028 Hof, Germany
e-mail: tobias.wirth@iisys.de

R. Peinl
e-mail: rene.peinl@iisys.de

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_43

consumers alike. However, if we look at scientific literature, studies about presence and immersion in VR are still often conducted with specialized hardware instead of off-the-shelf devices. Additionally, although many factors influencing presence were studied, there is no analysis on the effort needed to incorporate these factors into a VR experience at a given level of fidelity. It is, e.g., clear, that visual quality positively affects presence both regarding the display of the HMD and the rendered contents, but how does effort put into render quality relate to perceived presence stemming from visual quality? Is it a linear relation, are there plateaus or does the cost–benefit function converge against an upper limit that is far below being "indistinguishable from reality"? In this study, the effects of multiple factors are measured empirically and compared to the effort needed to implement them. To do that, we carefully chose a hardware setup that has a good cost–immersion-ratio and conducted several experiments to measure the factors with the highest impact and find a good compromise between effort and perceived presence.

2 Related Work

Presence can be seen as the sensation of "being there" in the virtual environment [1]. It is defined in literature as the psychological state where a user is feeling lost or immersed in the mediated environment, the degree to which he or she feels physically "present" in a virtual environment [2]. More recent interpretations make a distinction between the illusion of being in a place ("place illusion"), and experiencing events as if they were real ("plausibility") [3]. The former relies on head (and ideally eye) tracking to enable perception of the environment using the body in a natural way (bending down, looking around, looking past, listening by turning the head towards the source). The latter is the extent to which the events within the virtual environment are perceived as really happening (ibid) and is said to be influenced by three factors: "(a) the extent to which events in the environment refer specifically to the participant (b) the extent to which there are events that respond to the actions of the person (e.g., the participant smiles at a virtual human that smiles back), and (c) the overall credibility of the environment in comparison to expectations" [3]. Where presence describes the subjective feeling, immersion is the objective characteristic of the technology, referring to the extent to which the computer displays are capable of delivering an inclusive, extensive, surrounding, and vivid illusion of reality to the senses of a human participant [4].

Reference [5] studied display and interaction fidelity of a VR first-person shooter in a CAVE environment. They found that low display fidelity and low interaction fidelity together led to the quickest completion time of the task, whereas "mixed fidelity" conditions (low/high or high/low) led to the highest completion times. They concluded that both display and interaction fidelity had significant positive impact on presence, engagement, and usability. Reference [6] use an Oculus Rift DK1 to study immersion in VR environments. They use the Unity game engine to visualize engineering data. However, judging visual quality based on the photos in the paper, the visual quality was much lower than the one in our bath scenario. Their workflow from the CAD software to Unity was similar to ours. The body ownership illusion is another important factor [3]. Experiments show that different virtual bodies even influence behavior of the user in the virtual world (ibid). Reference [7] found that using body tracking with a Kinect is leading to significantly higher presence values than with a Wii controller. Reference [8] analyze the effect of getting haptic feedback from substitutional objects that are tracked in the real world to be accordingly represented at the same position in the virtual world. They found that substitutional objects do not have to be identical to virtual ones to enhance presence. It is sufficient if they share some common characteristics like size and shape. Others like material and weight can vary up to a certain degree. Reference [9] report about presence in traditional and immersive virtual environments. Although this was already two years after the availability of Oculus Rift DK2, they were using a non-standard HMD with unknown resolution for the traditional VR environment and a CAVE-like room with projections on the wall as immersive environment. They found the immersive environment to enhance presence compared to the HMD. Reference [10] found evidence that a self-avatar had a positive effect on presence and embodiment. In a different experiment with a singer inviting the participant to tap along it turned out to have a negative effect on embodiment compared to the singer not directly addressing the user. However, the results must be considered with caution, since data was collected "in the wild" and not under controlled laboratory settings. Reference [11] report about an experiment with intentionally coarse graphics in order to decrease hardware requirements for presence research. They use a VR HMD, but use it for mixed reality, where real objects are filmed with a camera and are incorporated into the virtual world in real time. They found that intentionally low visual quality but providing a coherent experience across the whole scene with full body tracking allowed for high presence values. Reference [12] analyze the impact of a HMD on spatial presence compared to a regular computer monitor. They find that a HMD increases perceived presence. However, higher presence is not a value in itself, but leads to higher liking of the virtual place and therefore also a stronger intention to visit the place physically [13].

3 Choice of Hard and Software

Regarding hardware, we know from literature that factors affecting immersion and therefore presence in VR are visual quality, latency of reaction on user activities, and the degree of freedom. Looking at off-the-shelf VR headsets we find three distinct categories. Smartphone-powered headsets are very low cost (~100€) if we already own a compatible smartphone, but provide only three degrees of freedom (3DoF). Samsung GearVR and Google Daydream are examples of this category. PC-based headsets do have 6DoF, but are mostly comparatively expensive (500–1500€) and are usually restricted by cables connecting them to the PC, which affects mobility and perceived freedom. However, they are able to provide enough processing power for high fidelity rendering of contents, which is limited for Smartphone-powered headsets. Oculus Rift and HTC Vive are the most used devices in this category. The last category, which is currently emerging is standalone devices. That means that they do not need external trackers like the HTC lighthouse system and also do not need a PC for rendering contents, but still provide 6DoF. Oculus Quest is an example of this new device category. It is the only one that has both a 6DoF HMD and 6DoF hand controllers, which are crucial for certain types of VR experiences. Regarding display quality, there is no principle distinction between the categories of devices. We have lower quality on PC-based HMDs like Oculus Rift than on GearVR with a Galaxy S9 smartphone, but higher quality on newly available PC-based HMDs like Pimax 8k or Varjo VR1. It has to be noted that visual quality of HMDs is not only affected by the resolution, but also by display technology since the most important aspect is the absence of the screen door effect (SDE). Samsung Odyssey+, for example, achieves a lower SDE than HTC Vive due to a special light distribution foil despite having a similar resolution. SDE stems from the black areas between pixels that get visible with the large magnification of the HMD's lenses. Finally, the field of view (FoV) is important for immersion. It is similar for most HMDs on the market ranging between 90 and 110°. However, recently the Pimax 8k became available which offers 200° FoV at the cost of some distortions at the border of the visible field. Regarding latency, there is no great difference between the headsets, so the factor is negligible for the choice of hardware.

Due to the unavailability of the Pimax 8k during the time of the study (December 2018), we chose the HTC Vive Pro with two trackers for the lighthouse system. It was powered by an AMD Ryzen 2700X CPU, 32 GB of RAM, and an Nvidia GeForce 1080 Ti GPU so that the overall hardware investment was around 3.000€. We did not use the wireless adapter which could enhance the perceived freedom of moving.

Regarding the software needed for creating the VR experience, there are two main tools needed. The 3D modeling software and the VR engine. For modeling, we preferred the open-source software Blender due to its free availability over the more capable professional tools like Maya or 3D Studio max. For the VR engine, the choice is mainly between Unity and the Unreal engine, which both combine a powerful game engine for 3D rendering with a development environment with a good graphical editor. A detailed comparison of those is beyond the scope of this paper and since features are very similar, it might be at the end a subjective personal choice. We chose Unity 2018.3 together with the SteamVR plugin due to existing knowledge about it and many resources available for training. Reference [14] showed that the Unreal Engine is working well, too.

4 Study Setup

Two different VR scenarios were created, both being in the context of interior design or architecture. Commercial usages we consider are virtual exhibitions of baths, kitchens, and furniture in general as well as visualization of the results of bath makeovers or interior design. Therefore, the first scenario was a bath. We took 360° photos of a bath exhibition in Hof, Germany and recreated one of the rooms as a 3D model. The photos were taken with an Insta360 Pro camera in 8k 3D and then stitched with Insta Stitcher and exported in RAW format. Exposure was enhanced using Adobe Camera Raw.

In the first experiment, we compared presence with photos in full resolution (8k 3D = 7680 × 8640), half resolution (4k 3D = 3840 × 4320) as well as both resolutions in 2D. In the second experiment, we compared the photos with the 3D model. We did not put additional effort into making the 3D model look identical to the photo, but did aim for a high similarity at the least possible effort. We chose freely available 3D models[1] as a basis and assembled them into a room created by ourselves. The goal was to find out whether perceived visual fidelity of the contents were higher in 360° photos than in our 3D model and how that affects presence.

The second scenario was a living room, which served as a basis to study the effects of different levels of interaction on presence. The 3D models used there were much simpler and not photo-realistic compared to the bath scenario. The task for participants was to look for books that were distributed in the room and count those with the title "Virtual Reality".

In the first experiment, the participants could only move through the room on foot or by using the hand controllers to teleport. The second experiment allowed participants to open drawers and doors of cabinets, as well as pick-up books to better look at them. However, the interaction by intention was made very simplistic like in most computer games. Once the controller was near an object, which was indicated by putting a yellow frame around the object of interaction, participants only had to press a button on the controller to grab it. A prerecorded animation was played then that showed how the door or drawer opened automatically and completely. In contrast to that, the third experiment allows for more natural interaction. The button on the controller was used for grabbing, but the movement needed to be induced by moving the arm and pulling the object. Once released the object was behaving physically correct maintaining its impulse for some time until being stopped by gravity and air resistance or friction. Additionally, books could be put everywhere around the room or thrown and also showed effects from gravity, whereas in experiment two they "magically" appeared back in their original location, once the controller button was released.

There were 28 participants between 18 and 52 years old (mean 30 years, sd 9.42) which were all acquired at the university. They did not receive any payment or other benefits. There were 20 male and 8 female participants. Eight participants had at least intermediate VR experiences. Four of those owned a VR headset. The study lasted for about 30 min. After each experiment, participants were asked to fill out a questionnaire to state their personal impressions on presence in various aspects. The questionnaire was mainly built with questions stemming from the presence questionnaire by Witmer and Singer with a few own additions. We did a pretest with three people to eliminate mistakes. The questionnaire consisted of 13 questions for

[1] free3dbase.com, cgtrader.com, turbosquid.com, blendswap.com.

Table 1 Aggregation of questions to presence aspects

	Questions bath	Questions living room
Involvement	6, 7, 8, 12	3, 4, 5, 6, 7, 8, 10, 14
Visual fidelity	10, 11	12, 13
Adaption/immersion	9, 13	9, 11, 15, 16, 17, 19
Interface quality		18

the bath and 19 questions for the living room scenario. They were summed up into the categories involvement, visual fidelity, adaption/immersion, and interface quality (see Table 1). The scores per category were calculated as a sum of single question scores, divided by the number of questions. The single questions were judged on a 7-point Likert scale from 1 being the lowest presence to 7 representing the highest presence score.

5 Results

A reliability analysis of the different questionnaires resulted in a Cronbach α of 0.83 for I_1, 0.88 for I_2 and 0.84 for I_3. Since all of these values are above 0.7, we can assume consistency of the answers. The same applies to the other scenario with Cronbachs α of 0.91 for B_1, 0.84 for B_2, and 0.72 for B_3. The data did not follow a normal distribution and therefore we used the Wilcoxon–Mann–Whitney test (U test) to test the significance of hypothesis. The resulting p-values from the U test are considered significant if they are below 5% and considered highly significant if they are below 1%.

The first hypothesis tested was

H_1: *presence values of 360° photos (B_1) are higher than those of the respective 3D model in the bath scenario, if we artificially limited the degrees of freedom of the 3D model to three (B_2).*

The assumption was that the visual quality of the photo is higher than that of our textured 3D model, although we put some effort in enhancing the quality of it to get near photorealism. Leveling the freedom to 3DoF should make the photos more compelling than the 3D model. However, it turned out that H_1 has to be rejected. There were slight differences in the presence scores, but they are not significant and they are even slightly higher for the 3D model.

H_2: *presence of the 3D model limited to 3DoF (B_2) will be lower than that of the same model with 6DoF (B_3).*

In accordance with literature, 6DoF makes a big difference and therefore presence values increased by nearly two, leading to a highly significant deviation ($p = 0.5\%$). Therefore, H_2 was accepted (Tables 2, 3, 4 and 5).

Table 2 Results from bath scenario (SD = standard deviation)

	B_1 mean	B_1 SD	B_2 mean	B_2 SD	B_3 mean	B_3 SD
Involvement	4.259	0.286	4.490	0.534	6.241	0.103
Visual fidelity	2.910	0.227	2.942	0.299	5.893	0.657
Adaption/Imm.	5.179	1.566	5.135	1.713	6.536	0.354
Presence-score	4.152	1.068	4.264	1.140	6.228	0.378

Table 3 Visual fidelity of 360° photos in different resolutions

	4k 2D	4k 3D	8k 2D	8k 3D
Mean	3.000	4.679	4.000	5.429
SD	1.440	1.249	1.440	1.136

Table 4 Results of the living room scenario (SD = standard deviation)

	I_1 mean	I_1 SD	I_2 mean	I_2 SD	I_3 mean	I_3 SD
Involvement	4.714	1.118	5.246	0.585	5.768	0.449
Visual fidelity	4.018	0.278	5.411	0.177	6.143	0.050
Adaption/Immer	4.887	1.784	4.994	1.638	5.435	1.426
Interface quality	2.071	NA	2.214	NA	2.321	NA
Presence-score	4.538	1.421	4.998	1.235	5. 492	1.202

Table 5 Overview of tasks for creating VR experiences

Task	Learning effort	Implementation effort
(1) Create 360° photo	Very low (minutes)	Short setup per location plus ~15 min per photo
(2) Create 3D model		
a. Acquire	Finding the right platforms	~15 min per object
b. Create on your own	~1 day of learning blender basics	1 h up to days depending on complexity
(3) Create materials and textures		
a. Acquire	Finding the right platforms	15–30 in per material
b. Create on your own	~1 day learning 3D photo scanning basics	1–4 h per material
(4) Fix 3D models or materials	~4 h of learning UV mapping and normals	10 min–4 h per object or material
(5) Assemble the 3D scene	~1 day learning unity basics	2–8 h depending on the complexity of the scene
(6) Set up lighting and reflections	~4 h of learning basics	1–4 h depending on the complexity of the scene
(7) Apply the post-processing stack	~4 h of learning basics	~1 h experimenting with the settings to get a good result

The next hypothesis dealt with visual quality regarding resolution of the contents, while keeping resolution of the display constant.

H_3: *a higher image resolution will increase presence for 360° images.*

With p-values of 0.63 and 1.00% for 2D and 3D images respectively, H_3 can be accepted. An image resolution of 8k leads to significantly higher presence scores, compared to 4k.

H_4: *stereoscopic images (3D) will result in higher presence values for 360° photos than monoscopic.*

Again, the hypothesis can be accepted. Results were significant for both 8 images ($p = 0.033\%$) as well as 4 images ($p = 0.003\%$). However, some participants also reported problems with nausea that increased for 3D contents.

For the living room scenario, the assumption was that presence would increase with realism of interaction.

H_5: *presence values in the experiment with simplistic interaction (I_2) will be higher than in the experiment with no interaction (I_1).*

Although presence values did indeed increase for I_2, the difference was not significant ($p = 37.02\%$). Therefore, H_5 has to be rejected.

H_6: *presence values in the experiment with realistic interaction (I_3) will be higher than in the experiment with simplistic interaction (I_2).*

Again, presence values for I_3 did increase compared to I_2. With a p-value of 2.6% the difference is significant, but not highly significant. H_6 was therefore accepted.

6 Discussion of Results in Relation to Effort

Some of the findings were as expected, whereas others were contrary to intuition. While increasing the resolution of images as well as adding a stereoscopic effect increases presence for 360° images as expected, it turned out that despite a perceived higher photorealism of 360° images the 3D model with high visual fidelity was preferred by participants, even if we limited it to 3DoF. One explanation for this is that some participants felt that proportions in the 360° photo were not right, so it felt like the floor or ceiling were too far away. Another part of the explanation is, that current computing power and game engines are able to produce a visual quality that is getting closer and closer to photorealism. This is especially true for static scenes where lights, reflections, and shadows can be precomputed. Unity makes it possible to achieve good results even for rather unexperienced 3D designers like us. We did not even use the post-processing stack, which can further enhance visual quality. The effort is also not too high overall (see **Fehler! Verweisquelle konnte nicht gefunden werden.**), although it is significantly higher than producing a 360° photo. The effort pays off, however, if we look at the increase in presence of 2.

Although the scale is ordinal and not metric, this is significant. The importance of visual quality can also be seen when comparing B_3 with I_1. Although the comparison is not perfectly adequate since the environment differs not only in visual fidelity, it shows a clear trend. Interaction, however, is also an important aspect. Despite a lower rise in presence scores of "only" ~1.5 between I_1 and I_3 in relation to ~2.0 from B_1 to B_3, it can be stated that interaction should not be neglected. This is especially true in relation to efforts, since using the SteamVR plugin together with its PreFabs makes implementing realistic interaction including physics easy. The feature "throwable" makes any 3D object immediately pickable and behaving according to physic laws (esp. gravity). Also friction can be simulated very easy. Both only need additional effort of a few minutes.

For drawers and cabinet doors, we were using the prefabs "linear drive" and "circular drive" from SteamVR. Again, they only need a few minutes to be added to 3D models that are rigged[2] properly. Astonishingly, the simplified interaction tested in I_2 is not only less effective in terms of increasing presence, but also causes more effort than the realistic interaction implemented in I_3. Therefore, it is a clear advice to use the latter kind of interaction. On the other hand, it was often harder than expected to integrate existing 3D models downloaded from the internet. In contrast to the PreFabs mentioned, 3D models in obj or fbx format often needed a considerable amount of fixing before they could be used in Unity. This applied to sizes as well as normals and materials including textures. Unity uses sizes where 1 unit should relate to 1 m in the real world. Imported 3D models often were either factor 100 or even 1000 too large, or factor 10 or 100 too small. This was however easy to fix. Normals facing in the wrong direction led to invisible surfaces. This was fixed in Blender. We can either flip the normals to the other side there, or we can let Blender recalculate them, which usually leads to the desired results. Reference [6] report about similar issues when exporting their models from the CAD tool. The most effort went into fixing materials and textures. There seems to be no working standard that lets one smoothly transport those from a 3D modeling tool to Unity. It is unclear, whether it is a Unity problem or existent in Unreal Engine as well. Basically, all materials have to be recreated. The only benefit we have is that the assignment of materials to objects or faces already exists. This problem even exists for Blender files, which are explicitly supported in Unity, so there is one less conversion. The problem is amplified by the fact that both in Blender and in Unity, and there are two different render engines for normal and high quality that are not directly compatible. In Blender, we have the normal renderer and Cycles, the physically based renderer. Similarly, we can create either a normal 3D project in Unity or one with the HD render pipeline. In any case, we cannot directly use materials from the other version.

[2] Prepared to be movable in the game.

7 Conclusion and Outlook

Our empirical study has shown that it is possible to create VR experiences with a high level of presence with comparatively low effort. 6DoF, visual quality of the content and realistic interaction with virtual objects are important factors affecting presence. Current off-the-shelf VR headsets are a good basis for VR experiences, since they provide a high level of immersion. The screen door effect and the low field of view were the main concerns people had regarding the hardware. Interesting side aspects discovered via free text comments of participants or oral utterances were that the cable of the Vive Pro did bother multiple people. We also found that 6DoF can reduce nausea compared to 3DoF, whereas 3D photos tend to increase it. This is however only anecdotal evidence and needs to be confirmed in further studies. Another aspect not studied enough in literature is breaking the feeling of presence by having virtual objects obstruct the real body of the user. Multiple people reported that they were pushed out of the presence illusion when they opened a cabinet door and it easily moved through their knee, as they were in a kneeling position. This happened although we did not use a virtual avatar, so it could not be seen directly, that the door was moving through the virtual knee. For a follow-up study, we envision to use an avatar animated on the basis of human pose detection and implement collision detection between virtual objects and this avatar. This should help avoiding these kinds of irritations. However, it cannot be avoided completely since no physical resistance will hinder users to still walk through virtual objects. Maybe the avatar could be kept in front of the virtual obstacle, but it would still be hard to show the real movement of feet and legs, while doing so. Another interesting aspect we completely neglected for this study is audio and especially positional audio. For the next study, we plan on analyzing the effects of simple audio and spatial audio effects on presence and relate it to the respective efforts. In order to make results from single experiments more comparable, we envision a single scenario for all different experiments, no matter whether it's audio, visual, or interaction experiments. Related work indicates that VR experiences benefit from spatial sound [15]. With the availability of Pimax 8k X and HP Reverb G2 after completion of the study, it would also be interesting to investigate the effects of different hardware with differences in screen door effect, field of view, and controller quality on presence and interaction.

References

1. Slater M (1999) Measuring presence: a response to the Witmer and Singer presence questionnaire. Presence Teleoperators Virtual Environ 8:560–565
2. Schubert T, Friedmann F, Regenbrecht H (2001) The experience of presence: factor analytic insights. Presence 10:266–281
3. Slater M, Sanchez-Vives MV (2014) Transcending the self in immersive virtual reality. Computer 47:24–30
4. Slater M, Wilbur S (1997) A framework for immersive virtual environments (FIVE): speculations on the role of presence in virtual environments. Presence Teleoperators Virtual Environ

6:603–616
5. McMahan RP, Bowman DA, Zielinski DJ, Brady RB (2012) Evaluating display fidelity and interaction fidelity in a virtual reality game. IEEE Trans Vis Comput Graph 626–633
6. Marks S, Estevez JE, Connor AM (2014) Towards the Holodeck: fully immersive virtual reality visualisation of scientific and engineering data. In: Proceedings of the 29th international conference on image and vision computing New Zealand, ACM, pp 42–47
7. Lorenz M, Busch M, Rentzos L, Tscheligi M, Klimant P, Fröhlich P (2015) I'm There! The influence of virtual reality and mixed reality environments combined with two different navigation methods on presence. In: 2015 IEEE virtual reality (VR), IEEE, pp 223–224
8. Simeone AL, Velloso E, Gellersen H (2015) Substitutional reality: using the physical environment to design virtual reality experiences. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, ACM, pp 3307–3316
9. North MM, North SM (2016) A comparative study of sense of presence of traditional virtual reality and immersive environments. Australas J Inf Syst 20
10. Steed A, Frlston S, Lopez MM, Drummond J, Pan Y, Swapp D (2016) An 'In the Wild' experiment on presence and embodiment using consumer virtual reality equipment. IEEE Trans Vis Comput Graph 22:1406–1414
11. Regenbrecht H, Meng K, Reepen A, Beck S, Langlotz T (2017) Mixed voxel reality: presence and embodiment in low fidelity, visually coherent, mixed reality environments. In: 2017 IEEE international symposium on mixed and augmented reality (ISMAR), IEEE, pp 90–99
12. Seibert J, Shafer DM (2018) Control mapping in virtual reality: effects on spatial presence and controller naturalness. Virtual Real 22:79–88
13. Tussyadiah IP, Wang D, Jung TH, tom Dieck MC (2018) Virtual reality, presence, and attitude change: empirical evidence from tourism. Tour Manag 66:140–154
14. Hilfert T, König M (2016) Low-cost virtual reality environment for engineering and construction. Vis Eng 4
15. Poeschl S, Wall K, Doering N (2016) Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. In: 2013 IEEE virtual reality (VR), IEEE, pp 129–130

A New Simple Computational Method of Simultaneous Constructing and Comparing Confidence Intervals of Shortest Length and Equal Tails for Making Efficient Decisions Under Parametric Uncertainty

Nicholas Nechval, Gundars Berzins, and Konstantin Nechval

Abstract A confidence interval is a range of values that provides the user with useful information about how accurately a statistic estimates a parameter. In the present paper, a new simple computational method is proposed for simultaneous constructing and comparing confidence intervals of shortest length and equal tails in order to make efficient decisions under parametric uncertainty. This unified computational method provides intervals in several situations that previously required separate analysis using more advanced methods and tables for numerical solutions. In contrast to the Bayesian approach, the proposed approach does not depend on the choice of priors and is a novelty in the theory of statistical decisions. It allows one to exclude unknown (nuisance) parameters from the problem using the technique of invariant statistical embedding and averaging in terms of pivotal quantities (ISE & APQ). It should be noted that the well-known classical approach to constructing confidence intervals of the shortest length considers at least three versions of possible solutions and is in need of information about the forms of probability distributions of pivotal quantities in order to determine an adequate version of the correct solution. The proposed method does not need such information. It receives this information through the quantiles of the probability distribution of the pivotal quantity. Therefore, the proposed method automatically recognizes an adequate version of the correct solution. To illustrate this method, numerical example is given.

Keywords Pivotal quantity · Isolating unknown (nuisance) parameters · Quantile function · Shortest length or equal tails confidence intervals

N. Nechval (✉) · G. Berzins
University of Latvia, Riga 1586, LV, Latvia
e-mail: nechval@junik.lv

K. Nechval
Transport and Telecommunication Institute, Riga 1019, Latvia

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_44

1 Introduction

To make a statistical inference in many problems under parametric uncertainty, the experimenter is interested in constructing a confidence interval that contains the true (unknown) value of the parameter (say, λ) with a given probability [1–3]. If we are given a random sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ from a density $f_\lambda(y)$ and a pivotal quantity $V(S(\mathbf{Y}), \lambda)$ (which is developed from either maximum likelihood estimate or sufficient statistic $S(\mathbf{Y})$), whose distribution does not depend on λ, then confidence interval for a single unknown parameter λ is often derived by using a pivotal quantity $V(S(\mathbf{Y}), \lambda)$. To clarify the problem, let $V = V(S(\mathbf{Y}), \lambda)$ be a pivotal quantity with the probability distribution function (PDF) $q(v)$ and cumulative distribution function (CDF) $Q(v)$. Then, the problem of constructing shortest length or equal tails confidence intervals in terms of pivotal quantities can be formulated more generally as follows.

1.1 Classical Approach to Constructing Shortest-Length Confidence Intervals

It is assumed that the length of the statistical confidence interval is given by

$$L(q_1, q_2|S(\mathbf{Y})) \propto L(q_1, q_2) = \int\limits_{q_1}^{q_2} \zeta(\tau)d\tau, \tag{1}$$

The length of the expected confidence interval is given by

$$E_\lambda\{L(q_1, q_2|S(\mathbf{Y}))\} \propto L(q_1, q_2) = \int\limits_{q_2}^{q_1} \zeta(\tau)d\tau, \tag{2}$$

In order to find the $100(1-\alpha)\%$ statistical (or expected) shortest-length confidence interval for λ, we should find a pair of decision variables q_1 and q_2 such that $L(q_1, q_2)$ is minimum.

Problem Statement in Terms of the Pivot V and Decision Variables $\mathbf{q_1}$and $\mathbf{q_2}$: Minimize

$$L(q_1, q_2) = \int\limits_{q_1}^{q_2} \zeta(\tau)d\tau \tag{3}$$

subject to

$$\int_{q_1}^{q_2} q(v)dv = Q(q_2) - Q(q_1) = 1 - \alpha. \tag{4}$$

Classical Analytical Approach to Solution of the Problem. Differentiating $L(q_1, q_2)$ with respect to q_1, we get

$$\frac{dL(q_1, q_2)}{dq_1} = \frac{d}{dq_1}\int_{q_1}^{q_2} \zeta(\tau)d\tau = \zeta(q_2)\frac{dq_2}{dq_1} - \zeta(q_1) \tag{5}$$

From (4) we find the derivative of q_2 with respect to q_1 as follows:

$$\frac{d}{dq_1}\int_{q_1}^{q_2} q(v)dv = \frac{d}{dq_1}(1-\alpha) \tag{6}$$

that is

$$q(q_2)\frac{dq_2}{dq_1} - q(q_1) = 0. \tag{7}$$

Thus, we have

$$\frac{dq_2}{dq_1} = \frac{q(q_1)}{q(q_2)}. \tag{8}$$

Substituting this into (5), we obtain

$$\frac{dL(q_1, q_2)}{dq_1} = \zeta(q_2)\frac{q(q_1)}{q(q_2)} - \zeta(q_1). \tag{9}$$

Now consider the following three versions of possible decision-making.
Version 1 of Possible Decision-Making:

$$\frac{dL(q_1, q_2)}{dq_1} = \zeta(q_2)\frac{q(q_1)}{q(q_2)} - \zeta(q_1) = 0. \tag{10}$$

Then, the optimal analytical solution of the problem is given by

$$\frac{q(q_1)}{\zeta(q_1)} = \frac{q(q_2)}{\zeta(q_2)}. \tag{11}$$

Version 2 of Possible Decision-Making:

$$\frac{dL(q_1, q_2)}{dq_1} = \zeta(q_2)\frac{q(q_1)}{q(q_2)} - \zeta(q_1) < 0. \tag{12}$$

It follows from (4) that

$$1 - \alpha < Q(q_2) \leq 1. \tag{13}$$

Then, the optimal analytical solution of the problem is given by

$$q_2 = \arg[Q(q_2) = 1], \; q_1 = \arg[Q(q_1) = \alpha]. \tag{14}$$

Version 3 of Possible Decision-Making. Let us assume that

$$Q(q_2) = 1 - \bar{Q}(q_2), \; Q(q_1) = 1 - \bar{Q}(q_1). \tag{15}$$

It follows from (4) that

$$\int_{q_1}^{q_2} q(v)dv = Q(q_2) - Q(q_1) = \bar{Q}(q_1) - \bar{Q}(q_2) = 1 - \alpha. \tag{16}$$

It follows from (16) that

$$\frac{d}{dq_1}\int_{q_1}^{q_2} q(v)dv = \frac{d}{dq_1}\left(\bar{Q}(q_1) - \bar{Q}(q_2)\right) = \frac{d}{dq_1}(1 - \alpha) \tag{17}$$

that is

$$\bar{Q}'(q_1) - \bar{Q}'(q_2)\frac{dq_2}{dq_1} = 0. \tag{18}$$

Thus, we have

$$\frac{dq_2}{dq_1} = \frac{\bar{Q}'(q_1)}{\bar{Q}'(q_2)}. \tag{19}$$

Substituting this into (5), we obtain

$$\frac{dL(q_1, q_2)}{dq_1} = \zeta(q_2)\frac{\bar{Q}'(q_1)}{\bar{Q}'(q_2)} - \zeta(q_1) > 0. \tag{20}$$

It follows from (16) that

$$1 - \alpha < \bar{Q}(q_1) \leq 1. \tag{21}$$

Then, the optimal analytical solution of the problem is given by

$$q_1 = \arg[\bar{Q}(q_1) = 1],\; q_2 = \arg[\bar{Q}(q_2) = \alpha]. \tag{22}$$

2 New Simple Computational Method of Constructing Shortest-Length Confidence Intervals

It is assumed that the length of the confidence interval is given by (3). The proposed approach is based on the use of the numerical values of the quantile functions q_1 and q_2. In order to find the $100(1-\alpha)\%$ shortest-length confidence interval for λ, we should find a pair of numerical values of the quantile functions q_1 and q_2 such that $L(q_1, q_2)$ is minimum.

Problem Statement in Terms of the Decision Variable p (Probability) and Quantile Functions $\mathbf{q_1}$and $\mathbf{q_2}$. Minimize

$$L^2(q_1, q_2) = \left(\int_{q_1}^{q_2} \zeta(\tau) d\tau \right)^2, \tag{23}$$

where the quantile function q_1 is given by

$$q_1 = Q^{-1}(p), \tag{24}$$

the quantile function q_2 is given by

$$q_2 = Q^{-1}(1 - \alpha + p), \tag{25}$$

subject to

$$0 \leq p \leq \alpha. \tag{26}$$

The decision variable to be determined is p (probability).

New Simple Computational Method for Numerical Solution of the Problem. The optimal numerical values of the quantile functions q_1 and q_2, which minimize $L(q_1, q_2)$, can be obtained from (23) to (26) using computer software "Solver".

3 Illustrative Numerical Example

3.1 Classical Analytical Approach

Problem Statement in Terms of the Pivot V and Decision Variables $\mathbf{q_1}$and $\mathbf{q_2}$. Minimize

$$L(q_1, q_2) = \int_{q_1}^{q_2} \zeta(\tau)d\tau = \int_{q_1}^{q_2} \tau^{-2}d\tau = q_1^{-1} - q_2^{-1}. \tag{27}$$

subject to

$$\int_{q_1}^{q_2} q(v)dv = Q(q_2) - Q(q_1) = 1 - \alpha, \tag{28}$$

where

$$q(v) = \frac{1}{2^{n/2}\Gamma(n/2)} v^{n/2-1} \exp\left(-\frac{v}{2}\right), \ v > 0, \ n > 0, \tag{29}$$

$$Q(u) = \int_{-\infty}^{u} q(v)dv. \tag{30}$$

Analytical Solution of the Problem. It follows from (8) to (27) that

$$\frac{dL(q_1, q_2)}{dq_1} = \zeta(q_2)\frac{d(q_2)}{d(q_1)} - \zeta(q_1) = q_2^{-2}\frac{q(q_1)}{q(q_2)} - q_1^{-2}, \tag{31}$$

which vanishes if

$$q_2^{-2}\frac{q(q_1)}{q(q_2)} = q_1^{-2}. \tag{32}$$

For this example, the version 1 is an adequate version of possible decision-making. It follows from (32) that the optimal solution is given by

$$q_1^2 q(q_1) = q_2^2 q(q_2). \tag{33}$$

Numerical results giving values of q_1 and q_2 to four significant places of decimals are available (see Tate and Klett [4]).

3.2 New Simple Computational Method

Problem Statement in Terms of the Decision Variable p (Probability) and Quantile Functions $\mathbf{q_1}$and $\mathbf{q_2}$. Minimize

$$L^2(q_1, q_2) = \left(\int_{q_1}^{q_2} \zeta(\tau) d\tau \right)^2 = \left(\int_{q_1}^{q_2} \tau^{-2} d\tau \right)^2 = \left(q_1^{-1} - q_2^{-1}\right)^2, \quad (34)$$

where the p—quantile function q_1 of V is given (via Excel software: CHISQ.INV (probability p, deg freedom n)) by

$$q_1 = Q^{-1}(p), \quad (35)$$

and the $(1-\alpha+p)$-quantile function q_2 of V is given (via Excel software: CHISQ.INV (probability $1 - \alpha + p$, deg freedom n)) by

$$q_2 = Q^{-1}(1 - \alpha + p), \quad (36)$$

subject to

$$0 \leq p < \alpha. \quad (37)$$

Numerical Solutions. The optimal numerical solution minimizing $L(q_1,q_2)$ can be obtained using the computer software “Solver". If, for example, $n = 3$, $\alpha = 0.1$, then the optimal numerical solution is given by

$$p = 0.099478, \ q_1 = 0.58208, \ q_2 = 17.63810464, \ q(q_1) = 0.227512, \ q(q_2) = 0.000248 \quad (38)$$

with

$$q_1^2 q(q_1) = q_2^2 q(q_2) = 0.077085 \quad (39)$$

and the $100(1 - \alpha)\%$ shortest-length confidence interval

$$L(q_1, q_2) = 1.661282. \quad (40)$$

The $100(1 - \alpha)\%$ equal tails confidence interval is given by

$$L(q_1, q_2 | p = \alpha/2) = 2.714186 \quad (41)$$

with

$$p = 0.05,\ q_1 = 0.351846,\ q_2 = 7.814728,\ q(q_1) = 0.198465,\ q(q_2) = 0.022409. \tag{42}$$

The proposed approach correctly recognized the adequate version 1 of possible decision-making and gave accurate numerical results.

Relative Efficiency. The relative efficiency of $L(q_1, q_2|p = \alpha/2)$ as compared with $L(q_1, q_2)$ is given by

$$rel.eff._L\{L(q_1, q_2|p = a/2), L(q_1, q_2)\} = \frac{L(q_1, q_2)}{L(q_1, q_2|p = a/2)} = \frac{1.661282}{2.714186} = 0.612074. \tag{43}$$

4 Conclusion

The novel unified computational approach proposed in this paper represents the conceptually simple, efficient, and useful method for constructing exact statistical (or expected) shortest length or equal tails confidence intervals in terms of pivotal quantities and quantile functions. The exact confidence intervals with the shortest length or equal tails can be found easily and quickly. Applying the proposed novel unified computational approach, we are not in need to use the following: (1) analytical recognition and computational confirmation of adequate versions of possible solutions, (2) tables for numerical solutions, (3) more advanced methods, and (4) special computer programs. For example, the special computer program for (33) is given below

Minimize

$$z = (q_1^2 q(q_1) - q_2^2 q(q_2))^2 \tag{44}$$

where the p—quantile function q_1 of V is given (via Excel software: CHISQ.INV (probability p, deg freedom n)) by

$$q_1 = Q^{-1}(p), \tag{45}$$

and the $(1-\alpha+p)$-quantile function q_2 of V is given (via Excel software: CHISQ.INV (probability $1 - \alpha + p$, deg freedom n)) by

$$q_2 = Q^{-1}(1 - \alpha + p), \tag{46}$$

subject to

$$0 \leq p < \alpha. \tag{47}$$

Optimal Numerical Solution. The optimal numerical solution minimizing z can be obtained using the computer software "Solver". If, for example, $n = 3$, $\alpha = 0.1$, then the optimal numerical solution is given by

$$p = 0.099478,\ q_1 = 0.58208,\ q_2 = 17.63813,\ q(q_1) = 0.227512,\ q(q_2) = 0.000248,\ q_1^2 q(q_1) = q_2^2 q(q_2) = 0.077085. \tag{48}$$

with the $100(1-\alpha)\%$ shortest-length confidence interval

$$L(q_1, q_2) = q_1^{-1} - q_2^{-1} = 1.661282. \tag{49}$$

The main advantage of the proposed approach is that it includes only one decision variable (probability p) for making decisions under constraints. In other words, the two decision variables q_1 and q_2 are reduced to one decision variable (probability p). This approach greatly simplifies the problem of constructing shortest length or equal tails confidence intervals for unknown parameters of various distributions and is a novelty in the theory of statistical decisions regarding confidence intervals. It allows one to exclude unknown (nuisance) parameters from the problem using the technique of invariant statistical embedding and averaging in terms of pivotal quantities (ISE & APQ) [5–14].

The unified computational method described in the paper is illustrated in detail for a certain general case. Applications of this method to construct shortest length or equal tails confidence intervals for unknown parameters of log-location-scale or other probability distributions can follow directly.

References

1. Juola RC (1993) More on shortest confidence intervals. Am Stat 47:117–119
2. Nechval KN, Nechval NA, Vasermanis EK, Makeev VY (2002) Constructing shortest-length confidence intervals. Transp Telecommun 3:95–103
3. Ferentinos KK, Karakostas KX (2006) More on shortest and equal tails confidence intervals. Commun Stat Theory and Methods 35:821–829
4. Tate RF, Klett GW (1959) Optimum confidence intervals for the variance of a normal distribution. J Am Stat Assoc 54:674–682
5. Nechval NA, Berzins G, Purgailis M, Nechval KN (2008) Improved estimation of state of stochastic systems via invariant embedding technique. WSEAS Trans Math 7:141–159
6. Nechval NA, Nechval KN, Danovich V, Liepins T (2011) Optimization of new-sample and within-sample prediction intervals for order statistics. In: Proceedings of the 2011 world congress in computer science, computer engineering, and applied computing, WORLD-COMP'11, Las Vegas, Nevada, USA, pp 91–97
7. Nechval NA, Berzins G, Danovics V (2016) Optimization of statistical decisions for age replacement problems via a new pivotal quantity averaging approach. Am J Theor Appl Stat (AJTAS) 5:21–28

8. Nechval NA, Berzins G, Balina S, Steinbuka I, Nechval KN (2017) Constructing unbiased prediction limits on future outcomes under parametric uncertainty of underlying models via pivotal quantity averaging approach. Autom Control Comput Sci 51:331–346
9. Nechval NA, Berzins G, Nechval KN (2018) Intelligent planning reliability-based inspections of fatigued structures for the crack initiation period in the Weibull case under parametric uncertainty. Autom Control Comput Sci 52:184–197
10. Nechval NA, Nechval KN, Berzins G (2018) A new technique for intelligent constructing exact γ-content tolerance limits with expected $(1-\alpha)$-confidence on future outcomes in the Weibull case using complete or type II censored data. Autom Control Comput Sci 52:476–488
11. Nechval NA, Berzins G, Nechval KN, Krasts J (2019) A new technique of intelligent constructing unbiased prediction limits on future order statistics coming from an inverse Gaussian distribution under parametric uncertainty. Autom Control Comput Sci 53:223–235
12. Nechval NA, Berzins G, Nechval KN (2019) A novel intelligent technique for product acceptance process optimization on the basis of misclassification probability in the case of log-location-scale distributions. In: Wotawa F et al (eds) Advances and trends in artificial intelligence. From theory to practice, IEA/AIE 2019, LNCS, vol. 11606. Springer Nature Switzerland AG, pp 801–818
13. Nechval NA, Berzins G, Nechval KN (2020) A novel intelligent technique of invariant statistical embedding and averaging via pivotal quantities for optimization or improvement of statistical decision rules under parametric uncertainty. WSEAS Trans Math 19:17–38
14. Nechval NA, Berzins G, Nechval KN (2020) A new technique of invariant statistical embedding and averaging via pivotal quantities for intelligent constructing efficient statistical decisions under parametric uncertainty. Autom Control Comput Sci 54:191–206

UMUSE: User Monitoring of the US Presidential Election

Christoph Glauser, Loris Schmid, and Jacques Savoy

Abstract This paper presents the results of a new monitoring project of the US presidential election with the aim of establishing computer-based tools to track the popularity of the two main candidates. The innovative methods that have designed and developed allow us to extract the frequency of search queries sent to numerous search engines, social media, and e-shops. Based on these data, this paper demonstrates that Donald Trump is significantly more frequently searched than Joe Biden. When analyzing various political topics, it is observed that the US Internet users have shown a remarkable interest in subjects like *Coronavirus*, and *Jobs* in 2020. Other subjects such as *Education* or *Healthcare* were given less attention. Finally, some "flame" topics such as *Black lives matter* appeared to be very popular for a few days before returning to a low level of interest.

Keywords Search engine · Online searches · Online presence · Data for politics · Political demand of citizens

1 Introduction

For more than three decades, search engines have been used to find the most pertinent web pages in response to user requests. To better understand users' needs, search engines and social media inspect the frequency distribution of submitted queries. Over the course of a year, some requests are sent very frequently (e.g.,

C. Glauser (✉) · L. Schmid
IFAAR, Bern, Switzerland
e-mail: Glauser@ifaar.ch

L. Schmid
e-mail: Schmid@ifaar.ch; Loris.Schmid@unine.ch

L. Schmid · J. Savoy
University of Neuchatel, Neuchatel, Switzerland
e-mail: Jacques.Savoy@unine.ch

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_45

"Facebook", "YouTube", as well as names of celebrities such as "Shakira", "Kardashian") while other queries occur often during a short period of time (e.g., in July with "Tour de France" or "Weight Watchers" during the first weeks of the new year). An overview of those frequent requests can be found on various platforms offering trends and keyword tools, as, for example, in the Google Trends service (`trends.google.com`).

Following a similar approach, the main objective of our project is to produce a systematic monitoring of the 2020 US presidential election, considering both the popularity of the two main candidates and the political demand of the citizens (by monitoring the most frequently searched political topics). These measurements are based on the occurrence frequencies of queries sent to numerous search engines (e.g., Google, Bing, Yahoo!, Ask, Lycos, Alexa, etc.) as well as social networks (e.g., Facebook, Twitter, Linkedin, etc.), and other platforms (e.g., ReutersNews, Amazon, eBay, etc.). Our underlying hypothesis is, the higher the frequency of a given query (i.e., "Donald Trump," "education," "jobs," or "coronavirus"), the more popular or interesting it is in the eyes of the users that are also citizens.

Based on this innovative Application Program Interface (API) technology, our first research question is to detect significant differences between the two candidates in the US presidential election [1]. According to a recent Gallup poll (September 2020), 30% of Americans identified as Democrats, 29% as Republicans, and 40% as Independents [2]. These values indicate that both parties enjoy an equal percentage of support among citizens. Does such an equilibrium also appear in the search queries of US users?

For our second research question, we want to monitor the evolution of the main concerns of the US citizens during this electoral campaign. For the 2016 presidential election, Ehrenfreud & Tankersley [3] singled out seven issues which are immigration, taxes, family, healthcare, trade, climate change, and foreign policy. During the 2020 presidential election, are these political themes still important and relevant, or can we observe other recurrent topics and how their attracted interest evolves during the campaign? Knowing this political demand and its intensity could be useful for candidates.

To provide an answer to these questions, the rest of this paper is organized as follows. Section 2 presents the technology and methods developed and their motivations while Sect. 3 describes their application to the candidates' names. Section 4 illustrates the monitoring system we applied to a set of selected topics reflecting the main concerns of the US digitally active citizens. Section 5 exposes a general discussion and draws the main findings of this study and a comparison between the monitoring of the two campaigns of 2016 and 2020.

2 Method and Motivations

Thanks to recent advances in natural language processing [4, 5], and API technology, data can be extracted from 14,103 search engines, social networks, and other electronic platforms such as e-shops and media. Each of these sources has its designated API and hence needs to be handled separately. Moreover, the data format may change without prior notification which requires ongoing daily monitoring of the integrity of results acquired from that source. The harvested data are then cross analyzed systematically, comparing similar search patterns on all the intermediates based on the frequency of search queries submitted by Internet users. To validate this data, the effect is considered that search inquiries are either high, medium, or low (in comparison) on almost all of the digital intermediates or channels.

Some additional parameters must be specified. First, the crawl can be parametrized to take account of only a predefined geographical country such as the US, the UK, or Canada. In this study, the targeted users are located in the United States. Second, it is important to examine and compare numerous search engines, social networks, and shops to assure a high level of coverage. Focusing only on a single (or a few) service(s) would run the risk of obtaining biased or inaccurate data. Third, the gathered data were validated and tested for reliability. Therefore, patterns can be identified and topics can be reasonably compared across most of the popular digital channels. The best scope for measuring valid data on various digital channels is the arithmetic average of scores over the previous 30 days at any given time. Reliability testing is done by comparing the multi-channel results of all search queries, with the entire number of active users in a given country (by domain). In the case of the US for this universe of active users, it is based on the ITU figures which identify approximately 284 million users.

Our innovative *finding* technology (developed and improved at the IFAAR institute in Bern between 2010 and 2020) allows us to examine, monitor, and systematically compare the evolution of the candidates' digital demand throughout the entire US electoral timespan. In prior work, some preliminary studies have been conducted on a smaller scale when analyzing campaigns or online behavior about COVID-19 in Switzerland [6].

To observe the evolution of each candidate's popularity, their names were used as queries (namely "Donald Trump," "Joe Biden"). Then, the data were gathered daily throughout the entire period of the elections from February 1, 2020 until October 28, 2020. Due to intermittent technical issues, there are no data available for only a few days in this whole period. We decided to use the full name of each candidate as a query to avoid possible false matches with other public figures carrying the same surname (e.g., Ivanka Trump) or corresponding to a name related to different named entities (e.g., Trump Tower).

By our hypothesis, user requests with a proper name are potentially indicating the awareness or popularity of the respective presidential runner. We must note that not all submitted queries denote an interest in favor of the corresponding candidate, but the majority thereof are generally assumed to be in favor of the candidate. Having a larger awareness for a given candidate during a longer period of an election campaign, means being the center of interest of the media, appearing more often in the news, both in the newspapers and TV. A candidate not getting enough awareness would not win an election. Moreover, we estimate that the proportion of unfavorable queries of a candidate would be similar for both runners. But this aspect is not an essential point. The favorable and unfavorable, or positive and negative, connotations in the digital age obviously play a much less important role if awareness is the main objective.

On the Web, various tools are using popularity-based information, under the assumption that popularity is positively correlated with usefulness and pertinence. For example, the Google search engine ranking algorithm favors web pages receiving a higher number of incoming links. In this case, a link to a page is viewed as a *vote* in favor of the target page, indicating its relevancy and usefulness [7]. As another example, Cleverdon et al. [8] found that only 5% of the bibliographic references in a scientific paper indicate a negative endorsement. References to other entities (authors, webpages, etc.) correspond, in general, to a positive perception of the target entity.

Finally, compared to traditional surveys, our approach can reach a significantly larger number of individuals of a population, including the 284 million active internet users in the US, which corresponds to 85.8% of the United States population. Additionally, it does not have the typical drawbacks and other situational influential effects of traditional surveys, such as people not telling the truth [9] or including only a very specific part of the population that tends to participate in surveys [10] (e.g., the erroneous predictions during the 2016 US election and for the Brexit vote). After these wrong predictions, the question of trust in public polls has been intensively discussed in the social- and market research communities, also on an international level.

3 Follow-Up on the Popularity of the Candidates

Based on the harvested data, Fig. 1 shows the number of queries (in millions) sent by US users to 14,103 search engines during the US 2020 election campaign. Each point in this graph represents the total number of queries submitted during the previous 30 days (moving average). Thus, the first point in Fig. 1 denoted "01 Mar 2020" represents the total number of requests sent during February 2020. Using an arithmetic average over one month smooths out the curve as opposed to having a sequence of sharp highs and lows. On the other hand, with such a smoothing approach, one cannot immediately detect the crowds' reaction to a given event. However, it may well show secondary meta-effects of certain events during the campaign or even toward the end of a campaign.

In this graph, vertical lines mark notable events, namely, Super Tuesday (March 3), state of emergency for COVID-19 (March 13), Sanders dropping out (April 16), death of G. Floyd (May, 25), the Democratic National Convention (DNC, Aug. 17–20), the Republican National Convention (RNC, Aug. 24–27), the death of R. Bader Ginsburg (September 18), and the tweets announcing Trump's infection (October 1). The horizontal axis is marked with increments of 20 million requests.

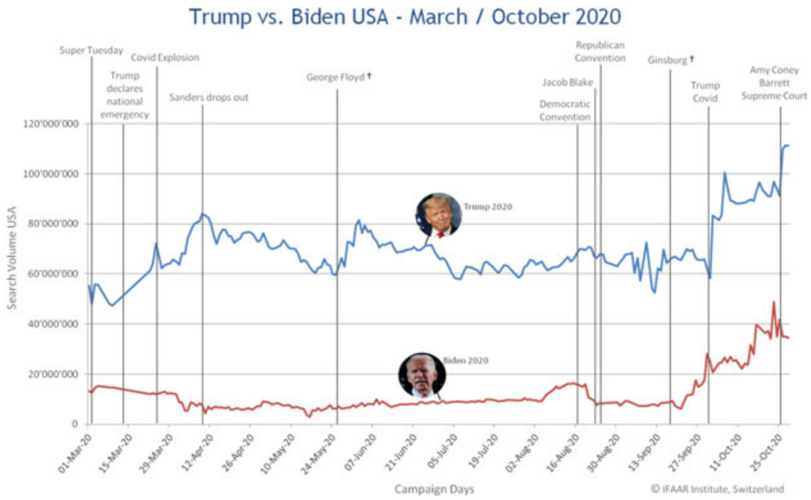


Fig. 1 Evolution of the number of search queries with candidates' names from US Internet users

From March 1st up to G. Floyd's death, the interest for Joe Biden tends to diminish slowly. This trend turns around at the end of May with a peak before and during the Democratic National Convention (mid-August). During the first two weeks of September, Biden's popularity was stable but rather low. A clear increase appears mid-September, certainly boosted by the first TV debate (September 29) and the closeness of the election. The final level reaches around 40 million requests per month.

On the other hand, Donald Trump clearly attracts more queries (translating to interest according to our hypothesis) than the Democratic Party nominee. During the entire campaign, the interest difference is always significant (t-test, 1% significance level). More precisely, this graph depicts an increasing interest in Trump from March to mid-April (Sanders' exit), then a decrease can be observed up to Floyd's death (May 25) where Joe Biden showed his empathy for the family at the funeral by video call. The important demonstrations and outbursts related to this death also boost the number of requests related to Donald Trump. His reaction to BLM by promising "Law and Order" to the Americans basically caused this upturn for Donald Trump. From the end of June to September, one can see an erratic behavior inside a slightly decreasing trend. But the first TV debate raised the interest for both Trump and Biden. In addition, the tweet sent on October 1st announcing Trump's infection with COVID-19 produced an incredible boost in the number of queries about Trump, reaching more than 100 million requests in a monthly average for the last weeks of the electoral campaign.

To provide a comparative basis, Fig. 2 displays the number of queries for both candidates during the US presidential election campaign in 2016 and 2020 (March to October). As one can see, Trump's curve in 2020 is always higher than those of 2016. Furthermore, the current US president is able to attract around five times

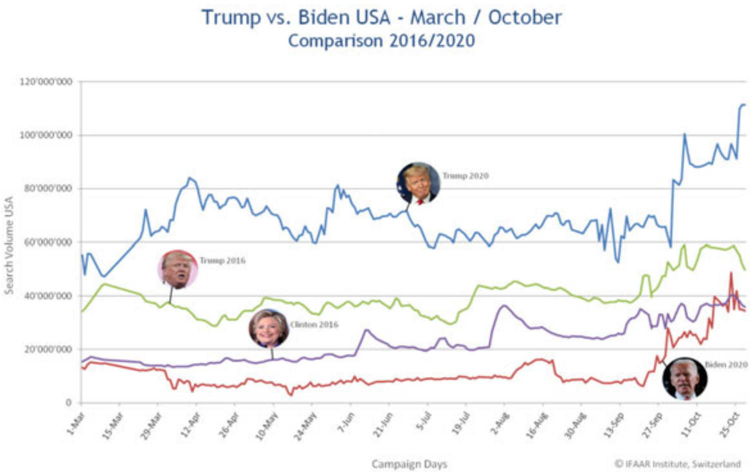


Fig. 2 Comparing the evolution of candidates' popularity by search requests in 2016 and in 2020

more interest from the active US internet users than the Democratic candidate. Also, Hillary Clinton was usually able to obtain more queries than Joe Biden, indicating a real concern for the Democratic Party in 2020. Only during the last weeks of this campaign, Joe Biden was capable to reach Hillary's level.

Of course, one can assume that the president in office can be the target of a number of requests, independently of the fact that he is also candidate for his own succession. However, this fact cannot explain the large difference in the number of queries between Trump in 2016 and in 2020.

4 Evolution of Main Topics Throughout the Election

During an electoral campaign, a set of topics can be defined to represent the most important concerns of the general public and to characterize the demand for political discussions and solutions. For a presidential election, one can retain issues such as jobs and unemployment, family, education, healthcare, economy, immigration, taxes, climate change, and foreign policy. Of course, this is not an exhaustive list. For example, gun control or defense issues are missing. In our project, the final selection of studied topics was also made on the basis of issues from former US election campaigns and website analyses of former candidates from 2000 to 2016. In addition, we have included the coronavirus and Supreme Court themes. Then, the search counts of the designated topics were collected via search engines, social media, and e-shops starting February 1 to October 28.

However, each topic cannot be defined by a single query. We must recognize that a given subject can be expressed in various words or formulations [11] without

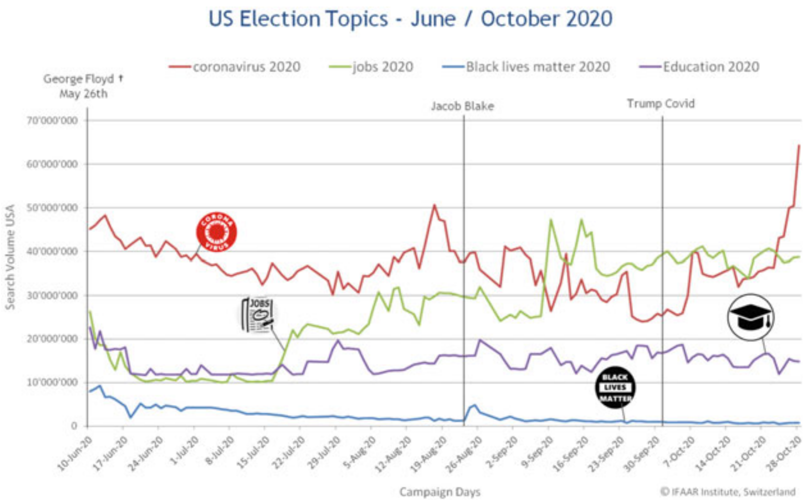


Fig. 3 Evolution of some topics during the 2020 US electoral campaign (June–October)

one being predominantly used more frequently than the others. This vocabulary problem is well known in various applications such as information retrieval [12]. To generate the most appropriate expressions, we analyzed the pertinent wordings used in previous US presidential elections. Therefore, each topic corresponds to many terms as, for instance, the topic *Immigration*[1] is mapped to the queries "illegal immigration," "immigration reform," "immigrant," "Latin immigration," "immigration raids," while the topic *Healthcare* maps to "Obamacare," "healthcare reform," "medicare," and "health care". A similar strategy has been applied for all other topics. This solution is similar to the technique applied by the DICTION [13] or LIWC [14] systems.

Displaying all main political topics on the same graph will render it unreadable. Therefore, Fig. 3 depicts only four main issues, namely *Coronavirus*, *Jobs* (and employment), *Education*, and *Black Lives Matter* (BLM). Usually, the crowd interest in a given theme is relatively stable during an electoral campaign. For instance, the topic *Education* or *Economy* (not shown in Fig. 3) corresponds to mainly a horizontal line, with some variations, around 15 million queries per month. For others, the underlying demand pattern can vary greatly as, for example, with *Small business* exhibiting constant fluctuations reflecting the topic's volatility.

More pertinent for this 2020 presidential election, the topic *Coronavirus* is overall stable around 40 million requests per month. One can however see a decrease leading into July, reaching a monthly level around 30 million. Then, one can see an increase to 40 million (August) followed once again by a decline to 30 million. This decrease could indicate a certain *Corona fatigue* after the lockdowns in several other countries

[1] Each topic is printed in italics to signal that such a name corresponds to a list of possible queries and not just one.

during the summer. With the second wave beginning in October, the interest for this subject raised up to a mean of 80 million requests per month. Of course, this question was a real concern for Trump's administration and an issue regularly used by the Democrats to attack him.

During this campaign, a second interesting topic is *Jobs,* having an average of 10 million inquiries during June, but with a distinctive increase from mid-July achieving 20 million queries in July, then to 30 in August, and 40 million on average since September. The high level of the unemployment rate during this period is clearly reflected in the rise of the frequency of searches for this topic. At its highest (April), the unemployment rate reached a value of 14.7%, and then tended to decrease (10.2% in July, 7.9% in September) as the year progressed. This is good news for Trump's reelection, especially when we consider that US citizens trust him more on this issue than his opponent Joe Biden [15].

The topics *Immigration* and *Government spending* received relatively little attention from the US Internet users. For the first subject, we can put forward the argument that these topics are discussed more intensively in social networks and do not completely correspond to a classic Internet search.

For a relatively short time span (a few weeks), US online users demonstrated interest in specific flame topics such as *Black lives matter.* As shown in Fig. 3, this theme was searched intensively during a few weeks, then return to a low level of attention. The two highest frequencies occur after the death of G. Floyd (May 25), and the shooting of J. Blake (August 23).

5 Discussion and Conclusion

In this study, we monitored the interest of the US internet users for the two main candidates to the 2020 US presidential election by analyzing search volumes on most digital channels. To monitor this, a new innovative technology was applied to repeatedly crawl numerous search engines and various online channels in order to determine the frequency of search queries associated with the candidates' name as well as with related important political topics.

Even if the percentage of Republicans and Democrats is similar (around 30%), the popularity of Donald Trump measured by the number of queries submitted to search engines is, in mean, five times higher than for Joe Biden (75 million vs. 15 million). For the Democratic Party nominee, the number of search requests with his name was relatively stable during most of the campaign, with an increased interest during the last five weeks (after the first TV debate, see Fig. 1). The level achieved by Joe Biden is however significantly lower than the mean number obtained by Hillary Clinton in 2016 (23 million). In addition, the Republican Party nominee in 2020 obtained a higher popularity than in 2016 (75 million vs. 40 million as shown in Fig. 2).

Second, through a selection of topics, we have observed the evolution of people's interest showing a major interest in *Jobs* (and unemployment) followed by *Coronavirus.* This last subject is new in an electoral campaign and could favor Biden's election. The direct effect of this pandemic was an impact on the labor market in the US with an increase of the unemployment rate up to 14.7% in April 2020. This aspect is very well reflected in the data by the *Jobs* topic depicting a clear growth in July followed by an increasing trend up to October. Other topics, such as *Economy*, *Business*, or *Education* were less important (see Fig. 3). For the candidates, those themes were also less significant. For example, one can count only five occurrences of the word "education" in tweets send by both candidates from August to October.

Our model is not a predictive one [16, 17], but can be applied to measure the general active demand of the citizens for some political topics and the popularity of candidates fully independently of any survey situation. We claim that based on our data, Donald Trump has clearly dominated the digital election campaign. A predictive model must take various other factors into account. For example, a leader must be a charismatic person, a strong and tough guy able to defend the US (and Trump is considered "as the toughest guy I've ever met" said R. Stone). The emotional component [18] is also an essential factor on citizens' decisions. During the 2016 campaign, Trump had focused his campaign on the fear and anger against immigrants, China, the political establishment in Washington, and traditional media carrying fake news and ignoring the (white) blue-collars [18, 19]. Thus, as expressed by R. Stone: "Those who are outraged will vote." Therefore, both the unemployment concern and identity crisis could be key factors for this 2020 US election (this paper was written before the election day in US (November 3, 2020)).

Acknowledgements This study was supported by Hasler Foundation (Bern, Switzerland).

References

1. Boller PF (2004) Presidential campaigns. From George Washington to George W. Bush. Oxford University Press, Oxford
2. Gallup historical trends: party affiliation. Gallup news. https://news.gallup.com/poll/15370/party-affiliation.aspx. Accessed Oct 15 2020
3. Ehrenfreud M, Tankersley J (2016) Seven things to know about Hillary Clinton and Donald Trump before you vote. Washington post, Nov 7 2016
4. Eisenstein J (2019) Introduction to natural language processing. The MIT Press, Cambridge
5. Savoy J (2020) Machine learning methods for stylometry. Springer, Cham
6. Glauser C, Schmid L, Kurpicz-Briki M (2020) Identifying the public interest in COVID-19 contact tracing apps in Switzerland based on online search behaviour. In: Proceedings of the 13th international conference on theory and practice of electronic governance (ICEGOV2020), The ACM Press, New York
7. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1–7):107–117
8. Cleverdon C, Mills J, Keen M (1966) Factors determining the performance of indexing systems. Technical report, ASLIB Cranfield research project, Cranfield, UK
9. Stephens-Davidowitz S (2017) Everybody lies: big data, new data, and what the internet can tell us about who we really are. Dey Street Books, New York
10. Wright KB (2005) Researching internet-based populations: advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. J Comput Mediat Commun 10(3):JCMC1034

11. Furnas G, Landauer T, Gomez L, Dumais S (1987) The vocabulary problem in human-system communication. Commun ACM 30(11):964–971
12. Manning C, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
13. Hart R, Childers J, Lind C (2013) Political tone. How leaders talk and why. The University of Chicago Press, Chicago
14. Tausczik Y, Pennebaker J (2010) The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 29(1):24–54
15. Cage M (2020) Americans rate Trump better on the economy. Washington post, Sept 2 2020
16. Lichtman AJ (2008) The keys to the White House: an index forecast for 2008. Int J Forecast 24(2):301–309
17. Gayo-Avello D (2012) No, you cannot predict elections with Twitter. IEEE Internet Comput 16(6):91–94
18. Hart RP (2020) Trump and us. What he says and why people listen. Cambridge University Press, Cambridge
19. Sides J, Tesler M, Vavreck L (2018) Identity crisis. The 2016 presidential campaign and the battle for the meaning of America. Princeton University Press, Princeton

TextRank Keyword Extraction Method Based on Multi-feature Fusion

Chenwei Wu, Lyuchao Liao, Francis Afedzie Kwofie, Fumin Zou, Yongqiang Wang, and Maolin Zhang

Abstract Keyword extraction technology is widely used in text mining tasks such as text classification, text clustering, and abstract extraction. There are many ways to extract keywords, but the efficiency of keyword extraction still needs to be improved. In response to this issue, we proposed a multi-feature fusion TextRank (MFFTR). After constructing the text into a word network graph, the frequency features and position features of the words were merged as the initial weights of word nodes, and the dual-core network node contribution calculation method was used to assign weights to edges. By doing so, a new node weight calculation method based on the traditional TextRank was introduced. Experiments were conducted on two different datasets with significant differences in average text lengths. The results showed that the multi-feature fusion TextRank proposed in this paper was significantly better than the five keyword extraction methods. We concluded that the use of multi-feature fusion can improve the efficiency of automatic keyword extraction.

Keywords Graph model · TextRank · Multi-feature fusion · Keyword extraction · Dual-core Network

C. Wu (✉) · L. Liao · F. Afedzie Kwofie · F. Zou · Y. Wang · M. Zhang
Fujian Key Laboratory of Automotive Electronics and Electric Drive, Fujian University of Technology, Fujian 350118, China
e-mail: 943204646@qq.com

F. Zou
e-mail: fmzou@fjut.edu.cn

Y. Wang
e-mail: yqwang0374@163.com

M. Zhang
e-mail: mailzml@163.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_46

1 Introduction

Keywords unveil the main purpose of an article. Before the popularity of the Internet, keywords were mostly obtained through manual labeling. With the growth of text data on the Internet, the method of manual labeling to obtain keywords is failing to meet the current demands. This opened the door for keyword extraction technology.

Keyword extraction technology is divided into supervised and unsupervised methods. Since the supervised methods require manual labeling in advance, the cost of using such methods is too large today with the increasing amount of data. Therefore, most of the current researches adopt unsupervised methods. The keyword extraction method based on the graph model is a kind of unsupervised method. It converts a text into a word graph and measures the significance of word nodes. TextRank [1], a kind of word network graph, can realize keyword extraction through the word co-occurrence relationship. However, in the process of keyword extraction, it ignores the attributes of the word itself, sets the initial weight of the word node to 1, and does not consider the processing method when the influence of the word on adjacent nodes is different. Basing on the word graph, we proposed a multi-feature fusion TextRank (MFFTR).

The main contributions of this paper are as follows: After fusing word position features and word frequency features as the initial weights of nodes, we proposed a method to calculate the dual-core network nodes contribution (DCNNC) and used it to assign weights to edges in the word graph. By using the statistical features of word graph, that is, the number of adjacent nodes of word nodes, we improved the iterative formula of the original TextRank. Comparing with five keyword extraction methods on two datasets, the experimental results showed that the MFFTR algorithm has better performance.

The rest of the paper is organized as follows: Section 2 introduces the research results of predecessors. Section 3 describes the keyword extraction scheme proposed in this paper. Section 4 shows the performance of this method on different datasets. Section 5 summarizes this research and introduces future research plans.

2 Related Works

Currently, the unsupervised keyword extraction method is divided into keyword extraction methods based on multi-feature fusion, topic models, and graph models.

The multi-feature fusion method often constructs a scoring function for each word by fusing a variety of lexical features and then selects several words with higher scores as keywords. For instance, the TAKE algorithm [2] extracts keywords by fusing different features. WAN et al. [3] improved the traditional word frequency feature acquisition scheme and merged the part-of-speech feature of words to propose a multi-feature fusion keyword extraction method. Based on the features of the subject, part-of-speech, and sentence component of words, Liu et al. [4] constructed a word weight calculation function to extract keywords.

The topic model holds that a document is composed of different topics and each topic contains different words. Blei [5] proposed the latent Dirichlet allocation (LDA) topic model which introduced prior distributions based on the PLSA topic model [6]. It formed a three-layer Bayesian model of document-topic-words, which improved the effect of topic mining. Due to the poor performance of the LDA on short text, scholars have proposed improved models such as BTM [7] and Twitter-LDA [8].

The graph-based keyword extraction method has achieved better results because it can be fused with a variety of features. The TextRank [1] is a popular graph model keyword extraction method and is based on the PageRank [9]. It modifies the transition probability of the edges between nodes in the PageRank based on the number of word co-occurrences, but ignores the semantic relevance of words, and does not consider context and auxiliary information [10]. As a result, the keyword extraction effect was not ideal. Due to this, many scholars have proposed a series of improved algorithms. For example, Wang et al. [11] proposed WordAttractionRank by using word embedding and Bougouin et al. [12] proposed TopicRank using a hierarchical clustering method.

3 Methodology

We proposed an improved algorithm based on the traditional TextRank. The multi-feature fusion method was used as the initial weight of the nodes in the word network graph, and the edge weights in the word graph were redefined through the DCNNC calculation method proposed in this paper. At the same time, the statistical features of word nodes in the word network graph were used to further improve the node transition probability of the original TextRank.

3.1 Basic Concepts and Definitions

Definition 1: Word Graph Construction We represent the text as a graph. The nodes in the graph represent the words in the document, and the edges represent the co-occurrence relationships between nodes.

Definition 2: The Initial Node Weight We use the product of the word TF-IDF value and the word position weight as the initial weight of the word node. The calculation method of word position weight $\boldsymbol{PWv_i}$ follows the method mentioned in the paper [13]. The reciprocal sum of the index positions of words in the document is the word position weight. For example, if a word is found in the following positions: 3rd and 6th, then the position weight is: $\frac{1}{3} + \frac{1}{6} = \frac{1}{2}$. The formula is as follows:

Fig. 1 Dual-core network

$$tfidf_{vi} = tf_{vi} \times idf_{vi} \quad (1)$$

$$W_{vi} = tfidf_{vi} \times pw_{vi} \quad (2)$$

Definition 3: Dual Core Network The network composed of nodes $\boldsymbol{v_i}$ and $\boldsymbol{v_j}$ and their respective set of adjacent nodes $\boldsymbol{N(v_i)}$ and $\boldsymbol{N(v_j)}$ is called a dual-core network. Node $\boldsymbol{v_i}$ and node $\boldsymbol{v_j}$ are the two cores of the dual-core network. As shown in Fig. 1, node 0 and node 1 are the two cores of the dual-core network.

Definition 4: Dual-Core Network Nodes Contribution In the dual-core network, as shown in Fig. 1, the contribution of core nodes v_i to v_j or v_j to v_i is recorded as DCNNC (dual-core network nodes contribution). The formula is as follows:

$$DCNNC(v_i, v_j) = \frac{W_{v_i}}{\sum_{v_k \in N(v_i)} W_{v_k}} \times \frac{W_{v_j}}{\sum_{v_h \in N(v_j)} W_{v_h}} \quad (3)$$

In the formula, Wv_i represents the initial weight of node v_i, and $N(v_i)$ and $N(v_j)$ represent the adjacent node set of nodes v_i and v_j, respectively.

3.2 *Multi-feature Fusion TextRank Algorithm*

The TextRank converts a document into a word graph and takes each word as a node, and the edges represent the co-occurrence relationships between nodes. After constructing the word graph, the scores of each node are calculated by the following formula:

$$S(v_i) = (1 - d) + d \times \sum_{v_j \in In(v_i)} \frac{W_{ji}}{\sum_{v_k \in Out(v_j)} W_{jk}} S(v_j) \quad (4)$$

where $In(v_i)$ represents the set of nodes pointed to node v_i, and $Out(v_j)$ represents the set of nodes pointed to by node v_j. W_{ji} represents the number of co-occurrences of the two-word node, and $S(v_i)$ represents the score of node v_i. $\boldsymbol{d}$ is the damping coefficient.

The TextRank does not consider the attributes of words themselves leading to poor results. To improve its accuracy, we incorporated the initial weight of the word node into the iterative process of the TextRank. By fusing the position and frequency feature as the initial weight of the word node, and using the calculation method of the dual-core network nodes contribution, the edge weight in the word graph was redefined:

$$E_{ij} = w_{ij} \times DCNNC\left(v_i, v_j\right) \tag{5}$$

where $DCNNC\left(v_i, v_j\right)$ is the result of node contribution of dual-core network. The proposed MFFTR was based on the undirected weighted word graph. Therefore, $In(v_i)$ and $Out(v_j)$ in the original formula (4) were both replaced by $N(v_i)$, which represents adjacent nodes of node v_i. Also, by integrating the statistical features of the word graph into the iterative process of the TextRank, we proposed a new method for calculating the weight of the word node. The improved TextRank formula is as follows:

$$S(v_i) = (1-d) + NW_{v_i} \times d \times \sum_{v_j \in N(v_i)} \frac{E_{ji}}{\sum_{v_k \in N(v_j)} E_{jk}} S\left(v_j\right) \tag{6}$$

where NW_{v_i} in the formula represented the number of adjacent nodes of each word node v_i. d was the damping coefficient, which was 0.85 in this paper.

4 The Experiment

The keyword extraction scheme proposed in this paper was made of seven steps. The specific steps are shown in the flowchart in Fig. 2:

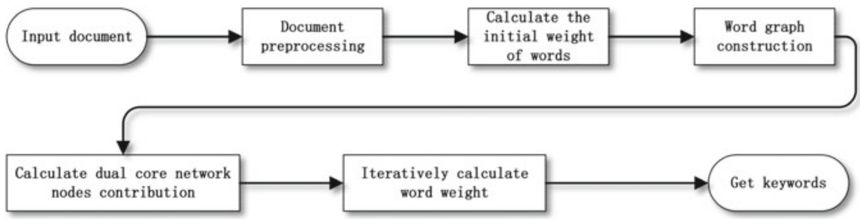


Fig. 2 Keyword extraction procedures

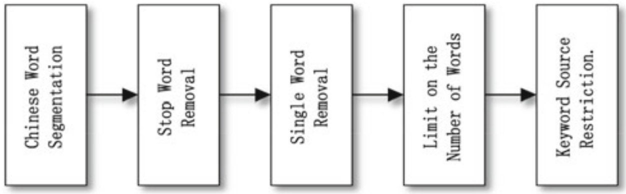


Fig. 3 Data preprocessing steps

4.1 Datasets

To effectively evaluate the keyword extraction efficiency of the proposed method, the study also made use of the Chinese datasets used in the works [14 and 15] by Liu Zhiyuan and Xiatian, respectively.

4.2 Data Preprocessing

In order to extract keywords accurately, text data needed to be preprocessed to obtain a suitable set of candidate keywords. The preprocessing steps used in this experiment mainly included five steps. The detailed steps were given in Fig. 3.

4.3 Evaluation Criteria

We evaluated the results of the keyword extraction method through precision, recall, and F1-measure. Precision stands for the percentage of correctly extracted keywords in automatically extracted keywords. Recall stands for the percentage of correctly extracted keywords in manually assigned keywords. F1-measure was the harmonic mean of precision and recall.

$$\text{precision} = \frac{|A \cap B|}{|B|}$$

$$\text{recall} = \frac{|A \cap B|}{|A|}$$

$$\text{F1 - measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In the formula, A stands for the handily assigned keywords and B stands for the automatically extracted keywords.

4.4 Experimental Results and Discussion

In order to evaluate the effect of the keyword extraction algorithm proposed in this paper, we compared it with five keyword extraction methods, which were TF-IDF, TextRank, PositionRank [13], Degree centrality, and Closeness centrality [16], on two datasets. We evaluated the variation of precision, recall, and F1-measure of various keyword extraction methods when extracting 1 to 6 keywords. The results were provided in Figs. 4 and 5.

The degree centrality and the closeness centrality are common methods for sorting the significance of nodes in a complex network. However, the keyword extraction effect of these two methods is not ideal, because they only rely on word graph features and do not consider other features of words when sorting nodes. TF-IDF is a popularly used keyword extraction method, but the chance of a word to be chosen as a keyword is only determined by word frequency. As a result, the effect of keyword extraction is not ideal. The PositionRank added the position features of the word to the iterative process of the PageRank. However, only considering the position weight of words was not comprehensive enough. TextRank only considered the word co-occurrence and neglected the attribute of the word itself in the process of iterative calculation of the weight of the word node, which led to the inability to effectively extract the keywords.

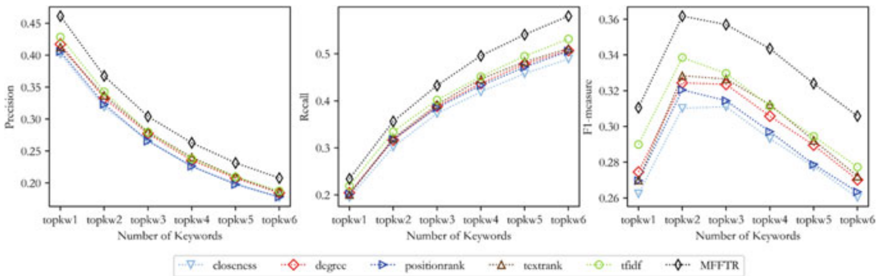


Fig. 4 Comparison results on the Liu Zhiyuan dataset

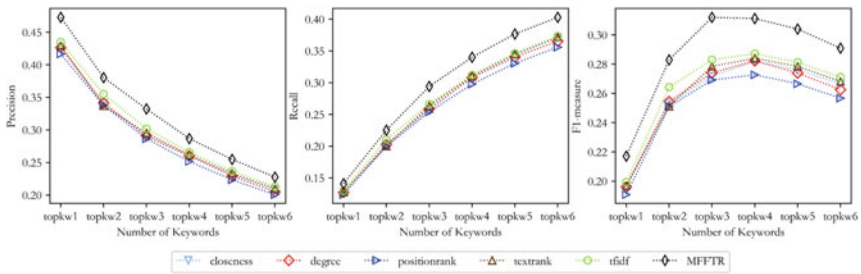


Fig. 5 Comparison results on the Xiatian dataset

This paper presented an improved algorithm based on the TextRank. The frequency and position features of words were used as the initial weights of the word nodes, and the calculation method of the DCNNC was used to redefine the edge weights in the word graph. The statistical features of the word graph were integrated into the iterative process of the TextRank to further improve the calculation method of the transition probability of the node. Results showed that the precision of the MFFTR and the five comparison algorithms were relatively high when extracting a small number of keywords. However, as the number of keywords extracted increased, the precision showed a downward trend. The recall was the opposite. When extracting a small number of keywords, the recall was generally low, and it showed an upward trend as the number of keywords extracted increased. Also, the F1-measure, the harmonic mean value of precision and recall, first showed a rising trend and then fell as the number of keywords extracted increased. The MFFTR proved to be significantly better than the five comparison algorithms in terms of precision, recall, and F1-measure.

5 Conclusion

To improve the accuracy of keyword extraction, we combined word frequency and word position features as the initial weights of word nodes. According to the dual-core network structure in the word graph, an algorithm for calculating the DCNNC was deduced, by depending on the node jump probability in the original TextRank, based on which a new method was proposed to assign weights to edges. Also, using the statistical features of the word graph, a new TextRank calculation formula was derived to calculate the node weight. Results showed that the proposed MFFTR was superior to the five traditional keyword extraction methods on two different datasets. Therefore, we concluded that the appropriate fusion of different word features into the TextRank can effectively improve the efficiency of keyword extraction. However, it is worth mentioning that this kind of improvement also increases the complexity of the algorithm.

Therefore, in future research, in addition to the reasonable integration of lexical features, it is also necessary to pay attention to the balance between the change of algorithm complexity and the keyword extraction effects after the introduction of features. This will be necessary to maximize the keyword extraction effect while sacrificing the complexity of the algorithm to the minimum.

Acknowledgments This work was supported in part by Projects of the National Science Foundation of China (41971340), projects of Fujian Provincial Department of Science and Technology (2018Y3001, 2018H0001, 2019I0019), and projects of Fujian Provincial Department of Education (FBJG20180049).

References

1. Mihalcea R, Tarau P (2004) TextRank: bringing order into texts. In: Proc. of the EMNLP, Stroudsburg: ACL, pp 404–411
2. Pay T (2016) Totally automated keyword extraction. In: Proceedings of 2016 IEEE International Conference on Big Data (Big Data). Doi: https://doi.org/10.1109/BigData.2016.7841059
3. Wan J, Zheng P, Si H, Xiong NN, Zhang W, Vasilakos AV (2019) An artificial intelligence driven multi-feature extraction scheme for big data detection. IEEE Access 7:80122–80132. https://doi.org/10.1109/access.2019.2923583
4. Liu X, Zhang Z, Li B, Zhang F (2019) Keywords extraction method for technological demands of small and medium-sized enterprises based on LDA. In: 2019 Chinese Automation Congress (CAC), pp 2855–2860. Doi: https://doi.org/10.1109/cac48633.2019.8996936
5. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Machine Learn Res 993–1022
6. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc, pp 289–296
7. Yan X, Guo J, Lan Y, Cheng X (2013) A biterm topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web—WWW 13, 1445–1456
8. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: Proceedings of European conference on information retrieval. Springer, Berlin, Heidelberg, pp 338–349. Doi: https://doi.org/10.1007/978-3-642-20161-5_34
9. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Stanford InfoLab
10. Jingsheng Z, Qiaoming Z, Guodong Z, Li Z (2017) Review of research in automatic keyword extraction. J Soft 28(9):2431–2449
11. Wang R, Liu W, McDonald C (2014) Corpus-independent generic keyphrase extraction using word embedding vectors. Soft Eng Res Conf. 39:1–8
12. Bougouin A, Boudin F, Daille B (2020) Topicrank: graph-based topic ranking for keyphrase extraction. In: Proceedings of international joint conference on natural language processing (IJCNLP). Nagoya, Japan, pp 543–551
13. Florescu C, Caragea C (2017) PositionRank: an unsupervised approach to keyphrase extraction from scholarly documents. In: Proceedings of the 55th annual meeting of the association for computational linguistics (vol. 1: Long Papers), pp 1105–1115
14. Liu Z, Chen X, Zheng Y, Sun M (2011) Automatic keyphrase extraction by bridging vocabulary gap. In: Proceedings of the fifteenth conference on computational natural language learning, pp 135–144
15. Tian X (2017) Extracting keywords with modified TextRank model. Data Anal Knowl Disc 1(2):28–34
16. Boudin F (2013) A comparison of centrality measures for graph-based keyphrase extraction. In: Proceedings of the sixth international joint conference on natural language processing, pp 834–838

Forming the System with the Functionality of Clinical Pharmacist for Personalized Treatment Strategy Searching

Vitalii Babenko, Olena Nosovets, Ievgen Nastenko, Volodymyr Pavlov, Viktoria Iakymchuk, Oleksandr Matviichuk, and Maksym Suvorov

Abstract Today, in Ukraine the task of creating an informative system for personalized treatment strategy searching is quite relevant. Its appearance will help to solve the long-burning issue of the absence of clinical pharmacists in medical institutions. These specialists are necessary because doctors cannot make an independent decision regarding the patient's treatment strategy, since even their full compliance with the clinical protocols does not ensure a fully optimal human state. This work describes how the informative system can be formed, which will not only serve as a decision support system for doctors but also will effectively find the necessary solutions. The personalized treatment strategy searching is a multi-objective optimization problem since the final state of the patient is described by several indicators. To solve this kind of problem, the authors developed an approach using the principles of the genetic algorithm and the analytic hierarchy process. This approach was used in the practical task of finding personalized treatment strategies for patients with congenital heart defects to demonstrate the difference between the decision made by doctors in real life and the decision produced by the algorithm. Predictive models of indicators after treatment were obtained by the random forest classifier algorithm. Most of the models had 100% accuracy on the testing sample, which indicates the high efficiency of the used classification method. The promoted approach will be the foundation for the informative system developed in the future, and medical institutions can use it for any type of task regardless of the disease types.

Keywords Treatment strategy · Clinical protocol · Clinical pharmacist · Patient's state · Multi-objective optimization · Genetic algorithm · Analytic hierarchy process · Random forest classifier

V. Babenko (✉) · O. Nosovets · I. Nastenko · V. Pavlov · V. Iakymchuk · O. Matviichuk
Igor Sikorsky Kyiv Polytechnic Institute, National Technical University of Ukraine, Kyiv 03056, Ukraine

M. Suvorov
National University of Pharmacy, Kharkiv 61002, Ukraine

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_47

1 Introduction

Treatment prescribing is certainly an important procedure in the fields of medicine and healthcare [1]. Slightest mistakes in no way can be made while choosing a treatment strategy for a patient, since they can cause irreparable harm to a person, and can also lead to unpredictable consequences for his further condition. There are clinical protocols [2], which are based on evidence-based medicine and must be followed by all healthcare specialists without any exception. They are the main medical technology documents, which are constantly updated and contain the best practices of recognized experts. Clinical protocols allow specialists to make the most effective decisions regarding treatment strategies.

If we talk about clinical treatment (which is necessary, for example, to eliminate postoperative complications), usually a specialist called a "clinical pharmacist" (or "clinical provisor") [3] deals with the selection of treatment (to be more precise with the choice of drugs and their dosages). This kind of specialist advises the doctors in a rational, comprehensive, and most adequate selection of therapeutic drugs, taking into account their pharmacokinetics, pharmacodynamics, interactions with other drugs, as well as taking into account the characteristics of the patient's physiology and pathology. Combining their wide knowledge of pharmaceuticals with the use of clinical protocols, clinical pharmacists can anticipate the direction of the disease, and, thanks to the drugs, predetermine and correct its further development.

Nevertheless, despite the obvious benefits of these specialists, at the moment Ukraine suffers from their shortage. It is quite rare to find clinical pharmacists in medical institutions, which may be due to financial constraints, so it makes it more profitable for institutions to provide a complete "carte blanche" to the doctors. Although this decision is quite understandable, do not forget that people's lives are at stake, and the doctor's lack of full knowledge of the types of drugs and their correct use at any time can lead to fatal consequences. Thus, it becomes necessary to create an informative system with the functionality of a clinical pharmacist for the optimal treatment strategy search, which will be personalized due to the physiological characteristics of the patient.

2 Problem Statement

The patient's treatment process covers two states: initial and final. These states are characterized by certain medical indicators of a person. It can be represented as the following vectors:

$X = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}^T$—describes the initial state of the patient (x_1, x_2, …, x_{n} are the patient's indicators before treatment).

$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix}^T$—describes the final state of the patient (y_1, y_2, …, y_m are the patient's indicators after treatment).

The treatment strategy, which is applied to the patient, can be described by a vector $I = \begin{pmatrix} i_1 \\ i_2 \\ \dots \\ i_k \end{pmatrix}^T$, where $i_1, i_2, \dots, i_k$ are the different types of drugs (the influence of parameters on the patient's state). Y directly depends on X and I and it can be represented by the following equation:

$$Y = X \times I \Leftrightarrow \begin{cases} y_1 = \sum_{j=1}^{n} \sum_{z=1}^{k} x_j i_z \\ y_2 = \sum_{j=1}^{n} \sum_{z=1}^{k} x_j i_z \\ \dots \\ y_m = \sum_{j=1}^{n} \sum_{z=1}^{k} x_j i_z \end{cases} \tag{1}$$

Equations for $y_1, y_2, \dots, y_m$ can be both linear and nonlinear. With their usage, it seems possible to simulate (modeling) the treatment process. Thus, a multi-criteria optimization problem arises, where it is necessary to find such values of vector I that will give the optimum of vector Y, values of vector X are set by default. It means that the personalized solution search will be made by taking into account the patient's initial state indicators. This idea was proposed in [4, 5] for a single-criteria problem. In the context of current work, this task can be formulated as "Personalized treatment strategy searching for patient's final state optimization".

It is necessary to create an algorithm that will be solving this kind of task, and the programmed computer will be producing the necessary solutions in a matter of seconds by algorithm usage. Its solution will make it possible to create an informative system based on this algorithm and doctors will be able to use the system in the process of prescribing the treatment for a patient.

3 Personalized Treatment Strategy Search Algorithm

In the process of creating the algorithm for solving the multi-criteria optimization problem, the following problems arise:

Table 1 The general form of criteria pairwise comparison

	y_1	y_2	...	y_m
y_1	$\frac{v_1}{v_1}$	$\frac{v_2}{v_1}$	...	$\frac{v_m}{v_1}$
y_2	$\frac{v_1}{v_2}$	$\frac{v_2}{v_2}$	...	$\frac{v_m}{v_2}$
...	...	...	...	...
y_m	$\frac{v_1}{v_m}$	$\frac{v_2}{v_m}$	...	$\frac{v_m}{v_m}$

1. The simultaneous optimization of the parameters of vector Y.
2. The search of values of vector I, which give the global optimum (NP-complete problem).

The first problem can be solved by using multi-criteria decision-making methods [6]. The main idea of these methods is to get a convolution of several criteria into one so-called “supercriterion”. This supercriterion in the process of solving the problem can be used to describe the final state of the patient.

One of the simplest and most easily interpreted methods of multi-criteria decision-making is the analytic hierarchy process (AHP) [7], invented by Thomas L. Saati in the 1970s. This method allows obtaining an additive convolution function by pairwise comparison of criteria priorities. The comparison mechanism by AHP in a general form is shown in Table 1.

where v_i is the sequential number in the list of criteria of vector Y, ranked by importance.

To obtain the additive convolution function, it is necessary to calculate the geometric mean for each row of the pairwise comparison matrix, and then normalize them. The calculated values will be the “weights” of each criterion, so the additive convolution function can be represented as follows:

$$F_{ac} = w_1 y_1 \pm w_2 y_2 \pm \ldots \pm w_m y_m \tag{2}$$

The signs in the formula are placed depending on whether it is necessary to maximize (then the “$+$” sign) or minimize (then the “$-$” sign) y_i. Therefore, if instead of $y_1, y_2, \ldots, y_m$ we substitute their form from Eq. (1), then the problem of multi-objective optimization is reduced to finding the maximum of F_{ac}, i.e. supercriterion.

If the matrix of pairwise comparisons has zero elements, and decisions are made in conditions of uncertainty, then to obtain the weighting factors we can use the mechanism of the four-level model of planning and decision-making, which was proposed in [8].

Searching the values of vector I, which give the global optimum, or more precisely, the maximum of F_{ac} function is now the single-criteria optimization problem. There are many ways to solve it. One of the most famous approaches is the genetic algorithm [9]—a stochastic method for finding the required solution. The ideas of natural selection and genetics provide a fast search for the global optimum, thus solving the NP-complete problem. The algorithm is shown schematically in Fig. 1.

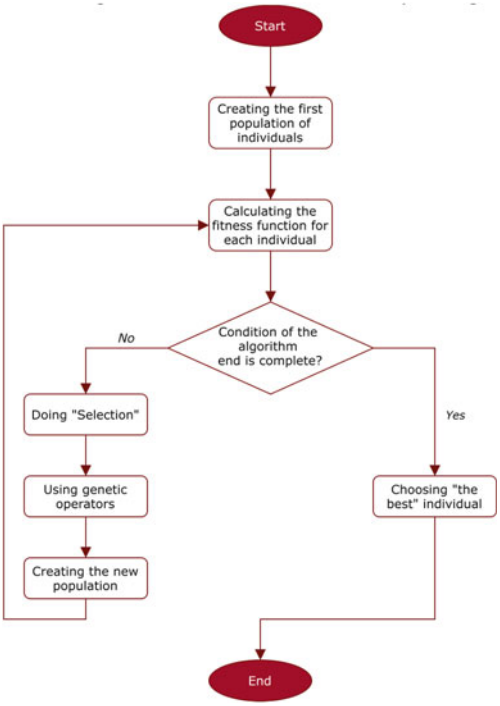


Fig. 1 The scheme of genetic algorithm

To describe the algorithm in more detail:

1. A random sample ("population") of N arrays ("individuals") that contain values of vector I is created.
2. "Fitness function" for each individual is calculated. In the context of current work, the fitness function is F_{ac} function (supercriterion).
3. After obtaining the values of F_{ac} function for each individual, the condition of the algorithm end is checked (it can be the presence of the preassigned value of F_{ac} function or exceeding the time limit of the algorithm).

   3.1 If the condition is complete, the algorithm returns the "best" individual (optimal treatment strategy).

   3.2 If the condition is not complete, the formation of a new population begins.

   3.2.1 The "selection" of individuals of the current population is carried out. The idea of this procedure is to select individuals for the creation of the next generation, and the chance of selecting each individual directly depends on the value of his fitness function. The selected individuals form N pairs, which will then give "new" individuals.

   3.2.2 The "crossover" (one of the genetic operators) for crossing the resulting pairs of individuals is used. A mixing of "genes" (parameters of vector I) occurs between a pair of individuals, thus forming a new individual that stores the information of his "ancestors". In general case, randomly the "crossover point" is determined, which allows mixing a pair of "parents": the genes of the first parent are before the crossover point, and after it—the genes of the second parent. That creates a new individual. To ensure the diversity of the population during the entire algorithm's operation, another genetic operator called "mutation" is also used. This operator can be triggered with low probability instead of crossover, and its main purpose is to replace randomly selected genes of individuals with completely new ones.

3.2.3 Back to point number 2.

As a result, the genetic algorithm has been obtained, where the additive convolution function F_{ac} is used as a fitness function, which is obtained from the ideology of the AHP. The idea of using the convolution function (obtained by the method of multi-criteria decision making) as a fitness function of the genetic algorithm is not new. In [10] the authors used the weight-sum approach and the Tchebycheff approach to get the convolution function. Authors of [11] were comparing non-dominated sorting genetic algorithm II (NSGA-II), multi-objective differential evolution (MODE), and multi-objective particle swarm optimization (MOPSO) algorithms. Nevertheless, such approaches are rather difficult to interpret, which makes it more difficult to explain to the doctor the principle of the algorithm for finding treatment strategies. That is why it was decided to choose AHP to obtain a convolution function.

4 Using the Proposed Approach in Practical Task

This section shows the application of the created algorithm to solve the next problem. We had a clinical database, which describes the treatment of patients with congenital heart defects. It contains 144 patients from 1 to 18 years old. They underwent surgical operations, after which they were prescribed medication (conservative treatment) to eliminate postoperative complications. In total, the database has the next 29 features: 10 *patient's indicators before conservative treatment* (there are both quantitative and qualitative features); 10 *types of drugs* that were used to treat the patient; 9 *patient's indicators after conservative treatment*, which are presented in the form of binary features that show the presence of certain postoperative complications.

To be able to solve the problem of searching personalized treatment strategies for a specifically selected base, it is necessary to obtain models of patient indicators after treatment by using the patient indicators before treatment and drugs as predictors. Since the indicators are binary (0 means the absence of postoperative complication, 1 means the opposite), the best option for obtaining models is to use the binary

Table 2 Evaluation of resulting models

Model	Training sample (85%)			Testing sample (15%)		
	Positive predictive value	Negative predictive value	Accuracy (%)	Positive predictive value	Negative predictive value	Accuracy (%)
y_1	1	1	100	1	1	100
y_2	1	1	100	1	1	100
y_3	1	1	100	1	1	100
y_4	1	1	100	1	1	100
y_5	1	1	100	1	1	100
y_6	1	0.992	99.7	1	0.913	96.9
y_7	1	1	100	1	1	100
y_8	1	1	100	1	1	100
y_9	1	1	100	1	1	100

classification algorithm. For today it is considered that the most optimal classification algorithm is the random forest classifier [12–14], which is one of the best choices for obtaining highly accurate prediction models. As a result, we got 9 models of the patient's indicators after treatment by splitting the general data sample into training (85%) and testing (15%) samples. Their evaluation, namely accuracy of predictions, positive (proportion of correctly predicted first class object) and negative (proportion of correctly predicted second class object) predictive values are represented in Table 2.

It should be noted that $y_1, \ldots, y_9$ are according to the importance of patient's indicators after treatment.

After obtaining models, the pairwise comparison of $y_1, \ldots, y_9$ was performed by using Table 1, which made it possible to obtain the additive convolution function F_{ac} (2) in the following form:

$$F_{ac} = 0.353 \cdot y_1 + 0.177 \cdot y_2 + 0.118 \cdot y_3 + 0.088 \cdot y_4 + 0.071 \cdot y_5 + \\ +0.059 \cdot y_6 + 0.05 \cdot y_7 + 0.044 \cdot y_8 + 0.039 \cdot y_9 \tag{3}$$

In this equation, we have " + " signs because it is necessary to maximize the value of each model which is the probability of the first (required) class given by predictors.

Using the obtained models and function F_{ac} (3) we can start the personalized treatment strategy search by launching the genetic algorithm (Fig. 1). To do this, it is enough to substitute the values of the indicators before the treatment of any patient into the resulting models, after which the algorithm will do the job.

The result of using the genetic algorithm is shown in Fig. 2.

It demonstrates a clear difference between the strategy that was adopted by doctors in real life and the strategy proposed by the algorithm. The adoption of the computer

Treatment strategy

	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
In real life	3	3	2	0	Not use	Not use	Not use	Use	Not use	Use
Genetic algorithm	5.2	1.7	1.7	0.54	Use	Use	Use	Use	Use	Not use

Patient's postoperative complications

	y1	y2	y3	y4	y5	y6	y7	y8	y9
In real life	Absence	Absence	Absence	Absence	Absence	Absence	Presence	Absence	Absence
Genetic algorithm	Absence	Presence	Presence	Presence	Absence	Absence	Presence	Presence	Presence

Fig. 2 Result of using the genetic algorithm and comparing with the real-life result

treatment strategy made it possible to hypothetically improve the quality value of the obtained result. It means that if doctors used the strategy suggested by the computer, then in theory the patient's final state would be better. However, it is worth considering that the algorithm of finding treatment strategies is only an auxiliary tool for the doctor in running through the options, which allows a comprehensive assessment of them.

That result was received in 1045 iterations (roughly 5 min) of the genetic algorithm. This is demonstrated in Fig. 3.

The fitness function (3) has a value between 0 (presence of all complications) and 1 (absence of all complications). The mission of the genetic algorithm to get a treatment strategy that gives fitness function value as big as possible. The ceiling value of the fitness function depends on the initial state of the patient. In this case (Fig. 3) it is 0.588 (only 3 complications out of 9). In total, the genetic algorithm worked for 1 h.

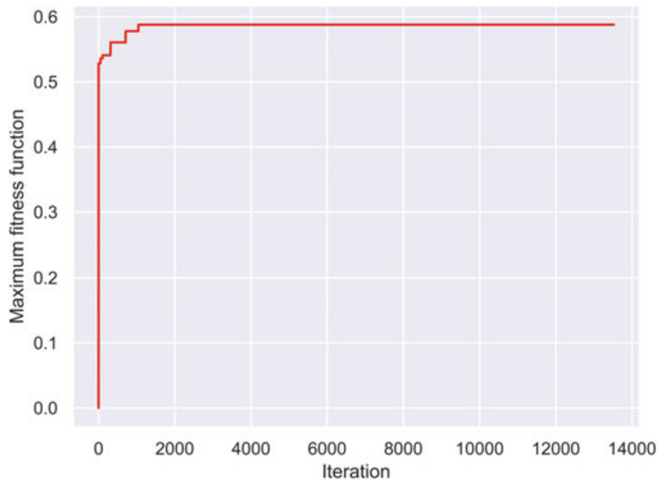


Fig. 3 Changing of fitness function maximum in process of genetic algorithm

5 Conclusion

The study describes the development of an algorithm for personalized treatment strategy searching by using the principles of genetic algorithm (for quick searching) and analytic hierarchy process (for criteria simultaneous optimization). The algorithm work stages and testing results were presented. Also, nine models of patient indicators after treatment were obtained by random forest classifier on training (85%) and testing data samples (15%). Models were evaluated by values of accuracy, positive predictive value, and negative predictive value. These results can be used as the beginning of creating an informative system with the functionality of a clinical pharmacist for personalized treatment strategy searching. It should be added that it is necessary to carefully elaborate on all stages of the system to provide effective support for the doctor in deciding on a treatment strategy. In this way, the system will be able to fill the absence of clinical pharmacists and optimize the work of medical institutions.

References

1. Chen Y-F, Neil KE, Avery AJ, Dewey ME, Johnson C (2005) Prescribing errors and other problems reported by community pharmacists. Ther Clin Risk Manag 1(4):333–342
2. Bucchi C, Valdivia-Gandur I, Sánchez-Bizjak R, Tallón-Walton V, Manzanares-Céspedes C (2017) Regenerative endodontic therapy: a systematic review of clinical protocols. Int J Clin Exp Med 10:2006–2015
3. Khan FU, Waqas N, Ihsan AU, Khongorzul P, Wazir J, Gang W, Mengqi Y, Xiaoqian L, Han L, Xiaohui Z (2019) Analysis of the qualities matching new classification of clinical pharmacist. Indian J Pharm Sci 81(1):2–10. https://doi.org/10.4172/pharmaceutical-sciences.1000473
4. Nastenko I, Pavlov V, Nosovets O, Zelensky K, Davidko O, Pavlov O (2019) Optimal complexity models in individual control strategy task for objects that cannot be re-trialed. In: Proceedings of IEEE 2019 14th international scientific and technical conference on computer sciences and information technologies, CSIT 2019—Proceedings, pp 207–210 https://doi.org/10.1109/STC-CSIT.2019.8929831
5. Nastenko I, Pavlov V, Nosovets O, Zelensky K, Davidko O, Pavlov O (2020) Solving the individual control strategy tasks using the optimal complexity models built on the class of similar objects. In: Proceedings of Advances in intelligent systems and computing. Springer, pp 535–546 https://doi.org/10.1007/978-3-030-33695-0_36
6. Panjwani S, Kumar SN, Ahuja L (2019) Multi-criteria decision making and its applications. Int J Innov Technol Explor Eng 8(9 Special Issue 4). https://doi.org/10.35940/ijitee.I1122.0789S419
7. Saaty TL (1990) Decision Making for Leaders: The analytic hierarchy process for decisions in a complex world. RWS Publications. https://doi.org/10.1016/0377-2217(89)90066-0
8. Zgurovsky MZ, Pavlov AA (2019) The four-level model of planning and decision making. In: Studies in Systems, Decision and Control, pp 347–406. https://doi.org/10.1007/978-3-319-98977-8_8
9. García JM, Acosta CA, Mesa MJ (2020) Genetic algorithms for mathematical optimization. In: Proceedings of Journal of Physics: Conference Series, p 5. https://doi.org/10.1088/1742-6596/1448/1/012020
10. Zhang Q, Li H (2007) MOEA/D: a multiobjective evolutionary algorithm based on decomposition. IEEE Trans Evol Comput 11(6):712–731. https://doi.org/10.1109/TEVC.2007.892759

11. Monsef H, Naghashzadegan M, Jamali A, Farmani R (2019) Comparison of evolutionary multi objective optimization algorithms in optimum design of water distribution network. Ain Shams Eng J 10(1):103–111. https://doi.org/10.1016/j.asej.2018.04.003
12. Biau G (2012) Analysis of a random forests model. J Mach Learn Res 13:1063–1095
13. Cheng L, Li L, Wang L, Li X, Xing H, Zhou J (2018) A random forest classifier predicts recurrence risk in patients with ovarian cancer. Mol Med Rep 18:3289–3297. https://doi.org/10.3892/mmr.2018.9300
14. Utkin LV (2020) An imprecise deep forest for classification. Expert Syst Appl 141:31. https://doi.org/10.1016/j.eswa.2019.112978

Clinical Text Classification of Alzheimer's Drugs' Mechanism of Action

Mina Esmail Zadeh Nojoo Kambar, Pouyan Nahed, Jorge Ramón Fonseca Cacho, Garam Lee, Jeffrey Cummings, and Kazem Taghva

Abstract The Alzheimer's disease Drug Development Pipeline [1, 2] delivers updates on potential AD-treatment, as well as drug development ongoing in clinical trials. To create these reports, researchers manually extract information from several resources like ClinicalTrials.gov and drug manufacturer websites; however, some of these items require expert review, such as when predicting a drug's Mechanism of Action (MOA). In this paper, we aim to assist researchers by predicting and suggesting a drug's MOA using Machine Learning. We test Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), XGBoost, and Decision Tree (DT) models. The latter showing the most promising results, with 95% accuracy, 100% recall, and a 0.92 F1-score.

Keywords Alzheimer's drug · Clinical text classification · Machine learning · Mechanism of Action

M. E. Z. N. Kambar · P. Nahed · J. R. F. Cacho (✉) · K. Taghva
Department of Computer Science, University of Nevada Las Vegas, Las Vegas, NV89154, USA
e-mail: Jorge.FonsecaCacho@unlv.edu

M. E. Z. N. Kambar
e-mail: Esmailza@unlv.nevada.edu

P. Nahed
e-mail: Nahed@unlv.nevada.edu

K. Taghva
e-mail: Kazem.Taghva@unlv.edu

G. Lee · J. Cummings
Department of Brain Health, University of Nevada Las Vegas, Las Vegas, NV89154, USA
e-mail: Cumminj@ccf.org

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_48

1 Introduction

The introduction of online medical databases, reports, and general use websites that store and display data from various diseases and their associated treatments has resulted in chaotic (unstructured, unlabeled, and small data), repository of medical information. Sometimes, even if data is properly organized, it may not be stored in one location, due to multiple clinical trials run in different organizations, or it is not easily available, due to sensitive data. While a lot of work has been done to organize some of this data, the sheer amount makes this feat very difficult, and has resulted in experts having to apply a processing, classification, and extraction pipeline to acquire information relevant to their research (e.g., adverse drug reaction [3] and cancer stage detection [4, 5]).

Attempts to solve this problem include Clinicaltrials.gov, "a Web-based resource that provides patients, their family members, health care professionals, researchers, and the public with easy access to information on publicly and privately supported clinical studies on a wide range of diseases and conditions." See Reference [6]; however, trying to browse through such a website to investigate trials related to a specific subject can be challenging and requires investigators to read through free text content if read directly form the website. Furthermore, when developing reports on treatment and drug development for specific diseases, researchers usually have to acquire additional information from other resources. For example, for the Alzheimer's Disease (AD) Drug Development Pipeline report [1, 2], which delivers updates on AD-treatment and drug development, the information found on ClinicalTrials.gov proves useful for cataloging information such as sponsors, disease phases, agents' criteria, and key trial dates, but does not always provide other information such as a drug's Mechanism of Action (MOA) or therapeutic purpose. Therefore, researchers must read free-text content from multiple resources, such as drug companies' websites, to make up for incomplete data. Even then, researchers need to use their expertise to make decisions on how to label some of these columns like MOA or therapeutic purpose of a drug.

In our research, we attempt to automate AD-Drug text analysis and classification for information that is not found in ClinicalTrials.gov. Specifically, we focus on the MOA of a drug, an item that is ever changing, depending on experimental data found beyond ClinicalTrials.gov and Drug Manufacturer websites, among other resources and datasets. In order to find the best AD-Drug MOA classifier, we explored different classification machine learning algorithms such as Random Forest(RF), XGBoost, Logistic Regression (LR), Support Vector Machines (SVM), and Decision Tree (DT). Preliminary results on the AD-Drugs' free-text illustrate that DT has the best performance among all tested algorithms, but there is room for improvement. The main contributions of our study can be summarized as

1. Different information is available related to AD-drugs in various resources. Therefore, we attempt to collect data and merge them to make a contiguous dataset.
2. In our research, different machine learning methods are evaluated to find the best model for classifying the AD-drugs' texts.

3. Our result shows Decision Tree is the best model for the classification task. This model can play an essential role in the medical group making an annual report on the AD-drug progression [7].

2 Background

Researchers have conducted different methods, using Machine Learning (ML), Deep Learning (DL), and knowledge-based systems in the clinical text domain [8]. For instance, statistical ML methods, such as Naive Bayes (NB), Support Vector Machine (SVM), Decision Trees, and AdaBoost, have been developed to detect and classify death caused by cancer from overall free-text certificates [9]. Mujtaba et al. [10] elaborated on various engineering features, using six decision models, SVM, NB, K Nearest Neighbors (KNN), J48 DT, Random Forest (RF), and Voted for forensic autopsy free-text to predict the reasons for death class.

Developing rule-based systems is complicated, tedious, and requires specialists; however, several applications employing them have been built. Classifications using the Unified Medical Language System (UMLS) [11]and trigger phrases feed into a Convolutional Neural Network (CNN) approach [12]. Similarly, a system [13] designed a binary rule-based classifier to categorize patients who had influenza like illness (ILI) or not. Wang et al. [14]proposed a paradigm for the weakly labeled data problem. They designed a rule-based algorithm that automatically labels the training data of clinical texts and represents deep pre-trained word embeddings. To show the algorithm's performance in classification, they tested the algorithm with four different ML models, such as SVM, RF, Multilayer Perceptron Networks (MLPNN), and CNN.

Various DL models are more convenient to develop because features are not selected automatically. However, they need masses of labeled data to optimize their performances. In our project, the drug dataset is not large enough, so we skip applying DL models in our approach.

3 Methods

We applied five different machine learning models, including RF, XGBoost, LR, SVM, and DT, to classify the AD-drugs' MOA. This section will explain the dataset, models, and discuss the experimental results achieved by the methods.

3.1 Dataset

We collected the background and finding texts of each drug from ALZFORUM (http://www.alzforum.org.) Therapeutics dataset. By using the drugs' targets from the website and the MOA classes of medications provided by [7, 15], our team obtained 233 labeled records of data. The AD-drugs' MOA is classified into two main categories, called small molecules and Disease-Modifying Therapies (DMT) Biologics. We used 75% (186 records) of the data as the training set and 25% (47 records) of the data as the test set. As one can see, the dataset is imbalanced for each of the classes, with 175 records of the first class and 59 records of the second class. Thus, we applied the Synthetic Minority Over-sampling Technique (SMOTE) [16] to the training set to make 141 records of each class. The overall text part of the dataset has 124,180 words and 7,693 sentences.

3.2 Data Preparation

First, we perform data preprocessing by removing stop words, punctuation, extra white spaces, undesired characters, and words with no meaningful information. Next, we used the Term Frequency-Inverse Document Frequency (TF-IDF) to produce the words' matrix representation text. The approach utilizes two factors: TF—term frequency of term i in the text j, and IDF—inverse text frequency of term i. In our research and testing, five ML algorithms using this method showed promising results. TF-IDF calculated as [17]

$$tf_{ij}idf_i = tf_{ij} \times log_2\left(\frac{N}{df_i}\right) , \tag{1}$$

where N is the number of text records in the dataset, tf_{ij} is the term frequency of word i in the text j, and df_i is the text frequency of the word i in the dataset.

3.3 Selecting Machine Learning Algorithms

In order to achieve the best result, we experimented with several ML algorithms to determine which one performs better. The following algorithms have been chosen:

1. **Decision Tree**: DT is a reliable and practical approach in biomedical tasks that can reach a high accuracy, especially when the represented data is simple [18]. In our study case, DT can choose important words and classify the document, respectively.

2. **XGBoost**: It is a Decision Tree-based algorithm that uses Gradient Boosting learning [19], which is recognized for its high speed and good performance.
3. **Random Forest**: RF consist of several different DT. An individual DT predicts a label, and the most predicted label is considered as the final prediction for that label. One of the features of RF is that it corrects the overfitting problem of [20].
4. **Logistic Regression**: LR is a Linear Classifier in which an event occurrence is modeled as a linear function of a set of predictor variables [21]. The outcome of LR is a value between 0 and 1, which shows the predicted probability.
5. **Support Vector Machine**: SVM provides reliable performance in text data classification and can linearly separate data in higher dimensions by using appropriate kernel functions and is resistant to overfitting [22]. In this case, we use the RBF kernel.

3.4 Results

We utilized the dataset discussed earlier with 141 and 47 records as the training/testing sets for experimental purposes. We used library packages provided by Python such as Gensim [23] to clean the clinical text, along with Scikit-learn [24] and Pandas [25] to implement our models.

The results across models are shown in Table 1. We measured Precision, Recall, F1-Score, Accuracy, and ROC Area for all five models. Figure 1 provides a visual comparison between the performance of the different algorithms on the ROC area metric. In our results, although DT reaches a better accuracy than SVM and Logistic Regression, these two models still perform much better in precision, which means they are predicting the classes with more caution. These two models are more precise because their results are closer than the predicted classes by DT in different case classification. Using the term TF-IDF produces better performance and illustrates that the models are sensitive to the presence or absence of the words instead of their meaning in the context. Figure 2 illustrates the right DT procedure on the dataset. We measured the Gini Index as the cost criteria to split the DT tree. This index calculates the overall variance over the tree class, and it is the appropriate criterion of node

Table 1 MOA classification results

Algorithms	Precision	Recall	F1-score	Accuracy	ROC area
Random forest	0.857	0.46	0.6	0.82	0.95
XGBoost	0.76	0.76	0.76	0.87	0.96
Logistic regression	0.9	0.62	0.78	0.89	0.96
SVM	0.9	0.69	0.78	0.89	0.97
Decision Tree	0.86	1.00	0.92	0.95	0.97

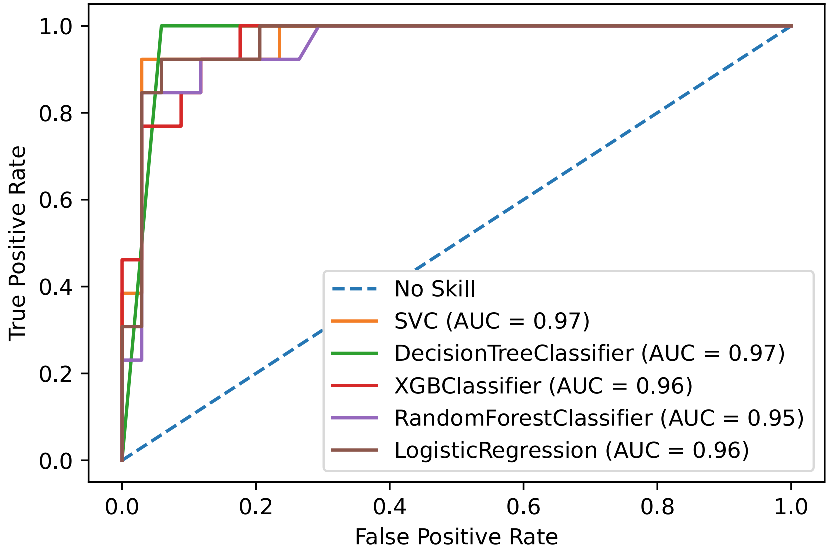


Fig. 1 ROC curve comparison between algorithms

pureness [26].

$$G = \sum_{k=1}^{K} q_{mk}(1 - q_{mk}), \tag{2}$$

where q_{mk} is the part of training set in the mth region from the kth class. As you can see the trained DT in Fig. 2, the height of the tree is 4 and the word "antibody" is chosen by the algorithm to separate the classes in the first level. While using TF-IDF as feature extraction, when the score of "antibody" is more than 0.017, the DT was able to distinguish 34 out of 40 total samples of the second class. The other words selected by the DT are "delivered", "mild", "mimics", and "intracerebral".

4 Conclusion

The initial results with the multiple machine learning models indicate that Decision Tree are the best choice for predicting the MOA; however, XGBoost or SVM may match Decision Tree with the correct hyperparameters, such as trying different Kernel functions with SVM. We intend to continue exploring these results. We also understand that the dataset is limited in size, making it difficult to evaluate how well this would generalize. We intend to explore this with the next pipeline [7] release. In the future, we intend to apply ensemble learning or other ways of combining models

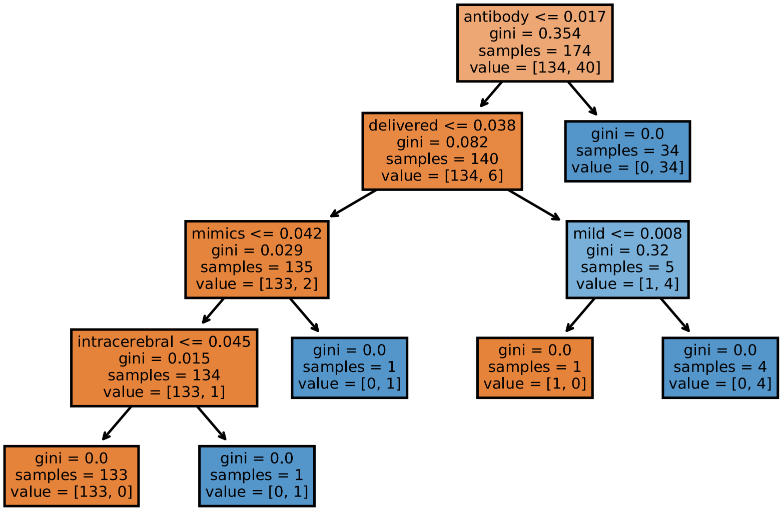


Fig. 2 Decision tree for AD-drugs' dataset

to improve results, along with attempting other models. This project is a subtask of a larger project in automating information retrieval from clinicaltrials.gov to generate potential candidates that can then be verified by an expert such as the MOA for a drug. We intend to use the created models to predict other categories usually done by an expert. Our goal is not to replace that work, but to complement it by suggesting the ideal candidates, similar to a spell checker. Should our work suggest links that may have otherwise been missed or save experts' time when suggesting an MOA for drugs, then we will consider ourselves satisfied as they can now use that time to get us closer to finding new treatments, or even a cure.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. 1625677.

References

1. Cummings JL, Morstorf T, Zhong K (2014) Cummings, Jeffrey L_Alzheimer's_drug development candidates failures_2014, pp 1–7
2. Cummings J, Lee G, Ritter A, Sabbagh M, Zhong K (2020) Alzheimer's disease drug development pipeline: 2020. Alzheimer Dement Transl Res Clin Interv 6(1):e12050

3. Sarker A, Gonzalez G (2015) Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 53:196–207. http://dx.doi.org/10.1016/j.jbi.2014.11.002
4. AAlAbdulsalam AK, Garvin JH, Redd A, Carter ME, Sweeny C, Meystre SM (2018) Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry. In: AMIA Joint summits on translational science proceedings. AMIA Joint Summits on Translational Science, vol 2017, pp 16–25
5. Chen VW, Ruiz BA, Hsieh MC, Wu XC, Ries LA, Lewis DR (2014) Analysis of stage and clinical/prognostic factors for lung cancer from SEER registries: AJCC staging and collaborative stage data collection system. Cancer 120(S23):3781–3792
6. Clinicaltrials.gov (2020) Information on clinical trials and human research studies. https://clinicaltrials.gov. Accessed 29 Oct 2020
7. Cummings J, Lee G, Ritter A, Sabbagh M, Zhong K (2019) Alzheimer's disease drug development pipeline: 2019. Alzheimer Dement Transl Res Clin Interv 5:272–293. https://doi.org/10.1016/j.trci.2019.05.008
8. Bozorgi M (2018) Application of machine learning in cancer research
9. Butt L, Zuccon G, Nguyen A, Bergheim A, Grayson N, Butt L (2013) Classification of cancer-related death certificates using machine learning what this study adds, pp 292–299
10. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K (2018) Prediction of cause of death from forensic autopsy reports using text classification techniques: a comparative study. J Forensic Legal Med 57:41–50.https://doi.org/10.1016/j.jflm.2017.07.001
11. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32(DATABASE):267–270
12. Yao L, Mao C, Luo Y (2019) Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Mak 19(3)
13. MacRae J, Love T, Baker MG, Dowell A, Carnachan M, Stubbe M, McBain L (2015) Identifying influenza-like illness presentation from unstructured general practice clinical narrative using a text classifier rule-based expert system versus a clinical expert. BMC Med Inform Decis Mak 15(1):1–11
14. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, Amin S, Liu H (2019) A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak 19(1):1
15. Cummings J, Lee G, Ritter A, Zhong K (2018) Alzheimer's disease drug development pipeline: 2018. Alzheimer Dement Transl Res Clin Interv 4(2018):195–214. https://doi.org/10.1016/j.trci.2018.03.009
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
17. Trstenjak B, Mikac S, Donko D (2014) KNN with TF-IDF based framework for text categorization. Proc Eng 69:1356–1364
18. Podgorelec V, Kokol P, Stiglic B, Rozman I (2002) Decision trees: an overview and their use in medicine. J Med Syst 26(5):445–463
19. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining, pp 785–794
20. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
21. Hand DJ, Adams NM (2014) Data Mining. Wiley StatsRef: Statistics Reference Online, pp 1–7
22. Fonseca Cacho JR (2019) Improving OCR post processing with machine learning tools
23. Rehurek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. ELRA, Valletta, Malta, May 2010, pp 45–50

24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
25. McKinney W, Others (2010) Data structures for statistical computing in python. In: Proceedings of the 9th Python in science conference, vol 445. Austin, TX, pp 51–56
26. Baboota R, Kaur H (2019) Predictive analysis and modelling football results using machine learning approach for english premier league. Int J Forecast 35(2):741–755

Construction of Some Good Binary Linear Codes Using Hadamard Matrix and BCH Codes

Driss Khebbou, Reda Benkhouya, and Idriss Chana

Abstract This paper presents a new method of constructing good binary linear codes. The proposed method is first based on the BCH codes parity matrix and Hadamard's matrix, and after on a GUAVA-based computer search, which allowed us to find several good codes with rates >0.5.

Keywords Code construction · Binary linear codes · BCH codes · Error correcting codes · Coding theory · Minimum distance

1 Introduction

After Shannon's work in 1948 [1], the Hamming codes [2] and Golay codes [3] became the foundation of the theory of error-correcting codes since 1974. Using these error-correcting codes, in the storage system and telecommunication, improves the reliability of information transmitted through communication channels, mainly when the information is corrupted due to channel noise. The concept for detecting and correcting errors is based on adding redundancy information to the original information.

The information source is segmented into message blocks of the length k. The processing of each block through the encoder generates additional bits (redundancy information), to provide a codeword of length n. If the set of 2^k codewords forms a subspace with dimension k of the vector space with dimension n over the field $GF(q)$ where q is prime, the code $C(n, k)$ is called a linear code.

According to the way the redundancy information is added to the information message block, two types of codes exist: block codes and convolutional codes. In this work, we are interested in designing some good block codes, in particular linear codes.

D. Khebbou (✉) · I. Chana
Moulay Ismail University, Meknes, Morocco

R. Benkhouya
Ibn Tofail University, Kenitra, Morocco

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_49

In the binary case and as described above, while $C(n, k)$ is a vector subspace of all n-tuples over the field $GF(2)$, it is possible to discover k linearly independent codewords of size n from the subspace, in order that each codeword from $C(n, k)$ is a linear combination of those k codewords. To be more specific, if $c_0, c_1, ..., c_{k-1}$ are linearly independent codewords from 2^k existing codewords in $C(n, k)$, and $u = (u_0, u_1, ..., u_{k-1}) \in \{0, 1\}^k$ is a message block of k bits, then the codeword v related to message block u is defined as follows:

$$v = u_0.c_0 + u_1.c_1 + \cdots + u_{k-1}.c_{k-1}$$

where $c_0, c_1, ..., c_{k-1}$ can be put in a matrix form, called generator matrix of the code C as follows:

$$G = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{k-1} \end{bmatrix} = \begin{bmatrix} c_{0,0} & c_{0,1} & \cdots & c_{0,n-1} \\ c_{1,0} & c_{1,1} & \cdots & c_{1,n-1} \\ \vdots & \vdots & \cdots & \vdots \\ c_{k-1,0} & c_{k-1,1} & \cdots & c_{k-1,n-1} \end{bmatrix}$$

$$\boldsymbol{v} = \boldsymbol{u}.\boldsymbol{G}$$

In addition to the length and dimension of linear code $C(n, k)$, an additional important parameter exists. It is the minimum hamming distance d which will be detailed in the next section. It is used for calculating the distance between two codewords (number of bits in which two codewords differ). In this context, one of the main problems of the coding theory is to find the best values of parameters n, k and d compared to the existing results.

A code is named “good” if its parameters n, k and d allow reaching the existing results, and it is “record-breaking” if its parameters allow getting better results than the existing ones and it is “optimal” if it reaches the theoretical bound [4]. However, searching for error-correcting codes is still difficult.

The remainder of this work is structured as follows. In the next section, we describe BCH codes, Hadamard matrices and minimum hamming distance for a linear code. In Sect. 3, we illustrate a method for finding some good binary linear codes from the BCH codes and Hadamard matrices. In Sect. 4, we present the results of the good linear codes found using the proposed method. Finally, a conclusion and future topics for reflection are given in the last section.

2 Background Knowledge

2.1 BCH Codes

Explored by D.K. Ray-Chaudhuri, R.C. Bose, and individually by A. Hocquenghem, BCH codes are of considerable importance for error correction [5–7], particularly if the number of errors planned is low, compared to the length. These codes are linear and of fixed-length types such as there is a BCH code for any positive integer (m > 3 and t < 2^{m-1}), whose parameters are:

$$n = 2^m - 1, \quad n - k \le mt, \quad d_{\min} \ge 2t + 1$$

whereas t is the number of errors that can be corrected in a codeword of length n. This form of BCH code is known as primitive BCH codes.

More specifically, as demonstrated in 1960 by Peterson [8], the BCH code are cyclic codes so that if $v = v_1 v_2 v_3 \ldots v_n$ is a codeword, then every cyclic shift of v will also be a codeword (i.e., $v_2 v_3 v_4 \ldots v_n v_1$ is a codeword).

Since these codes are linear, they can be outlined by a generator matrix defined by k rows and n columns, thus any codewords equate to the linear vector combinations in the generator matrix [9]. This means that any codeword $v \in BCH(n, k, d)$ can be defined by the coefficients of the basis vectors b_i for i 1,2,3, …, k, and an information vector $m = \{m_i\}$ for $i = 1,2,3, \ldots, k$ and $m_i = \{0,1\}$ represents a codeword $\sum \{m_i * b_i\}$ where addition and multiplication are modulo 2 in the binary case. These codes are able to extend to non-binary BCH codes [10].

2.2 Hadamard Matrices

As defined by the mathematician M.J.Hadamard in 1893 [11], the Hadamard matrix is a $n*n$ matrix with entries $\{+1, -1\}$, whose rows and columns are mutually orthogonal. For this, Hadamard matrices are used in the construction of error-correcting codes. A matrix (+1, -1) of size $n*n$ is a Hadamard matrix of order n if

$$H^T H = H H^T = nI$$

(This implies H is invertible, $H^{-1} = \frac{1}{n} H^T$ and the rows and columns are mutually orthogonal).

To express it using the binary elements $\{0, 1\}$, we can replace "$+1$" with "0" and "-1" with "1". Any exchange of rows or columns transforms the matrix into another Hadamard matrix [12]. Hadamard matrix of order n is generated by the following method:

$$H_1 = [1] \equiv [0]$$

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \equiv \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$H_n = H_2 \otimes H_{n/2}$$

wher\e $\otimes$ signifies the direct product or the Kronecker product, also known as the tensor product of two matrices M and $M^{'}$ defined by:

$$M \otimes M' = \begin{pmatrix} m_{11}M' & m_{11}M' & \cdots & m_{11}M' \\ m_{21}M' & m_{22}M' & \cdots & m_{2n}M' \\ & \vdots & \ddots & \vdots \\ m_{k1}M' & m_{k1}M' & \cdots & m_{kn}M' \end{pmatrix}$$

2.3 Minimum Distance of Linear Codes

In order to create an error-correcting code, it is necessary to ensure that two valid codewords are considerably distinct so that a single-bit corruption (or possibly a small number of bits) does not transform one valid codeword into another. This means that the hamming distance (the third critical error-correcting code parameter) equals the number of single-bit errors to transform one of them to the other.

Let x be an element over $GF(2)$, the weight (hamming weight W_h) of this element is described as follows:

$$W_h(x) = \begin{cases} 0 \; if \;\; x = 0 \\ 1 \; if \;\; x \neq 0 \end{cases}$$

The weight of a vector element $X = (x_1, x_2, \ldots, x_n) \in \mathrm{GF}(2)^n$ is defined as:

$$W_h(X) = \sum_{i=1}^{n} W_h(x_i)$$

The minimum weight (or the minimum hamming weight) of an error-correcting code C over GF(q) refers to the smallest positive integer as the weight of a non-zero codeword. Otherwise, it can be put as the following expression:

$$\min\{W_h(X) | X \in C \; and \; X \neq 0\}$$

If the code is linear, the minimum hamming distance (or the minimum distance) d is the minimum weight of this code [10]:

$$d_{\min} = \min_{X \in C,\, X \neq 0} W_h(X)$$

3 The Proposed Method to Search Good Linear Codes

In this section, we describe the method to construct some good binary linear codes from the Hadamard matrix and the redundant part of the generator matrix of some BCH codes in the systematic form. The idea is based on the extraction of linear codes from the resulting Kronecker product matrix between the redundant part of the generator matrix of the chosen BCH code on one hand, and the Hadamard matrix on the other.

Let G be a $k{*}n$ generator matrix in the systematic form (elementary operations on the rows and/or columns of the generator matrix) of the chosen $BCH(n, k, d)$ code as follows:

$$G = [I_k|P] = \begin{pmatrix} 1 & 0 & \cdots & 0 & p_{0,0} & \cdots & p_{0,n-k-1} \\ 0 & 1 & \cdots & 0 & p_{1,0} & \cdots & p_{1,n-k-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & p_{n,0} & \cdots & p_{n,n-k-1} \end{pmatrix}$$

where I_k is an identity matrix of size $k * k$ and P is a redundant matrix of size $k * (n - k)$. In addition, the Hadamard matrix is

$$H_2 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \; or \; H_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

The result of the Kronecker product between the redundant part P and H_2 (or H_4) is a matrix whose dimensions are $2k * 2(n - k)$ (or $4k * (n - k)$).

Let A be the result matrix of the Kronecker product. The next step is to try to extract linear codes C(n', k', d') of length $n' = 2(n - k)$ and different values of dimension k' from A. Since the rows of the generator matrix of a linear code are codewords and the minimum distance of any linear code is the minimal weight of codewords belonging to the code. So, it is necessary to remove the rows from the matrix A, whose weight is less than the lower bound given in Brouwer's database [13] according to the parameters n' and k'. The lower bound is the best established minimum distance preceding found in all existing works [4].

For example, taking $n' = 12$ and $k' = 6$. The lower bound of these parameters (from Brouwer's database) is 4, so for search improvement, it is important to eliminate all rows whose weight is less than 4 in order to extract $A_{filtred}$ from A.

Finally, the program tries to construct linear codes in the form of a generator matrix, by combining k' rows from the filtered matrix $A_{filtred}$. It is trivial to make sure that the k' rows of the generator matrix of a code are linearly independent and test if the minimum distance is greater than or equal to the lower bound.

As mentioned before, the problem of finding the lowest weight codewords is an NP-hard problem and it is similar to the minimum distance research problem for linear codes [14, 15]. It is clear that this approach becomes infeasible as k grows larger. In this work, we have calculated the minimum distance using Zimmermann's algorithm [16], which is implemented in GAP through the GUAVA package [17]. The said package includes several functions that allow computations related to the theory of error-correcting codes to be carried out, such as the generation of codes; the construction of a code from others; decoding/error correction; and computation of essential code data.

Since it is equal to the minimum weight for a linear code, the minimum distance was determined using the GUAVA minimum weight function, based on an external program developed in C, which is based on Chen's algorithm for cyclic code and Zimmermann's algorithm for linear code [17]. As a result, the minimum weight function execution time is faster than the minimum distance function.

For example, we will try to construct a good (n', k', d') linear code from the Hadamard matrix H_2 and a generator matrix of BCH code with length $n = 7$ and dimension $k = 4$. The matrix of the $BCH(7, 4, 3)$ code is taken from the website of the art of error-correcting code [18]. The systematic form of the chosen BCH code is defined as follows:

$$G = [I_4|P] = \begin{pmatrix} 1\,0\,0\,0 \mid 1\,1\,0 \\ 0\,1\,0\,0 \mid 0\,1\,1 \\ 0\,0\,1\,0 \mid 1\,1\,1 \\ 0\,0\,0\,1 \mid 1\,0\,1 \end{pmatrix}$$

Then we perform the Kronecker product between redundant matrix P and Hadamard matrix H_2 as follows:

$$A = P \otimes H_2 = \begin{pmatrix} 1\,1\,0 \\ 0\,1\,1 \\ 1\,1\,1 \\ 1\,0\,1 \end{pmatrix} \otimes \begin{pmatrix} 1\,1 \\ 1\,0 \end{pmatrix}$$

$$A = \begin{pmatrix} 1\,1\,1\,1\,0\,0 \\ 1\,0\,1\,0\,0\,0 \\ 0\,0\,1\,1\,1\,1 \\ 0\,0\,1\,0\,1\,0 \\ 1\,1\,1\,1\,1\,1 \\ 1\,0\,1\,0\,1\,0 \\ 1\,1\,0\,0\,1\,1 \\ 1\,0\,0\,0\,1\,0 \end{pmatrix}$$

We notice that the number of result columns is $n' = 2 * (7 - 4) = 6$. Let's choose $k' = 3$. Since the lower bound of length 6 and dimension is 3 from [13], so we need to eliminate all rows whose weight is less than the lower bound or equal to *n'*. The filtered Kronecker product result is as follows:

$$A_{filtred} = \begin{pmatrix} 1\,1\,1\,1\,0\,0 \\ 0\,0\,1\,1\,1\,1 \\ 1\,0\,1\,0\,1\,0 \\ 1\,1\,0\,0\,1\,1 \end{pmatrix}$$

We can now create combinations of k' rows from $A_{filtred}$, to have matrices of *n'* * *k'* dimensions as generator matrices of our suspect good binary linear codes. If the generator matrix does not contain a complete row rank, one of the linearly dependent rows will be excluded. So *k*' decreases in order to create a generator matrix whose rows are linearly independent.

Then, good binary linear codes designed from matrices extracted from $A_{filtred}$ are (6, 3, 3) as follows:

$$G_1 = \begin{pmatrix} 1\,1\,1\,1\,0\,0 \\ 0\,0\,1\,1\,1\,1 \\ 1\,0\,1\,0\,1\,0 \end{pmatrix}$$

$$G_2 = \begin{pmatrix} 1\,1\,1\,1\,0\,0 \\ 1\,0\,1\,0\,1\,0 \\ 1\,1\,0\,0\,1\,1 \end{pmatrix}$$

In the previous example, the program also produced (6, 2, 4) good binary linear code due to linear dependency between some rows, whose matrix is defined as follows:

$$G_3 = \begin{pmatrix} 1\,1\,1\,1\,0\,0 \\ 0\,0\,1\,1\,1\,1 \end{pmatrix}$$

The algorithm steps are defined below. Let's use the following notations:

– ***G*** : $n{*}k$ generator matrix in the systematic form of chosen BCH code;

– $\boldsymbol{H_2}$ (***or*** $\boldsymbol{H_4}$): Hadamard matrix;
– ***K'***: dimension of the desired binary linear code ($k^{'} \leq 2k$);
– ***LB***: The lower bound of $(2(n-k), k', d')$ desired binary linear code.

Algorithm steps

Inputs: $\boldsymbol{G, H_2, k', LB}$
Outputs: List off n' = 2(*n*–*k*), *k'*, *d'* binary linear codes

Begin
Step 1: Extract a redundant matrix of size $\boldsymbol{k*(n-k)}$ from $\boldsymbol{G}$
Step 2: Perform the Kronecker product between the result of step1 and $\boldsymbol{H_2}$ (***or*** $\boldsymbol{H_4}$)
Step 3: Parse the Kronecker product result of step 2 so that the weight of each result row is greater than or equal to $\boldsymbol{LB}$ and is less then ***n***'
Step 4: Construct from the result of step 3, matrices by combinations of ***k'*** rows
Step 5: for each matrix of step 4:
• Construct a linear code by the matrix whose rows are linearly independent;
• Calculate the minimum distance $\boldsymbol{d}^{'}$ of the linear code;
• If ***d'*** $\geq$ ***LB*** then add the code to the list of (**n' = 2(*n*–*k*), *k', d'***) linear codes
End

4 Results and Interpretation

4.1 *Results of the Proposed Algorithm*

For large dimensions, the exhaustive search of the generator matrices of linear code becomes impossible. The results obtained in this work, given in Table 1, are founded by a program built under the GUAVA package for computing with codes and a computer with an Intel(R) Core(TM) i5-4210U RAM 4 CPU @1.70 GHz configuration.

4.2 *Interpretation of Results*

Channel capacity is one of the most significant parameters of a data communication system. It reflects the maximum rate over which data can be transmitted in a theoretically low error probability system between two points. Shannon has shown theoretically that information can be transmitted at a coding rate $R = \frac{k}{n}$ lower than the capacity of the channel, with a very low probability of error. It also showed that transmission at a coding rate R greater than or equal to the capacity of the channel will certainly incorporate transmission errors.

Recently, the construction of error correction codes is focusing on higher rate codes. In this work, to take full advantage of the channel's capacity, we are often

Table 1 Results of the proposed search method

Rate	n	k	d	Lower bound	Upper bound
0.6	10	6	3	3	3
0.75	12	9	2	2	2
0.66	12	8	3	3	3
0.5	12	6	4	4	4
0.33	12	4	6	6	6
8/13	13	8	4	4	4
0.61	16	13	2	2	2
0.75	16	12	2	2	2
0.62	16	10	4	4	4
0.56	16	9	4	4	4
0.5	16	8	5	5	5
0.72	18	13	2	3	3
0.66	18	12	3	4	4
0.61	18	11	4	4	4
0.55	18	10	4	4	4
0.63	19	12	4	4	4
0.75	20	15	3	3	3
0.8	20	16	2	2	2
0.55	20	11	4	5	5
0.6	20	12	4	4	4
0.65	20	13	4	4	4
0.81	22	18	2	2	2
0.68	22	15	4	4	4
0.63	22	14	4	4	4
0.59	22	13	4	5	5
0.72	22	16	4	4	4
0.86	30	26	2	2	2
0.66	30	20	4	5	5

interested in codes with a rate higher than 0.5. Furthermore, the minimum distance of the codes obtained in the majority of cases is equal to the lower bound in the majority of cases.

The exhaustive search becomes impossible from a certain dimension of code, given the limitations of the computer used. The focus of this paper is to illustrate that there is a way to search for good linear codes from popular codes. Our future vision aims to improve our methodology by using heuristic algorithms to search for high-dimensional codes, because, for high dimensions, a total enumeration of all matrices constructed from $A_{filtred}$ and the search for minimum distance is impracticable.

5 Conclusion

In this work, we have tried to build good codes from BCH codes and Hadamard matrices. These good codes are constructed by extraction from the result of the Kronecker product between the redundant part of BCH code and the Hadamard matrices. Since the search for good codes is exhaustive and it cannot be carried out if the dimension is high, we have the vision to take advantage of the mathematical properties of BCH codes, as well as artificial intelligence including genetic algorithms to broaden our search for good codes.

References

1. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
2. Hamming RW (1950) Error detecting and error correcting codes. Bell Syst Tech J 29:147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x
3. Marcel Golay (1949) Notes on digital coding
4. Tomlinson M, Tjhai CJ, Ambroze MA, Ahmed M, Jibril M (2017) Bounds on error-correction coding performance. In: Tomlinson M, Tjhai CJ, Ambroze MA, Ahmed M, Jibril M (eds) Error-correction coding and decoding: bounds, codes, decoders, analysis and applications. Springer International Publishing, Cham, pp 3–23
5. Augot D, Charpin P, Sendrier N (1992) Studying the locator polynomials of minimum weight codewords of BCH codes. IEEE Trans Inf Theory 38:960–973. https://doi.org/10.1109/18.135638
6. Augot D, Sendrier N (1994) Idempotents and the BCH bound. IEEE Trans Inf Theory 40:204–207. https://doi.org/10.1109/18.272483
7. Dianwu Y, Zhengming H (1996) On the dimension and minimum distance of BCH codes over GF(q). J Electron China 13:216–221. https://doi.org/10.1007/BF02685830
8. Peterson W (1960) Encoding and error-correction procedures for the Bose-Chaudhuri codes. IRE Trans Inf Theory 6:459–470. https://doi.org/10.1109/TIT.1960.1057586
9. Spence LE, Insel AJ, Friedberg S (1999) Elementary Linear algebra: a matrix approach. presented at the (1999)
10. Morelos-Zaragoza RH (2006) The art of error correcting coding. John Wiley & Sons Inc., Hoboken, NJ, USA
11. Hadamard J (1893) RESOLUTION D'UNE QUESTION RELATIVE AUX DETERMINANTS—in Bulletin des Sciences Mathematiques, Septembre 1893. See Description (1893)
12. Hadamard Matrix—an overview | ScienceDirect Topics, https://www.sciencedirect.com/topics/computer-science/hadamard-matrix
13. Grassl, Markus Bounds on the minimum distance of linear codes and quantum codes. http://www.codetables.de
14. Vardy A (1997) The intractability of computing the minimum distance of a code. IEEE Trans Inf Theory 43:1757–1766. https://doi.org/10.1109/18.641542
15. Bozzano Luisa. Handbook of Coding Theory, Volume II—1st Edition, https://www.elsevier.com/books/handbook-of-coding-theory/luisa/978-0-444-50087-8
16. Zimmermann K-H (1996) Integral Hecke modules, integral generalized reed-Muller codes, and linear codes. Techn Univ, Hamburg-Harburg
17. The GAP Group, http://www.gap-system.org
18. Robert H. Morelos-Zaragoza: The Art of Error Correcting Coding, http://the-art-of-ecc.com/

Clinical Note Section Identification Using Transfer Learning

Namrata Nair, Sankaran Narayanan, Pradeep Achan, and K. P. Soman

Abstract In the healthcare industry, clinical records written by doctors, during the consultation with patients, are often in an unstructured fashion. The standard clinical practice recommends the notes to be written in a structured format such as SOAP. Through NLP Transfer Learning, clinical note-taking application can be trained to intelligently recognize sections and subsections within the clinical notes. In this study, we propose several Transfer Learning models based on clinical contextual embeddings for the classification of clinical notes into several major SOAP sections. We validate our models using the clinical notes from the i2b2 2010 dataset. We show that intelligent note-taking applications can be easily developed using the proposed Transfer Learning models, and their integration into Hospital Information Systems will pave the way for effective clinical note generation and alleviation of physician burnout.

Keywords Electronic health record · Machine learning · Natural language processing · Section identification · Clinical narrative · Free text

N. Nair (✉) · S. Narayanan · P. Achan · K. P. Soman
Amrita Technologies, Amritapuri, India
e-mail: namratanair@am.students.amrita.edu

S. Narayanan
e-mail: nsankaran@am.amrita.edu

P. Achan
e-mail: pradeep@amritatech.com

K. P. Soman
e-mail: kp_soman@amrita.edu

N. Nair · S. Narayanan
Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

K. P. Soman
Amrita Center for Computational Engineering and Networking (CEN), Amrita Vishwa Vidyapeetham, Coimbatore, India

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_50

1 Introduction

Health providers, in recent time, started using Electronic Health Records (EHR) [1] increasingly resulting in enormous amounts of unstructured natural language clinical notes, in EHR systems. Although structured formats such as SOAP [2] are available for EHRs, some parts of an EHR like lab results and physical examination are often captured in a structured tone; many other parts like patient history, and hospital course, are captured in an unstructured natural language, as a continuous stream of words, forming a paragraph [3].

Unstructured notes pose a formidable challenge in the process of extraction of knowledge for vital purposes like clinical research, medical education, insurance claim substantiation, and clinical decision support systems [4]. Converting these notes into a structured form will enable medical students and doctors to learn from the collective experiences of the medical community. Clinical queries can be more effective focusing on specified sections. Precise identification of relevant disease conditions can aid patients and providers receive apt reimbursements on insurance claims. Comprehension of the context of prescribed medications can help simplify downstream workflows (such as re-ordering of prescriptions) [5–7].

Intelligent user-friendly interfaces can be developed to decipher the clinical text as the Physician types in his findings and prognosis. For example, the system can prop up auto-completion cues to the medical professional such as prescription orders based on the symptoms and the findings. It can also assign, in real-time, the typed-in entries to relevant SOAP sections and supplement it with the requisite best-practice information. For instance, when the Physician types *She is a seventy-one-year-old female with a history of hypertension*, the system can automatically recognize it as a symptom [2] and assign it to the *Patient History* section. Physicians only need to rectify if there are any errors in the assignment.

Part of the challenge in building such a platform arises from the variability in the natural language used by doctors: specialty-specific terminology, medical jargon, spelling errors, abbreviations, and ambiguous terminology [8, 9]. Common approaches to solve this problem rely on rules-based textual clues and heuristics [10–12], semi-automated construction of datasets, manual label verification, and construction of machine learning models [6, 7]. These methods require deep expertise in NLP, sufficient availability of labeled datasets, and long training times which make the task of developing an intelligent note-taking application a major challenge.

Modern NLP Transfer Learning methods have proven ability to handle complex linguistic dependencies, resilient to variability in usage, do not require hand-generated features, and adopt a data-driven approach. These factors make it attractive to leverage Transfer Learning as the basis for an intelligent clinical note application. So far no detailed work has been done to investigate leveraging these methods for the section identification task. In this context, we developed several transfer learning models for the section note classification task and evaluated their performance using the i2b2 2010 clinical notes' dataset comprised of 427 clinical notes.

2 Related Work

Based on the exhaustive review of section identification approaches by Pomares-Quimbaya et al. [1], two broad categories of existing techniques were identified:

Rule-based Approaches A common method to parse clinical notes is to identify a clinical section from the section header, followed by tagging sentences contained within that section, based on sub-concepts. Exact pattern matches, regular expressions, and probabilistic heuristics are some of the common techniques adopted by rule-based systems. Examples include popular NLP software such as cTAKES [10], CLAMP [11], and ClarityNLP [12]. For example, *History of Present Illness* identifies *Patient History*.

Machine Learning Approaches Section identification using machine learning has used traditional algorithms like Hidden Markov Models [13], Conditional Random Fields [14], and Support Vector Machines [15]. Li et al. [13] built a classifier that can predict 15 sections with a model containing 15 states. Their label generation method is rule-based, similar to the method we follow. Dia et al. [14] used a dataset composed of discharge summaries and procedural notes; SecTag was used to identify the plausible section headers, together with appropriate manual verification. Next, the CRF algorithm was adopted for the learning model. Mowery et al. [15] used support vector machines to classify emergency department reports into their corresponding SOAP sections. All these approaches call for manual intervention—hand-generated features and manually annotated datasets—to achieve good performance.

Deep Learning Techniques The second category of algorithms lean on neural networks such as LSTM [6] for section tagging. Sadoughi et al. [6] used a recurrent neural network, with a pre-trained, Word2vec word embedding layer, for section identification during medical dictations. Jeblee et al. [7] used a neural network for SOAP classification, as part of their system for automated note-taking. Ni et al. [16] used active learning and distant supervision to reduce annotation cost and fast model adaptation. They showed that History and Physical Exams are similar to lexical distribution present in the reports, using 50 reports for their study.

3 Transfer Learning Methodology

NLP Transfer Learning involves training a language model on a large-scale textual corpus, followed by using that pre-trained model as the starting point for another downstream task. We leveraged a combination of the following pre-trained models using a feature-fusion approach [17] for transfer learning:

Classical Word Embeddings Word Embeddings are effective in capturing latent syntactic and semantic similarities among words. They transform an input word into a fixed vector, in a multi-dimensional embedding space. They are created by unsupervised learning on a large general-purpose textual corpus.

(a) *GloVe*: GloVe was adopted as the baseline model for this paper's study because of its prevalence.
(b) *Cui2Vec*: Cui2Vec [18] is a clinical domain-specific semantic embedding [19] with 108477 medical concepts. Given a UMLS Concept Unique Identifier (CUI) [20], Cui2Vec returns a 500-dimensional semantic embedding vector.

Contextual Word Embeddings Contextual Word Embeddings, as the name implies, take into account the context of a word and generate different embeddings, depending on the context of the sentence in which the word appears. These *Language Models*, pre-trained on a large textual corpus, are enabled to predict the next word in a sentence. They are thus able to capture a wide range of linguistic phenomena, such as long-term dependencies, and negation. This knowledge can then be transferred to initialize and quickly train the downstream model.

(a) *BERT* [21] achieved state-of-the-art results based on a transformer-language model. ClinicalBERT [22] is a pre-trained version of BERT created from a corpus of 150000 clinical notes, composed of generic clinical text and discharge summaries. It has been shown to enhance clinical concept extraction performance.
(b) *ELMo* [23] extracts context-sensitive features, using a bi-directional (left-to-right and right-to-left) character-level language model. For each word, the concatenation of both of these representations constitutes the word embedding. The ELMo-PubMed model, trained on the Pubmed corpus, was adopted for the present study.
(c) *Flair* [24] is a contextual string embedding model. Essentially, the sentence is passed to a character-language model to generate the embeddings. The Flair-PubMed model, also trained on the Pubmed corpus, was adopted for the present study.

4 Dataset

Using 427 clinical notes contained in the i2b2 2010 dataset [25], a custom dataset was generated comprised of 17398 sentences, divided into the following four major SOAP sections (Table 1) with a 80/10/10 train/dev/test split:

(a) History of Present Illness (*HISTORY*): Subjective sections such as *HISTORY* describe the condition of the patient and reason for seeking care.
(b) Physical Examination (*PE*): Objective sections such as *PE* document-specific measurements, test results, and exam findings done by the Physician.

Table 1 Dataset characteristics

SOAP section	Label	Samples	Example
Subjective	HISTORY	2941	She is a seventy-one-year-old female with a history of hypertension
Objective	PE	2804	She had 5/5 strength in bilateral upper and lower extremities
Assessment and Plan	COURSE	7105	A Neurosurgery Service consultation was obtained and they felt that no surgical intervention was required
Assessment and Plan	MEDS	4548	1. Colace 100 mg PO bid
Total		17398	

(c) Hospital Course (*COURSE*): Assessment and Plan sections such as *COURSE* synthesize the Subjective and Objective sections to arrive at a diagnosis and the steps taken to treat the patient.
(d) Medications (*MEDS*): As part of the Plan section, the *MEDS* section outlines the prescribed Medications.

5 Results

5.1 Model Performance

The Flair framework [24] was used for model construction. A standard single-layer bi-directional Long Short Term Memory (LSTM) with 512 hidden layers was used to refine the embedding layer generated representations followed by a SoftMax layer for classification. Table 2 shows the performance of the proposed transfer learning models. The F1 value at the end of 10 epochs is reported in Column 3. The final performance at the end of 30 epochs is reported in Column 4.

(a) The baseline model leveraged GloVe alone with an overall classification performance of 86.78%.
(b) Adoption of Cui2Vec enhanced final performance marginally, despite exhibiting a pronounced early ramp-up (Row 2).
(c) Language-model fine-tuning of Flair-PubMed models [26] using the i2b2 2010 text corpora helped enhance performance (Row 3).
(d) Stacking of classical and contextual embeddings helped enhance performance marginally (Rows 5, 7, 9).

Table 2 Model performance

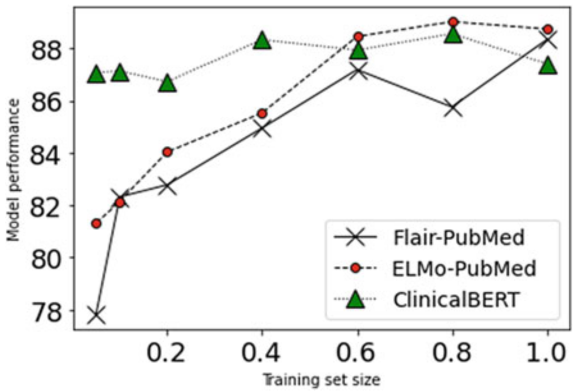
Exp. no.	Architecture	F1	F1
		10 epochs	30 epochs
1.	GloVe (G)	81.09	86.78
2.	G + Cui2Vec (C)	83.0	87.07
3.	Flair-PubMed (Default)	83.62	85.17
4.	Flair-PubMed (Tuned)	87.01	88.33
5.	G + C + Flair-PubMed (T)	87.3	88.39
6.	ELMo-PubMed	88.74	88.74
7.	G + C + ELMo-PubMed	88.91	88.91
8.	ClinicalBERT	87.41	88.50
9.	G + C + ClinicalBERT	87.7	88.68

5.2 *Effect of Reduction in Training Set Sizes*

The three contextual embedding-only models were evaluated under varying training set sizes (5 to 100 %) using random down-sampling. The test dataset was kept the same. Figure 1 shows the final performance.

(a) ClinicalBERTs out-performed at lower training set sizes delivering the best performance at 87.07 % demonstrating the influence of pre-training on a domain-specific corpus.
(b) ELMo-PubMed out-performed above 60% training set sizes but exhibited wider variability (81.32 to 89.02 %, Fig. 2).
(c) Both ELMo-PubMed and Flair-PubMed exhibited an 8 % deviation in performance for the *COURSE* section while ClinicalBERT remained stable (Fig. 3). The minimum number of samples used was 579 with the highest being 7100.

Fig. 1 Overall performance (F1)



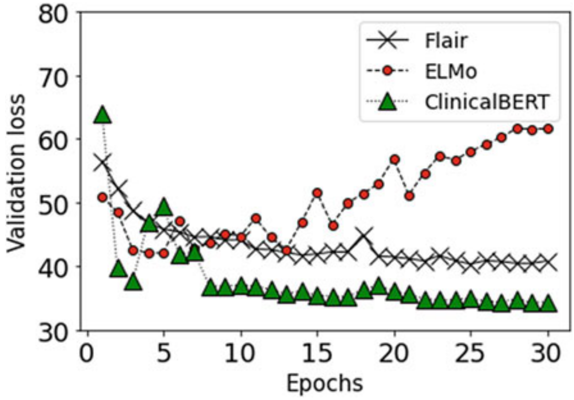


Fig. 2 Validation loss micro-F1

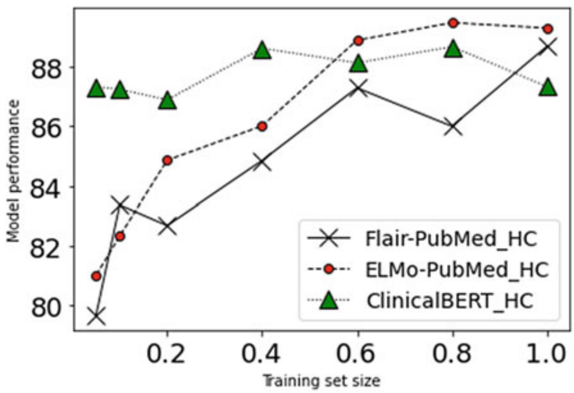


Fig. 3 COURSE micro-F1

(d) The *HISTORY* section exhibited the maximum deviation in performance between the models (Fig. 4). The lowest number of samples used was 230, with the highest being 2941. For this range, ClinicalBERT exhibited a 2.5 % difference, while ELMo-PubMed degraded by 14.42 %. Flair-PubMed degraded by a similar amount while there were 500 or more samples, but underperformed when there were fewer than 500 samples with F1 = 36.54 %. We analyze this in the next section.

5.3 Label-Wise F1 Analysis

Table 3 details the label-wise performance for the three contextual embedding-only models. *MEDS* and *PE* section exhibited relatively straight-forward classification while *HISTORY* presented a challenge. A text such as 'Head circumference was 34.5 cm 75th percentile and length was 50.7 cm also the 75th percentile' was correctly

Fig. 4 HISTORY micro-F1

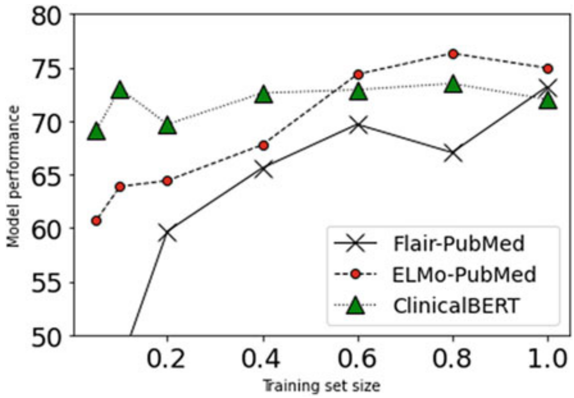


Table 3 Label-wise F1 (30 epochs)

S.No	Label	ClinicalBERT	ELMo-PubMed	Flair-PubMed
1.	HISTORY	71.62	75.11	73.19
2.	COURSE	89.21	89.17	88.68
3.	PE	91.90	90.27	91.33
4.	MEDS	97.03	96.2	96.18

identified as PE. A text like 'A tumor was noted which was biopsied and revealed squamous cell carcinoma in situ.' was incorrectly classified as *COURSE* although it occurred in the *HISTORY* section. Our Subject Matter Expert noted that without proper context given, such a sentence can occur in both *HISTORY* and *COURSE*.

6 Limitations and Future Work

The following are few limitations uncovered in this study:

(a) We have considered only four SOAP subsections for this work. Accounting for all the subsections of SOAP is part of our investigations.
(b) We have not considered the context sensitivity of clinical sentences as explained in the previous section. Context determination is part of our future work.

7 Conclusion

In this paper, we showed that contextual embeddings provide an efficient solution to the task of clinical note section identification. We showed that around 500 samples per section would be a sufficient starting point for a well-performing model. We established that transfer learning provides an effective solution to the task of section classification. Our future work will look into validation with diverse clinical specialties and integration into an intelligent note-taking interface for effective generation of clinical notes and assuage of Physician burnout.

8 Availability of Data and Materials

The i2b2 2010 dataset is not public but can be obtained from the n2c2 NLP Research datasets repository [27] after executing the requisite legal agreements. The code used for this project is available in Git.

Acknowledgements We thank Dr. Smriti Nair, Amrita Institute of Medical Sciences, Kochi, India, for reviewing the test datasets and providing valuable suggestions to help understand the results.

References

1. Pomares-Quimbaya A, Kreuzthaler M, Schulz S (2019) Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. BMC Med Res Methodol 19(1):155
2. Podder V, Lew V, Ghassemzadeh S (2020) SOAP Notes. In: StatPearls [Internet]. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK482263/
3. Hayrinen Kristiina, Saranto Kaija, Nykanen Pirkko (2008) Definition, structure, content, use and impacts of electronic health records: a review of the research literature. Int J Med Inform 77(5):291–304
4. Wang Yanshan et al (2018) Clinical information extraction applications: a literature review. J Biomed Inform 77:34–49
5. Denny JC et al (2008) Development and evaluation of a clinical note section header terminology. In: AMIA annual symposium proceedings, vol 2008. American Medical Informatics Association, 156p
6. Sadoughi N et al (2018) Detecting section boundaries in medical dictations: Toward real-time conversion of medical dictations to clinical reports. In: International conference on speech and computer. Springer, pp 563–573
7. Jeblee S et al (2019) Extracting relevant information from physician-patient dialogues for automated clinical note taking. In: Proceedings of the tenth international workshop on health text mining and information analysis (LOUHI 2019), pp 65–74
8. Jonnalagadda SR et al (2017) Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFPEF patients for clinical trials. J Cardiovasc Transl Res 10(3):313–321
9. Honghan Wu et al (2018) SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 25(5):530–537
10. Savova GK et al (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17(5):507–513
11. Soysal Ergin et al (2018) CLAMP-a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 25(3):331–336

12. Georgia Tech Research Institute (2018) ClarityNLP Section Tagging. https://claritynlp.readthedocs.io/en/latest/developerguide/algorithms/sectiontagger.html
13. Li Y, Lipsky Gorman S, Elhadad N (2010) Section classification in clinical notes using supervised hidden markov model. In: Proceedings of the 1st ACM international health informatics symposium (2010), pp 744–750
14. Dai HJ et al (2015) Recognition and evaluation of clinical section headings in clinical documents using token-based formulation with conditional random fields. BioMed Res Int
15. Mowery Danielle et al (2012) Building an automated SOAP classifier for emergency department reports. J Biomed Inform 45(1):71–81
16. Ni Jian, Delaney Brian, Florian Radu (2015) Fast model adaptation for automated section classification in electronic medical records. Stud Health Technol Inform 216:35–39
17. Barathi Ganesh HB et al (2020) MedNLU: natural language understander for medical texts. In: Deep learning techniques for biomedical and health informatics. Springer, pp 3–21
18. Beam AL et al (2018) Clinical concept embeddings learned from massive sources of multimodal medical data arXiv:1804.01486
19. Yu M, Dredze M (2014) Improving lexical embeddings with semantic knowledge. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, vol 2: Short Papers, pp 545–550
20. Bodenreider Olivier (2004) The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 32(suppl 1):D267–D270
21. Devlin J et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
22. Alsentzer E et al (2019) Publicly available clinical BERT embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. Association for Computational Linguistics, Minneapolis, Minnesota, USA, June 2019, pp 72–78. https://doi.org/10.18653/v1/W19-1909 https://www.aclweb.org/anthology/W19-1909
23. Peters ME et al (2018) Deep contextualized word representations. In: Proc. of NAACL
24. Akbik A, Blythe D, Vollgraf R (2018) Contextual string embeddings for sequence labeling. In: Proceedings of the 27th international conference on computational linguistics, pp 1638–1649
25. Uzuner Ozlem et al (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 18(5):552–556
26. Narayanan S et al (2020) Evaluation of transfer learning for Adverse Drug Event (ADE) and medication entity extraction. In: Proceedings of the 3rd clinical natural language processing workshop. Association for Computational Linguistics, pp 55–64. https://www.aclweb.org/anthology/2020.clinicalnlp-1.6
27. Harvard University Medical School (2020) n2c2 NLP Research Datasets https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Multi-Agent-Based Recommender Systems: A Literature Review

Joaquim Neto, A. Jorge Morais, Ramiro Gonçalves, and António Leça Coelho

Abstract Considering the growing volume of information and services available on the web, it has become essential to provide websites and applications with tools, such as recommender systems, capable of helping users to obtain the information and services appropriate to their interests. Due to the complexity of web adaptation and the ability of multi-agent systems to deal with complex problems, the use of multi-agent approaches in recommender systems has been increasing. In the present work, we make a thorough review of the use of multi-agent-based recommender systems. The review shows the diversity of applications of multi-agent systems in recommender systems, namely on what concerns the diversity of domains, different types of approaches and contribution to the performance improvement of the recommender systems.

Keywords Recommender systems · Multi-agent systems · Survey · Literature review · State-of-the-art

J. Neto (✉) · A. J. Morais
Universidade Aberta, Lisboa, Portugal
e-mail: jfn@lnec.pt

J. Neto · A. L. Coelho
Laboratório Nacional de Engenharia Civil, Lisboa, Portugal
e-mail: alcoelho@lnec.pt

A. J. Morais
LIAAD - INESC TEC, Porto, Portugal
e-mail: jorge.Morais@uab.pt

R. Gonçalves
UTAD (Universidade de Trás-Os-Montes e Alto Douro), Vila Real, Portugal
e-mail: ramiro@utad.pt

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_51

1 Introduction

The increasing volume of information available to users on the web makes it difficult to obtain the information they need at any given time. To help users getting the appropriate information, websites and web applications are using recommender systems for recommending web pages, learning objects, web repository documents or other types of digital resources. The complexity of the problem, the distributed nature of the web and, as defended by [1], the propensity of multi-agent approaches to address complex problems, such as the case of web adaptation, justify the use of multi-agent systems technology approaches.

Considering that the areas of recommender systems and multi-agent systems play an essential role in our research interests, it becomes fundamental to perceive the state-of-the-art of the research domain that brings together the two technologies. As the first step in our research, we started by searching for literature reviews, surveys or state-of-the-art that summarised work already done involving recommender systems and multi-agent systems. To do so, we started our search process by searching google scholar for publications that, in their title, satisfied the following query: *((“recommender systems” OR “recommending systems”) AND (review OR survey OR “state of the art”)).* As a result, we obtained several dozens of publications, also stored in libraries like *SpringerLink, Science Direct, IEEE Xplore and ACM Digital Library,* satisfying that query. However, if we add to the above query the terms: *(multi-agent OR “multi-agent” OR “agent-based”)*, we found no publications. Therefore, we consider it helpful and relevant to develop a research work that presents the state-of-the-art of multi-agent recommender systems field of research.

In Sect. 2, we introduce the multi-agent and recommender systems basic concepts. In Sect. 3, we present the literature review, starting by explaining the research methodology, following with the search findings and concluding with a review of the selected papers. In Sect. 4, a discussion, comparative analysis and main contributions of the articles are presented. Finally, in Sect. 5, we write the conclusions of the present survey.

2 Multi-Agent Systems and Recommender Systems: Concepts and Definitions

2.1 Defining Agent and Multi-Agent System

According to [2], there is no consensus definition for the concept of an agent. In the literature, it is possible to find multiple definitions [3] and different authors define agent according to the use that each one makes of the term [4]. Reference [4] defines an autonomous agent as a system that can perceive its environment and acts on it according to its purposes. Wooldridge [2] defines an agent as a computer system capable of reaching the goals delegated to it, by autonomously acting in the environment it is located. According to Jennings [5] agents: (i) are entities capable of solving problems with well-defined limits and interfaces; (ii) receive as input the state of the environment in which they are located through sensors, acting in that same environment through actuators; (iii) are designed to serve a specific purpose and have determined goals to achieve; (iv) exhibit flexibility and have control over their behaviour in pursuit of their goals; (v) have to be proactive and reactive. The concept of autonomy, which we can understand as the ability of the agent to fulfil the objectives delegated to him, is shared by all definitions.

As referred by [6] a multi-agent system can be defined as a system composed of multiple intelligent agents that are capable of working together to achieve objectives and are more difficult to achieve by an individual agent. Being part of a multi-agent system, agents can cooperate or compete with each other. When interacting with each other cooperatively, they do so directly or through the environment to reach common goals. However, agents inserted in a multi-agent system may also have their own objectives, so they will have to negotiate to achieve the system goals.

2.2 Recommender Systems

The exponential growth of information and data available on the web complicates the user's task of obtaining, in time, the information he needs. To help users in this challenging task websites have been incorporating tools, such as recommender systems, that allow the website adaptation to provide users with improved user experience, facilitating obtaining the information that is more accurate and in line with their interests. Recommender systems have two purposes. They can be used to stimulate users to do something, to make a decision, such as the acquisition of a particular good. They can also be used to alleviate information overload in users, making available to them the items that are most in line with their interests.

In the study of Wei et al. [7], a recommendation can be seen as a reference to an item (a webpage, for example) that is made available to the user looking for the information. For the same authors, a typical recommender system aggregates and directs the recommendations to the appropriate recipients. Wei et al. [7] consider that recommender systems can provide quality recommendations to users because such recommendations are based on their past preferences or the preferences of other users with similar interests.

The way recommendations are generated are used to classify the different types of recommender systems. As referred by [8], three types of approaches may be identified: (i) *content-based*—the recommended items are those with similar content to the ones preferred by the user in the past; (ii) *collaborative*—the recommended items are those that users with preferences similar to the active user have liked in the past, and (iii) *hybrid* approaches, which allow the combination of techniques used in both types of approaches. According to [8] hybrid approaches, by incorporating "the advantages of both methods while inheriting the disadvantages of neither" [8, p. 66], "can provide more accurate recommendations than pure collaborative and content-based approaches" [9, p. 9]. Collaborative and content-based algorithms can be combined in two different ways. The algorithms are applied separately, and afterwards, the results are combined to produce the recommendations set (linear way). One algorithm is applied first, and the results of this algorithm are filtered by the other algorithm (sequential way).

As referred by [9], collaborative and content-based approaches only consider the item-user pair, not taking into account the circumstances in which the recommendations occur or other contextual information. According to Rahman [10], many recent studies have been demonstrating that ignoring context information may put the results of the recommendations under suspicion, pointing out that an improvement in the performance of a recommendation system begins with the collection of context-related data. The approach, which, besides the item-user pair, also considers the context information to generate the recommendations, is called a *context-based approach.*

3 Multi-Agent Recommender Systems: A Survey

As referred above, adapting and personalising websites and web applications to users' interests is a complex problem that justifies the use of multi-agent systems technology approaches in recommender systems. In this section, it is our purpose to carry out a thorough state-of-the-art in the area of multi-agent-based recommender systems.

3.1 Research Methodology

For our research work, we searched in *Google Scholar, Science Direct, ACM Digital Library, IEEE Xplore and SpringerLink* publication databases to build a comprehensive bibliography of research papers on the area of multi-agent-based recommender systems. For this research, we follow three main steps:

1. As the first and preliminary step, we searched in Google Scholar for literature reviews on recommender systems, using the following query for title search: *((“recommender systems” OR “recommending systems”) AND (review OR survey OR “state of the art”))*, checking, from the obtained results, which of the founded works are about literature reviews on multi-agent recommender systems approaches.
2. In the second step we searched the referred publication databases for publications that, in their title, met the criteria: *((recommender OR recommending) AND (multi-agent OR “multi-agent” OR “agent-based”)).*
3. Based on the search results, we identified and selected a set of papers that may represent the state-of-the-art in the area of multi-agent-based recommender systems, according to purpose and domain application, types of approaches and main contributions.

3.2 Findings

Searching for literature reviews

As already mentioned, we started this work by looking for published works focused on recommender systems reviews, and among them, those that focused their research on multi-agent approaches. Based on the searches carried out, it was not possible to find literature reviews focused on multi-agent-based recommender systems. However, concerning recommender systems, there was an enormous quantity of research work available. From those literature reviews oriented to the recommender systems in general, deserves particular attention, by its meaning and impact, the works [11–14]. We also found many literature reviews targeting specific domains such as tourism [15] or e-learning [16]; about particular approaches, such as content-based [17], collaborative filtering [18], hybrid [19] and context-based [20]. In more than 500 articles analysed by those literature reviews, less than 20 of them rely on recommendation solutions based on the technology of multi-agent systems.

Searching for publications on multi-agent-based recommender systems

In the second step of our research work, we found more than 150 publications that meet the search criteria referred to in Sect. 3.1. After a thorough analysis, we select a set of 15 papers that we consider representative of the state-of-the-art in the area of multi-agent-based recommender systems.

E-commerce

E-commerce was one of the areas whose need for recommendation systems was most felt. In [21] two intelligent agent-based systems are proposed, to address two kinds of recommendations: (i) based on customer's purchasing history; and (ii) based on customer's current preferences obtained from customer interactions. For both types of recommendations, the authors proposed two different multi-agent approaches.

Considering the need to provide personalised recommendation on e-commerce, e-service and e-business sites, [22] proposed a solution based on fuzzy cognitive agents. Those agents follow an agent knowledge model based on the fuzzy cognitive map (FCM) theory [23], and they support the following behaviours: communication with users; perception of the environment; learning from most recent users' behaviour; and making inferences based on the current personal preferences of the users, as well as based on expert's domain knowledge.

General-purpose

To deal with information overload and help users find the needed information, [24] proposed *Implicit*, an agent-based recommender system, where personal agents, representing the user in the system, use data mining techniques to discover the user's behaviours. The implicit knowledge that personal agents are capable of extracting from the user's behaviours is used to suggest web page links to the user, or other personal agents.

In their work, [25] proposed a solution to overcome the limitations related to the similarity-based recommendation solutions, where a similar opinion of the users are collected by agents to make predictions about a new item. The authors proposed a trust-based agent community that considers trust in the item recommendations.

Recommender systems are being used to create personalised information based on personal user data. Concerned with users' privacy, [26] proposed an agent-based privacy-preserving recommender system that allows the generation of recommendations while preserving the participant's privacy. The authors justify the use of a multi-agent system (MAS) technology due to agents' typical properties: autonomy, ability to communicate and adaptability.

In [27], the existing recommender systems lack initiative, intelligence and self-adaptation. To address these disadvantages, the authors presented *APRS*, a personalised recommender system based on a multi-agent system. The recommendations to the users are generated based on explicit information, through user's feedback, and implicit information.

In their research work, [28] and [6] proposed a multi-agent-based recommender system where recommender agents, running different algorithms, compete with each other to provide the most adequate set of recommendations to website visitors. Based on an auction, the *user agent* selects the best set of recommendations and sends them to the user. The offline test phase results were later confirmed in real-time tests [29].

Social networking websites (SNSs) are bringing people together. SNSs are using recommender systems to provide information to users, namely for product advertising and marketing, most of the times of no interest for the users. To address that problem, [30] designed and developed an adaptable, intelligent agent-based interface to make recommendations based on users friend's interests on different SNS sites.

E-tourism

Recommender systems are increasingly used in e-tourism to recommend appropriate destinations to tourists. In their work, [31] present *Turist@*, an agent-based recommender system to address a less explored problem in the literature: recommend leisure and cultural activities to the tourist, when he has already arrived at the destination. The recommender agent follows a hybrid approach, incorporating both content-based and collaborative recommendation strategies. Another significant contribution of this work was the system modularity. That modularity allows the addition of new activity agents in runtime, and the possibility of splitting the system into several parts, making it possible to run in lightweight devices such as smartphones.

E-learning

E-learning is another area where the use of recommender systems is growing. In their work, [32] presented a knowledge-based recommender system (KRS) based on the K-means clustering technique. According to the authors, the experiments show that the KRS can retrieve relevant LO as well as improve the recommendation precision.

Ubiquitous learning (u-learning) allows learning to take place independently of time and place. In [33] a multi-agent context-aware u-learning system is proposed, aimed at providing context-aware learning planning, access to current context-aware collaborative learning activities, personalised course evaluation, search and selection of learning objects according to student profile, and search for thematic learning assistants.

Financial markets

Financial markets are another well-known area of application for multi-agent systems and recommender systems. In their approach [34] proposed a multi-agent recommender system for computational investing. The solution uses a hybrid filtering technique to recommend the most profitable stocks at the right time, according to the investor's interest.

Internet of Things (IoT)

In the Internet of Things (IoT) environments, the recommendation of useful things is an essential task for the most diverse applications such as urban computing, smart cities or healthcare. In their paper, [35] present MAS that generates personalised recommendations on mobile devices, based on contextual data acquired from the IoT.

Energy Management

The adoption of solutions based on multi-agent systems is common in power and energy systems, so [36] proposed a case-based reasoning (CBR) multi-agent recommender system for intelligent management of energy in buildings. In their approach, the system recommends the energy reduction that should be applied at the moment, in an amount based on similar past cases. The complexity and dynamics of power and energy systems have led researchers to use an agent-based solution. The proposed architecture relies on multiple independent MAS, each dealing with its part of the energy system. Interoperability between the different multi-agent systems is ensured through ontologies.

4 Discussion

The research work carried out showed the increasing presence of recommender systems in websites and web applications, as well as the increasing number of recommendation solutions based on intelligent agent technology and multi-agent systems. As already mentioned, when we searched for published literature reviews on recommender systems, we found plenty of them. However, none is found exclusively dedicated to multi-agent-based recommender systems, which seems to be a demonstration of the interest of this work.

In Table 1, we present a comparing overview of the reviewed research papers according to the following dimensions: application domain; main contributions; recommender approaches; and theoretical vs practical papers. Concerning application domains, most of them were categorised as "general purpose" in the sense that the proposed solutions address problems like information overload or website personalisation, and are domain application-independent. E-commerce, e-learning, and e-tourism are areas of application where we found many research works, and we select some of them for our review. Most of the works reviewed are practical

Table 1 Comparison of reviewed publications according to the domain of application, contributions, type of approach and whether they concern practical applications or theoretical approaches

Paper	Domain	Main contribution	Recommender approach	Theoretical or Practical
Lee et al. [21]	E-commerce	One system uses implicit information (users' past behaviour); the other uses explicit information (users' feedback)	Collaborative: Content-based	Practical with experimental results
Birukov et al. [24]	General-purpose	Implicit knowledge based on data mining technique: combines search results with suggestions from community members	Collaborative	A prototype developed and experimental results showed. MAS framework: JADE
Huang et al. [27]	General-purpose	Based on explicit (users feedback) and implicit (users behaviour information)	hybrid	Theoretical
Weng et al.[25]	General-purpose	Incorporation of trust; contributes with a trust metric that allows the possibility of an agent to quantify the degree of trust on others	Collaborative	Simulation-based on the MovieLens dataset
Miao et al. [22]	E-commerce	Implicit and explicit information; uses fuzzy cognitive maps; personalized recommendations based on: user's preferences; expert's knowledge	Collaborative	A prototype developed and experimental results showed. MAS framework: JADE
Cissée and Albayrak [26]	General-purpose	Addresses privacy-preserving in recommender systems	Collaborative	A prototype developed and experimental results showed. MAS Framework: JIAC IV[1]
Batet et al. [31]	E-tourism	System modularity; addition of new activities agents in runtime; possibility of splitting the system into several parts, ability to run in lightweight devices such as smartphones	Hybrid	Practical with experimental results MAS framework: JADE

(continued)

[1] http://www.jiac.de/.

Table 1 (continued)

Paper	Domain	Main contribution	Recommender approach	Theoretical or Practical
Morais et al. [6] Neto and Morais [29]	General-purpose	Market-based approach; auction (sealed-bid auction protocol); item-based collaborative filtering; implicit information; accepts new recommender agents in runtime	Collaborative	Prototype developed: (1) simulation based on real data; and (2) real-time experiments. MAS Framework: JADE
Taghavi et al. [34]	Financial markets	A new way of combining collaborative and content-based algorithms	Hybrid	Theoretical
Moin et al. [30]	General-purpose	Recommendations based on social circle	Collaborative	Android App developed and experimental results showed. MAS Framework: JADE
Rodríguez et al. [32]	E-learning	Based on the K-means clustering technique;	Knowledge-based	Prototype developed and experimental results showed. MAS Framework: JADE
Salazar et al. [33]	E-learning	Context-aware; u-learning; allows students' interaction; allows teachers to know the students' interest within the virtual course	Context-based	Prototype developed, tested and experimental results showed. MAS Framework: JADE
Twardowski and Ryzko [35]	IoT	Based on contextual data acquired on IoT	Context-based	Theoretical
Pinto et al. [36]	Energy Management	Proposes a MASs society, with different MAS. Ontologies support the interoperability among the MAS. Uses K-NN clustering algorithm	Collaborative	Tested and evaluated model. MAS Framework: JADE

applications, most of them being developed with JADE[2] framework, and present results of the experimental tests carried out. Concerning the types of recommendation approaches, most of the works follow collaborative approaches, being significant the presence of hybrid approaches, which is not surprising given the typical characteristics of multi-agent systems. Regarding the contributions of each paper, Table 1 is exhaustive in this matter. However, it is essential to highlight the aspects related to the modularity of the systems, resulting from the use of multi-agent systems technology, as well as the proposed solutions to overcome issues related to privacy and trust.

5 Conclusion

Helping users to obtain the information or services they need and when they need, or to alleviate them from information overload are complex tasks that may be overcome by incorporating recommender systems in websites and web applications. Considering the distributed nature of the internet and its propensity to assist in solving complex problems, multi-agent systems appear as support for recommender systems. In our survey, we identify and describe several multi-agent-based approaches applied to recommender systems in areas of application as diverse as the financial markets, energy management, e-commerce, e-learning, social networks, tourism and IoT, as shown in Table 1.

References

1. Jennings NR (2001) An agent-based approach for building complex software systems. Auton Agent Multi Agent Syst 24(1):141–174
2. Wooldridge M (2009) An Introduction to MultiAgent Systems—Second Edition by Michael Wooldridge Published May 2009 by John Wiley & Sons ISBN-10: 0470519460 ISBN-13: 9780470519462. 2nd Edition. John Wiley & Sons, Inc.
3. Dorri A, Kanhere S, Jurdak R (2018) Multi-agent systems: a survey. IEEE Access 6:28573–28593
4. Franklin S, Graesser A (1996) Is it an agent, or just a program?: A Taxonomy for Autonomous Agents. In: Proceedings of the third international workshop on agent theories, architectures, and languages, pp 21–36
5. Jennings NR (2000) On agent-based software engineering. Artif Intell 117(2):277–296
6. Morais AJ, Oliveira E, Jorge AM (2012) A multi-agent recommender system. In: Advances in Intelligent and Soft Computing, vol 151, AISC, pp 281–288
7. Wei YANZ, Moreau LUC, Jennings NR (2005) A Market-based approach to recommender systems, vol. 23, no. 3, pp 227–266
8. Balabanović M, Shoham Y (1997) Fab: content-based, collaborative recommendation. Commun ACM 40(3):66–72

[2] http://jade.tilab.com/.

9. Adomavicius G, Sankaranarayanan R, Sen S, Tuzhilin A (2005) Incorporating contextual information in recommender systems using a multidimensional approach. ACM Trans Inf Syst 23(1):103–145
10. Rahman MM (2013) Contextual recommender systems using a multidimensional approach. Int J Intell Inf Syst 2(4):55
11. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state of the art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749
12. Bobadilla J, Ortega F, Hernando A, Gutiérrez A (2013) Recommender systems survey. Knowledge-Based Syst 46:109–132
13. Lu J, Wu D, Mao M, Wang W, Zhang G (2015) Recommender system application developments: a survey. Decis Support Syst, vol. 74
14. Beel J, Gipp B, Langer S, Breitinger C (Nov. 2016) Research-paper recommender systems: a literature survey. Int J Digit Libr 17(4):305–338
15. Borràs J, Moreno A, Valls A (Nov. 2014) Intelligent tourism recommender systems: a survey. Expert Syst Appl 41(16):7370–7389
16. Klašnja-Milićević A, Ivanović M, Nanopoulos A (2015) Recommender systems in e-learning environments: a survey of the state-of-the-art and possible extensions. Artif Intell Rev 44(4):571–604
17. Lops P, de Gemmis M, Semeraro G (2011) Content-based Recommender Systems: state of the Art and Trends. In: Recommender Systems Handbook, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, pp 73–105
18. Yang X, Guo Y, Liu Y, Steck H (Mar. 2014) A survey of collaborative filtering based social recommender systems. Comput Commun 41:1–10
19. Burke R (Nov. 2002) Hybrid recommender systems: survey and experiments. User Model. User-adapt. Interact. 12(4):331–370
20. Verbert K et al (2012) Context-aware recommender systems for learning: a survey and future challenges. IEEE Trans Learn Technol 5(4):318–335
21. Lee W-P, Liu C-H, Lu C-C (May 2002) Intelligent agent-based systems for personalized recommendations in Internet commerce. Expert Syst Appl 22(4):275–284
22. Miao C, Yang Q, Fang H, Goh A (May 2007) A cognitive approach for agent-based personalized recommendation. Knowledge-Based Syst 20(4):397–405
23. Kosko B (1986) Fuzzy cognitive maps. Int J Man Mach Stud 24(1):65–75
24. Birukov A, Blanzieri E, Giorgini P (2005) Implicit: an agent-based recommendation system for web search. In: Proceedings of the fourth international joint conference on autonomous agents and multiagent systems, pp 618–624
25. Weng J, Miao C, Goh A, Shen Z, Gay R (2006) Trust-based Agent Community for Collaborative Recommendation. In: Proceedings of the fifth international joint conference on autonomous agents and multiagent systems, pp 1260–1262
26. Cissée R, Albayrak S (2007) An agent-based approach for privacy-preserving recommender systems. In: Proceedings of the 6th international joint conference on autonomous agents and multiagent systems, pp 182:1–182:8
27. Huang L, Dai L, Wei Y, Huang M (2008) A personalized recommendation system based on multi-agent. In: 2008 Second International Conference on Genetic and Evolutionary Computing, 2008, pp 223–226
28. Morais AJ (2009) A multi-agent approach for web adaptation. In: Proceedings of 7th international conference on practical applications of agents and multi-agent systems (PAAMS 2009), pp 349–355
29. Neto J, Morais AJ (2014) Multi-agent web recommendations. Adv Intell Syst Comput 290:235–242
30. Moin S, Muhamamd A, Martinez-Enriquez AM (2014) Agent based mobile recommender system. In: Proceedings of 2014 11th international conference on Electrical Engineering, Computing Science and Automatic Control (CCE), pp 1–6

31. Batet M, Moreno A, Sánchez D, Isern D, Valls A (2012) Turist@: Agent-based personalised recommendation of tourist activities. Expert Syst Appl 39(8):7319–7329
32. Rodríguez P, Duque N, Ovalle DA (2015) Multi-agent system for knowledge-based recommendation of learning objects using metadata clustering. In: Highlights of Practical Applications of Agents, Multi-Agent Systems, and Sustainability—The PAAMS Collection, pp 356–364
33. Salazar OM, Ovalle DA, Duque ND (2015) Adaptive and personalized educational ubiquitous multi-agent system using context-awareness services and mobile devices. In: Learning and Collaboration Technologies, pp 301–312
34. Taghavi M, Bakhtiyari K, Scavino E (2013) "Agent-based computational investing recommender system. In: Proceedings of the 7th ACM conference on recommender systems, pp 455–458
35. Twardowski B, Ryzko D (2015) IoT and Context-aware mobile recommendations using multi-agent systems. In: Proceedings of 2015 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT), vol. 1, pp 33–40
36. Pinto T, Faia R, Navarro-Caceres M, Santos G, Corchado JM, Vale Z (Mar. 2019) Multi-agent-based CBR recommender system for intelligent energy management in buildings. IEEE Syst J 13(1):1084–1095

The Impact of Cloud Computing and Artificial Intelligence in Digital Agriculture

Kohei Dozono, Sagaya Amalathas, and Ravan Saravanan

Abstract Cloud computing technology and Artificial Intelligence (AI) techniques are becoming significant components in many applications across multiple domains, including agriculture. Cloud computing allows the mass storage of various types of data readily available for descriptive, predictive, and prescriptive analytics. AI techniques, such as deep neural networks, have achieved promising disease detection and yield prediction outcomes. Not many agricultural systems have applied cloud computing and AI and have been made available for farmers to use practically. This study discusses the design and development of a digital agriculture platform using AI and cloud technology. The proposed platform provides an end-to-end digital agriculture system to farmers. This study discusses the benefits of applying cloud computing and AI according to the proposed system and existing studies. The achieved outcomes delivered a significant edge over the conventional platform for digital agriculture. We developed a plant disease diagnosis model to enable farmers to identify diseases through a mobile application on the proposed platform. Our proposed model showed significant results for identifying and diagnosing plant diseases with high accuracy.

Keywords Cloud computing · Artificial intelligence · Digital agriculture · Deep learning · Convolutional neural network

K. Dozono (✉)
Taylor's University, Jalan Taylors, Subang Jaya, Malaysia
e-mail: koheidozono@sd.taylors.edu.my

R. Saravanan
RC Buminiaga Sdn Bhd, Kampung Sungai Kayu Ara, Petaling Jaya, Malaysia
e-mail: ravan@rcbuminiaga.com.my

S. Amalathas
University of Southampton, Nusajaya, Johore, Malaysia
e-mail: S.S.Amalathas@soton.ac.uk

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_52

1 Introduction

Nowadays, cloud computing and Artificial Intelligence (AI) technology are becoming more significant for agricultural productivity. Cloud computing technology contributes to the growth of agriculture by enabling the storage and processing of mass amounts of agricultural data. Cloud computing technology can bring about data availability by storing soil-related, weather-related, crop, and farmer-related data all in one place. End-users such as farmers, experts, consultants, and researchers can easily access those data anywhere and anytime through their devices connected to the cloud system [1]. In addition, the cloud-based application is capable of providing a precision agriculture system without high upfront investments, and that allows efficient use of computational resources and cost control based on usage [2]. Additionally, cloud computing empowers distributed wireless sensor networks to improve the storage as well as capabilities of information processing capabilities. From an information security standpoint, computing resources' abstraction through cloud computing technology and efficient access control techniques enhances system security [3].

The AI-based digital agricultural system has taken agriculture to another level. Applying this technology has enhanced crop production and enable real-time crop monitoring, yield prediction, and disease identification [4]. For example, disease and pest detection based on deep convolutional neural networks and weed spraying automation using computer vision technology have significantly increased work efficiency in a farm [5]. Also, AI has provided a capability for monitoring crop health issues and possible nutrient deficiencies in the soil efficiently [6]. AI has enabled farmers to utilize the local data, such as weather conditions, temperature, water usage, soil conditions, and many other factors collected from farms, to analyze them in real-time. These analyses help to grow healthier crops with optimum natural resources. Furthermore, AI delivers large-scale transformation through advanced approaches that redefine the traditional patterns and limitations of agriculture. AI helps lead the agricultural revolution in an era where we must produce more food with fewer resources [7].

Furthermore, Agri-technology and precision agriculture emerged as a new data-driven science field to improve agricultural productivity with minimal environmental impact [8]. Besides, digital agriculture has encouraged the modernization of agriculture to increase food production significantly [9].

From an economic perspective, improvements in the agricultural system using cloud computing technology and AI would deliver farmers' benefits since the agriculture industry plays a significant role in the global economy [8]. In particular, the significant sub-regions of Asia (South Asia, Southeast Asia, and East Asia) accounted for about 55% of the total global population and 73% of the world's agricultural population in 2012. Therefore, agriculture could continue to play an essential role in economic development [10].

Though the potential of cloud computing technology and AI is vast, related research has assessed effectiveness in the agricultural system's domain. However, not many systems have been available for farmers to use in a practical manner. This study presents the design and development of a digital agriculture platform using AI and cloud technology. We describe the benefits of using them based on our study and existing studies. Our contributions to this study are as follows:

1. The existing digital agriculture platform applied to AI and cloud technology has been analyzed and reviewed.
2. The proposed platform has been developed, taking into account AI and cloud technology for digital smart agriculture.
3. The benefits of using AI and cloud technology for the digital agricultural domain have been enumerated.

The rest of the paper is organized as follows: Sect. 2 delivers background studies and related works. Section 3 discusses the benefits of using cloud computing and AI-technology for digital agricultural platforms. Section 4 proposes the framework and presents the features of the plant disease diagnosis function. Finally, Sect. 5 concludes the paper.

2 Background Study

Roopaei et al. proposed the cloud-based smart agriculture system for water irrigation automation. The author's proposed system was based on the Cloud of Things networks that were capable of containing the Internet of Things (IoT) as well as a cyber-physical system. It provided data analysis and smart decision making. This system also delivered analyses and performance results to farmers with real-time data on the operations and made precise decisions. This system utilized edge devices and smartphones for data distribution storage, which processed offline data [9].

Singh et al. researched a cloud-based smart agriculture system with a rule-based decision module. This system was to collect and centralize various data from many IoT devices, the agriculture expert's input, and the user's input. Acquired data was stored in the database, and it would be used for analyzing and modeling for future decisions, which was agricultural productivity, especially in this study. Also, this system utilized the mobile application to facilitate access to the system for users. The application enabled users to input parameters for using the intelligent model and submit queries to obtain a professional response from agriculture domain experts. This study achieved a more timely and costly efficient system proposal compared to existing solutions [11].

Jinbo et al. proposed a modern agricultural platform based on a networking application with IoT devices, cloud computing, agricultural analysis, and big data technology. This proposed application provided the management function to agricultural products, especially for production components, agricultural crop growth, agriculture cultivation products based on the knowledge-based approach, and the sales and safety of agriculture products. Also, the data module in this platform allowed users to know climate conditions to help make decisions. Furthermore, the decision-making module provided product price forecasting, management decision analysis for productions, and market intelligence's multidimensional study [12].

Zamora-Izquierdo et al. designed a fully automated Precision Agriculture System distributed into three main components: a crop Cyber-Physical Systems (CPS), an edge computing layer, and data analytics and smart management at the cloud. This system utilized cloud computing for computer-intensive or global-view tasks performed on elements. Also, it enabled us to analyze collected data to make agricultural decisions with a distributed file system [13].

Ampatzidis et al. developed a cloud-based precision agricultural application with Artificial Intelligence (AI) applied data by Unmanned Aerial Vehicles (UAVs). The system delivered efficient processing, analysis, and visualization of UAV collected data, including a user-friendly and interactive interface. The AWS platform has been utilized to achieve a precise cloud-based application. First, the system invoked a tree detection algorithm based on Convolutional Neural Network (CNN) to detect tree locations on UAV image data. Also, this system allowed users to select the region of interest of uploaded images manually. The second layer with AI algorithms was then executed to obtain the total number of trees and tree gap counts, tree height and canopy area, and plant health indices for each detected tree. Those outputs were shown to users through the user-interface on the application [14].

Rajeswari et al. described smart agriculture systems' work collaborating with IoT technology, cloud computing, big data analysis, and mobile technology. This system stored data by IoT devices, crop information, and crop and fertilizer prices in the cloud database. In the significant data analysis component, MapReduce was applied for data processing and predictive analysis, especially for crop sentences, the schedule of cultivation for next crops, total crop production, and fertilizer requirement in the area of interest. This system was to give outputs and information through the mobile app [15].

Hori et al. proposed an application based on cloud computing, which helped improve agricultural production processes. They used web and mobile phone applications to provide four functions: sales and production planning, operation planning/result management, patrolling support, and cultivated land data management. Two data management technologies served all four features. One of the components was data storage, which contains environmental data, image, audio data, and acquired data by the edge and mobile devices. Another component was the data analysis, which was to find a value among data, and provided an analysis that helped business. The above functions allowed features to address various on-site staff requirements [16].

Foughali et al. developed a decision support system based on cloud computing and IoT-sensor network. Their system provided data visualization and notified the user of the thread of the late blight occurrence when an output made by the mathematical model exceeded a certain threshold. The proposed cloud platform visualized the data from different sensors and stored the collected data. Those stored data, such as location-specific weather data, were used for late blight disease forecasts as input parameters [17].

3 Benefits of Using Cloud Computing and AI-Technology

This section will discuss how it is beneficial to use cloud technology and AI for digital agriculture platforms based on the existing works and how benefits can be applied to our proposed system.

3.1 Benefits of Using Cloud Computing for Digital Agriculture Platform

Using cloud technology delivers three benefits: efficient data storage, strong platform management with external devices, and lastly, powerful computing capability for complex tasks.

A digital agriculture platform based on the cloud has enabled us to quickly scale up data storage and management. Also, this benefit is able to bring a boon for storing vast amounts of and various types of data on the platform. In addition, the cloud-based system allows us to design a distributed-file system to deal with big data analysis and provide efficient data management with tremendous amounts of data.

Second, cloud technology can benefit efficient and robust platform management with external devices, such as IoT devices, edge devices, remote-sensing devices, and mobile devices. The cloud can act as the core of the application and smooth those devices' management. For example, IoT devices acquire environment data then transmit it to the cloud without any complicated configuration, edge devices connected to the cloud will be executed by the cloud operations. Remote sensing devices, such as drones, can collect geospatial data then send it to the cloud to process it.

Third, cloud computing is capable of handling complex tasks, for example, big data analysis using various data types with a massive amount of data. The cloud-based platform can be configured with computational resources to suit the application needs; hence the cost contribution can be considered. Additionally, the cloud platform with the mobile app is able to utilize APIs controls for functions. Therefore, the mobile device's performance does not need to worry about since all computational tasks can proceed in the cloud. For example, executing crop yield prediction using the model based on a big data analysis on the cloud, then the only result will be sent to the mobile application. Furthermore, processing with multispectral or hyperspectral data requires a relatively high-performance environment, even so, cloud computing can handle this, and only outputs of the processing will be needed to send to the mobile application.

3.2 Benefits of Using AI-Technology for Digital Agriculture Platform

AI-powered digital agriculture platform has a considerable potential to improve productivity, solve agriculture issues, and reduce labor costs.

First, applying AI-technology would be able to deal with complex decision-making tasks as parametric estimation. For example, image recognition using deep learning, a part of AI technology, has been applied for computer vision-based disease diagnosis. Besides, complicated tasks that could not be handled in the past can be accomplished by using AI-technology.

Second, more accurate decision making or prediction could be provided by applied AI technology than conventional methods. In general, the intelligent system's performance depends on the data volume, and AI comes into its own when there is many data. As the previous section described, cloud technology enables us to handle massive volumes of data. Therefore, intelligent systems using AI technology have the potential to outperform traditional systems.

Third, AI-based systems and applications could help reduce labor costs and increase productivity. Labor shortages have always been a problem in agricultural industries. At this point, AI cannot replace all human operations, but it can help or be a part of those operations. For example, yield management with AI technology could contribute to farmers' productivity so as to let them know what action they should take to maximize their yield.

AI-based crop health monitoring and diagnosis would help farmers inform those crop's conditions before becoming a bigger problem. An irrigation system with IoT devices based on AI technology would bring smart environmental control to maximize productivity.

However, AI technology and cloud computing, used independently of each other, cannot provide a modern digital agricultural system. Thus, collaboration with AI technology and cloud computing is needed. In other words, AI supports the intelligent part, such as the human brain, and the cloud acts as the core part of the application for smooth and expansible system management.

3.3 Benefits of Using Cloud Computing and AI-Technology for Our Current Work

Our proposed system has utilized those benefits mentioned above. In terms of data storage and management, lots of different data types will be transmitted, such as images, sensor data, weather data, multispectral/hyperspectral data, et cetera. Therefore, applying storage containers and SQL-like data tables to enhance data storage and management within our system will help us manage data. Also, using a cloud environment means that it is easier to exchange data between components, such as pulling data for modeling. Our proposed digital agricultural platform involves many external devices. Besides, smooth interaction with those devices is required. For example, mobile devices will transmit data. Also, those devices have deployed APIs by the cloud. IoT devices and remote sensing devices will push environmental data to the system. Centralized device management is an essential factor to provide efficient system flow, especially for the agricultural platform. Our proposed platform needs sufficient computing resources to analyze data and make a decision-making function for users. Especially for agricultural domain decision-making, model development involves many kinds of data, such as images, sensor data, and multispectral data. The cloud-computing will provide enough computing resources to conduct them.

Moreover, using AI technology helps deliver more accurate and generalized models to users in our proposed platform. Our proposed system includes disease diagnosis function. Plant diseases could not be correctly identified by conventional technology. However, it can now be made possible by using AI technology. Another significant advantage of using AI is the ability to store large amounts of data on the platform. Furthermore, this benefit helps reduce labor costs, as the AI-based function provides smart decision-making. Besides, our system benefits farmers in planning preventive measures and taking action before the damage spreads.

4 Current Work

Our proposed framework provides the end-to-end and iterative digital agriculture system to farmers that collaborate with a mobile application and a cloud-based platform. Also, it involves a computer-vision based recognition using deep convolutional neural networks (DNNs) in our system. Those proposed functions are described below.

4.1 System Functionality

i. Disease recognition and nutrient deficiencies:

Deep learning has recently attracted much attention intending to develop an automatic and accurate image identification system and classification. Furthermore, DCNNs have delivered breakthroughs for image classification. Those algorithms have also been applied to diagnose and recognize plant diseases as well as nutrient deficiency. Our proposed system works to diagnose plant diseases and recognize nutrient deficiencies with DNNs-based algorithms using a smartphone camera. How to treat those diseases and deficiencies will be shown to users through an interface based on the diagnosis.

ii. Data Storage:

Collecting data is crucial to optimize DCNN's models. Even nowadays, many companies are focusing on getting more data to improve their AI-based systems for profit. Efficient data collection and storage must be considered in the system to facilitate model developments. The proposed system handles three kinds of data: images, time-series data, and metadata such as plant type, weather, and location. Those data will be sent from the user's smartphones, IoT devices, and remote sensing devices properly. The storage helps develop models and manage them efficiently.

iii. Data Annotation:

In order to develop DCNNs based models, using a vast amount of appropriately annotated datasets is required. Primarily, the development of an application with computer-vision based plant disease recognition models must involve the domain experts to obtain the proper labels. However, annotation tasks are quite time-consuming, so that our proposed system has implemented a competent and productive data annotation system. It enables the domain experts to conduct annotations smoothly as well as productively. Also, those annotated data can cooperate with the model development component.

iv. Model Development and Deployment:

Plant diseases cause yield losses and degradation of the quality of farmers' crops [18]. Also, a new plant disease would appear in accordance with the times and the environment. Therefore, to protect farmers from the threat, preparing the end-to-end model development and deployment system is vital to provide precise plant disease detection models to customers in the shortest possible time. The model development component has connected to the model deployment component seamlessly. Developed models will be deployed as REST API to provide desired functions to the users.

4.2 System Components

The main feature of our proposed system is that each component is connected iteratively so that data and model management has been streamlined, and it provides users with more accurate models. Figure 1 represents a proposed system architecture that applies the functions discussed above.

Furthermore, the details of each component with the flow of the system are as follows.

i. Data acquisition:

First, our developed mobile app has a function where users send images to diagnose them on our server. Simultaneously, metadata such as location, specific site, plant entity, and crop type will come with uploaded images. Moreover, the weather API has been implemented into the application to collect meteorological data according to the user's location. Our system also involves agriculture scientists

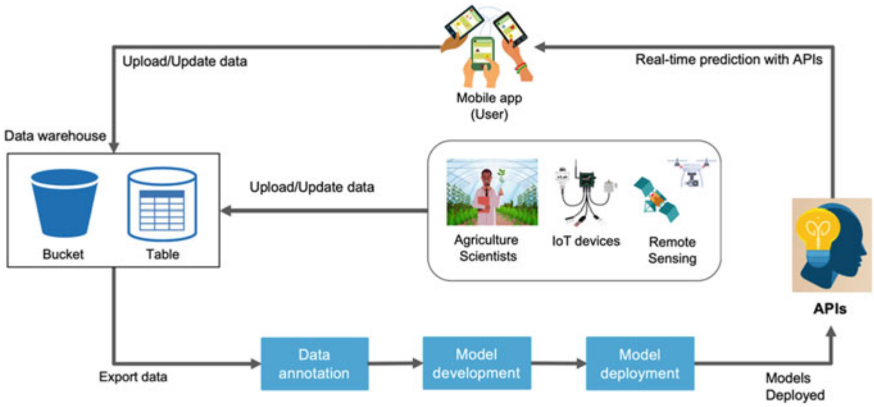


Fig. 1 The system architecture of digital agriculture platform

who can collect data in their actual experimental field and push them to our storage. IoT devices will collect environmental data such as soil conditions, temperature, humidity, leaves' wetness, et cetera. Moreover, remote sensing devices will obtain geological details and send it to our data storage as well.

ii. Data warehouse:

Our data warehouse consists of two types of storage: a bucket and a table. A bucket is called a storage container, and it can hold one or more records. In our proposed framework, the bucket is used for storing acquired images. Those images are stored by date in PNG format. Additionally, to manage non-image data, the NoSQL database has been adopted. Metadata, location data, and weather data will go into a specific table in the database. Also, each row refers to each image uploaded to the bucket. Therefore, it enables us to access and manage each data efficiently to prepare datasets for modeling. Furthermore, the bucket can be connected to the modeling environment without any configuration to download smoothly, and the NoSQL database is able to conduct model development.

iii. Data annotation:

The objective of this component is to conduct labeling precisely and accurately in our framework. In order to develop models, the labeled dataset must be necessary, and data labeling needs to be done by plant disease experts. Our system can facilitate annotation tasks and synchronize with the buckets that contain uploaded images smoothly. The annotation system can configure an admin and annotators accordingly, and an admin can assign labeling tasks to them. The annotated label's data will be saved in JSON format, and it will be able to be used for model development afterward.

iv. Model development and deployment:

The objective of this function is to provide an efficient environment to develop models for disease recognition and nutrient deficiency recognition. Also, this enables us to deploy developed models as APIs without any particular configuration. First, the annotated dataset is retrieved from the buckets. Then the dataset will be fed into the training scripts. In this job, TensorFlow will be used for developing models. GPU is available in this environment that provides fast-moving model training. Those trained model's outputs and trained weight parameters will be saved in the specific buckets. Finally, the trained model with saved weight parameters is pulled from the bucket. Afterward, the model will be deployed into API to publish models in order to enable users to recognize diseases and nutrient deficiencies using their mobile phones.

4.3 Results and Discussion

This section will discuss an application that takes into account the components and features mentioned above. First, a farmer submits images through the mobile application with the farmer's site information, crop type, plant entity, and crop cycle (Fig. 2a). Submitted images and metadata will be transmitted to the cloud server for

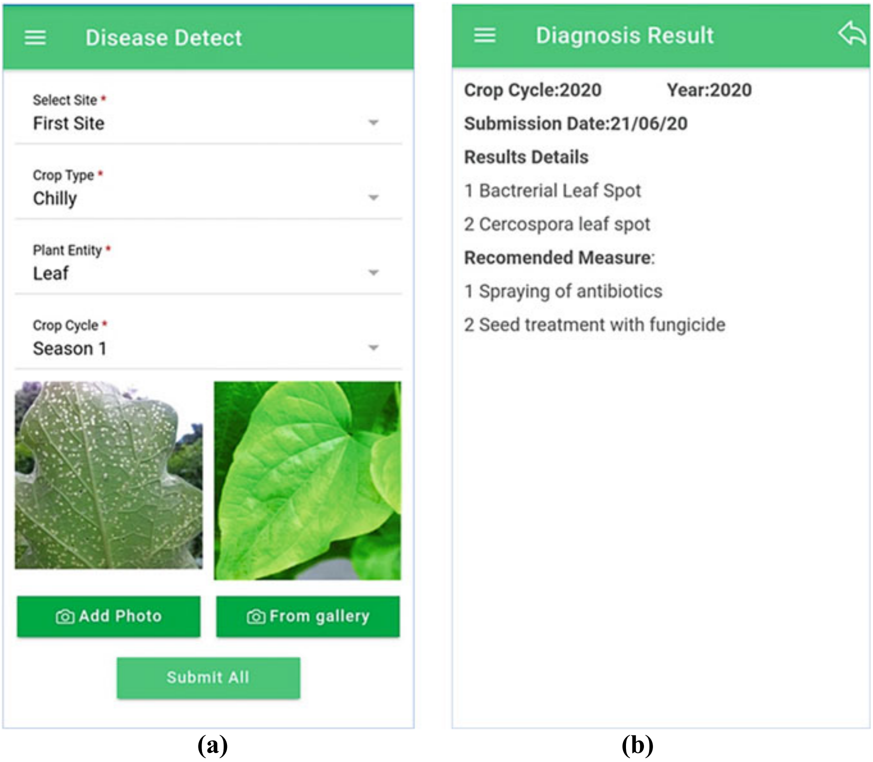


Fig. 2 **a** Images and metadata submission. **b** Diagnosis results report

Table 1 AUC Score of training and test on the model development

Models	Training AUC score	Test AUC score
MobileNet	0.9976	0.9061
MobileNet V2	0.9972	0.7907
NasNet Mobile	0.9985	0.8989
DenseNet 121	0.9994	0.9421
DenseNet 169	0.9970	0.8630

disease diagnosis. In the cloud, deployed disease recognition API will be invoked and diagnose uploaded images in there.

We developed a plant disease diagnosis model based on DNNs in terms of the API. For this study, we focused on chili diseases. Models classified four diseases: Bacterial Leaf Spot, Cercospora Leaf Spot, Mosaic Disease, Powdery Mildew, and one category for healthy images. A total of 1652 images was split into 80% for training and 20% for testing. This development involved 5 DNNs: MobileNet, MobileNet V2, NasNet Mobile, DenseNet 121, and DenseNet 169. These were compared to select a model with the highest evaluation score against unseen data. AUC-ROC score was adapted to evaluate the models for this time. Table 1 shows the AUC score of all trained models.

The system selected and deployed DenseNet 121, which obtained the highest AUC score against a test dataset.

Next, a diagnosis result and recommended measures will appear to a user based on the disease diagnosis. A diagnosis result shows the possible disease, and recommended measures indicate how to treat each disease. Diagnosis API derives each likelihood for each disease and then provide a disease name based on the likelihood. The mobile app provides diagnosis results to users through an interface (Fig. 2b).

Implementing the diagnosis function in API provides several benefits. First, it achieves independence on the users' device performance. The cloud instance will execute the plant disease diagnosis instead of the user's device when the API is invoked. Therefore, the proposed system can provide the same high-performance and high-speed diagnosis to users. Second, it enables us to manage diagnosis functions efficiently. The diagnosis models are developed and stored in the system. The system deploys those developed models to APIs. Thus, central management of models and APIs allows us to control their version and crop domains seamlessly rather than compile the models and implement them directly to the mobile device.

5 Conclusion

Applying cloud computing and Artificial Intelligence has the potential of increasing efficiency and productivity in digital agriculture systems. Cloud computing has provided mass storage to store various data types, computing resources, and efficient external device management. Artificial Intelligence technology has allowed us to develop smart decision-making functions, such as disease detection. The proposed platform provided the end-to-end and iterative digital agriculture system to farmers. This study utilized these technologies to achieve significant real-time disease detection outcomes and publish recommended measures for each disease to farmers. There were two limitations in this study. First, the discussed plant disease diagnosis models have involved only images. Second, other crop disease models could not be implemented in the proposed system. In the next stage of our research, we will further improve the diagnosis model's accuracy involving various data and attempt to carry out disease identification for other identified crops. We will also attempt to develop and implement disease prediction models for our digital agriculture platform.

Acknowledgements The first and second authors thank RC Buminiaga for financial, domain expertise, and technical support for this research.

References

1. Choudhary S, Jadoun R, Mandoriya H (2016) Role of cloud computing technology in agriculture fields, Computing 7(3)
2. Jinesh V (2011) Best practices in architecting cloud applications in the AWS cloud. Cloud Comput Princ Parad 18:459–490
3. Ojha T, Misra S, Raghuwanshi NS (2017) Sensing-cloud: Leveraging the benefits for agricultural applications. Comput Electron Agric 135:96–107
4. Talaviya T, Shah D, Patel N, Yagnik H, Shah M (2020) Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artificial Intelligence in Agriculture, 4, 58–73
5. Partel V, Charan Kakarla S, Ampatzidis Y (2019) Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence. Comput Electron Agric 157(January):339–350
6. Sudduth KA, Woodward-Greene MJ, Penning BW, Locke MA, Rivers AR, Veum KS (2020) AI down on the Farm. IT Profess 22(3):22–26
7. Liu SY (2020) Artificial intelligence (AI) in agriculture. IT Profess 22(3):14–15
8. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D (2018) Machine learning in agriculture: a review. Sensors (Switzerland) 18(8):1–29
9. Roopaei M, Rad P, Choo KKR (2017) Cloud of things in smart agriculture: intelligent irrigation monitoring by thermal imaging. IEEE Cloud Comput 4(1):10–15
10. Amrinto LE (2014) The worlds of agriculture in Asia : agricultural and economic development. May, p 148
11. Singh S, Chana I, Buyya R (2020) Agri-info: cloud based autonomic system for delivering agriculture as a service, Internet of Things, 9, 100131
12. Jinbo C, Yu Z, Lam A (2018) Research on monitoring platform of agricultural product circulation efficiency supported by cloud computing. Wireless Pers Commun 102(4):3573–3587
13. Zamora-Izquierdo MA, Santa J, Martínez JA, Martínez V, Skarmeta AF (2019) Smart farming IoT platform based on edge and cloud computing. Biosys Eng 177:4–17
14. Ampatzidis Y, Partel V, Costa L (2020) Agroview: cloud-based application to process, analyze and visualize UAV-collected data for precision agriculture applications utilizing artificial intelligence. Comp Elect Agricul 174(February), 105457
15. Rajeswari S, Suthendran K, Rajakumar K (2018) A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics. In: Proceedings of 2017 international conference on intelligent computing and control, *I2C2 2017*, *2018-Janua*, pp 1–5

16. Hori M, Kawashima E, Yamazaki T (2010) Application of cloud computing to agriculture and prospects in other fields. Fujitsu Sci Techn J 46(4):446–454
17. Foughali K, Fathallah K, Frihida A (2018) Using Cloud IOT for disease prevention in precision agriculture. Proc Comp Sci 130:575–582
18. Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A (2019) The global burden of pathogens and pests on major food crops. Nat Ecol Evol 3(3):430–439

Deciphering Consumer Behavior Through Emotions Using Neuromarketing

Alma Casas-Frausto, Bogart Yail Márquez, Samantha Jiménez, and Arnulfo Alanís

Abstract The main objective of marketing is undoubtedly to help companies link their objectives with the needs and preferences of current and potential customers, and learn more about customer or consumer behavior. At present, with the help of new technologies, there has been a greater interest in applying the methods of neurology in various areas other than medicine, which are related to human behavior, neuromarketing being one of the ways to investigate and know how the consumer reacts to a product, thus supporting us to better understand psychological behavior and dictate more precise actions for marketing based on knowledge of the brain reactions.

Keywords Neuromarketing · Emotions · Neural networks · Neuroscience · Encephalogram · Bioelectrical signals · Fuzzy systems

1 Introduction

Marketing helps companies link their objectives with current and potential customers' needs and preferences to learn more about customer or consumer behavior. It is therefore to know the reactions of the market to the different marketing actions, companies have traditionally resorted to their analysis by conducting surveys, the following being the most popular: market research, customer satisfaction surveys,

A. Casas-Frausto (✉) · B. Y. Márquez · S. Jiménez · A. Alanís
Tecnológico Nacional de México, Instituto Tecnológico de Tijuana, Mexico City, Mexico
e-mail: alma.casas@tectijuana.edu.mx

B. Y. Márquez
e-mail: bogart@tectijuana.edu.mx

S. Jiménez
e-mail: samantha.jimenez@tectijuana.edu.mx

A. Alanís
e-mail: alanis@tectijuana.edu.mx

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_53

purchase analysis, lead generation surveys, product inspection, advertising inspection, and price elasticity. Of product segmentation, product testing and market, in this technique the most common is that the potential customer tests the product without commitment; this helps to facilitate the decision in the realization of group dynamics or focus; this consists of bringing together a group of people (either homogeneous or heterogeneous). It is directed by a moderator and with a structure developed by him. Its principal function is to evaluate various aspects of a product/service. The different uses or possibilities that it could have are based on the perception and observations of the group participants, among other research activities, but these methodologies focus on knowing the conscious behavior, while the consumer's buying behavior is a fundamentally unconscious [1] part of the fundamental principle that all behavior is the result of a neurological process and manifests itself through the neurological system, to the way we use our five senses, to turn our experience into processes of both conscious and unconscious thoughts, because the ability to programming lies in our neurological activity. All humans have a brain and a nervous system that allow them to perceive their surroundings. These systems allow to think and feel, to select behaviors [2]. Emotions play a fundamental role in the purchase decision process. It is very evident that our purchases daily do not mark a definite and clear emotional state since they are significantly more evident in complex purchases and high involvement by consumers, or at least they are more perceptible [3]. Ekman found that there are a series of expressions that are universal and occur in both Western and Eastern world cultures. Ekman stated six basic expressions in 1972: anger, joy, surprise, disgust, sadness, and fear. Currently, new technologies developed a greater interest in applying the methods of neurology in various areas other than medicine, which relate to the human behavior. Neuromarketing is the study of the functioning of the brain and the neuron in the nervous system. It provides the knowledge and essential tools to understand diseases of the system. It is one of the ways to investigate and understand consumer behavior and make purchasing decisions, thus supporting us to better understand psychological behavior and dictate actions more accurate for marketing based on knowledge of brain reactions [1].

2 State of the Art

The nervous system is defined as the system through which the human being receives information about his environment, through its senses, processing said information and issuing responses [4]. There are methods to measure emotions in the processes of consumer purchasing decisions: Body language, Emphatic design, and Eye tracking [5].

2.1 Neuromarketing

Braidot determines that Neuroscience seeks to understand the functioning of the sensory system of the brain according to the encoding of information from an external panorama, that is, how the nervous system performs the conversion of large amounts stimuli to which an individual is exposed to brain language: activation and deactivation of neurons, interaction among neurons, emission of information manifestation, and neuroplasticity [4]. Neuromarketing is commercial marketing communication field that applies neuropsychology and neuroscience to market research [4]. Although neuromarketing does not offset traditional marketing perspectives, there is great certainty that functional magnetic resonance is a tool needed.

It will allow increasing the highest performance in the formulation and implementation of brand strategies. Of the brand and the diffusion is recognized that have a significant impact on consumer preferences for products.

2.2 10–20 System

The 10/20 system is a prestigious standard method used for extracranial electrode placement. The 10/20 numbers make reference to the 10 and 20 percent of distance in which the electrodes are on the front, occipital, right, and left sides of the skull. Each point has a letter to identify the lobe and a number to identify the location of the hemisphere. The letters mark the area (Fp, prefrontal; F, frontal; C, central; P, parietal; T, temporal, and O, occipital), while the the numbers designate the hemisphere (right even, left odd) and the midline electrodes are marked with a "z"; then Fz is found frontally in the midline.

2.3 Magnetic Resonance Imaging (FMRI)

The term "magnetic resonance imaging" from FMRI refers to a method that performs an anatomical sample of the brain through magnets (Postma 2012; cited by Zara and Tuta 2013). It is carried out using an FMRI scanner to calculate the degree of oxygen in the blood, indicating an increase in brain activity in some specific regions. The measurement is carried out as follows: the magnetic field can identify the blood's oxygen content in the brain. Therefore, while neuronal activity is increasing in a particular area of the brain, the blood increases, mainly because the brain needs oxygen to perform its function [4].

2.4 Consumer Buying Behavior

Regularly, the consumer purchase process is understood as the set of developed steps, with eventual backward returns, until reaching the final decision, translated into a purchase. There are numerous proposals in this regard. However, most coincide with the following phases: awakening, identifying, or recognizing the needs; collection and processing of information; formulation and implementation of the election; including assessment of consequences [3].

2.5 Emotion: A Key Factor for Marketing

The purchase process constitutes a complex of relationships and actions that until now continue to constitute a real mystery and an exciting challenge for those of us who dedicate ourselves to the noble and arduous task of research, in both directions, considering internal factors and those external to the consumer buying process itself. The internal elements are the most important, for example, emotions as a high impact axis for marketing purposes, neurosciences without this meaning that it is the only factor that influences the final consumer decision [2].

2.6 Neurology and Neuromarketing

Braidot relates brain's division into three systems (reptilian, cortex, and limbic system); each specializes in different tasks. The reptilian system specializes in regulating instinctual behaviors (breathing) and the most basic needs and emotions (eating). For its part, the limbic system focuses on learning, memory, and a large part of emotions. Finally, the cortex manages the processes associated with thinking and reasoning [5].

2.7 Neuroscience in Neuromarketing

There are various definitions of neuromarketing, but they all deal with how the brain reacts to deciding on a product/service. The rise of neuromarketing has advanced significantly in predictable market research, showing how feelings and unconscious responses influence customer decisions' perceptions and determination. Neuromarketing carries out a combination of marketing and neuroscience techniques with the goal of watching the nervous and emotional process that controls an individual's selection and actions, and examining these procedures would help clarify customer response to marketing. Moreover, adding neuroscientific approaches can help academics detect a more robust understanding of marketing, including customer behavior. Neuromarketing can entail several perceptible subjects. Daugherty and Homan implemented a restructured and upgraded taxonomy, including six different classifications, to frame current studies on desired marketing outcomes. The categories are shopper attention/activation, brand extensions, item/mark evaluations, shopping behavior, memory, and product favorite [6].

2.8 Main Techniques Used

Neuroimaging procedures are used in this area to test hypotheses, refine existing knowledge, and test the outcome of marketing stimuli in the consumer's brain. Research has already determined that brain activity patterns are closely related to behavior and reasoning [5].

Research using neuromarketing is carried out with the help of equipment that, until recently, was only used by medical science. The following are the most frequently used procedures: Positron emission tomography (PET), Functional Magnetic Resonance Imaging (fMRI), Transcranial Magnetic Stimulation, Steady State Topography, analysis of emotions through micro-expressions, Galvanic response of the foot, Eye-tracking, and Facial electromyography [5].

2.9 Neural Network

Artificial neural networks (RN) are a model of processing and machine learning inspired by the way the nervous system processes. It is a system of linkage of neurons that participate among themselves to execute an output stimulus [7].

3 Materials

The MUSE headband is a tool that measures brain signals as a heart rate monitor detects heartbeat. It has seven sensors finely calibrated that detect and measure brain activity, located in two, on the forehead, two behind the ears, and three reference sensors. Broadband MUSE is a headband which will help us perform an encephalogram and obtain bioelectrical signals from the brain. The headband consists of seven dry sensors, four input channels, and five output channels. It facilitates access and use of brain wave data [8].

3.1 Detections in an Encephalogram

An electroencephalogram (EEG) can detect a series of specific frequencies grouped into frequency bands associated with a particular area of the brain or with particular activities. The types of signals emitted by the brain capable of being detected by an encephalogram are the following [9].

3.2 Encephalography Protocol

3.2.1 Process

It is carried out by collecting the electrical process of the cerebral cortex using surface electrodes. The received signal is minimal, that it becomes necessary to use various amplification systems. The amplifiers used are differential, obtain an electrical impulse of 2 points, and amplify the difference of the way the electrodes are placed on the head which is subject to the international system [10].

3.2.2 Preparation and Performance for Performing Encephalography

- The patient should be relaxed.
- The level of alertness or confusion is noted before, during, and after the process.
- Take data from the patient, giving greater importance to age.
- Information is collected about the pathological process, so the test and prescribed medication are performed.
- Explain the dynamics of the test and its absolute innocuousness, requesting the collaboration, without which it is not possible to carry it out.
- Inform the patient that the EEG is not contraindicated in any case, nor does it have any side effects, so no sign of any kind is necessary for consent and/or authorization.
- It is highly recommended to clean the hair and the absence of foams or gummies to avoid or minimize artifacts.
- Place the hat having as a reference, that the Frontopolar electrodes are located on frontal prominences so that, in this way, the occipital electrodes rest on the occipital region and not over the cerebellum.
- Introduce conductive gel on each electrode, fixing them as much as possible, to avoid movement artifacts.
- Check that the impedances are adequate (good conductivity at the registration point, between the electrode and skin) and record them for later review.
- Use the appropriate filters, both high and low; low frequency allow us to correct, some the fast rhythms and others slow ones; always looking to obtain a more optimal signal for later interpretation.

- Annotations of each moment are required, which occur throughout the record, movements, blinking, drowsiness, sleepiness, nervousness, swallowing, etc., to be taken into consideration when interpreting [10].

4 Emotion Classification Methodology

The methodology implemented is based on neuronal networks and simplifications of the biological model to propose its mathematical development. To carry out the classification process, it is constituted of 4 inputs, five outputs, and the weights for each input [7].

The inputs are captured using electrodes capturing the bio-electrical signals in which it consists of collecting by electrodes, the electrical activity of the cerebral cortex. The signal obtained at the moment is averaged. The amplifiers used are differential, that is, they take in the electrical signal from 2 points and amplify the capacity difference among them [7]. Recognition of brain operation is calculated using generated potentials, said encephalogram components arising in response to an optical impulse of 3 s per picture and 4 s of counteraction each with three images, and utilization the internationally standardized 10–20 positioning method [7].

4.1 Learning Rules

For the knowledge rule, the network structure has to be determined. We have four inputs and five outputs. For the provided neural network, there are hidden layers where there are also nodes. The output and function of the node

$$i_l \; (i_l = 1, \ldots, N_l) \tag{1}$$

in the cape

$$l \tag{2}$$

are represented as

$$a^l_{i_l} \tag{3}$$

and

$$f_l \tag{4}$$

This class of learning is still controlled since it uses a loop of propagation and adjustment in phases. A pattern has been executed to the input of the network as an impulse. It spreads through the upper layers of the network, generating an output. The output signal is likened to the wanted output, and the error for a piece of the outputs is evaluated. The error outputs spread backward, starting from the output layer, to all the neurons in the hidden layer that directly provide the output. Nevertheless, the neurons of the hidden layer only obtain a segment of the total error signal.

$$\varepsilon^2 = -2 \cdot e \tag{5}$$

The data for training is made up of several pairs of input and output training patterns. The input is the stimuli that are captured from pictures in a range of 9 s; the fact of knowing the output entails that the training benefits from monitoring. Moreover, it gives a new preparation pattern. The weights are tailored next:

$$W_{\text{new}}^{l,l-1} = W_{\text{old}}^{l,l-1} - n^l . \delta^l . \left[\overline{a}^{l-1} \right]^T \tag{6}$$

For learning weight

$$w_{i_l, i_{l-1}}^{\sim l,l-1} \tag{7}$$

updates only after presentation of the entire dataset or only after an epoch where

$$\delta_{il}^l \tag{8}$$

is the designate error.

The error thrown by a neural network as a function of its weights generates a space of n dimensions, where n is the number of connection weights. When evaluating the error gradient at a point on this superficies, the direction where the error works will have a maximum growth, as the purpose of the learning process is to minimize the error [11].

5 Medium Quadratic Error

The learning problem in a neural network is search trouble. Space is allocated both by the structure of the network and the value space of the synaptic weights. Given the structure of the network, it is intended to upgrade operation. This is done by minimizing the mean square error [11].

$$mse \equiv \frac{sse}{qN_2} = \frac{\sum_{p=1}^{q} \sum_{i_2=1}^{N_2} \left(t_{2,p} - a_{t_2,p}^2 \right)^2}{qN_2} \tag{9}$$

6 Conclusions

Without a doubt, neuroscience offers us the possibility of optimizing market research results through the application of neuromarketing. However, those responsible for marketing are not yet convinced to implement this method because they are not sure about the capacity it offers before collecting data that can help them define consumer behavior. Previous research in neuromarketing shows us that this technique will give us accurate data that will help make decisions. Neuromarketing's best application predicts consumer behavior, which is the most significant challenge facing traditional marketing: the relationship between mind and behavior. Advances in Neuromarketing will make it possible to choose the format that best suits the market study. The consumer says things that do not always coincide; therefore, if artificial neurons are applied in neuromarketing for market research, this will generate reliable results.

References

1. Ronda PC (2013) Neuromarketing, el futuro ya esta aqui? 3c Empresa: investigación y pensamiento critico 2(7):3
2. Requena A (2005) Ramiro el neuromarketing. Una visión diferente para entender al cliente perspectivas, vol 8, no 2. Universidad Catálica Boliviana San Pablo Cochabamba, Bolivia, pp 83–97
3. Coca Carasila AM (2010) Neuromarketing: las emociones y el comportamiento de compra perspectivas, núm. 25, enerojunio. Universidad Católica Boliviana San Pablo Cochabamba, Bolivia, pp 9–24
4. Muñoz Osores M (2015) Conceptualización del neuromarketing: su relación con el mix de marketing y el comportamiento del consumidor Conceptualisation of neuromarketing: its relationship with the mix of marketing and consumer behavior por qué es importante el neuromarketing 1(2):91–104
5. Camacho V (2016) Oscar Neuromarketing y la ética en la investigación de mercados Ingenieria Industrial. Actualidad y Nuevas Tendencias, vol V, no 16. Universidad de Carabobo Carabobo, Venezuela, pp 69–82
6. Nilashi M, Yadegaridehkordi E, Samad S, Mardani A, Ahani A, Aljojo N, Razali NS, Tajuddin T (2020) Decision to adopt neuromarketing techniques for sustainable product marketing: a fuzzy decision-making approach. Symmetry 12. https://doi.org/10.3390/sym12020305.5285
7. Marquez Lobato BY, Alanis Garza A (2017) Classification algorithm for measuring human emotion: 'i like it' and 'i do not like' in Neuromarketing. IEEE Lat Am Trans 15(11):2177–2184
8. Lee GJ, Suhr JA (2019) Expectancy effects on self-reported attention-deficit/hyperactivity disorder symptoms in simulated neurofeedback: a pilot study. Arch Clin Neuropsychol
9. Alejandro R, Gimeno A, Cucarella EQ (2018) Procesamiento de artefactos en EEG para aplicaciones de comunicación y control
10. Talamillo García T (2011) Manual básico para enfermeros en electroencefalografía," Enfermería Docente, no 94, pp 29–33

11. Márquez BY, Garza AA, Palencia JSM, Romero Alvarado K (2018) Neural network algorithm to measure the human emotion. In: Proceedings of the 2018 international conference on algorithms, computing and artificial intelligence—ACAI 2018
12. Olivas N, Guitierrez C, Bribiesca LJ (2010) Mapeo electroencefalográfico y neurofeedback

A Review of Internet Topology Research at the Autonomous System Level

Timotius Witono and Setiadi Yazid

Abstract Internet topology research is often done at the autonomous system level because of its highest level of granularity and the availability of several public datasets. The purpose of this literature review is to discover the reasons why Internet topology research at the level of the autonomous system needs to be carried out, to observe how related research is carried out, and to list the limitations of current research and the possibility of future research in this field. Literature search as sources of literature review was conducted on three digital libraries with the publication time of the last 10 years; the search results were then selected to fit the inclusion criteria. The results of the literature study show that the main reasons for research in this area are to measure the growth and development of the evolving Internet topology, to map the structure of the Internet topology that is not centralized and managed by various parties, and to model the Internet topology in order to improve its performance, while the weakness that remains a major problem in research in this field is the problem of the incompleteness of data sources due to limited public access to private resources on the Internet topology.

Keywords Internet topology · Autonomous system · Internet modeling · Internet mapping · Internet measurement

1 Introduction

Internet topology can be inferred as the connection structure of many components, such as routers, hosts, and autonomous systems. Research in the area of Internet

T. Witono · S. Yazid (✉)
Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
e-mail: setiadi@cs.ui.ac.id

T. Witono
e-mail: timotius.witono@ui.ac.id; timotius.witono@it.maranatha.edu

T. Witono
Faculty of Information Technology, Maranatha Christian University, Bandung, Indonesia

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_54

topology is driven by the huge advancement of the Internet and also many problems that follow the progress. Since its involvement in many aspects of people's daily life, research for improving the Internet is very important, so is the case with research in the field of Internet topology. Internet topology is a field of research that attracts many researchers to gain further understanding of the Internet. The division of the level of the Internet topology can be divided into interface level, router level, point of presence (POP) level, and autonomous system (AS) level. One of the most prominent Internet topology studies is at the level of the AS; this is due to the highest level of granularity and its ease of gathering data sources [1–3]. AS also can be defined as *network of networks*; AS continues to grow with independent administration for each of them. The growth and the interconnection of AS are done independently by interacting AS administrators and not by a single central authority; in this viewpoint, the Internet is an independently organized complex system. Ignorance of this independent structure can cause trouble in the future because of unforeseen problems; therefore researchers analyze the Internet in order to learn about its structure and development [4].

There are several motivations that stimulate researchers in the field of Internet topology at the AS level, e.g., to see how the Internet looks like, to get information about certain topological properties, to predict the future progress, and to generate Internet topology for simulation purposes [1]. In general, motivation in network topology research is divided into four categories as follows: scientific *(to understand the generic properties of specified network)*, adversarial *(to obtain proprietary information about rivals)*, managerial *(to ensure that their network has been recorded completely and accurately)*, and informational *(to get general information about their provider's network)* [5]. Internet topology studies at the AS level often use data derived from the Border Gateway Protocol (BGP), which is usually referred to as BGP dumps, as sources of research data. BGP is used by ASes to communicate and exchange routing information to each other. In regard to BGP, there are two long-term projects which continually collect BGP table dump from many collectors in various vantage points in the world; they are Route Views (RV) Project by University of Oregon and Routing Information Service (RIS) by RIPE Network Coordination Center (RIPE NCC). Various studies show that getting a complete Internet topology is very difficult; while getting a complete set of ASes is quite easy, the hardest part is to obtain the relationship between them. There are two main things that hinder the creation of a complete Internet topology, namely the decentralized nature of Internet management and the nature of proprietary network management from commercial Internet Service Providers (ISPs); these things make the goal of obtaining an accurate and complete Internet topology remain an open problem in research in the related field. This problem of lack of accuracy and completeness in the representation of Internet topology can have a worse impact, given that much further research uses this resulting topology as the basis for their experiments [6].

The purpose of this literature review is to find out the various contributions in research on Internet topology at the AS level. The second section discusses the method of literature searching in several digital libraries; initially, the research questions were set as the driving direction of the research, then the search for the literature with appropriate keywords was carried out, and finally the selection process was car-

Table 1 Literature search results on digital libraries

Digital library	Result
ACM DL (https://dl.acm.org/search/advanced)	90
IEEE Xplore (https://ieeexplore.ieee.org/search/advanced)	52
Scopus (https://www.scopus.com/search/form.uri?display=advanced)	518
Total	660
Records after duplicates removed	582
Records after screened	79
Studies included in qualitative synthesis[a]	61

[a]Inclusion criteria: full-article can be accessed, written in English, the contents of the article are related to the research questions

ried out in stages. The third section explains the results obtained from the literature review and discussion of these results, divided into four subsections according to the research questions related to Internet topology at the AS level. The fourth section tells about the conclusions obtained from the literature review process that has been done related to Internet topology at the AS level.

2 Method

The search for various literature is based on research questions that have been set to be the basis of this literature review research. The following are established research questions (RQ) related to the Internet topology studies at the AS level:

1. *(RQ1) Why do they need to be studied?*
2. *(RQ2) How are they carried out?*
3. *(RQ3) What are the limitations of current studies?*
4. *(RQ4) What are the future studies of them?*

Three digital libraries that have been determined to search literature are Association for Computing Machinery Digital Library (ACM DL), Institute of Electrical and Electronics Engineers Xplore (IEEE Xplore), and Scopus. Searches in digital libraries are performed with the following keywords: `( ALL ( "internet topology" ) AND ALL ( "autonomous system" ) ) AND PUBYEAR > 2009 AND PUBYEAR < 2020`. Search keywords are structured to provide literature results containing the keywords *"internet topology"* and *"autonomous system"* in all fields of literature, with publication time from 2010 to 2019. The process of searching and screening the literature is carried out from August 2019 to December 2019.

The search results on the ACM DL produced 90 literature that matched the search criteria, while the search results on the IEEE Xplore obtained 52 literature results. The search results using the Scopus digital library obtained results totaling 518 literature for the period of publication 2010–2019. The total number of search results on the

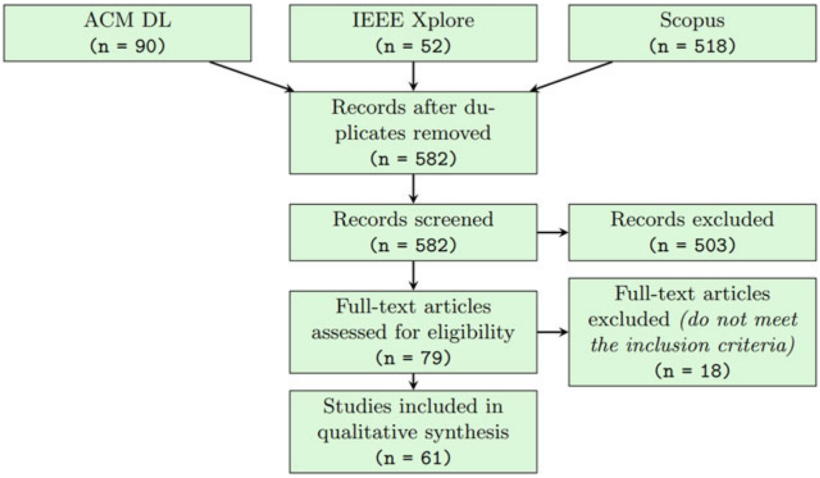


Fig. 1 The process of searching and selecting literature from three digital libraries [7]

three digital libraries is a total of 660 literature, as can be seen in Table 1; after discarding duplicated results, the total results were 582 literature that matched the search criteria. 582 literature search results are then filtered based on the suitability of their titles and abstracts with the aim of the study, resulting in 79 literature that are ready to be assessed for eligibility. Full-text reading was carried out on 79 selected articles and reselected based on inclusion criteria that have been set; inclusion criteria include "*full-article can be accessed*", "*written in English*", and "*the contents of the article are related to the research questions*". In accordance with the flowchart in Fig. 1, the final results of the search and selection process yielded 61 literature which became reference sources for the literature review in this study.

3 Result and Discussion

The results of the literature review are discussed in four subsections according to the four research questions that have been set; these are research questions related to Internet topology studies at the AS level: *(RQ1) Why do they need to be studied?*, *(RQ2) How are they carried out?*, *(RQ3) What are the limitations of current studies?*, and *(RQ4) What are the future studies of them?*. In each subsection, the results of the literature review are also divided into three parts. The first two parts are for the results of a literature review on "*Internet topology modeling*" and "*Internet topology mapping*", while the last part is for the results of a literature review on "*Internet topology measurement*". This subdivision was chosen by considering three broad categories of research in the field of Internet topology at the AS level, namely modeling, mapping, and measurement. Figure 2 shows the grouping of reference

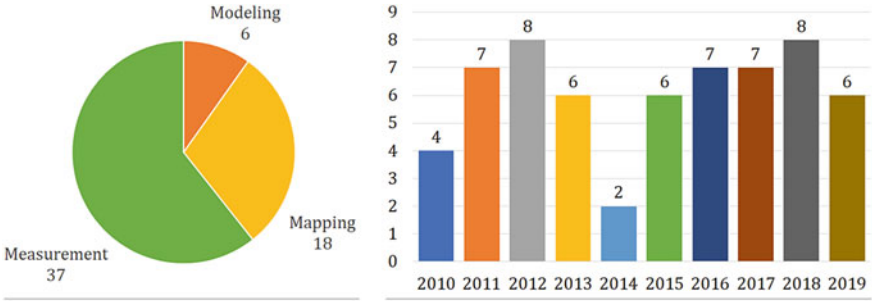


Fig. 2 The number of reference literature grouped into three research categories in the field of Internet topology—modeling, mapping, and measurement—*(left-side box)* and the number of reference literature classified in the year of their publication *(right-side box)*

literature based on the three Internet topology research categories and the grouping based on the year of publication of these literature. There are 37 reference literature that fall into the category of Internet topology measurement research, while for the categories of Internet topology mapping and modeling, there are 18 and 6 reference literature, respectively.

3.1 (RQ1) Why Do They Need to Be Studied?

The first research question is "*Why do they—Internet topology modeling, mapping, and measurement—need to be studied?*". A summary of the results of the literature review for this question is written in Table 2, while the discussion about the reasons for Internet topology research is written in the paragraphs in this subsection, divided into three parts, namely each for Internet topology modeling, mapping, and measurement.

3.1.1 Reasons Why Internet Topology Modeling Needs to Be Studied

The main reason for the study of Internet topology modeling is the aim of developing the Internet to a better level [8]. This reason is still relevant to current conditions; the Internet is still growing to a better level supported by research in the field of Internet topology. With ongoing demands for the development of the Internet, this reason will continue to be one of the main reasons for research in the field of modeling Internet topology. Another reason is to measure the performance of the Internet topology, with many components that make up the Internet [9]. Various points of view can be carried out for this kind of research, starting from the perspective of complex networks, or with graphs and other mathematical models. Modeling with a mathematical approach makes it possible for researchers to calculate the performance of certain components on the Internet. With the dynamic nature of components on the Internet, this reason

Table 2 Why do Internet topology modeling, mapping, and measurement need to be studied?

Reasons why Internet topology modeling needs to be studied
1. To develop the Internet to a better level [8]
2. To be able to measure the performance of the Internet topology and its related components from the perspective of complex networks, graphs, or other mathematical models [9]
Reasons why Internet topology mapping needs to be studied
1. To find out significant changes that occur in the Internet architecture and how the Internet is regulated in each region and how it relates to various factors in the region [10, 11]
2. To detect and analyze problems, and to do network optimization [12]
3. To improve understanding of the Internet economy and also improve routing models [13]
4. To understand the structure of the Internet and its development [14–17]
5. To provide an overview of the Internet ecosystem [18–20]
6. To understand Internet topology regionally [21]
Reasons why Internet topology measurement needs to be studied
1. To quantify the evolution and trends of the Internet topology from various points of view over a certain period of time, because this will greatly impact the Internet ecosystem and future Internet architecture design [22–33]
2. To analyze changes in the Internet ecosystem in a certain area in a certain time period for various needs [34–39]
3. To find out the quality of the Internet topology that can be inferred and provide a strategy to increase it, because an accurate Internet topology is critical for future evaluation and development of the Internet [6, 40–43]
4. To analyze the impact and development of Internet eXchange Point (IXP) on regional and global Internet topology ecosystems, because IXP plays an important role in interconnection and can be a representation of the Internet [44–47]
5. To identify critical points, serious problems, and the impact of attacks on Internet topology [48–50]
6. To provide insight into complex Internet structures and insights for the decision-making needs of certain business or policy decisions [51–53]

can still be the basis for researchers to conduct research in the field of modeling Internet topology.

3.1.2 Reasons Why Internet Topology Mapping Needs to Be Studied

One of the most important reasons for Internet topology mapping research is to understand the structure and development of the Internet [14–17]. The results of the mapping will be able to provide an overview and understanding of the Internet topology ecosystem, both globally and regionally [18–21]. The reasons for this Internet topology mapping will continue to encourage researchers, due to the dynamic development of Internet topology. The results of Internet topology mapping produced today are not necessarily relevant in the future, so researchers can still expect new findings from the results of similar studies conducted at different times. Internet topology mapping results can also be used to understand the Internet economy and also for technical matters related to the development of routing protocols, problem detection and analysis, and network optimization [12, 13]. Likewise, recent research uses the results of Internet topology mapping to see significant changes in the architecture of the Internet in a particular geographical area. Even mapping results can also be used to analyze Internet regulation in an area, as well as correlation analysis with various factors in that area [10, 11].

3.1.3 Reasons Why Internet Topology Measurement Needs to Be Studied

One of the most continuous researches in the field of Internet topology measurement conducted over the past decade is to measure the evolution and trends of Internet topology. The researchers took their measurements from various viewpoints and certain time frames. The researchers conducted these studies because the results of measuring trends, and evolution of the Internet topology can have a significant impact on the running Internet ecosystem, or on the development of Internet architecture design in the future [22–33]. Measurement analysis of changes that occur in the Internet ecosystem at certain time periods is very useful for various parties, which is why this is also a reason for researchers to take measurements of Internet topology [34–39]. For example, for decision makers, the results of the measurement of Internet topology as a complex system can be used to assist the decision-making process of certain policies or taking business decisions [51–53]. Likewise, recent research uses the results of Internet topology measurements to identify the impact of attacks on Internet topology, along with the identification of critical points and serious problems caused [48–50]. Measurements of specific components on the Internet are also carried out by researchers, for example, measurements of the impact of Internet eXchange Point (IXP) on the Internet ecosystem, both globally and regionally. These studies are considered important, because IXP plays an important role in Internet interconnection and at the same time can be a representation of the Internet ecosystem [44–47]. Measurement of the quality of Internet topology inference results is also still a reason for researchers to conduct research in the field of Internet topology measurement. This is due to the need for an accurate Internet topology for the process of future Internet development [6, 40–43].

3.2 *(RQ2) How Are They Carried Out?*

The second research question is "*How are they—Internet topology modeling, mapping, and measurement—carried out?*". A discussion of the methods used in research in the field of Internet topology will be presented in this subsection, while Table 3 summarizes them.

Table 3 How are Internet topology modeling, mapping, and measurement carried out?

Ways to do Internet topology modeling
1. By applying Triangular Connecting Mechanism (TCM) to an evolving network model [8]
2. By creating a level structured network model on Internet Content Provider (ICP) and IXP using a complex network perspective [9]
3. By using a network graph model that represents IXP with its AS members [54]
4. By utilizing a time series mathematical model to dynamic Internet resources for the purposes of forecasting and event detection [55]
Ways to do Internet topology mapping
1. By using the geolocation of Internet Protocol (IP) prefixes to estimate the geographical location of ASes to meet certain purposes [11, 12, 21]
2. By using certain methods or proposing new algorithms to infer various types of AS relationships—as completely as possible—based on various BGP data, traceroute, geolocation, traffic flow, WHOIS, looking glass, and other supporting data [13–20, 56–59]
3. By using the results of Internet topology mapping in the form of AS graph as a basis for certain experiments and analyses [10, 60]
4. By surveying Internet topology data collection techniques and their limitations, as well as presenting the latest efforts on Internet topology modeling and mapping [2]
Ways to do Internet topology measurement
1. By applying and proposing certain theories or algorithms as a means to investigate certain behaviors in Internet topology, such as trends and evolution, comparison and classification, as well as structural changes [22, 23, 25, 27, 29]
2. By analyzing historical data from various sources over a period of time, with the aim of measuring evolution, trends, behavior, or other specific things from the Internet topology ecosystem at the AS level [24, 26, 28, 30–33]

3.2.1 Ways to Do Internet Topology Modeling

The most common way to do Internet topology modeling is to make a mathematical model. One that is widely used is to use a network graph model to represent nodes and their interconnections, for example to represent interconnections between members of IXP. Researchers also use a mathematical time series model to represent dynamic Internet resources, so that they can make certain predictions or detect certain events on the Internet topology [54, 55]. Sometimes, researchers also use specific mathematical modeling, for example, the use of Triangular Connecting Mechanism (TCM) to model the development of Internet topology networks [8]. Other modeling approaches use complex network perspective, for example, in research that builds a level structured network model on Internet Content Provider (ICP) and IXP [9]. Mathematical model approaches are still relevant ways for modeling Internet topology in subsequent studies, whereas the two theories that are by far the most suitable for modeling and analyzing the results of Internet topology are graph theory and complex network theory.

3.2.2 Ways to Do Internet Topology Mapping

The most common way to do research in the field of Internet topology mapping is to infer various types of AS relationships. The method for inference is often done by proposing new algorithms for relationship inference—or sometimes using certain other methods—to get the AS relationships as complete as possible on the Internet topology map. The dataset used in these studies varies, ranging from various BGP data, traceroute, looking glass, WHOIS, geolocation, traffic flow, or other supporting data [13–20, 56–59]. The accuracy and completeness of the AS relationships inference algorithm is still an open problem, because there are still inaccuracies or incomplete relationships between ASes from various methods that have been proposed. Further analysis of the results of Internet topology mapping was also carried out by several researchers, by conducting certain experiments and analyses using AS graph as the results of Internet topology mapping [10, 60]. There are also classic studies that carry out a survey of the latest efforts of Internet topology mapping studies, or surveys of ways to collect Internet topology data and their limitations [2]. Some recent studies involve geolocation of Internet Protocol (IP) prefixes as a basis for determining the location of ASes [11, 12, 21]. The direction of research involving geolocation is estimated to still be something that will be done by researchers in the future. The involvement of several recent research trends—such as big data or deep learning—is also likely to be involved by researchers in the process of data analysis in subsequent Internet topology mapping studies.

3.2.3 Ways to Do Internet Topology Measurement

Internet topology measurement methods are often carried out by implementing certain theories or algorithms as a means of investigating the behavior of the Internet topology. Some behaviors are often measured in research, such as structural changes, comparisons, and classifications, as well as trends and evolution of Internet topology [22, 23, 25, 27, 29]. The mechanisms of measuring some technical things on the Internet topology—such as its evolution, trends, behavior, and several other specific things—are often done by analyzing historical data from various sources within a certain time period [24, 26, 28, 30–33].

3.3 (RQ3) What Are the Limitations of Current Studies?

The third research question is "*What are the limitations of current studies—on Internet topology modeling, mapping, and measurement—?*". Limitations of studies that have been carried out by researchers in the field of Internet topology will be discussed in the paragraphs of this subsection, and Table 4 summarizes these limitations.

Table 4 What are the limitations of current studies on Internet topology modeling, mapping, and measurement?

Limitations of current studies on Internet topology modeling
1. On the deficiencies in accuracy, completeness, ease, and application [54]
Limitations of current studies on Internet topology mapping
1. On the incompleteness of dataset for mapping [11, 13, 57, 58]
2. On the inaccuracy of the mapping results [11, 12]
3. On the limitations of the scope of the mapping results [12, 13, 57, 58]
4. On specific limitations in the mapping effort, such as BGP community requirements on the route server and the inability to associate ASes with their parent organization [20, 57, 58]
Limitations of current studies on Internet topology measurement
1. On the incompleteness of dataset for measurement [3, 47]
2. On the limitations of data obtained from looking glass servers [43, 46]
3. On specific limitations in the measurement effort, such as ambiguity to multigraphs, mismatch of simulation results with real conditions on the Internet, limited number of ISP data, and data glitch in the dataset [3, 23, 61, 62]

3.3.1 Limitations of Current Studies on Internet Topology Modeling

Research in the area of Internet topology modeling still has several major deficiencies. The first limitation is the lack of accuracy, so that the resulting Internet topology model is not accurate enough to represent the actual Internet topology. The next limitation is the problem of completeness of the resulting Internet topology model, so it cannot provide an overall representation of the real Internet topology. Both of these limitations, namely the lack of accuracy and completeness, still open opportunities for researchers to conduct further research in order to improve the accuracy and completeness of the Internet topology model. Another limitation that also remains an open problem in this area of research is the lack of ease of the Internet topology model when it will be applied applicatively [54].

3.3.2 Limitations of Current Studies on Internet Topology Mapping

Research in the field of mapping Internet topology still has some prominent limitations. One limitation that remains a classic issue is the incompleteness of the data sources used for mapping; because the dataset used is incomplete, the results of the mapping are also incomplete [11, 13, 57, 58]. Another classic problem is regarding the inaccuracy of the results of the Internet topology mapping, so that the Internet topology map from the research results cannot accurately represent the actual topology of the Internet [11, 12]. These two classic problems can cause other problems, such as the limited scope of the resulting Internet topology maps [12, 13, 57, 58]. These problems are still open issues that are still relevant and can be responded to by researchers in the field of mapping Internet topology in subsequent studies.

However, there are also limitations that are specific in some studies, for example, the requirement for certain features such as BGP communities on route servers for the research process, or limitations of research due to the inability to do certain things such as associating ASes with their parent organizations [20, 57, 58].

3.3.3 Limitations of Current Studies on Internet Topology Measurement

Research in the field of Internet topology measurement still presents several limitations that arise from the process or research results. One of the main things is the limited availability of data sources for measuring Internet topology; the incompleteness of the dataset for the process of measuring Internet topology is a problem that has not been fully resolved [3, 47]. Likewise, the strategy to add measurement data from other sources, for example, by adding data from looking glass servers, has not completely resolved the problem [43, 46]. There are also some specific research limitations, such as data errors in datasets, data ambiguity, and specific data limitations (for example, ISP data). The limitations possessed by previous research in the field of Internet topology measurement can result in a mismatch of simulation results with real conditions on the Internet, which are often complained by subsequent researchers [3, 23, 61, 62].

3.4 (RQ4) What Are the Future Studies of Them?

The final research question is "*What are the future studies of them—Internet topology modeling, mapping, and measurement—?*". Further research opportunities in the field of Internet topology modeling, mapping, and measurement will be discussed in this subsection. Further research opportunities are discussed in paragraph narratives and a summary of them is included in Table 5.

3.4.1 Future Studies of Internet Topology Modeling

There are several further research directions for researchers in the field of Internet topology modeling. One of them is from the article which suggests that research on the structure and evolution of ASes must also take into account critical elements on the Internet, such as economic aspects, technological aspects, and social aspects [64]. Internet topology evolution modeling at AS level generally uses graphs; research using other general complex networks besides graphs is something that can be done by future researchers [63]. There are also specific research proposals in the field of Internet topology modeling, for example, analysis of stochastic variations on global AS data reachability and also on event detection [55].

Table 5 What are the future studies of Internet topology modeling, mapping, and measurement?

Future studies of Internet topology modeling
1. Analysis of stochastic variations on global AS reachability data and also on event detection [55]
2. Evaluation of the AS-level Internet model uses other general complex networks besides using graphs [63]
3. Study of the structure and evolution of ASes by also taking into account the social, technological, and economic aspects—as critical elements on the Internet—[64]
Future studies of Internet topology mapping
1. Improvement of the AS geolocation accuracy with more detailed resolution of geographic tagging algorithm [11, 12, 21]
2. Development of inference relationship methods to improve the valid inference results of various types of AS relationships in a wider scope [13, 14, 16, 17]
3. Specific future developments, such as data integration and feedback for active measurement improvement, efficient algorithm for IXP peering finding, implications of the ongoing Internet revolution, and a more informed AS clustering algorithm [13, 15, 20, 58]
Future studies of Internet topology measurement
1. Analysis of the evolution and correlation between parts of the Internet ecosystem [22, 32, 36, 46]
2. Measurement of the security risks in Internet topology [49, 52, 61]
3. Measurement of the structure of ASes or measurement of the correlation between ASes and other metrics [35, 38, 53]
4. Studies that provide recommendations to Internet stakeholders, with the aim that research in the field of Internet topology can be done better in the future [26, 43]
5. Various other advanced studies in the field of Internet topology measurement, such as the correlation between ISP characteristics with stock values, providing a public tool to store and monitor the routing policy changes, analysis of the global impact of BGP information, and utility to determine peering intelligently [37, 41, 62, 65]

3.4.2 Future Studies of Internet Topology Mapping

Research on Internet topological mapping has two main directions for further research. The first direction is to improve the geolocation accuracy of ASes with a geographic tagging algorithm that has a higher resolution [11, 12, 21]. Research involving geolocation has recently become a hot issue in the field of topological mapping of the Internet. Meanwhile, the second research direction is to improve the inference relationship methods in order to increase the validity of the inference results of various types of AS relationships in a wider scope [13, 14, 16, 17]. Subsequent research in order to improve inference relationship methods may involve the use of the latest trend methods and technologies, such as machine learning, big data, and deep learning. Apart from those two main directions, there are also specific ideas for future developments, such as a more informed AS clustering algorithm, implications of the ongoing Internet revolution, efficient algorithm for IXP peering finding, as well as data integration and feedback for active measurement improvement [13, 15, 20, 58].

3.4.3 Future Studies of Internet Topology Measurement

Research in the field of Internet topology measurement has several directions for further research. One of the research directions that have been done recently is the measurement of security risks in Internet topology [49, 52, 61]. Research which analyzes the evolution of the Internet ecosystem and the relationships between components is an ongoing study in the field of Internet topology measurement [22, 32, 36, 46]. Likewise, researches that provide recommendations to Internet stakeholders are also continuously needed, with the aim that research in the field of Internet topology can be done better in the future [26, 43]. Other research directions are studies that lead to the measurement of AS structure, the relationship between ASes, and other related metrics [35, 38, 53]. Several specific directions of research can also be carried out, such as providing a public tool to store and monitor the routing policy changes, analysis of the global impact of BGP information, the correlation between ISP characteristics with stock values, and utility to determine peering intelligently [37, 41, 62, 65].

4 Conclusion

A literature review has been carried out on studies that contain keywords "*internet topology*" and "*autonomous systems*" involving 61 literature published in the last 10 years—from 2010 to 2019—. In general, Internet topology research at the AS level can be grouped into three broad groups, namely Internet topology modeling, Internet topology mapping, and Internet topology measurement. The most dominant reason in Internet topology modeling research is to develop the Internet to a better level; whereas the most widely used Internet modeling method is graph theory, although with some criticism that follows modeling in this way. The weaknesses that still exist in Internet topology modeling are problems of accuracy, completeness, and applicability. Mapping Internet topology at the AS level is mainly in the form of efforts to obtain maps of relationships between ASes on the Internet, which is generally confidential data belonging to each Internet provider. Understanding maps between ASes on the Internet is very important in order to understand the structure of the Internet topology and also for the benefit of future Internet development. The main problem that remains an open problem in mapping Internet topology is the problem of the incomplete mapping of links between ASes, especially for peer-to-peer links. Measurement of Internet topology is often done to see trends and evolution of Internet topology over a certain period of time; the measurement process is carried out globally and regionally, with the measurement process generally involving complex network theory. The trends and evolution of the Internet topology are very important to measure, because they can provide an overview of the growth of the Internet topology and can provide predictions of its development in the future. Various components in the Internet topology ecosystem are also measured in various studies on the measurement of Internet topology; one component that is widely measured is IXP.

Common weaknesses in measuring Internet topology are the limited number of vantage points and their uneven placement in all parts of the Internet topology globally, as well as the limited public access to Internet topology measurement points which are mostly private. The use of methods and technologies that can provide insight through processing large amounts of data—such as machine learning, big data, and deep learning—can be involved in research in the field of Internet topology.

Acknowledgements This research was funded by Universitas Indonesia (PUTI 2020 No:NKB-862/UN2.RST/HKP.05.00/2020).

References

1. He Y, Siganos G, Faloutsos M (2012) Internet topology. In: Meyers RA (ed) Computational complexity. Springer, New York, NY, pp 1663–1680. https://doi.org/10.1007/978-1-4614-1800-9_107
2. Motamedi R, Rejaie R, Willinger W (2015) A survey of techniques for internet topology discovery. IEEE Commun Surv Tutorials 17(2):1044–1065. https://ieeexplore.ieee.org/document/6970764/
3. Witono T, Yazid S (2020) Portrait of Indonesia's internet topology at the autonomous system level. In: Alfred R, Lim Y, Haviluddin H, On CK (eds) Computational science and technology, Lecture notes in electrical engineering, vol 603. Springer, Singapore, pp 237–246. http://link.springer.com/10.1007/978-981-15-0058-9_23
4. Accongiagioco G (2014) Modeling and analysis of the internet topology. PhD thesis, IMT Institute for Advanced Studies, Lucca
5. Willinger W, Roughan M (2013) Internet topology research redux, chap. 1. ACM SIGCOMM, pp 1–59
6. Oliveira R, Pei D, Willinger W, Zhang B, Zhang L (2010) The (In)completeness of the observed internet AS-level structure. IEEE/ACM Trans Networking 18(1):109–122. http://ieeexplore.ieee.org/document/5200324/
7. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP (2009) Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. PLOS Med 6(7):1–6 . https://doi.org/10.1371/journal.pmed.1000097
8. Tang T, Hu G (2019) An evolving network model based on a triangular connecting mechanism for the internet topology. Lecture notes in computer science, vol 11635. Springer International Publishing, pp 510–519. http://link.springer.com/10.1007/978-3-030-24268-8_47
9. Fan Z, Tang WKS (2018) Interplay of ICP and IXP over the internet with power-law features. Int J Modern Phys C 29(01):1850001. https://www.worldscientific.com/doi/abs/10.1142/S0129183118500018
10. Ting L, Qin D, Ge L (2019) Experimental research on internet ecosystem and AS hierarchy. In: Data science. ICPCSEE 2019. Communications in computer and information science, vol 1058. Springer International Publishing, pp 253–263. http://link.springer.com/10.1007/978-981-15-0118-0_20
11. Keller UY, Savin E, Fabian B, Ermakova T (2019) Towards geographical analysis of the autonomous system network. Int J Networking Virtual Organ 21(3):379. http://www.inderscience.com/link.php?id=103435
12. Yasin Nur A, Engin Tozal M (2018) Cross-AS (X-AS) internet topology mapping. Comput Netw 132:53–67. https://doi.org/10.1016/j.comnet.2018.01.011, https://linkinghub.elsevier.com/retrieve/pii/S1389128618300124
13. Giotsas V, Luckie M, Huffaker B, Claffy K (2014) Inferring complex AS relationships. In: Proceedings of the 2014 conference on internet measurement conference—IMC '14. ACM

Press, New York, New York, USA, pp 23–30. http://dl.acm.org/citation.cfm?doid=2663716.2663743

14. Luckie M, Huffaker B, Dhamdhere A, Giotsas V, Claffy K (2013) AS relationships, customer cones, and validation. In: Proceedings of the 2013 conference on internet measurement conference—IMC '13. ACM Press, New York, New York, USA, pp 243–256. http://dl.acm.org/citation.cfm?doid=2504730.2504735
15. Deng W, Muhlbauer W, Yang Y, Zhu P, Lu X, Plattner B (2012) Shedding light on the use of AS relationships for path inference. J Commun Netw 14(3):336–345. http://ieeexplore.ieee.org/document/6253094/
16. Gregori E, Improta A, Lenzini L, Rossi L, Sani L (2011) BGP and Inter-AS economic relationships. Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNCS(PART 2) 6641:54–67. http://link.springer.com/10.1007/978-3-642-20798-3_5
17. Giotsas V, Zhou S (2011) Detecting and assessing the hybrid IPv4/IPv6 as relationships. In: Proceedings of the ACM SIGCOMM 2011 conference on SIGCOMM—SIGCOMM '11. ACM Press, New York, New York, USA, p 424. http://dl.acm.org/citation.cfm?doid=2018436.2018501
18. Ager B, Chatzis N, Feldmann A, Sarrar N, Uhlig S, Willinger W (2012) Anatomy of a large european IXP. In: Proceedings of the ACM SIGCOMM 2012 conference on applications, technologies, architectures, and protocols for computer communication—SIGCOMM '12. vol 42. ACM Press, New York, New York, USA, p 163. http://dl.acm.org/citation.cfm?doid=2342356.2342393
19. Csernai M, Gulyas A, Retvari G, Heszberger Z, Csaszar A (2010) The skeleton of the internet. In: 2010 IEEE global telecommunications conference GLOBECOM 2010. IEEE, Miami, FL, pp 1–5. http://ieeexplore.ieee.org/document/5684319/
20. Cai X, Heidemann J, Krishnamurthy B, Willinger W (2010) Towards an AS-to-organization map. In: Proceedings of the 10th annual conference on internet measurement—IMC '10. ACM Press, New York, New York, USA, p 199. http://portal.acm.org/citation.cfm?doid=1879141.1879166
21. Gregori E, Improta A, Lenzini L, Rossi L, Sani L (2012) Inferring geography from BGP raw data. In: 2012 proceedings IEEE INFOCOM workshops. IEEE, Orlando, FL, pp 208–213. http://ieeexplore.ieee.org/document/6193493/
22. Liu X, Wang J, Jing W, de Jong M, Tummers JS, Zhao H (2018) Evolution of the internet AS-level topology: from nodes and edges to components. Chin Phys B 27(12):120501. http://stacks.iop.org/1674-1056/27/i=12/a=120501?key=crossref.8ef865e63be2a2127b07f3cf40db21aa
23. Xiao L, Zhong Z, Wang X, Wang X (2016) Evolution and characteristics of the internet nucleus. In: 2016 international conference on computer communication and informatics (ICCCI). IEEE, pp 1–5. http://ieeexplore.ieee.org/document/7480012/
24. Wang Y, Zhang K (2016) Quantifying the flattening of internet topology. In: Proceedings of the 11th international conference on future internet technologies—CFI '16. vol 15-17. ACM Press, New York, New York, USA, pp 113–117. http://dl.acm.org/citation.cfm?doid=2935663.2935682
25. Yang D, Rong Z (2015) Evolution of the internet at the autonomous system level. In: Zhao QLS (ed) 2015 34th Chinese control conference (CCC), vol 2015. IEEE, pp 1313–1317. http://ieeexplore.ieee.org/document/7259824/
26. Lodhi A, Larson N, Dhamdhere A, Dovrolis C, Claffy K (2014) Using peeringDB to understand the peering ecosystem. In: ACM SIGCOMM computer communication review, vol 44. Association for Computing Machinery, pp 20–27. http://dl.acm.org/citation.cfm?doid=2602204.2602208
27. Masoud MZ, Hei X, Cheng W (2013) A graph-theoretic study of the flattening Internet AS topology. In: 2013 19th IEEE international conference on networks (ICON). IEEE, Singapore, pp 1– 6. http://ieeexplore.ieee.org/document/6781991/
28. Zhao Z, Bi J (2013) Characterizing and analysis of the flattening Internet topology. In: 2013 IEEE symposium on computers and communications (ISCC). IEEE, Split, pp 000219–000225. http://ieeexplore.ieee.org/document/6754949/

29. Dhamdhere A, Cherukuru H, Dovrolis C, Claffy K (2012) Measuring the evolution of internet peering agreements. Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNCS(PART 2) 7290:136–148. http://link.springer.com/10.1007/978-3-642-30054-7_11
30. Shavitt Y, Weinsberg U (2012) Topological trends of internet content providers. In: Proceedings of the fourth annual workshop on simplifying complex networks for practitioners—SIMPLEX '12. ACM Press, New York, New York, USA, p 13. http://dl.acm.org/citation.cfm?doid=2184356.2184360
31. Xiang Y, Yin X, Wang Z, Wu J (2011) Internet flattening: monitoring and analysis of interdomain routing. In: 2011 IEEE international conference on communications (ICC). IEEE, Kyoto, pp 1–6. http://ieeexplore.ieee.org/document/5962885/
32. Zhang G, Quoitin B, Zhou S (2011) Phase changes in the evolution of the IPv4 and IPv6 AS-level internet topologies. Comput Commun 34(5):649–657. https://linkinghub.elsevier.com/retrieve/pii/S0140366410002653
33. Dhamdhere A, Dovrolis C (2011) Twelve years in the evolution of the internet ecosystem. IEEE/ACM Trans Networking 19(5):1420–1433. http://ieeexplore.ieee.org/document/5729851/
34. Gregori E, Improta A, Sani L (2018) On the African peering connectivity revealable via BGP route collectors. Lecture notes of the institute for computer sciences, social-informatics and telecommunications engineering. LNICST 250:368–376. http://link.springer.com/10.1007/978-3-319-98827-6_35
35. Silva Berenguer S, Valera Pintor F (2018) Radiography of internet autonomous systems interconnection in Latin America and the Caribbean. Comput Commun 119:15–28. https://linkinghub.elsevier.com/retrieve/pii/S0140366418300446
36. Cakmak G, Aydin MN (2017) A country-specific analysis on internet interconnection ecosystems. In: 2017 9th international congress on ultra modern telecommunications and control systems and workshops (ICUMT), vol 2017. IEEE, pp 232–237. http://ieeexplore.ieee.org/document/8255151/
37. Fanou R, Francois P, Aben E, Mwangi M, Goburdhan N, Valera F (2017) Four years tracking unrevealed topological changes in the African interdomain. Comput Commun 106:117–135. https://linkinghub.elsevier.com/retrieve/pii/S0140366417302074
38. Wählisch M, Schmidt TC, de Brün M, Häberlen T (2012) Exposing a nation-centric view on the German internet – a change in perspective on AS-level. Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNCS 7192:200–210. http://link.springer.com/10.1007/978-3-642-28537-0_20
39. Kajackas A, Rainys R (2010) Internet infrastructure topology assessment. Elektronika ir Elektrotechnika 103(7):91–94. http://eejournal.ktu.lt/index.php/elt/article/view/9284
40. Canbaz MA, Thom J, Gunes MH (2017) Comparative analysis of internet topology data sets. In: 2017 IEEE conference on computer communications workshops (INFOCOM WKSHPS). IEEE, pp 635–640. http://ieeexplore.ieee.org/document/8116451/
41. Berenguer SS, Carisimo E, Alvarez-Hamelin JI, Pintor FV (2016) Hidden internet topologies info. In: Proceedings of the 2016 workshop on fostering Latin-American research in data communication networks—LANCOMM '16. ACM Press, New York, New York, USA, pp 4–6. http://dl.acm.org/citation.cfm?doid=2940116.2940118
42. Gregori E, Improta A, Lenzini L, Rossi L, Sani L (2015) A novel methodology to address the internet AS-level data incompleteness. IEEE/ACM Trans Networking 23(4):1314–1327. http://ieeexplore.ieee.org/document/6821271/
43. Khan A, Kwon T, Kim HC, Choi Y (2013) AS-level topology collection through looking glass servers. In: Proceedings of the 2013 conference on internet measurement conference—IMC '13. ACM Press, New York, New York, USA, pp 235–242. http://dl.acm.org/citation.cfm?doid=2504730.2504758
44. Fan Z, Tang WKS, Lin D, Li D (2018) A comparative study of IXP in Europe and US from a complex network perspective. Adv Intell Syst Comput 566:242–252. http://link.springer.com/10.1007/978-3-319-60663-7_23

45. Masoud M, Jaradat Y, Jannoud I (2017) A measurement study of internet exchange points (IXPs): history and future prediction. Turkish J Electr Eng Comput Sci 25(1):376–389. http://online.journals.tubitak.gov.tr/openDoiPdf.htm?mKodu=elk-1412-23
46. Brito SHB, Santos MAS, Fontes RDR, Perez DAL, Rothenberg CE (2016) Dissecting the largest national ecosystem of public internet exchange points in Brazil. Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9631:333–345. http://link.springer.com/10.1007/978-3-319-30505-9_25
47. Fan Z, Tang WK (2016) Unraveling the impacts of IXP in internet ecosystem using bi-layered network. Physica A: Stat Mech Appl 456:327–339. https://linkinghub.elsevier.com/retrieve/pii/S0378437116300486
48. Yang Y, Yin X, Shi X, Wang Z, He J, Fu TZ, Winslett M (2019) Inter-domain routing bottlenecks and their aggravation. Comput Netw 162:106839. https://doi.org/10.1016/j.comnet.2019.06.017, https://linkinghub.elsevier.com/retrieve/pii/S1389128618310466
49. Phung CD, Silva BF, Nogueira M, Secci S (2019) MPTCP robustness against large-scale man-in-the-middle attacks. Comput Netw 164:106896. https://doi.org/10.1016/j.comnet.2019.106896, https://linkinghub.elsevier.com/retrieve/pii/S1389128619304566
50. Nur AY, Tozal ME (2018) Identifying critical autonomous systems in the internet. The J Supercomputing 74(10):4965–4985. http://link.springer.com/10.1007/s11227-018-2336-3
51. Tozal ME (2017) Autonomous system ranking by topological characteristics: a comparative study. In: 2017 annual IEEE international systems conference (SysCon). IEEE, pp 1–8. http://ieeexplore.ieee.org/document/7934814/
52. Acharya HB, Chakravarty S, Gosain D (2017) Few throats to choke: on the current structure of the internet. In: 2017 IEEE 42nd conference on local computer networks (LCN), vol 2017. IEEE, pp 339–346. http://ieeexplore.ieee.org/document/8109374/
53. Gregori E, Lenzini L, Orsini C (2011) k-dense communities in the internet AS-level topology. In: 2011 third international conference on communication systems and networks (COMSNETS 2011). IEEE, Bangalore, pp 1–10. http://ieeexplore.ieee.org/document/5716413/
54. Nomikos G, Sermpezis P, Dimitropoulos X (2017) Re-mapping the internet: bring the IXPs into play: www.inspire.edu.gr/ixp-map. In: 2017 IEEE conference on computer communications workshops (INFOCOM WKSHPS). IEEE, pp 910–915. http://ieeexplore.ieee.org/document/8116497/
55. Meenakshi SP, Raghavan SV (2015) Forecasting and event detection in internet resource dynamics using time series models. Eng Lett 23(4):245–257. http://www.engineeringletters.com/issues_v23/issue_4/EL_23_4_03.pdf
56. Giotsas V, Dhamdhere A, Claffy KC (2016) Periscope: unifying looking glass querying. Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9631:177–189. http://link.springer.com/10.1007/978-3-319-30505-9_14
57. Giotsas V, Zhou S (2013) Improving the discovery of IXP peering links through passive BGP measurements. In: 2013 IEEE conference on computer communications workshops (INFOCOM WKSHPS). IEEE, Turin, pp 121–126. http://ieeexplore.ieee.org/document/6562878/
58. Giotsas V, Zhou S, Luckie M, Claffy K (2013) Inferring multilateral peering. In: Proceedings of the ninth ACM conference on emerging networking experiments and technologies—CoNEXT ’13. ACM Press, New York, New York, USA, pp 247–258. http://dl.acm.org/citation.cfm?doid=2535372.2535390
59. Gregori E, Improta A, Lenzini L, Rossi L, Sani L (2012) On the incompleteness of the AS-level graph: a novel methodology for BGP route collector placement. In: Proceedings of the 2012 ACM conference on Internet measurement conference—IMC ’12. ACM Press, New York, New York, USA, p 253. http://dl.acm.org/citation.cfm?doid=2398776.2398803
60. Tozal ME (2016) The internet: a system of interconnected autonomous systems. In: 2016 annual IEEE systems conference (SysCon). IEEE, pp 1–8. http://ieeexplore.ieee.org/document/7490628/
61. Leyba KG, Edwards B, Freeman C, Crandall JR, Forrest S (2019) Borders and gateways: measuring and analyzing national AS chokepoints. In: Proceedings of the conference on computing

& sustainable societies—COMPASS 19. ACM Press, New York, New York, USA, pp 184–194. http://dl.acm.org/citation.cfm?doid=3314344.3332502
62. Dey PK, Canbaz MA, Yuksel M, Gunes MH (2018) On correlating ISP topologies to their businesses. In: 2018 IEEE international conference on communications (ICC), vol 2018. IEEE, pp 1–7. https://ieeexplore.ieee.org/document/8422620/
63. Fan ZP (2012) Evaluating the AS-level internet models: beyond topological characteristics. Chin Phys B 21(2):028902. http://stacks.iop.org/1674-1056/21/i=2/a=028902?key=crossref.19eff4bd832db40aa11afa82c10ca683
64. Roughan M, Willinger W, Maennel O, Perouli D, Bush R (2011) 10 lessons from 10 years of measuring and modeling the internet's autonomous systems. IEEE J Selected Areas Commun 29(9):1810–1821. http://ieeexplore.ieee.org/document/6027863/
65. Lodhi A, Laoutaris N, Dhamdhere A, Dovrolis C (2015) Complexities in internet peering: understanding the "Black" in the "Black Art". In: 2015 IEEE conference on computer communications (INFOCOM). IEEE, pp 1778–1786. http://ieeexplore.ieee.org/document/7218559/

Addressing the Effects of the Spectrum Sensing Data Falsification Attack Using the Enhanced Q-out-of-m Rule

Velempini Mthulisi, Ngomane Issah, and Mapunya Sekgoari Semaka

Abstract The spectrum sensing data falsification (SSDF) attack modifies the energy values received from the primary users (PUs). During cooperative spectrum sensing (CSS), it reports the incorrect data to its neighbouring secondary users (SUs). CSS is conducted by SUs which perform local sensing and then share observed data to achieve distributed global sensing. In this study, the effects of SSDF in cognitive radio ad hoc network (CRAHN) are evaluated. The enhanced q-out-of-m rule scheme is proposed to address the effects of SSDF in CRAHN. The proposed scheme was evaluated in MATLAB and the simulation results show that it outperformed the DBSD scheme.

Keywords Cognitive radio network · Cognitive radio ad hoc network · Cooperative spectrum sensing · Enhanced q-out-of-m rule · Spectrum sensing data falsification

1 Introduction

Cognitive radio networks (CRNs) address the underutilization of the spectrum by enabling secondary users (SUs) to use opportunistically the vacant licensed users' spectrum [1–3]. The SUs have to avoid interfering with the transmission of primary users (PUs) by cooperating in sensing the spectrum. SUs can either cooperate or sense the spectrum non-cooperatively [2, 4, 5]. SUs cooperatively sense the spectrum and share the sensing data and thereafter make some informed spectrum access decisions. Non-cooperatively, the SUs sense the band and make decisions independently.

V. Mthulisi (✉) · N. Issah · M. S. Semaka
Department of Computer Science, University of Limpopo, Limpopo, South Africa
e-mail: mthulisi.velempini@ul.ac.za

N. Issah
e-mail: Issah.ngomane@ul.ac.za

M. S. Semaka
e-mail: sekgoari.mapunya@ul.ac.za

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_55

Cooperative spectrum sensing (CSS) is superior to non-cooperative spectrum sensing. Non-cooperative spectrum sensing is susceptible to challenges such as the hidden terminal problem (HTP), signal and multi-path fading in the band [6]. Unfortunately, CSS is susceptible to the spectrum sensing data falsification (SSDF) attack security challenge. In SSDF attackers collaborate in sensing the band with legitimate SUs and then share incorrect observations. This can result in interference to PUs or denial of service to the SUs [7]. This study investigated the SSDF also known as the Byzantine attack. We propose the implementation of the enhanced q-out-of-m rule (q – m rule) scheme to address the Byzantine attack.

We implemented the enhanced q – m rule scheme to address the effects of the Byzantine attack. The q – m rule whereby a random 60% of m number of nodes is selected and the final transmission decision is based on q. If q is found to be 1, the spectrum band is considered to be occupied and when q is 0, the band is idle.

The study was motivated by the fact that CRN is a promising technology that can be utilized to address the spectrum underutilization challenge. Hence there is a need to improve the security of the network.

This paper is arranged as follows: Section 2 presents related works. Section 3 briefly describes the network environment in which the experiments were carried out. Section 4 analyses the results and the study is concluded in Sect. 5.

2 Related Works

SSDF defence schemes were designed and implemented by researchers to address the SSDF attack. The study in [8] assigns each cognitive user (CU) a weight value in an attempt to address the SSDF. It compares the binary output of SUs to the fusion centre (FC) output. The reputation value of the CU is incremented by one if the output of the CU matches FC's output, otherwise, it is decremented. Sensing helps the CUs to decide whether the spectrum is idle or not.

The reputation of CUs which unintentionally misbehaves is restored; however, it works best when there is only one attacker. We propose a scheme that does not require the services of an FC. The role of the CU is performed by each CU which fuses its own observations and the observations of its neighbours.

In [9], a statistical consensus-based scheme was proposed to address the SSDF attack. The scheme is CSS based and SU makes independent transmission decisions. SUs compute the difference between received data and the computed mean. A high deviation disqualifies SU's input and is classified as malicious. The limitation of this scheme was that it isolated the unintentionally misbehaving SUs or nodes. The unintentionally misbehaving nodes temporarily misbehave due to challenges such as the HTP and signal fading. Schemes should be able to accommodate the misbehaving nodes while preserving the integrity of cognitive radio ad hoc networks (CRAHN). In [10], the SSDF attack is addressed by using the modified Z-test and the q – m scheme in CRAHN. The results show that the scheme can isolate and address the SSDF.

A distributed scheme designed to address the SSDF in CSS is proposed in [11]. The reports from malicious nodes are dropped. The probability density of the random variable is estimated using the kernel density estimator. An incorrect report is isolated. Unfortunately, density-based SSDF detection (DBSD) isolates all incorrect reports, including reports of unintentionally misbehaving nodes. It then computes the average based on the remaining sensing reports in comparison to the LU detection threshold. The scheme assumes that malicious nodes are relatively few.

The scheme in [12] assigns each CU a weight value in an endeavour to address the SSDF. The binary output of each user is assessed to determine its closeness to the FC's output. If the output matches the one from FC, its reputation is incremented by one, otherwise, it is decremented. The CUs have to decide whether the LU present or not. The scheme restores the reputation of unintentionally misbehaving CUs; unfortunately, the scheme is only effective when the MUs are less than the CUs.

The work in [13] compares the effectiveness of a trust-based scheme against a statistical approach scheme in addressing the effects of the Byzantine attack. The authors observed that the trust-based scheme performed better compared to the statistical approach.

A lightweight multi-fusion-based distributed spectrum sensing scheme (MFDSS) designed to mitigate the SSDF in CRAHN is proposed in [14]. It detects an outlier and also implements a data fusion technique which is employed to evaluate the effects of the SSDF attack. The outliers are isolated using the A-modified z-test. In addition, reputation serves as a second detection layer designed to detect outliers that were missed. We propose an enhanced $\mathrm{q} - \mathrm{m}$ rule scheme to isolate outliers.

In [15], the MUs of the SSDF is investigated. A double-sided neighbour distance (DSND) scheme is proposed to detect the MUs. The scheme characterizes SUs malicious when their values differ significantly from other cooperative nodes' reports. Unfortunately, the scheme isolates nodes that report similar observations. It should be noted that even non-malicious nodes can occasionally report observations that are the same. In our scheme, we isolate outliers through the use of the enhanced $\mathrm{q} - \mathrm{m}$ rule.

3 Network Environment

The network considered for our study consists of SUs and PUs. The SUs make use of the cognitive capabilities to opportunistically access the spectrum while the PUs are licensed to utilize the spectrum band. Unfortunately, the network also has MUs which utilize the band either for selfish reasons and causes the spectrum unavailable to other SUs. This is a denial of service which starves the SUs as depicted in Fig. 1.

Figure 1 shows the implementation of CSS in a cognitive network. We demonstrate how the nodes cooperate to conduct reliable sensing. Unfortunately, the SSDF attack denoted by node 13 also collaborates with the SUs in sending incorrect spectrum observations. In Fig. 2, we also show the different types of SUs that are investigated in the work.

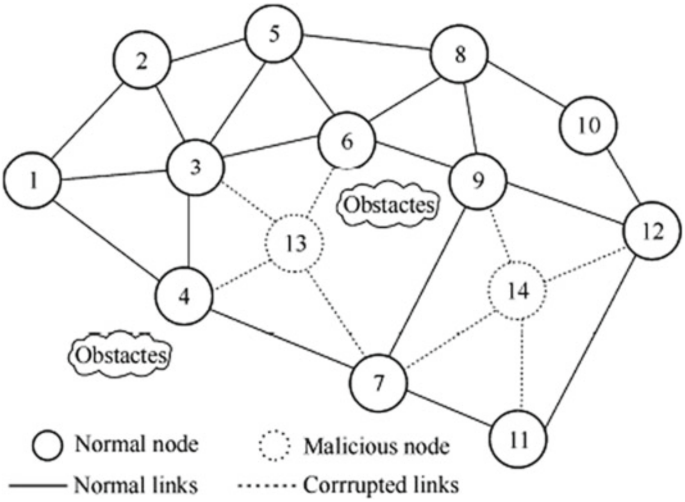


Fig. 1 Implementation of CSS in CRN [16]

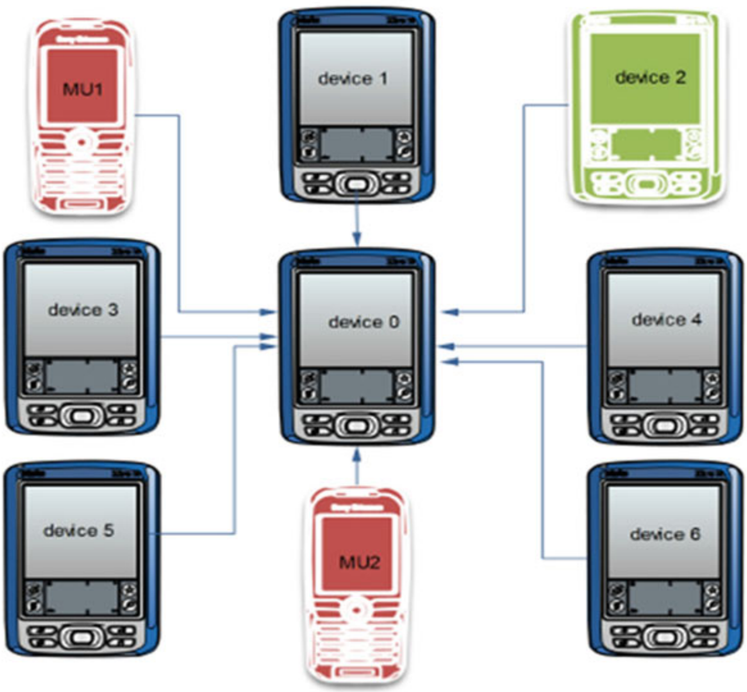


Fig. 2 Types of SSDF attacks [17]

We investigated the unintentionally misbehaving SUs denoted by node 2, the normal SUs denoted by nodes 0–6 and the malicious SUs denoted by MU1-MU2. Node 2 is said to be unintentionally misbehaving due to environmental obstructions in the network that cause the device to temporarily misbehave [18, 19]. The environmental challenges are the HTP, multipath and signal fading [20].

4 Results and Analysis

For the performance evaluations, we used the MATLAB R2015a version in Windows 10 environment to simulate our scheme. The signal-to-noise ratio was set to 10 decibels in the proposed enhanced $q - m$ rule scheme. Energy detection was implemented as the sensing technique. The detection threshold was set to 0. The SUs first perform CSS before computing binary values before they could share the observations. We set a number of nodes $N = 10$, 50, 100, 150 to 250 to evaluate the performance of schemes in different size of networks. We selected our MUs to be 20, 40 and 60%. The probability of detection (PoD), missed detection (MD) and false alarm probabilities (FAP) were selected for evaluation metrics. The proposed scheme was evaluated against the DBSD scheme in [11]. DBSD was designed to solve the same challenge in CRN.

Figure 3 shows success probability results when the percentage of MUs in the network is 20. With $N = 10$, we had 2 MUs which were detected by the schemes. Unfortunately, as we observe from the results, as the network size increased, the success probability of the schemes degraded gradually. This was caused by the attacks which were not detected in the first fusion step. The unintentionally misbehaving

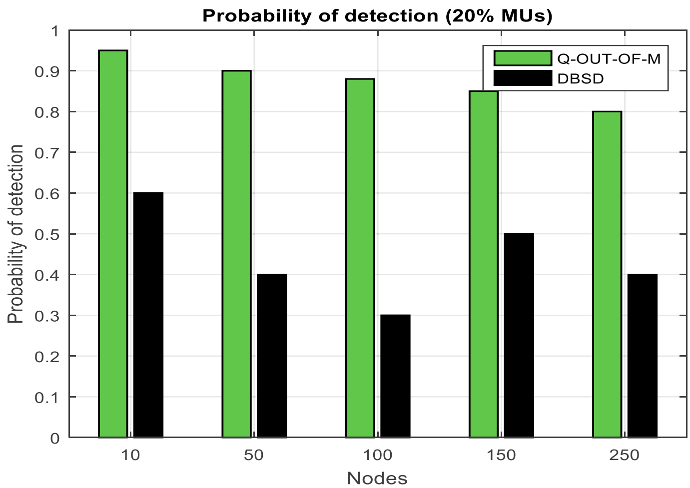


Fig. 3 Probability of detection with 20% malicious users

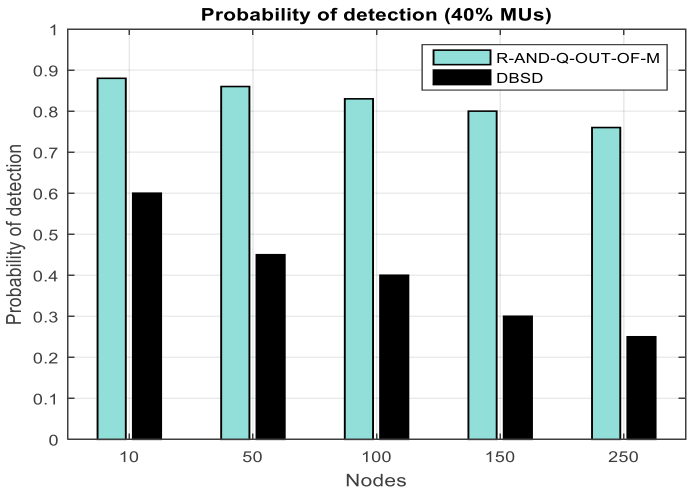


Fig. 4 Probability of detection with 40% malicious nodes

nodes were also detected as malicious nodes that have an impact on the success probability. However, the enhanced q − m rule scheme performs better in terms of the success probability. Figure 4 presents the PoD in a network consisting of 40% of nodes being malicious.

We increased the percentage of malicious users in Fig. 4 to 40% and observed a decrease in the PoD. We concluded that the increase in the percentage of MUs has an impact on the success probability. However, our enhanced q − m rule scheme performed better than the DBSD scheme. The scheme managed to detect and isolate all the malicious nodes.

Figure 5 shows the PoD results with 60% malicious users in the network. Both schemes managed to detect more than 30% of the attack in every network size. As the network size increased and the unintentionally misbehaving nodes also increased, which degraded the performance of the scheme.

Figure 6 shows the results of the missed detection probability (PMD) from a small network size of 10 to a large network size of 250 nodes with 20% MUs in each network size. According to the results, the q − m rule scheme performed better in missed detection as it has the lowest missed detection rate. This is because in implementing the q − m rule scheme, which was 60% of the nodes were selected randomly and the attack was isolated in the network. Figure 8 presents the missed detection in a network with 40% of MUs.

Figure 7 shows missed detection results with 40% MUs in the network. The increasing number of SUs increases the PMD. The PMD was degraded by the increasing number of MUs. These results prove that a q − m rule system performs better compared to the DBSD scheme because of the unintentionally misbehaving nodes which can be detected using the q − m rule. The DBSD approaches usually

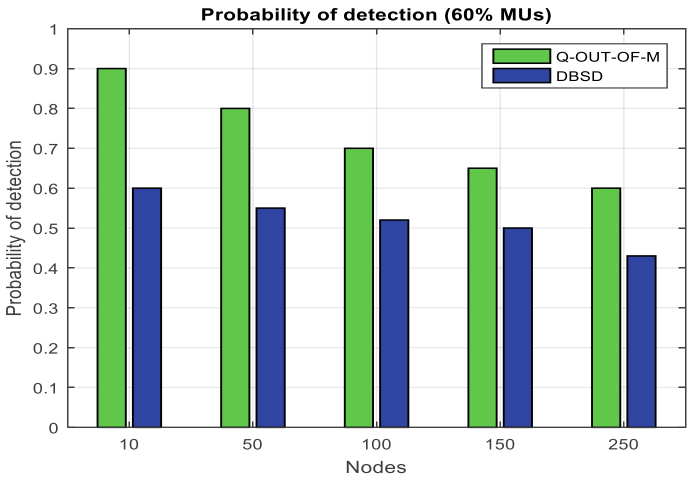


Fig. 5 Probability of detection with 60% malicious users

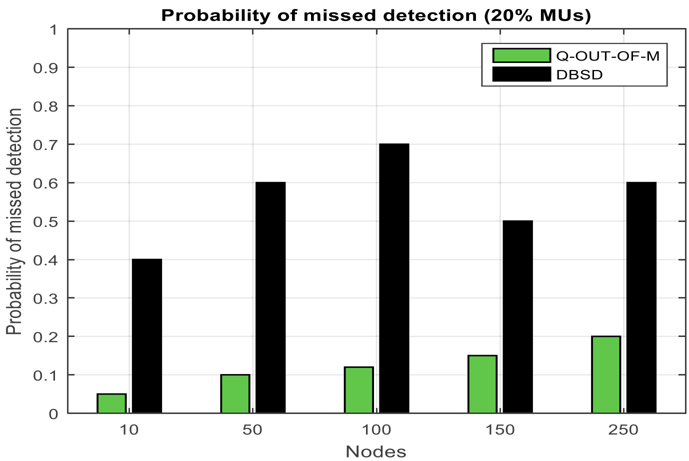


Fig. 6 Missed detection probability with 20% malicious nodes

detect and isolate the always yes and always no MUs only which is a limitation. Figure 8 presents the missed detection in a network with 60% of MUs.

The PMD of the schemes in detecting the SSDF attack increased in response to an increasing number of MUs. It was observed that the increasing number of MUs degraded the performance of the schemes. The enhanced q − m scheme managed to detect and isolate more attackers compared to the DBSD scheme. Both schemes implemented a statistically-based approach instead of a trust-based approach. The DBSD scheme considered the actual sensed values of the PUs received energy while the enhanced q − m rule scheme considered the binary computation of the PUs

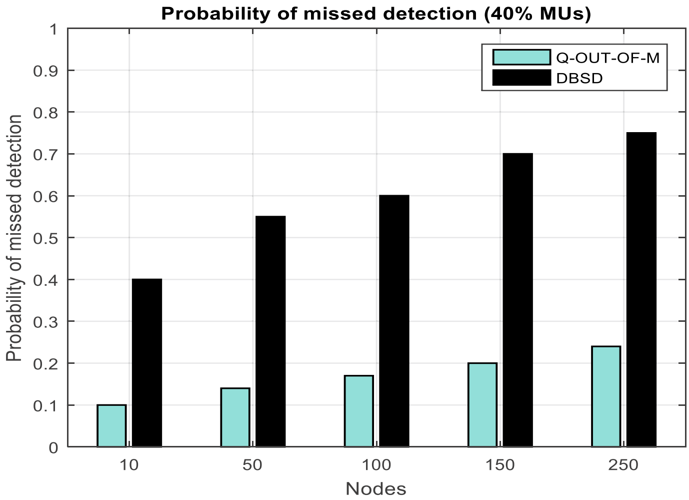


Fig. 7 Missed detection probability with 40% malicious nodes

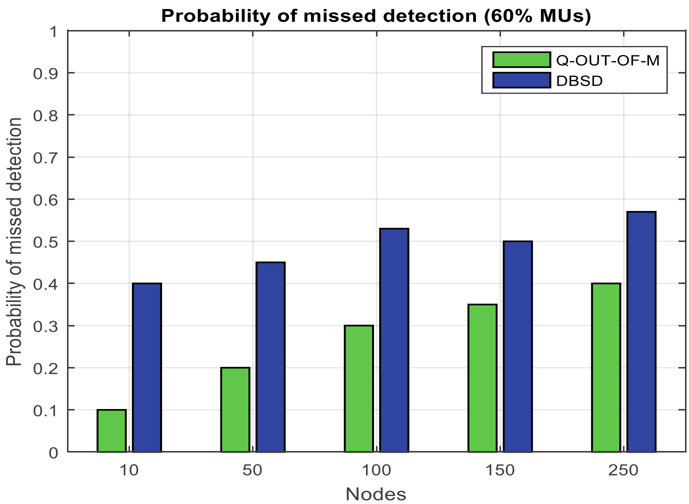


Fig. 8 Missed detection probability with 60% of malicious nodes

received energy levels. This caused the proposed scheme to detect and isolate binary values that differ from the majority of reports in the final transmission decision-making process. Figure 9 presents the PFA alarm results.

We present the PFA results in Fig. 9 in the following scenarios in N = 10, 50, 100, 150 to 250 with 20% MUs in each network size. The PFA is the probability that the schemes detect legitimate SUs as malicious. In N = 10 we had 8 SUs and 2 MUs. In order to find effective methods to accommodate the misbehaving SUs,

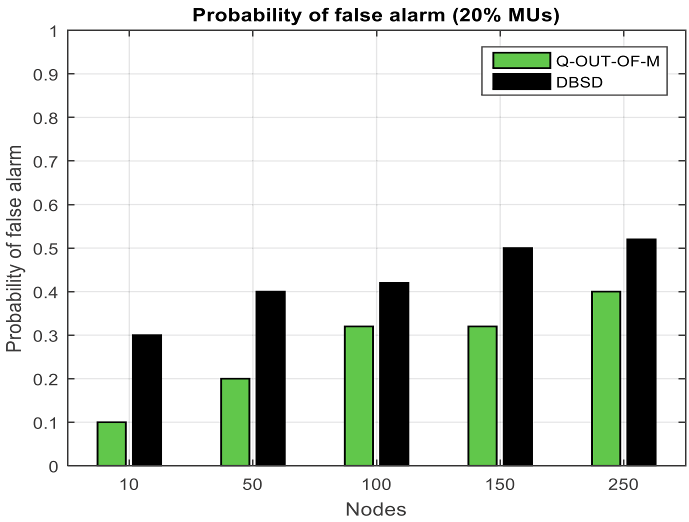


Fig. 9 False alarm probability with 20% malicious nodes

most of the nodes were set to be unintentionally misbehaving. We observed from the results that as we increased the size of the network and the size of the misbehaving nodes, the false alarm probability increased as well. We observed that the $q - m$ scheme performed better in accommodating the misbehaving SUs due to the 60% threshold, thus it had the lowest PFA. The DBSD scheme discarded most reports including reports from legitimate SUs which caused the increase in the PFA. The same behaviour is observed even in 40% MUs. Figure 10 presents the PFA results

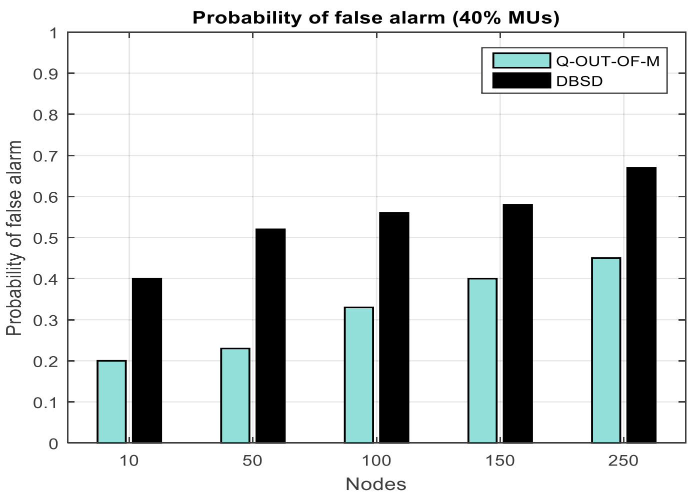


Fig. 10 False alarm probability with 40% intruders in the network

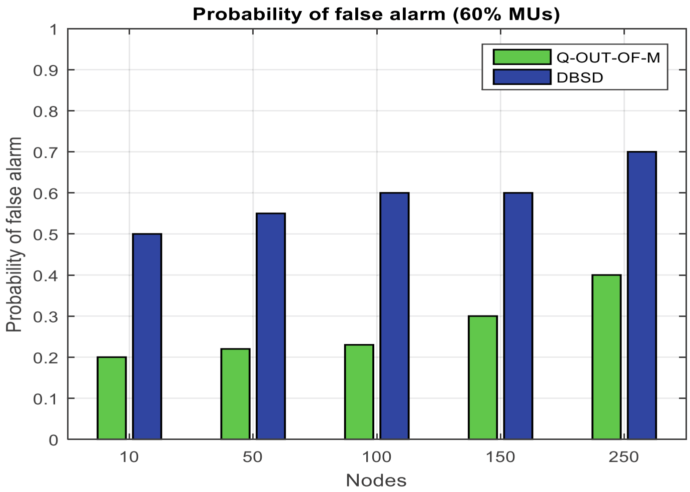


Fig. 11 False alarm probability with 60% intruders in the network

in a network with 40% of MUs.

As observed in Fig. 10, in which the results of PFA in a scenario with 40% intruders, given a high number of malicious users, the PFA was also high. It increased as we increased the number of SUs. We set different misbehaving nodes in each network size which caused the results to fluctuate. We conclude that the misbehaving nodes have a negative effect on the PFA of the schemes. Given the increasing number of misbehaving nodes in a network environment, PFA rises. The q − m scheme outperformed the DBSD in all the metrics. It had the highest PoD and the lowest PMD and PFA. Figure 10 presents the PFA results in a network with 60% of MUs.

Figure 11 depicts the PFA results of the schemes in a network with 60% MUs. The PFA is the probability of detecting unintentionally misbehaving nodes as malicious. We observed from the results that the q − m scheme has the lowest false alarm probability rate. We randomly selected *m* nodes which constitute 60% of the nodes making the final transmission decision. Our scheme accommodates the misbehaving SUs while preserving the integrity of CRAHN by assuring a correct transmission decision.

5 Conclusion

In this paper, we studied an attack known as the SSDF attack or Byzantine attack in CRAHN. We evaluated two approaches used in implementing a defence mechanism to counter the SSDF attack. We investigated the DBSD and compared it with the proposed enhanced q − m rule scheme. We studied two types of SUs, the legitimate SUs and the unintentionally misbehaving SUs. We simulated the schemes in

CRAHNs which do not require the services of a data FC. The schemes were simulated in MATLAB. The PoD, PMD and PFAs were used to compare the schemes. We observed that the enhanced $q - m$ rule system outperforms the DBSD scheme. The proposed scheme still need to be enhanced to alleviate the effects of primary user emulation attack where the attackers mimic the characteristics of the primary user in an attempt to mislead legitimate users.

References

1. Mapunya S, Velempini M (2019) The design of byzantine attack mitigation scheme in cognitive radio ad-hoc networks. In: Proceedings of 2018 international conference on intelligent and innovative computing applications (ICONIC), Holiday Inn Mauritius, Mon Trésor, Plaine Magnien, Mauritius
2. Wei Z-H, Hu B-J (2018) A fair multi-channel assignment algorithm with practical implementation in distributed cognitive radio networks. IEEE Access 6:14255–14267
3. Boddapati KH, Bhatnagar RM, Prakriya S (2018) Performance of incremental relaying protocols for cooperative multi-hop CRNs. IEEE Trans Veh Technol 67(7):6006–6022
4. Mapunya SS, Velempini M (2019) The design and implementation of a robust scheme to combat the effect of malicious nodes in cognitive radio ad hoc networks. South Afr Comput J 2(31):178–194
5. Tephillah S, Manickam ML (2020) An SETM Algorithm for Combating SSDF Attack in Cognitive Radio Networks. Wirel Commun Mob Comput 2020:1–9
6. Osama E, Maha E, Osamu M, Hiroshi F (2018) Game theoretic approaches for cooperative spectrum sensing in energy-harvesting cognitive radio networks. IEEE Access 6:11086–11100
7. Chen C, Song M, Xin C (2014) CoPD: a conjugate prior based detection scheme to countermeasure spectrum sensing data falsification attacks in cognitive radio networks. Wireless Netw 20(8):2521–2528
8. Ngomane I, Velempini M, Dlamini SV (2016) The design of a defence mechanism to mitigate the spectrum sensing data falsification attack in cognitive radio ad hoc networks. In: Proceedings of 2016 international conference on Advances in Computing and Communication Engineering (ICACCE), Durban, South Africa
9. Tong J, Jin M, Guo Q, Li Y (2018) Cooperative spectrum sensing: a blind and soft fusion detector. IEEE Trans Wireless Commun 17(4), 2726–2737
10. Chen H, Zhou M, Xie L, Li J (2017) Cooperative spectrum sensing with M-Ary quantized data in cognitive radio networks under SSDF attacks. IEEE Trans Wireless Commun 16(8):5244–5257
11. Sun M, Jin M, Guo Q, Li Y (2017) Cooperative spectrum sensing under ambient malicious interferences. IEEE Commun Lett 22(2):432–435
12. Jadhav A, Shinde SC (2015) Centralized cooperative and non-cooperative spectrum sensing using energy detection in cognitive radio. Int J Recent Sci Res 6(8):5764–5767
13. Akyildiz I, Lo B, Balakrishnan R (2011) Cooperative spectrum sensing in cognitive radio networks: a survey. Phy Commun 4(1):40–62
14. Chen Q, Motani M, Wong W-C, Nallanathan A (2010) Cooperative spectrum sensing strategies for cognitive radio mesh networks. IEEE J Select Topics Sign Proc 5(1):56–67
15. Nie G, Ding G, Zhang L, Wu Q (2017) Byzantine defense in collaborative spectrum sensing via bayesian learning. IEEE Access 5:20089–20098
16. Wang H, Yao Y-D, Peng S (2018) Prioritized secondary user access control in cognitive radio networks. IEEE Access 6:11007–11016
17. Seif M, Karmoose M, Youssef M (2015) Censoring for improved sensing performance in infrastructure-less cognitive radio networks. In: Proceedings of 2015 IEEE 81st vehicular technology conference, Glasgow, UK

18. Chen C, Song M, Xin C (2013) A density based scheme to countermeasure spectrum sensing data falsification attacks in cognitive radio networks. In: Proceedings of 2013 IEEE Global Communications Conference (GLOBECOM), Atlanta, GA, USA
19. Yu FR, Tang H, Huang M, Li Z, Mason P (2009) Defense against spectrum sensing data falsification attacks in mobile ad hoc networks with cognitive radios. In: MILCOM 2009–2009 IEEE Military Communications Conference. MA, USA, Boston
20. Bai P, Zhang X, Ye F (2017) Reputation-based Beta reputation system against SSDF attack in cognitive radio networks. In: Proceedings of 2017 Progress in Electromagnetics Research Symposium, PIERS

COVID-19 Severity Prediction in Patients Based on Anomaly Detection Approach

Alisher Ikramov, Fatima Adilova, Khikmat Anvarov, and Abduhakim Khadjibaev

Abstract **Background** As the course of COVID-19 varies dramatically, one of the critical challenges is detecting high-risk patients for early prevention and treatment. Lockdown and treatment regimen in Uzbekistan in Spring, 2020 made it possible to collect a significant amount of patient data, including asymptomatic, mild, and severe cases. Based on the data, we propose several models for the early prediction of severe COVID-19. **Results**: We compared supervised learning and anomaly detection algorithms for the task of severe illness prediction. We analyzed the performance and evaluated risk factors based on the proposed models. The best performing model achieves an F1 score of 0.928 and a C-index of 0.965 on the test set. We evaluated the robustness of the model and tested it on an external dataset. We used different techniques to build model interpretability. **Conclusions**: Machine learning methods can help in detecting patients that have a high risk of developing a severe course so that medical experts can start proper treatment earlier and, as a result, decrease the number of critical cases.

Keywords Artificial intelligence · Machine learning · Anomaly detection · Feature selection · COVID-19 · Medical prognosis · Retrospective analysis · K-means

1 Background

COVID-19 is a serious disease that quickly affected the entire world. The ratio of patients with severe COVID-19 was estimated as 2 in 10 people [1] back in March based on the information from China, while Oran et al. [2] estimated asymptomatic patients to be 40% of all positive cases.

A. Ikramov (✉) · F. Adilova
Institute of Mathematics, Mirzo Ulugbek street 81, Tashkent, Uzbekistan 100170
e-mail: a.ikramov@mathinst.uz

K. Anvarov · A. Khadjibaev
Republican Scientific Centre for Emergency Medicine, Kichik halqa yuli 2, Tashkent, Uzbekistan 100081

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_56

Garcia-Vidal et al. [3] demonstrated the advantages of using personal treatment advised by artificial intelligence to reduce lethal cases by roughly half. The high virulence [4] that led to the pandemic puts a heavy burden on the healthcare system, making it unable to function at the proper level. Healthcare systems in poor regions were affected the most, the overload of the hospitals lead to a rise in mortality rate [5].

1.1 Related Works

U.S. Centre for Disease Control [6] has evaluated risk factors based on clinical data from China, and in the United States, age and comorbidities are strong risk factors for severe illness. Wu et al. [7] had also analyzed data from the Chinese CDC and stated that 81% of the patients had a mild course. Docherty et al. [8] characterized clinical features of 20,133 inpatients in the UK. They found that 60% of patients were men, interquartile range of age was 58–82, the most common comorbidities were chronic cardiac disease (31%), uncomplicated diabetes (21%), chronic pulmonary disease (18%), and chronic kidney disease (16%).

Ortolan et al. [9] analyzed 85 articles to compare COVID-19 mortality, severity, and recovery rates. They found that the mortality ratio of men to women was 1.81, the severity ratio was 1.46, and the recovery chance ratio was 0.72. Klein et al. [10] analyzed data from New York City and found that the mortality rate of males is almost twice as high as female mortality. Wynants et al. [11] reviewed more than 100 articles describing 145 models. One of the drawbacks of all the studies was the lack of use of the TRIPOD Checklist [12]. They found 91 diagnostic models and 50 prognostic models of the mortality risk. All models were ranked as high-biased based on their selection and exclusion processes. C-index in those reviewed papers varied from 0.68 to 0.99.

De Lusignan et al. [13] built multivariable logistic regression models with multiple imputations to identify risk factors for positive SARS-CoV-2 tests. The research was aimed at finding key factors of becoming infected with COVID-19.

1.2 The Aim of the Work

The work aimed to develop an automated system to predict the risk of developing a severe course of COVID-19 in patients. The system should use only easily identifiable predictors that most medical personnel can measure or collect. The system should allow identifying patients with possible complications at the initial stage so that they could start their corresponding treatment as early as possible.

2 Material and Methods

2.1 Patients' Data

Uzbekistan saw its first case on March 15, 2020, and went to strict lockdown shortly afterward. All people who traveled from other countries had to be quarantined in special facilities for 14 days. Everyone who was tested positive for COVID-19 had to be treated in special hospitals only.

We collected the data on 668 patients with COVID-19 from 9 hospitals in Tashkent city (3), Tashkent Region (2), Samarkand (1), Fergana (1), Namangan (1), and Bukhara (1). The database contains daily information on each patient from the date of hospitalization until their discharge or death. The severity of the condition was decided by medical professionals based on daily observations of the patients. The start of the study was March 15, 2020, and the end of the study was June 1, 2020. We excluded from the database all censored patients. Consequently, we had 592 negative cases and 55 positive cases. Among positive cases, there were 15 lethal outcomes. The study used retrospective data for training, model selection, and testing.

Development, validation, and test sets

We randomly selected 80% of negative cases for the development set, 10% of negative cases for the validation set, and the rest 10% of negative cases for the test set. We randomly chose 50% of positive nonlethal cases for the validation set and the rest for the test set. All 15 lethal cases were included only in the test set.

Selected predictors

Based on the study of related work and availability of data, we used the following predictors for the prognosis: sex, age, comorbidities, complications, body temperature, breathing rate, heart rate, blood pressure systolic and diastolic, X-ray examination, auscultatory signs of pneumonia, SpO_2. We used fast kNN imputation [14] to impute the missing values for the selected predictors only.

External test

The Republican Scientific Centre for Emergency Medicine evaluated the resulting model on an external test set. It contained 130 patients with severe COVID-19 and 4 patients with moderate COVID-19. Of them, 25 patients had died, the rest had recovered.

2.2 Models

For this study, we used both classification methods and anomaly detection methods.

We used methods and metrics from the sklearn library version 0.23.2, python 3.7.

Classification models

We used logistic regression, random forest, extra trees, support vector machine (SVM), K-means, K-nearest neighbors. Classification methods require both classes to be represented in the training set. We combined development and validation sets as the training set and used fivefold cross-validation to select the best model of each algorithm. The criterion was the F1 score of the model.

Anomaly detection models

We used Gaussian mixture, isolation forest, local outlier factor, and one-class SVM. We trained each model on the development set and used the validation set to find the proper value of the threshold for the scores produced by each algorithm. The criterion was the F1 score on the best value of the threshold. Anomaly detection methods use the largest class for the training. Considering the high imbalance of the classes in our dataset, these methods are the most appropriate for the task.

Reduction of false negatives

As the treatment of COVID-19 changes, more people would have COVID-19, and the accuracy of the model would drop in real-time application. That is why we decided to modify the F1 score and calculated a harmonic mean of precision and recall squared:

$$\mathrm{F1}_m = \frac{2}{\frac{1}{\mathrm{prec}} + \frac{1}{\mathrm{rec}^2}} = \frac{2 \cdot \mathrm{prec} \cdot \mathrm{rec}^2}{\mathrm{prec} + \mathrm{rec}^2} \tag{1}$$

We used only the F1 score to measure the accuracy of the models on validation and test sets, while we used the modified F1 score only for the threshold selection.

Measurement of accuracy

We evaluated the best models on the test set. We used the F1 score and C-index as metrics of our models' performance. During the testing procedure, we did not differ lethal cases and counted them all as positives.

3 Results and Discussion

3.1 Prognostic Models

We have built a variety of prognostic models based on different machine learning algorithms using the same predictors for all models. Owing to the limitations on the number of pages, we provided all tables with statistical data in [15].

Table 1 Performance of the best classification and anomaly detection models on the test set

Algorithm	F1	C-index	Algorithm	F1	C-index
Logistic regression	**0.845**	**0.954**	Isolation forest	0.774	0.933
Random forest	0.627	0.729	Local outlier factor	0.771	0.938
Extra trees	0.627	0.729	One-Class SVM	0.780	0.928
Support vector machine	0.654	0.743	Gaussian mixture	**0.901**	**0.961**
K-means	0.615	0.720			
K-nearest neighbors	0.364	0.606			

Classification models

The best classification method for our task is logistic regression. It outperforms support vector machines by more than 28% in each metric used.

Anomaly detection models

All anomaly detection algorithms were trained on the development set using only negative samples. During the validation phase, we selected the best set of hyperparameters along with the best threshold for the models. We then evaluated the best models on the test set. The best anomaly detection algorithm is the Gaussian mixture (K-means). We chose the Gaussian mixtures as the best method for the task (Table 1).

Predictors selection

We used the best model as the baseline. We performed training of the Gaussian mixture models, excluding one predictor at a time and evaluating the model on combined validation and the baseline model's false positives from the development set. We assume that if excluding a predictor from the training procedure leads to a decrease in the model performance, then this predictor is important for the task at hand. Thus, the most valuable predictors (in descending order) for the selected algorithm were SpO_2, age, breath rate, heart rate, and pneumonia signs in X-ray examination.

Building the optimal model

Based on the predictor analysis, we decided to remove complications and auscultatory signs of pneumonia from the set of predictors. We trained a new model and evaluated the model on the test set. We presented the results in Table 2.

Table 2 The confusion matrix of the optimal Gaussian mixture model

Ground truth		Positive	Negative
Predicted	Positive	32	2
	Negative	3	58

Robustness analysis

We evaluated predictors using the optimal model. We flipped categorical data and counted the number of patients that had their risk level changed. For numerical data, we changed each predictor by small values with a step of 0.001. We normalized all numerical data and the standard deviation was 1 for all predictors. We plotted the ratios of patients for which the optimal model changed and the predicted risk level due to the change of the selected predictor on the graph shown in Fig. 1.

Using inverse transformation, we calculated the increment values that led to the altering of 50% of patients' risk levels predicted by the optimal model.

External test results

Medical experts used the model to test the risks of new patients after the end of the study. They found that in 130 cases out of 134 the model correctly assigned patients with their risk level.

The resulting program

We uploaded the optimal model onto the webserver. The prognostic model returns a risk value (from the range 0 to 100%). To transform scores returned by the model to

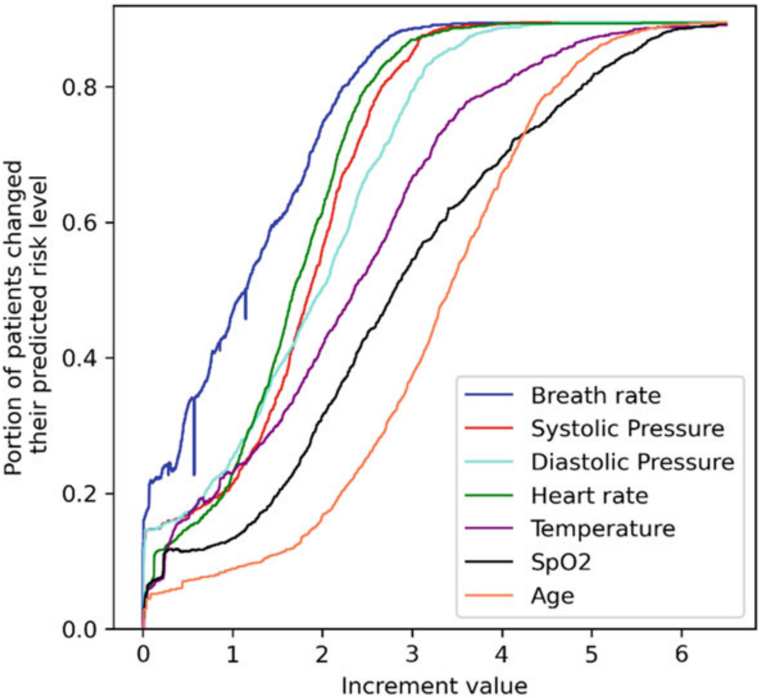


Fig. 1 The proportion of inpatients that had their risk level changed after the increase in the predictors of a given value according to the optimal model

the risk values, we used the sigmoid function:

$$\text{Risk} = S\left(\frac{T - \text{score}}{10}\right) * 100\%, \tag{2}$$

where $S =$ sigmoid function, $T =$ the threshold.

Based on the transformation method we assign the risk levels to the patients: Low $\leq 33\%$, $33\% \leq$ medium $< 50\%$, $50\% \leq$ high $\leq 66\%$, very high $> 66\%$.

3.2 Discussion

The resulted model demonstrated not only high metrics values but also the necessity for finding an optimal subset of predictors.

The first policy of Uzbekistan led to the collection of possibly unbiased data on COVID-19 patients as the study followed all those patients in the controlled environment regardless of the severity of the illness.

We used both feature exclusion and incremental individual conditional expectation to interpret the model.

4 Conclusion

We showed that the Gaussian mixture model is more suitable for our task than the widely used logistic regression and other machine learning algorithms. The resulting optimal model has the F1 score of 0.928 and C-index of 0.965.

We demonstrated the importance to the model of the predictors as well as stated increment values that led to change in the predicted risk level based on changes in the performance of the optimal model. The most valuable predictors for the selected model are SpO_2, age, breath rate, heart rate, and pneumonia in X-ray examination.

5 Availability and Requirements

The dataset is available for any interested parties upon request to the corresponding author via email. The model is implemented and uploaded on the webserver. It is freely available on http://185.74.5.185/main.html.

Acknowledgments The authors would like to thank the staff of the Medical Centre involved in the study for collecting patients' data. They would like to thank Dr. Khamid Kasimov for performing external testing, Maksim Bolonkin for the discussion, ideas of possible models to use, and language correction.

Authors Contributions AK and KA organized work on data collection, data verification, task formulation. FA analyzed the data, reviewed existing models and publications. AI developed, trained, and evaluated models, performed data processing, and programmed the resulting software.

Funding Funding is expected from the Ministry of Innovation Technologies of Uzbekistan.

Competing Interests The authors declare that they have no common competing interests.

References

1. CDC COVID-19 Response Team (2020) Severe outcomes among patients with coronavirus disease 2019 (COVID-19): United States, February 12–March 16, 2020. MMWR Morb Mortal Wkly Rep. 2020 Mar 27; 69(12):343–346
2. Oran D, Topol E (2020) Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. Ann Intern Med M20–3012
3. Garcia-Vidal C, Moreno-García E, Hernández-Meneses M et al (2020) Personalized therapy approach for hospitalized patients with COVID-19. Clin Infect Dis. https://doi.org/10.1093/cid/ciaa964
4. Liu Y, Gayle A, Wilder-Smith A et al (2020) The reproductive number of COVID-19 is higher compared to SARS coronavirus. J Travel Med 27(2)
5. Baqui P, Bica I, Marra V et al (2020) Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study. Lancet Glob Health 8(8):E1018–E1026. https://doi.org/10.1016/S2214-109X(20)30285-0
6. Management of Patients with Confirmed 2019-NCoV. Centres for Disease Control and Prevention, 6 Apr 2020. www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html
7. Wu Z, McGoogan J (2020) Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese centre for disease control and prevention. JAMA
8. Docherty A, Harrison E, Hardwick H et al (2020) Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO clinical characterization protocol: prospective observational cohort study. BMJ 2020:369. https://doi.org/10.1136/bmj.m1985 (Published 22 May 2020)
9. Ortolan A, Lorenzin M, Felicetti M et al (2020) Does gender influence clinical expression and disease outcomes in COVID-19? A systematic review and meta-analysis. Int J Infect Dis. https://doi.org/10.1016/j.ijid.2020.07.076
10. Klein S, Dhakal S, Ursin R et al (2020) Biological sex impacts COVID-19 outcomes. PLoS Pathog 16(6):e1008570.https://doi.org/10.1371/journal.ppat.1008570
11. Wynants L, Calster B et al (2020) Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 369:m1328. https://doi.org/10.1136/bmj.m1328
12. Collins G, Reitsma J, Altman D, Moons K (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med 13(1):1 PMID: 25563062
13. de Lusignan S, Dorward J, Correa A et al (2020) Risk factors for SARS-CoV-2 among patients in the Oxford royal college of general practitioners research and surveillance centre primary care network: a cross-sectional study. Lancet Infect Dis 20(9):1034–1042. https://doi.org/10.1016/S1473-3099(20)30371-6
14. Obadia Y, The use of KNN for missing values. https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c637
15. Tables with statistics. https://github.com/ikramov/covid19/blob/main/covid19tables.pdf

Transition Probability-Based Detection of Hardware Trojan in Digital Circuits

Usha Mehta and Jayesh Popat

Abstract Hardware security is becoming a major concern and threat due to the emerging hardware Trojan attacks. The threat poses due to malicious hardware Trojans during SoC life cycles are major causes of a security breach, financial theft and malfunctioning of SoCs. An attacker may mount such an attack by keeping the goal of operation failure or information leakage. Today's SoC design and fabrication processes involve untrusted parties at different stages of the IC life cycle. Therefore, it increases the vulnerability of such attacks. In this paper, we have inserted hardware Trojans to low-probability nodes of ISCAS combinational benchmark circuits to mimic the real-life scenario. We propose the novel transition probability-based technique to detect the insertion of malicious hardware Trojan. The technique is performed by calculating probabilities on each net before and after Trojan insertion. Experimental results show that a logical OR gate as a Trojan gate has a minimal effect on the transition probability of primary outputs compared to all other logic gates.

Keywords Hardware Trojan attacks · Side-channel analysis · Trojan detection · Transition probability · Probabilistic analysis · Hardware security

1 Introduction

In the era of nanotechnology and complex SoC designs, and short time-to-market, the third-party IP cores play a vital role in the VLSI industry. The SoC designs have become so much complex that fabless designs houses need to work with many parties to fulfill all the requirements of the market. These parties may include third-party IP suppliers, EDA tools, standard cell libraries and overseas manufacturing facilities [1].

The hardware Trojans are referred to as malicious modifications made during the design or fabrication process. The ICs can therefore contain secret functions that

U. Mehta (✉) · J. Popat
Institute of Technology, Nirma University, Ahmedabad, Gujarat, India
e-mail: usha.mehta@nirmauni.ac.in

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_57

trigger under rare conditions. These are termed as “Hardware Trojans”. Trojans are placed normally in the hard-to-control or observation points [2–5]. US military has published evidence of such malicious Trojans in ICs recently [5–7].

Hence, there is an utmost need of detecting, preventing or tolerating these unwanted changes in SoC. The detection of hardware Trojan is possible through pre-silicon verification, physical verification or manufacturing test [8].

This paper involves the transition probability-based method to detect hardware Trojans in combinational circuits. The ISCAS’85 benchmark combinational circuits are taken into consideration for experimental purposes.

The paper is divided into several sections. Section 2 includes the Trojan taxonomy and the most recent Trojan detection techniques. Sections 3 and 4 contain the proposed approach and its results correspondingly. Finally, Sect. 5 outlines the conclusion.

2 Trojan Taxonomy

The attacker may insert the Trojan in a circuit so that it can be difficult to detect such Trojan through a post-silicon test. In order to achieve such a situation, the triggering or activation of Trojan is only possible under very rare conditions. Hence, the erroneous output is only produced under the sequence of multiple rare conditions. The malfunctioning output node is called the payload node. Figure 1 illustrates hardware Trojan taxonomy in terms of trigger and payload [9, 10].

The trigger method is categorized into two types: digital and analog. Digital triggering may be done in two ways: combinational trigger or sequential trigger. During a combinational trigger, the rare triggering condition on the trigger node will produce an erroneous output value on the payload node. While in the case of a sequential trigger, the erroneous output value is produced by sequentially applying

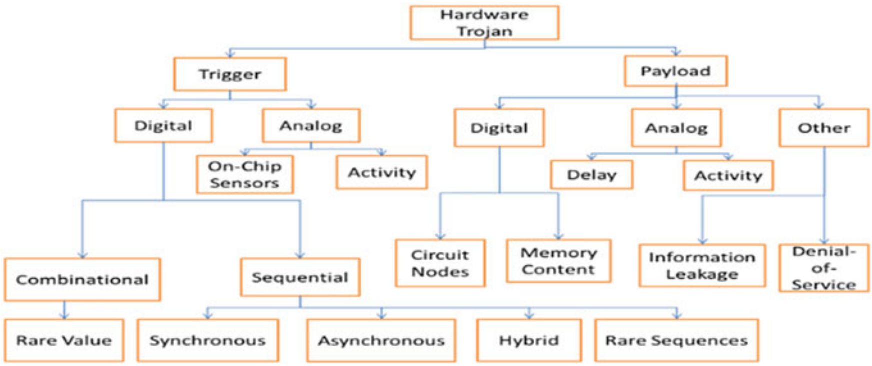


Fig. 1 Trojan taxonomy [9]

rare node conditions multiple times to trigger nodes synchronously or asynchronously [11].

Analog Trojans require on-chip sensors to be triggered. The ring oscillator and counter may raise the temperature of the chip due to high switching activity to perform malfunction [12].

Considering payload approaches, Trojans can be classified into two basic categories: digital and analog. In the case of digital payload Trojans, the circuit node may be affected and produce a wrong value, or tampering in the memory data may happen. However, analog payload Trojans are in the form of a pull-up resistor or pull-down capacitor. They may change analog parameters of circuits like power, noise, delay and switching activity.

The other types of payload Trojans include information leakage and denial-of-service. Information leakage involves the leaking of confidential and sensitive information through radio signals. A denial-of-service attack will not allow the users to access their own system functionality.

The most recent Trojan detection approaches are categorized in the following types: destructive and non-destructive. Destructive approaches require silicon delayering of IC, taking images of metal layers using a scanning electron microscope (SEM), comparing this netlist with the golden one. The non-destructive approach has five possible ways of detecting the presence of Trojan: 1. Logic testing [13]; 2. Side-channel analysis [14–16]; 3. IP trust verification [17]; 4. Design-for-security (DFS) [18]; 5. Run-time monitoring [10].

3 Our Approach

The proposed approach is introduced for the detection of undesired circuits inserted into the netlist. Our approach focuses on considering the fact that an attacker at the design site may introduce Trojan in the design, and to mimic such a situation, we are going to add different Trojan gates in the netlist. The transition probability on every net in the netlist is useful to detect extra malicious circuitry.

Transition probability can be defined as the multiplication of the probability of Logic 0 and Logic 1 on any net in the given circuit. We target low-probability areas in the netlist because those are the prime concerned areas for an adversary to insert extra hardware circuitry. The proposed approach algorithm is defined as below.

3.1 Algorithm

The algorithm for performing the approach is mentioned in the paper [19].

1. Primary inputs nets probabilities are set to (½, ½) for Logic 0 and Logic 1.
2. Set the netlist under consideration as the current design.

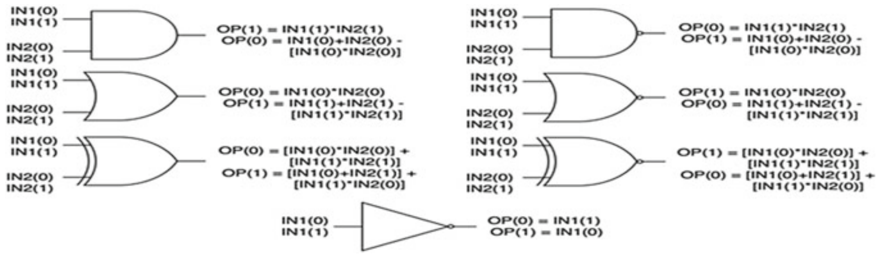


Fig. 2 Probability calculation of each Boolean logic gate

3. Evaluate Logic 1 and Logic 0 probabilities according to gate type on each net in the original netlist traversing from PI to PO.
4. Compute transition probability (P_t) for every net as per the below equation:

$$\text{Transition Probability } P_{ti} = P_i(0) \times P_i(1) \tag{1}$$

   where *i* indicates net number.
5. Arrange all nets probability in ascending order in the table and find out the lowest probability net where Trojan can be added in the netlist.
6. Insert an extra Trojan gate in the lowest probability area of the current design to mimic the actual Trojan insertion scenario. This netlist is called Trojan inserted netlist.
7. Evaluate again all the nets probability in the Trojan inserted netlist.
8. Compare the transition probability of all the nets in the original and Trojan inserted netlist until we find mismatches.
9. Mismatch in the transition probability value indicates Trojan affected area.

3.2 Types of Trojan

The different types of Trojan that we consider for the experimental purpose are the basic Boolean logic gates. We then analyze that which type of gate is likely to be inserted by an adversary as a Trojan at the low-probability area of the circuit. Figure 2 shows the Boolean logic probabilities of different logic gates.

3.3 Application to ISCAS'85 Benchmark Circuits

We apply the proposed detection strategy to very widely referenced ISCAS'85 combinational benchmark circuits. For initial experiments, C17 benchmark circuit is chosen. First, all the nets Logic 0 and Logic 1 probability in the original C17 benchmark circuit is evaluated, as illustrated in Fig. 3a. Then transition probability

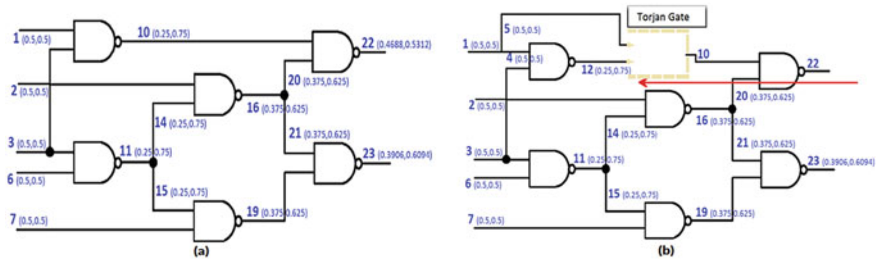


Fig. 3 ISCAS benchmark circuit (C17) with probability distribution at each net. (a) Original circuit (b) Trojan inserted circuit

Table 1 ISCAS benchmark circuit (C17) nets with lowest transition probabilities

Net number	Transition probability
10	0.1875
11	0.1875
14	0.1875
15	0.1875

is calculated on each net and stored in the array. The lowest transition probability nets (10, 11, 14, and 15) are identified from the array and given in Table 1.

We, therefore, consider one of the potential locations for our experimental purpose where Trojan has to be inserted. Since net 10 has the lowest probability, we have introduced a different logic gate as the Trojan gate at that location to mimic the real scenario.

Fig. 3b illustrates the Trojan-affected netlist. The algorithm is again run for this netlist and output net transition probabilities are computed for each net for Trojan inserted ISCAS benchmark circuit. Table 2 indicates a change in the transition probabilities at the primary outputs for corresponding types of inserted Trojan.

4 Experimental Setup, Results and Discussion

The proposed work is implemented using C language. All the experiments are conducted on a computing machine with a Core2Duo processor with 1.4 GHz speed and 1.5 GB of memory.

As mentioned in the previous section, the algorithm applies to the original as well as Trojan inserted netlist for the C17 ISCAS'85 benchmark circuit. Two tables, indicating transition probability at different nets, are produced.

The algorithm compares the original netlist transition probabilities with Trojan inserted transition probabilities at different nets considering Table 2. The comparison results in the nets differ in the transition probabilities values. Table 2 shows that the transition probability values at net 10 and primary output net 22 differ from the

Table 2 Transition probability calculation for original and trojan inserted ISCAC benchmark circuit (C17)

Net No	Original Netlist	AND Trojan	OR Trojan	NOR Trojan	NAND Trojan	XOR Trojan	XNOR Trojan
1	0.25	0.25	0.25	0.25	0.25	0.25	0.25
2	0.25	0.25	0.25	0.25	0.25	0.25	0.25
3	0.25	0.25	0.25	0.25	0.25	0.25	0.25
4	–	0.25	0.25	0.25	0.25	0.25	0.25
5	–	0.25	0.25	0.25	0.25	0.25	0.25
6	0.25	0.25	0.25	0.25	0.25	0.25	0.25
7	0.25	0.25	0.25	0.25	0.25	0.25	0.25
10	0.1875	0.2343	0.1093	0.1093	0.2343	0.6875	0.6875
11	0.1875	0.1875	0.1875	0.1875	0.1875	0.1875	0.1875
12	–	0.1875	0.1875	0.1875	0.1875	0.1875	0.1875
14	0.1875	0.1875	0.1875	0.1875	0.1875	0.1875	0.1875
15	0.1875	0.1875	0.1875	0.1875	0.1875	0.1875	0.1875
16	0.2343	0.2343	0.2343	0.2343	0.2343	0.2343	0.2343
19	0.2343	0.2343	0.2343	0.2343	0.2343	0.2343	0.2343
20	0.2343	0.2343	0.2343	0.2343	0.2343	0.2343	0.2343
21	0.2343	0.2343	0.2343	0.2343	0.2343	0.2343	0.2343
22	0.2490	0.1794	0.2477	0.072	0.2379	0.590	0.3857
23	0.2380	0.2380	0.2380	0.2380	0.2380	0.2380	0.2380

C17 benchmark original netlist values. Considering the primary output net first and moving backward in the design netlist, it had brought into notice that two fan-in nets (10 and 20) of the NAND gate cause the primary output (net 22) to be generated. Further investigation yields that net 20 transition probability in Trojan inserted netlist remains unchanged. This leads to further checking of all the fan-in nets from net 10 in the design netlist; however, it can be seen from both tables that no fan-in nets transition probability is altered from the original netlist values. Hence, we must say that the culprit has accomplished malicious alteration amid net 22 as well as net 10 in the original circuit netlist. This strategy provided the following information: Trojan insertion and its location in the design netlist.

It is evident from Table 3 that different types of Trojan gates produce different transition probability values on primary output. However, it is evident from the experimental results that Logic OR gate is the most likely choice for an adversary since it makes a negligible effect on the transition probability of primary output (PO) nets if the Logic 0 probability is very less than the Logic 1 probability at Trojan inserted site. All other Trojan gates have a comparatively high effect on the transition probability of PO nets.

For experiment purposes, we have considered different ISCAS'85 combinational benchmark circuits, their netlist and basic Boolean gates as Trojan. The netlist is fed

Table 3 Change in transition probability of primary output nets in case of different Trojans

	ISCAS'85 benchmark circuits							
	C17		C499		C880		C1355	
Types of Trojans	Original Netlist Tr*	Modified Netlist Tr*	Original Netlist Tr*	Modified Netlist Tr*	Original Netlist Tr*	Modified Netlist Tr*	Original Netlist Tr*	Modified Netlist Tr*
AND	0.249	0.179	0.0017	0.00039	0.1775	0.1776	0.2499	0.249
OR	0.249	0.247	0.0017	0.0015	0.1775	0.174	0.2499	0.179
NAND	0.249	0.072	0.0017	0.00023	0.1775	0.052	0.2499	0.068
NOR	0.249	0.237	0.0017	0.0025	0.1775	0.167	0.2499	0.489
XOR	0.249	0.590	0.0017	0.0036	0.1775	0.340	0.2499	0.741
XNOR	0.249	0.385	0.0017	0.0028	0.1775	0.296	0.2499	0.450

Tr* = Transition probability of primary output net to the proposed algorithm. After that, the probabilities of Logic 0 and Logic 1 are evaluated on different nets. The transition probabilities on each net are calculated by the algorithm. Next, low transition probability nets are the major focus for Trojan gate insertion. We have inserted different types of Trojan gates for experimental purposes. Table 3 includes the experimental results for widely used ISCAS'85 benchmark circuits as well as different types of Trojan gates. It contains original circuit transition probabilities on primary output as well as Trojan inserted netlist transition probabilities on the same output for different types of inserted Trojans.

Since malicious Trojan insertion changes the net probabilities, the path from the primary output (PO) to primary input (PI) in the original and Trojan netlist is traversed back while comparing all the net probabilities until a mismatch is found.

5 Conclusion

This paper mainly focuses on hardware Trojan detection using novel transition probabilistic techniques. The widely cited ISCAS'85 benchmark combinational circuits are taken into consideration. The proposed approach must be applied during the pre-silicon phase in VLSI flow. The Trojan gate coverage is better in our proposed schedule for all the cases. With our scheme, we get a difference in the transition probabilities at primary output nets, and traversing back the same net path in the design netlist leads to the identification of the Trojan circuit.

Acknowledgments The authors are thankful to Dr. K.S. Dasgupta and Dr. Virendra Singh for their guidance, reviews and constructive suggestions for this research work.

References

1. Semiconductor Industry Association (SIA) (2008). Global billings report history (3-month moving average). http://www.siaonline.org/galleries/Statistics/GSR1976-March09.xls
2. High Performance Microchip Supply (2008). Annual Report by the Defense Science Board. http://www.acq.osd.mil/dsb/
3. DARPA. TRUST in integrated circuits (TIC) (2007)
4. Anderson MS, North CJG, Yiu KK (2008) Towards countering the rise of the silicon Trojan, vol 11, pp 2–5. Australian Government DoD-DSTO
5. Adee S (2008) The hunt for the kill switch. IEEE Spectr 45(5):34–39
6. Koushanfar FA (2008) Extended abstract: designers hardware trojan horse. In: Proceedings of IEEE international workshop on hardware oriented security and trust. Anaheim, IEEE Press, USA
7. King ST, Tucek J, Cozzie A, Grier C, Jiang W, Zhou Y (2008) Designing and implementing malicious hardware. In: Proceedings of the 1st USENIX workshop on large-scale exploits and emergent threats (LEET), pp 1–8
8. Abramovici M, Bradley P (2009) Integrated circuit security: new threats and solutions. In: Proceedings of the 5th annual workshop on cyber security and information intelligence research: cyber security and information intelligence challenges and strategies, pp 55. ACM

9. Bhunia S, Hsiao MS, Banga M, Narasimhan S (2014) Hardware Trojan attacks: threat analysis and countermeasures. Proc IEEE 102(8):1229–1247. https://doi.org/10.1109/jproc.2014.2334493
10. Chakraborty R, Narasimhan S, Bhunia S (2009) Hardware Trojan: threats and emerging solutions. In: High level design validation and test workshop. IEEE International, pp. 166–171
11. Wolff F, Papachristou C, Bhunia S, Chakraborty RS (2008) Towards Trojan-free trusted ICs: Problem analysis and detection scheme. In: IEEE design automation and test in Europe, pp 1362–1365
12. Jin Y, Kupp N (2008) CSAW 2008 team report (Yale University). CSAW embedded system challenge. http://isis.poly.edu/vikram/yale.pdf
13. Alkabani Y, Koushanfar F (2009) Consistency-based characterization for IC Trojan detection. In: Proceedings of IEEE/ACM international conference on computer-aided design, pp 123–127
14. Baktir S, Güçlüoğlu T, Özmen A, Alsan HF, Macit MC (2012) Detection of Trojans in integrated circuits. In: IEEE international symposium on innovations in intelligent systems and applications, pp 1–5
15. Wang LW, Luo HW (2011) A power analysis based approach to detect Trojan circuits. In: IEEE international conference on quality, reliability, risk, maintenance, and safety engineering, pp 380–384
16. Jin Y, Makris Y (2008) Hardware Trojan detection using path delay fingerprint. In: IEEE international workshop on hardware-oriented security and trust, pp 51–57
17. Guo X, Dutta RG, Jin Y, Farahmandi F, Mishra P (2015) Pre-silicon security verification and validation: a formal perspective. In: Proceedings of the 52nd annual design automation conference, pp 1–6
18. Tehranipoor M, Cliff W (2011) Introduction to hardware security and trust. Springer
19. Popat J, Mehta U (2016) Transition probabilistic approach for detection and diagnosis of Hardware Trojan in combinational circuits. In: IEEE annual India conference, pp 1–6

Data Protection Based on Hidden Space in Windows Against Ransomware

Joon-Young Paik, GeunYong Kim, Seoyeon Kang, Rize Jin, and Eun-Sun Cho

Abstract As ransomware gets more sophisticated, it becomes increasingly challenging to protect user data from ransomware. Ransomware detection tends to entail file loss in spite of successful detection. A great file backup would be to perfectly protect user files from a ransomware attack. Infection of Windows systems has become the most urgent and serious problem, accounting for 99.7% of the systems infected by ransomware. In this paper, we propose a new file backup mechanism to protect user data against ransomware. For this, we use the stealth space, alternate data streams (ADS), in the Windows system. In our mechanism, original files are backed up to an ADS-based hidden secure area in a local system (i.e., a user system), and the recovery keys of the backup files are stored in a remote server. The use of the ADS property allows us to stealthily keep the backup files in a local system while going completely unnoticed by ransomware. The experimental results showed that the encrypted files by a realistic ransomware sample were perfectly protected by our backup mechanism with much smaller transferred data than a traditional remote file backup system.

Keywords Ransomware · Defense · Data protection · File backup · Alternate data streams · Windows

J.-Y. Paik (✉) · R. Jin
Tiangong University, Tianjin, China
e-mail: pjy2018@tiangong.edu.cn

R. Jin
e-mail: jinrize@tiangong.edu.cn

G. Kim · S. Kang · E.-S. Cho
Chungnam National University, Daejeon, South Korea
e-mail: gokenny21@cnu.ac.kr

S. Kang
e-mail: rkdtjdus9@cnu.ac.kr

E.-S. Cho
e-mail: eschough@cnu.ac.kr

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_58

1 Introduction

Ransomware is becoming one of the most dangerous threats in the cyber world. Ransomware encrypts victims' files and then demands a ransom in exchange for decryption keys for the encrypted files [1]. This simple strategy has affected vast numbers of individuals, industries, and government agencies. WannaCry ransomware caused approximately $4 billion in financial losses in a single day [1].

Ransomware targets machines running the Microsoft Windows operating system. WannaCry exploited a vulnerability in the Windows SMBv1 server. Windows operating systems accounted for 99.7% of the systems infected by ransomware in January 2019 [2]. The most urgent issue is to protect Windows systems against ransomware.

Ransomware detection has been studied extensively [3–7]. Most detection techniques use the behavioral features of ransomware. They detect ransomware samples reliably. However, they do take a certain period of time to make a decision on attacks, which usually results in the loss of a few files.

Another option for defense against ransomware is to use file backup to preserve original files against loss. The secure space for file backup can be in local or remote regions. To support a secure local space, an existing system may need to be modified (i.e., ShildFS [5] and RDS3 [6]). Alternatively, secure remote spaces can be used in external resources, as in CLDSafe [7]. Original files are transferred to a backup space on a remote server. The amount of transferred data depends on the size of the backup files. File backup mechanism, if properly configured, should efficiently store original files with little modification to existing systems.

In this paper, we propose a file backup mechanism against Windows ransomware using alternate data streams (ADS). Our mechanism manages secure local and remote spaces. Original files are backed up in the local space and their recovery keys are stored in the remote space. To preserve original files in the local space, we utilize ADS, which is supported by the new technology file system (NTFS) in Windows [8]. ADS functions as a hidden space because files linked with ADS are not normally visible (e.g., in Windows explorer). A secure local space is created using ADS to stealthily keep original files. The remote space is created on an external server. Only the recovery keys for the backup files are transferred to the remote space. Our backup mechanism allows us to securely backup user files without little modification to existing systems.

This paper makes the following contributions:

- As far as we know, this is the first work to use ADS to defend against ransomware.
- Our proposal is easily applied to the Windows operating system without any system-level modifications.
- Compared to normal backup systems with client–server architectures, our mechanism reduces the amount of transferred data via a network.

The remainder of this paper is organized as follows. Section 2 describes ADS. Previous efforts to defend against ransomware are explained in Sect. 3. We propose our file backup mechanism in Sect. 4. Section 5 evaluates our file backup method with a realistic ransomware sample. Section 6 contains the conclusion.

2 Alternative Data Streams

NTFS has been the default file system of the Windows operating system since Windows NT was released in 1993. It provides a functionality called ADS. ADS allows data to be attached to a regular file or directory; multiple pieces of data can be linked as *streams*. The data attached to a file or a directory are not exposed by normal interfaces. Therefore, ADS can serve as hidden spaces.

The stealthy character of ADS has attracted hackers interested in concealing malicious code. An example of how a hacker can insert malware.exe as a stream to ADS on the Windows calculator (i.e., calc.exe) at the command prompt is as below.

```
C:\> type c:\malware.exe > c:\windows\system32\calc.exe:mal-
     ware.exe
```

Then the malicious program is launched by typing this command: "start c:\windows\system32\calc.exe:malware.exe". The existence of the malware.exe in this normal file is not visible to users via Windows explorer, including administrators. Moreover, the size of the calc.exe does not change; the size of the calc.exe is 27,647 bytes before and after the malware.exe is attached to calc.exe. Malicious code can also be attached to the ADS of a directory.

In our proposal, ADS is exploited to protect original files against file encryption attacks of ransomware, which is different from the malicious use by hackers.

3 Related Works

Ransomware is a type of malware that encrypts user files and demands a ransom. To prevent encrypted files from being decrypted, ransomware developers use symmetric or asymmetric encryption algorithms or both. The keys used for encryption and decryption are managed by a command-and-control server (C&C), for which platforms such as Slack and Telegram can be used. Offline ransomware encrypts user files without C&C communication.

The ransomware attacks are detected by effective detection strategies [3, 4]. They usually leverage dynamic features revealed during ransomware execution; file entropy, file similarity, file access patterns, and so on. However, despite successful ransomware detection, they inevitably lose a few files because they make a decision based on accumulated information over a period of time.

On the other hand, file backup mechanisms focus on getting original user files pre-served when the systems are infected by ransomware. Original data that have been backed up are stored in local or remote areas.

RDS3 [6] locates two independent execution environments in a single system, using isolation techniques (i.e., Xen and VirtualBox): a virtualized system for user's daily use and another virtualized system for backup purposes. Data in the user operating system are periodically copied to the tiny operating system. Then, later, the backup data are able to be recovered to the user operating system by its recovery mechanism. ShieldFS [5] supports file recovery at a file system level. It constructs a shadow drive for backup. Whenever a file is modified or deleted, the file is copied to the shadow drive. SSD-Insider [9] has the functionality of data recovery at flash-based SSDs. It copies original data to a hidden space visible only to a storage device not to a host system. To do this, the functionality is implemented in the firmware in SSDs. To deploy those methods, a few modifications on existing systems are imposed; two different virtualization instances for RDS3, a new Windows driver for ShieldFS, and a special SSD firmware for SSD-Insider. On the contrary, the basic functionality, ADS, of Windows is used in our proposal.

On the other hand, there are backup methods based on remote servers. CLDSafe is executed on cloud storage servers [7]. A user computer (i.e., a client-side) has a synchronized folder. Files in this folder are automatically copied to the server when they are updated. Then the cloud server selectively backs them up when the similarity between a new file and its old one is below a threshold.

In such a client–server architecture, the amount of transferred data via network depends on that of backup files. That is, it could take non-negligible time for backing up a large file. In our proposal, only small metadata of backup files are transferred rather than the entire files.

4 File Backup Mechanism Using ADS

In this section, we first show the architecture of our backup system for Windows. Then we explain how it backs up original files by exploiting ADS.

4.1 Architecture

Our goal is to easily integrate the functionality of file backup into an existing system. To achieve this, the following considerations should be taken into account.

- File backup functionality is added without any modification to the existing system.
- Backup space is invisible to ransomware.
- Files are backed up, leaving no evidence that ransomware is able to identify.

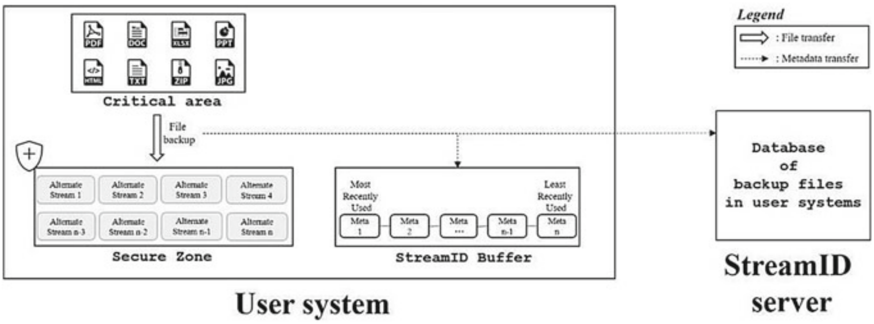


Fig. 1 Overall architecture

We handle the first consideration by exploiting ADS. To address the second point, we make an ADS-based hidden space for file backup which ransomware cannot readily find. The third consideration is related to creating and maintaining recovery keys for original files. The keys are temporarily kept in RAM on a user system to manage the backup space, which prohibits ransomware from finding the keys.

Figure 1 shows the architecture of our backup mechanism. *Critical* area indicates the directories being monitored. It can be expanded to cover the whole system. *Secure zone* is a hidden backup space that is built based on ADS. It is attached to ADS belonging to a file or directory. When an attempt is made to destroy files in the critical area by the update or deletion requests, the files are backed up to the Secure zone before any execution of such requests. *StreamID buffer* includes recovery keys of backup files in the Secure zone. The recovery keys are managed by a least recently used (LRU) replacement policy in RAM. *StreamID server* maintains the recovery keys of the backup files, which are synchronized with those in the StreamID buffer. The recovery keys are permanently stored for file recovery.

We recommend that the Secure zone is attached to the ADS of a vital file or a directory (i.e., Windows, Program Files, and Program Files (×86) directories). Ransomware does not usually touch such system files or directories because it must let victim systems work well so that users can pay ransom.

4.2 System Details

Our backup mechanism consists of two components: a backup manager and a server manager.

Backup Manager. A backup manager is responsible for backing up original files and managing their recovery keys. It backs up original files to the Secure zone with their recovery keys, which are the stream names to the Secure zone. Then it inserts the stream names into the StreamID buffer in a local system, and synchronously sends them to the StreamID server.

The manager monitors suspicious activities in the critical areas; we focus on write and deletion operations because read operations have no impact on file contents. When an existing file is requested to be modified or deleted, it is further analyzed; the format of the demanding file is checked. Our interest is restricted to the file formats that ransomware usually targets (.txt,.jpg,.doc, and many others). When the file format matches one in our file format list, it prepares to back up the file to the Secure zone.

Subsequently, it checks whether the Secure zone is large enough to accept the file. If there is not enough free space, the backup manager makes more space by eliminating a backup file from the Secure zone; the backup manager evicts its stream name from the StreamID buffer according to LRU replacement policy, requests the StreamID server to remove the same stream name from the database, and the backup file corresponding to the evicted stream is eliminated from the Secure zone. This is repeated until there is available space for the new file.

Then the file is moved to the Secure zone. Its contents are backed up to the Secure zone as a form of the stream whose name is its relative path. At the same time, its stream name is inserted into the StreamID buffer and is sent to the StreamID server.

Server Manager. The StreamID buffer is free from file encryption attacks of ransomware because it is a form of in-memory data structure in RAM. However, the data in the buffer disappear when a user system is powered off. Therefore, the StreamID server is designed to improve the reliability of our mechanism. A database, *StreamID_DB*, is created on the StreamID server and has the stream names of the backup files.

A server manager in the StreamID server interacts with the backup manager to update stream names in a synchronous manner and manages the StreamID_DB. When the StreamID server receives an insertion request for a stream name, it inserts the stream name into the StreamID_DB. When it receives a deletion request for a stream name, it deletes the stream name from the StreamID_DB.

5 Experiments

5.1 Experimental Setup

We implemented the proposed backup mechanism in two execution environments: a user system and a StreamID server. The user system was a Windows system, and the remote server was a Linux system. In the user system, we set a directory as the critical area. The Secure zone was created with the ADS of an arbitrary existing directory in C:\Windows. The size of the Secure zone was limited to 10 GB.

The critical area was monitored by leveraging the user-level hooking technique, Deviare [10]. We focused on APIs related to file writes and file deletions because they destroy the contents of files. When an API call in which we were interested was heading to a file in the critical area, the file was stored into the Secure zone. On the

Table 1 User files in the critical area

	Documents				Images		Compressed files	
	.txt	.docx	.pdf	.xlsx	.jpg	.gif	.zip	.rar
Number of file	100	100	100	100	100	100	100	100
Average file size (in bytes)	32,370	110,390	162,596	35,127	74,579	6,135	41,607	13,192

StreamID server, the MySQL database managed the stream names that the backup manager sent.

A sample, jigsaw, of ransomware was employed; it used the AES256 algorithm. Various file formats that ransomware mainly targets were put into the critical area. Table 1 shows the information about the victim files in the critical area. The files consisted of documents (.txt,.docx,.pdf, and.xlsx), images (.gif and.jpg), and compressed files (.zip, and.rar). The total number of files is 800.

5.2 *Experimental Results*

We first evaluated the effectiveness of the proposed mechanism using the realistic ransomware sample. The Jigsaw sample encrypted all the files in the critical area, and deleted their original files, while our mechanism backed up the original ones to the Secure zone. We confirmed that all the original 800 files were perfectly kept in the Secure zone when the system was infected by the ransomware sample, as shown in Table 2. It means that the Secure zone was protected against the attack and the recovery keys in RAM were correctly managed by the StreamID buffer.

Next, we compared our backup system with a client–server-based backup system. Table 3 lists the details of these two backup systems. The proposed system was Model

Table 2 Number of backup files by the proposed mechanism

Ransomware name	Number of encrypted files	Number of backup files
Jigsaw	800	800

Table 3 Comparison of the proposed backup system and a client–server-based backup system

Model	Backup system	Local backup	Remote backup
Model I	The proposed system	O (files)	O (stream names)
Model II	Client–server-based backup system	X	O (files, absolute paths of backup files)

Table 4 Comparison of the average amount of transferred data and the average time for a file backup

Model I		Model II
The average amount of transferred data via network (in bytes)	49.99	60,355.22
The average time to back up a file (in *ms*)	15.9	36.7

I. It utilized local and remote backups. Model II was developed based on a client–server backup approach. It backed up user files to a remote server by transferring their entire contents when files in the critical area are modified or deleted. Models I and II used a remote server. However, Model I transferred only the stream names of backup files, unlike Model II.

Both models perfectly backed up all the original files but caused different performance overheads. Table 4 lists the average amount of transferred data and the average time for a file backup. Model I transferred much less data than Model II because the stream names of the backup files are much smaller than their contents. In terms of the average time to back up a file in the critical area, Model I outperformed Model II. Model I backs up user files to a local area by minimizing the network overhead while Model II backs up them to a remote area via the network. These results imply that our backup system backs up original files fast by reducing the network overhead by sending only the recovery keys.

6 Conclusions

Our new backup mechanism for defending against ransomware exploits the ADS of a Windows operating system to create a Secure zone for file backup. The use of the ADS enables the protection of original files against ransomware without any system-level modifications. Under ransomware attacks, the mechanism backs up user files to the Secure zone in a local system and gets their recovery keys stored in a remote system, which reduces the network overhead. According to the experimental results, our mechanism perfectly protected the original files by backing them up to the Secure zone. In addition, our system reduced the backup time by decreasing network overhead, compared to a traditional client–server-based backup system.

Acknowledgments This work was supported, in part, by the National Natural Science Foundation of China (NSFC) under Grant 61806142, the Natural Science Foundation of Tianjin under Grant 18JCYBJC44000, and the Tianjin Science and Technology Program under Grant19PTZWHZ00020 and, in part, by the Institute for Information & communications Technology Planning&Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01343, Regional strategic industry convergence security core talent training business).

References

1. Symantec: WannaCry: lessons learned 1 year later. https://www.symantec.com/blogs/feature-stories/wannacry-lessons-learned-1-year-later. Accessed 2 Sep 2019
2. Analysis of ransomware epidemic in February 2019. https://blog.360totalsecu-rity.com/en/analysis-of-ransomware-epidemic-in-february-2019. Accessed 15 Sep 2019
3. Scaife N, Carter H, Traynor P, Butler K (2016) CryptoLock (and Drop It): stopping ransomware attacks on user data. In: IEEE international conference on distributed computing systems (ICDCS), pp 303–312
4. Kharaz A, Arshad S, Mulliner C, Robertson W, Kirda E (2016) UNVEIL: a large-scale, automated approach to detecting ransomware. In: USENIX security symposium, pp 757–772
5. Continella A, Guagnelli A, Zingaro G, Deasquale G, Barenghi A, Zanero S, Maggi F (2016) ShieldFS: a self-healing, ransomware-aware filesystem. In: Annual conference on computer security applications (ACSAC), pp 336–347
6. Subedi K, Budhathoki DR, Chen B, Dasgupta D (2017) RDS3: ransomware defense strategy by using stealthily space space. In: IEEE symposium series on computational intelligence (SSCI), pp 1–8
7. Yun J, Hur J, Shin Y, Koo D (2017) CLDSafe: an efficient file backup system in cloud storage against ransomware. IEICE Trans Inf Syst E100(9):2228–2231
8. Broomfield M (2006) NTFS alternate data streams: focused hacking. Netw Secur 2006(8):7–9
9. Baek S, Jung Y, Mohaisen A, Lee S, Nyang D (2018) SSD-insider: internal defense of solid-state drive against ransomware with perfect data recovery. In: IEEE conference on distributed computing systems (ICDCS), pp 875–884
10. Deviare2 Hook Homepage. https://github.com/nektra/Deviare2. Accessed 01 Oct 2020

Image Processing: Comparative Analysis of Face Processing by AI and Humans

Mikhail O. Matveev, A. Zhuravishkin, and Denis D. Yershov

Abstract We deal with dozens of human faces every day, including those of our relatives, friends, colleagues and the many familiar faces of media celebrities. Dealing with faces is quite a complex task involving a plethora of factors that affect their recognition such as viewing angle, illumination, face expressions affected by emotions and speech movements. Despite the considerable empirical study, the human face is still in the focus of researchers' attention, and the development of technical and methodological research methods has led to an increase in the number of fundamental and applied research. Face perception is a rather extensive area that requires an interdisciplinary approach uniting social, biological and technical sciences. A deep understanding of face perception will, in its turn, help us create more precise and applicable computer algorithms aimed at face processing. For this reason, experimental research is required to deeper understand the face perception mechanisms. Our short investigation showed that the main distinction between human and AI face processing lies in the way people and machines deal with the information they receive. While AI works with the image as a whole, humans distinguish certain aspects of the face image and work with them. It is clearly seen when memorizing something. We hope that our conclusions could be corroborated by creating algorithms that will imitate the human mind and conducting comparative studies involving humans and AI.

Keywords Face processing · AI · Psychology · Face perception · Mental images

M. O. Matveev (✉)
Moscow State Institute of International Relations (MGIMO University), 76, Prospect Vernadskogo, Moscow 119454, Russia
e-mail: mike-matveev@mail.ru

I.M. Sechenov First Moscow State Medical University, 8/2, Trubetskaya, Moscow 119048, Russia

A. Zhuravishkin
Institute of Linguistics of the Russian Academy of Sciences, Bolshoy Kislovsky Ln, 1/1, Moscow 125009, Russia

D. D. Yershov
Moscow State Linguistic University, 38/1, Ostozhenka, St. Moscow 119034, Russia
e-mail: Yershovden@yandex.ru

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_59

1 Introduction

We deal with dozens of human faces every day, including those of our relatives, friends, colleagues and the many familiar faces of media celebrities. Dealing with faces is quite a complex task involving a plethora of factors that affect their recognition, such as viewing angle, illumination, face expressions affected by emotions and speech movements. It may be possible to recognize faces last seen a long time ago. Another important factor that influences face perception and recognition is the change in the head structure and facial features that accompanies human growth and ageing with the latter affecting both skin and face contours, as well as hair and eyebrow colour.

Both AI researchers and psychologists are exploring the mechanisms underlying face representations called mental images. Understanding these mechanisms may one day offer insight into the critical problems that still challenge both psychologists and AI researchers: the mechanisms of processing and storing information. It is worth saying that even the best face recognition algorithms are well on their way to achieving the recognition ability humans show for unfamiliar faces, but have a long way to go before they can compete with humans in the best-case scenario.

Each human face is special and provides information about its owner's identity. Humans deal with hundreds (if not thousands) of individual faces. This far exceeds our ability to keep in mind individual objects of any type. Thus, the recognition and identification processes occur separately, which means that faces are coded perceptually, and this particular code can be stored in memory without referring to other semantic information (a person's name, position in society, country of origin, etc.). Because all types of information are interconnected, machines, unlike humans, always identify an individual's label or the tag when recognizing their face.

It should be noted that, despite the considerable empirical study, the human face is still in the focus of researchers' attention, and the development of technical and methodological research methods has led to an increase in the number of fundamental and applied research.

2 Methodology

Face studies are quite a complex multidisciplinary area that relies on a number of methods. On the one hand, psychological sciences deal with faces to deeper understand the psychological and neurophysiological mechanisms underlying the various aspects of face perception. Here we can mention cognitive studies of healthy participants, developmental studies of neonates to adolescents, impairments in neuropsychological, psychiatric, and developmental conditions, brain imaging techniques, computer simulations and comparative research identifying homologues in non-human species. Neuroimaging techniques have provided us with a better understanding of human brain structure and function. We are now seeing the impact of these techniques on theoretical frameworks of face perception. As for the technical sciences, many attempts are being made to use the psychological and neurological data to modify and optimize automated face recognition systems.

In recent decades, face studies have been intensively developed in general and social psychology, psychophysiology and clinical psychology. Informatics, neuroscience, anthropology, medicine, ethnography and aesthetics show an acute interest in the problem. It has been proved that a person's face is directly related to personality, identity, self-awareness as an individual; a face can express a complex range of personal, socio-cultural and emotional characteristics; there are specialized brain areas that are activated only when looking at people's faces; a special type of visual impairment—prosopagnosia—a face recognition disorder has been revealed; a connection between genetic abnormalities, psychopathic disorders and face morphology and structure has been established. Modern cosmetic surgery is capable of transforming a person's face; the development of automated face recognition systems makes it possible to use a face to identify a person, as well as his fingerprints; the socio-cultural dynamics of the portrait genre in painting and literature, etc.

Speaking about the object of face studies, it includes dozens of thousands of individual human faces.

Both computer vision researchers and psychologists are trying to understand how it is possible to create a representation of faces that are used by people when they know someone well.

Improving the quality of a face representation produced when a person gets acquainted with another individual may be useful for understanding the mechanisms underlying face recognition by both humans and artificial intelligence. This type of investigations may also be of use when trying to define the conditions under which human face recognition can be assumed reliable. Nowadays, there is a big room for improving even the best modern face recognition software to compete with the human's recognition ability for dealing with faces. For this reason, computer scientists should turn to the data obtained by psychologists to apply it in their work.

If we look at the face processing from the psychological viewpoint, we can say that object perception produces its mental image, which is then used to substitute the real-world object when dealing with it without physical access to it. Thus, this mental image can be called a kind of reflection of this object, its avatar. We can agree with N.J.T. Thomas, who defined mental imagery as a quasi-perceptual experience resembling perceptual experience; it occurs in the absence of the appropriate external stimuli. It is also generally understood to bear intentionality (i.e., mental images are always images of something or other), and thereby to function as a form of mental representation. Traditionally, visual mental imagery, the most discussed variety, was thought to be caused by the presence of picture-like representations (mental images) in the mind, soul or brain, but this is no longer universally accepted [1].

Very often, imagery experiences are understood by their subjects as echoes, copies or reconstructions of actual perceptual experiences from their past; at other times they may seem to anticipate possible, often desired or feared, future experiences.

Interest in the study of face perception and its expressions is associated with the fundamental psychological problem: the ratio of external and internal in the mind [2]. For example, in the course of communication, the main attention is paid to the person's face, as we assess the other individual's personality and his or her emotions. According to V. Barabanshchikov, the subject of perception sees their interlocutor as an independent unity not reducible to its constituent parts, signs of emotions and personal characteristics [3].

In addition, the person's inner world is constantly read out from the face during communication, so some researchers compare it with a text [ibid]. (It is also worth saying that our mental image of a person changes as we "read" them. Our perception of a person may change radically as we look at certain facial features.) In fact, it is a typical situation when a physical object substitutes another physical object or becomes a symbol of something. For example, a human face is a physical object that represents the person, thus playing a part of the avatar. The face is the first thing we see when we communicate and it is the main source of information about a person. Nevertheless, it is necessary to take into account the fact that the perception of the person's psychological portrait, as well as his or her emotions, does not only imply scanning their facial features or structures with the observer's mind eye; it is a much more complex process in which intellectual, emotional and volitional components are involved, as well as the personality of the subject of perception as a whole [ibid]. It happens because a face is a complex structure that provides interlocutors with a variety of information about a person. Thus, a face can be considered as a multidimensional, hierarchically organized entity consisting of different "physiognomic layers" [4; 2]. The surface layer consists of facial patterns, which reflect the emotional state of a person in a particular situation. The middle layer affects more stable patterns of the face, formed by its deformations, folds, characteristic squints etc., formed over many years and reflecting certain personality traits. The deep layer is represented by the face constitution, i.e., its shape, size, ratio of its parts and is associated primarily with the temperament and constitution of the human body [4]. Physiognomic layers are superimposed on each other, merging into a multidimensional whole, which is actually called a "facial expression" [5].

We conducted a psychological experiment to find out what mental mechanisms are involved in face perception. Even though our sample is based on human respondents, it would be of big scientific interest to compare our results to those achieved by computer scientists.

Participants

The experiment involved 110 respondents from two Russian universities (Moscow State Institute of International Relations, First Moscow State Medical University) aged between 18 and 24 with an approximately equal male-to-female ratio. All respondents were of Russian origin. The number of respondents was sufficient to use the semantic differential method.

Experimental material and procedure

The experimental procedure was the following. The participants were shown images of two men and two women. In the first case, the photographs were of a real male and female with a neutral facial expression sourced from the Internet so that the respondents were very unlikely to have been exposed to each image prior to the experiment. The images were edited to redact all parts except the face. The second case involved exposing the participants to collect images of a contemporary Russian male and female generated using a database that contains a large number of faces of people belonging to different ethnic groups, including the ethnic Russian. The choice of representatives of the Russian ethnic group was determined by the selection of respondents (students of Russian universities, Russian native speakers who indicated this in the questionnaire during the experiment) because the results of the experiment are also affected by the ethnicity of the depicted people and respondents (Brigham 2008) (Figs. 1, 2 and 3).

Fig. 1 Male portrait (left) and photograph (right)

Fig. 2 Female portrait (left) and photograph (right)

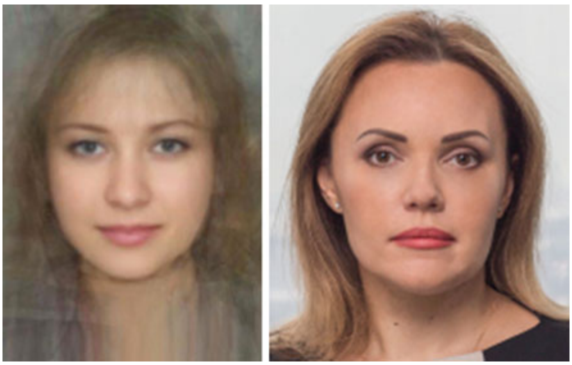


Fig. 3 Form

Form

Instructions: rate on a scale from -3 to 3 how each of the qualities offered to you characterizes this person.

Gender: Age Date: University: Faculty:

	-3 -2 -1 0 1 2 3	Text	Picture	Picture+text
1	evil – good			
2	irritable - calm			
3	stupid - clever			
4	reserved - sociable			
5	shy – easy-going			
6	deceitful - honest			
7	insidious - straightforward			
8	cunning - open			
9	humble - confident			
10	lazy – hard-working			
11	Laid-back - ceremonious			
12	selfish - altruistic			
13	greedy - generous			
14	coward - courageous			
15	rude - tender			
16	dangerous - safe			
17	undetermined - determined			
18	compliant - stubborn			
19	negligent - neat			
20	weak - strong			

After perceiving the portraits, students filled in the form assessing the stimuli received from −3 to 3, where −3 meant, for example, that the male in the picture is evil, while 3 meant that he was good, and 0 meant that he is neither good nor evil.

3 Results

The data obtained from the respondents were processed with the SPSS statistical package using the method of factor analysis with the subsequent transformation of the factors to a simple structure. As a result of data processing, four factors were identified (energy, character, potency and activity). Then, the results were presented in the form of semantic spaces and factors. The factors are named on the basis of their included scales and are presented in descending order of contribution to the total variance. This method originates from Osgood's works on semantic differential and it is used for measuring the word meaning [6–8]. Its versatility of uses with the bipolar adjectives and the simplicity of understanding them have made this method popular in psychological and psycholinguistic research.

As for the factor analysis method itself, it is worth saying that it implies special mathematical procedures for the simplification of interrelated measures to discover patterns in a set of variables [9]. Factor analysis originates back to the early 1900s when Charles Spearman created his two-factor theory that inspired other researchers

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \ldots\ldots a_{jm}F_m + e_j$$

Fig. 4 Factor analysis' mathematical formula

Table 1 Abbreviations used

Definition	Abbreviation
Memory image created by man's portrait	MPMem
Memory image created by man's photo	MPhMem
Perception image created by man's portrait	MPPerc
Perception image created by man's photo	MPhPerc
Memory image created by female portrait	FPMem
Memory image created by female photo	FPhMem
Perception image created by female portrait	FPPerc
Perception image created by female photo	FPhPerc

on creating a number of theories [10]. This type of analysis is used by specialists in social sciences to deal with the data they receive.

In the "classical factor analysis" mathematical model, p denotes the number of variables (X1, X2, …, Xp) and m denotes the number of underlying factors (F1, F2, …, Fm). Xj is the variable represented in latent factors. Hence, this model assumes that there are m underlying factors whereby each observed variables is a linear function of these factors together with a residual variate. This model intends to reproduce the maximum correlations (Fig. 4).

The graphs with the results of our experiments are given below, as well as their description and analysis.

The abbreviations used are given in Table 1.

First of all, we should note that there is a difference between the mental images formed in different manners, i.e. the perception images and the memory images. The former we formed in respondents as follows: they received the pictures and then they were asked to fill the questionnaire while still having the pictures before their eyes. The latter were formed as follows: the respondents were asked to examine the given pictures for 1 min and then they gave the pictures back to receive the questionnaire (provided above).

When analysing the answers, we can conclude that the characteristics of mental images differ depending on the way they were formed. The difference between them is shown in the following graphs.

It is worth saying that a number of scientists agree that the visual characteristics of a face in themselves have an influencing potential and can serve as a stimulus for the formation of mental images; moreover, respondents will judge not only the look of the faces they see but also the characters of the people depicted, based only on their appearance [11–19] (Fig. 5).

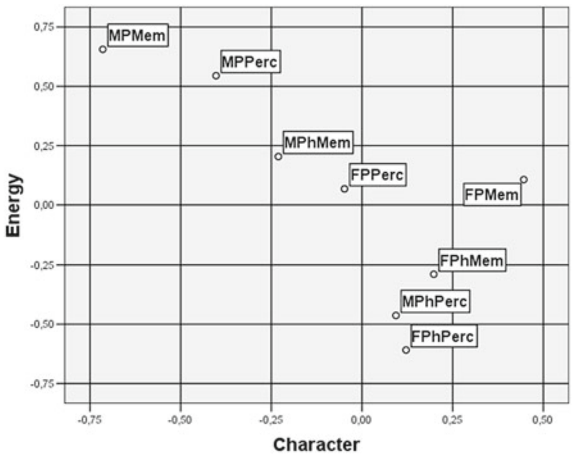


Fig. 5 Graph "Energy—Character"

Energy

It is worth mentioning that, when dealing with memory images, respondents evaluated the people on the pictures higher than when dealing with their perception images. This effect can be explained by the mere exposure effect [20, 21].

Character

If we compare the results, we can see the difference between male and female images: men are considered to have a better character when dealing with their perception images, while women's character is considered to be better when dealing with memory images.

Besides, the female character is considered positive, while the male character is considered negative. These two phenomena can be explained by the gender stereotypes that still exist in society [22]. When dealing with memory images, people work not only with the image of the person they have just seen but also with the stereotypes that exist in the society, while in the case with the perception images, their effect on respondents is weaker (Fig. 6).

Potency

If we compare the characteristics of perception and memory images in terms of this factor, we can come to the conclusion that perception images have less pronounced values. It means that memory images of people's faces tend to seem to us more vivid than those of perception. It may be explained by the fact that, as mentioned before, human memory does not just store the exact image of a person, but deals with combinations of human's features (both psychological and appearance) to optimize storage and easier extract this kind of information. This fact can also explain the following phenomenon: while the memory images of portraits are considered less potent than the perception images, the situation with the photographs is the opposite (Fig. 7).

Fig. 6 Graph "Energy—Potency"

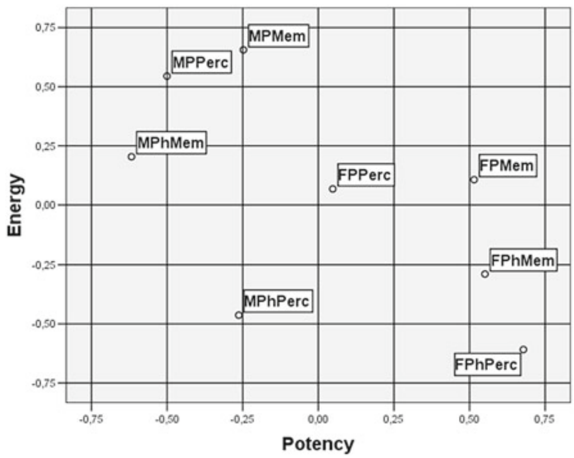
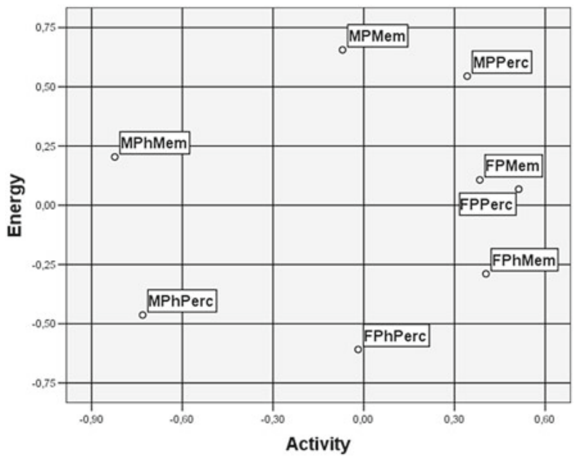


Fig. 7 Graph "Energy—Activity"



Activity

If we compare the mental images from the perspective of the fourth factor, we can conclude that men and women were considered more active and vivid when remembered than when perceived. This phenomenon can also be explained by the memory structure: the respondents created a more vivid image of the persons they had seen because they did not store a simple visual image in memory but a number of features can be combined together. To do so, they had first to analyse the person's characteristics based on their previous experience. Here, we can also see the effect of gender stereotypes: males are considered to be less active than females.

4 Discussion

Face perception is a rather extensive area that requires an interdisciplinary approach uniting social, biological and technical sciences. A deep understanding of face perception will, in its turn, help us create more precise and applicable computer algorithms aimed at face processing. These algorithms can be created using the principles the human mind uses to work with the information it receives. It can be similar to the way biologically inspired engineering uses biological methods and systems found in nature to the study and design of engineering systems and modern technology. For this reason, experimental research is required to deeper understand the face perception mechanisms [23].

Our short investigation showed that the main distinction between human and AI face processing lies in the way people and machines deal with the information they receive. While AI works with the image as a whole, humans distinguish certain aspects of the face image and work with them. It is clearly seen when memorizing something. We hope that our conclusions could be corroborated by creating algorithms that will imitate the human mind and conducting comparative studies involving humans and AI.

However, there are some limitations as well. First, even though our results may be interesting to computer vision and AI specialists, we still have not started an investigation to test our psychological hypotheses on AI and computer-based models. Second, face perception is an emerging area which, of course, has big potential, but lacks some theoretical base that will help us to work more effectively. We hope that using several scientific methods (psychological, neurological, biological and computer science) will help us to get more data and create a theory of face processing.

5 Conclusions

In this study, we examined the psychological mechanisms underlying face perception. We conducted an experiment to define the factors that influence this process by showing the respondents face images as stimuli and using the semantic differential method to create a model of face perception. Although the semantic differential method has been used since the 1950s, it is still reliable and worth using in different spheres, for example, face studies. It is worth saying that this very method has not been used to study faces (even though, it is widely used in other spheres, for example, when assessing the meaning of things and concepts [8]), that is why our investigation can be considered the first of this kind. The semantic differential technique allows one to get a schematic representation of respondents' mental images using mathematical methods. We suppose that combining psychological methods with those of the technical and natural sciences can provide us with the data we can use to create a model of cognitive processes. It is possible that results would vary if we take more stimuli and make this research more extensive. Future researchers should consider investigating face perception using various stimuli, i.e. more face types, providing respondents with different types of pictures (black-and-white vs. colour) different angles (full-face, half-face, etc.), various stimuli (photos, videos and paintings), etc. Regardless, our results point to the need for various researchers to use a combination of methods for gathering even more information about the face perception mechanism and using it in computer studies.

References

1. Thomas NJT (1999) Are theories of imagery theories of imagination? An active perception approach to conscious mental content. Cognitive Science (23):207–245
2. Rubinshteyn SL (2002) Osnovy obshchey psikhologii—Izdatel'stvo: Piter, p 720
3. Barabanshchikov VA, Zhegallo AV, Korol'kova OA (2016) Pertseptivnaya kategorizatsiya vyrazheniy litsa. M.: Kogito-tsentr, 376 p
4. Diveyev DA (2009) Rol' formy litsa v vospriyatii individual'no-psikhologicheskikh kharakteristik cheloveka: Dis. … kand. psikhol. nauk. M.: IP RAN
5. Barabanshchikov VA (2009) The perception of facial expressions [Vospriyatie vyrazhenij lica]. Moskva, IP RAN, p 360
6. Osgood CE (1962) Studies on the generality of affective meaning systems. Am Psychol 17(1):10–28. https://doi.org/10.1037/h0045146
7. Osgood CE, Suci GJ, Tannenbaum PH (1957) The measurement of meaning. University of Illinois press, p 47
8. Snider JG, Osgood CE (1969) Semantic differential technique: a sourcebook. Aldine, Chicago, p 681
9. Child D (2006) The essentials of factor analysis. A&C Black, p 200
10. Harman HH (1976) Modern factor analysis, 3rd edn. University of Chicago Press, Chicago, p 508
11. Bzdok D, Langner R, Caspers S, Kurth F, Habel U, Zilles K, Laird A, Eickhoff SB (2011) ALE meta-analysis on facial judgments of trustworthiness and attractiveness. Brain Struct Funct 215(3–4):209–223
12. Goodwin GP, Piazza J, Rozin P (2014) Moral character predominates in person perception and evaluation. J Pers Soc Psychol 106(1):148
13. Hassin R, Trope Y (2000) Facing faces: studies on the cognitive aspects of physiognomy. J Pers Soc Psychol 78(5):837
14. Hurwitz D, Wiggins NH, Jones LE (1975) A semantic differential for facial attribution: the face differential. Bull Psychon Soc 6(4):370–372
15. Olivera-La Rosa A, Arango-Tobón OE, Ingram GPD (2019) Swiping right: face perception in the age of Tinder. Heliyon 5(12):e02949. https://doi.org/10.1016/j.heliyon.2019.e02949
16. Oosterhof NN, Todorov A (2008) The functional basis of face evaluation. Proc Natl Acad Sci 105(32):11087–11092
17. Peterson MA, Rhodes G (ed) (2006) Perception of faces, objects, and scenes: analytic and holistic processes. Oxford, Oxford University Press, New York, p 406
18. Ranzini G, Lutz C (2017) Love at first swipe? Explaining Tinder self-presentation and motives. Mobile Media Commun 5(1):80–101
19. Rhodes G (2006) The evolutionary psychology of facial beauty. Annu Rev Psychol 57:199–226
20. Swap WC (1977) Interpersonal attraction and repeated exposure to rewarders and punishers. Pers Soc Psychol Bull 3(2):248–251. https://doi.org/10.1177/014616727700300219.ISSN0146-1672
21. Zajonc RB (2001) Mere exposure: a gateway to the subliminal. Curr Dir Psychol Sci 10(6):224–228. https://doi.org/10.1111/1467-8721.00154

22. Lemish D, Elias N (2019) Perpetuating gender stereotypes from birth: analysis of TV programs for viewers in diapers. In: Hermansson C, Zepernick J (eds) The Palgrave handbook of children's film and television. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-17620-4_27
23. Esomba DS (2012) Twenty-first century's fuel sufficience roadmap, 1st edn, p 659
24. Barabanshchikov VA, Nosulenko VN (2004) Sistemnost'. Vospriyatiye. Sistemnost'. M.: Institut psikhologii RAN, 480 p
25. Bruce V, Young A (2000) In the eye of beholder. The science of face perception. N. Y.: Oxford University Press, p 304

An Event-Driven Architecture (EDA) Adapted to Cloud-Based Hospital Information Systems (HIS)

Mohammed Amine Chenouf and Mohammed Aissaoui

Abstract Cloud computing is a new technology that grant a network access to shared pools of IT configurable resources. The solution can be quickly configured with less management efforts. Increasingly used by companies in all industries, the migration of their information systems to the cloud has allowed them to take advantage of its benefits such as cost optimization, easy provisioning, scalability, simplicity and elasticity. Health care is a critical and complex sector in a continuous evolution that inquires a novel architecture and technology in a goal to improve the performance and minimize the costs of Hospital Information Systems (HIS). This article defines an Event-Driven technical architecture adapted to the HIS, which respects HIS's functional and technical requirements, based on the Cloud Computing technologies. The main benefits are to minimize the implementation and maintenance time and costs of the hospital information system.

Keywords Hospital information system · Cloud computing · Provisioning · Technical architecture · Event-driven · Deployment · Microservices · Scalability and elasticity

1 Introduction

Conventionally, hospital information systems (HIS) come into play to manage all the medical and administrative data related to the diverse entities providing medical services. And recently with the rise of Big Data, the hospital information systems (HIS) becomes more data centric. Indeed, the data flow which circulates within a HIS is exponentially enormous and requires a very developed infrastructure and a very large storage and archiving capacity. Also, the process of data analysis and

M. A. Chenouf (✉) · M. Aissaoui
National School of Applied Sciences, Mohammed Premier University, Oujda, Morocco
e-mail: m.chenouf@ump.ac.ma

M. Aissaoui
e-mail: m.aissaoui@ump.ac.ma

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_60

processing is potentially slow and may not be very accurate. Hospital information systems provide many advantages for the overall management of the entire data flow in healthcare organizations.

But with the current challenges, namely the high digitalization of information systems and Big Data, actual HIS systems have some limitations that can hinder their proper development. In fact, current HIS are often commercial in nature with a closed source code. They present other limitations for handling, in a scalable and automatic way, the nature and the size the healthcare provider.

To overcome these challenges, we propose a novel design of HIS with the following principal characteristics:

- A Cloud-Based infrastructure;
- An Event-Driven Microservices Modular System.

The architecture that we offer allows hospitals to optimize implementation costs and increase the performance of information systems since it will rely on the advantages of the cloud. Cloud computing can significantly improve the quality of health services; it will be the appropriate solution of care coordination, collaboration and communication between the various stakeholders (doctors, patients, administrators and developers …) [11]. For this, we also need a cloud native architecture that will be compatible with cloud computing [7], open to evolution, which allows the technological renewal of the infrastructure while reducing both the time and cost of treatment.

In this article, we will present at first a technical architecture that combines and takes advantage of both the cloud computing technology and microservices architecture. Then we will make a simple test by hosting one instance to a cloud provider and using 3 microservices inside it to make a website of appointments in a hospital establishment.

The article consists of 6 essentials parts:

- Section 1: Global introduction to describe the problem and why the adoption of cloud technologies for our implementation.
- Section 2: Related works to frame our work.
- Section 3: We present the enabling technologies and describe the paradigms adopted for our architecture.
- Section 4: In this part, we will describe the technical architecture adopted and explain in more detail the different components of the solution.
- Section 5: Present the steps and the methodology for our lab test.
- Section 6: This section concludes the article and presents our perspective for future work.

2 Related Work

Many works were done about the building of hospital information system in the Cloud; the goal is the benefits from the advantages of the cloud and other new technologies to analyze, deploy and configure hospital information systems with less efforts and time. We will present some related works that try to discuss the same solution.

- The work in [16] introduces that Fujian department has begun to create a unified medical archive by the staff of Health. It is predicted to cover many hospitals to reach their prescriptions, medical advices, review findings and sharing other information.
- The work in [15] proposes a HIS constructed on the technology and method of XML application. It can realize integration and sharing of information between hospitals and other departments.
- Koufi et al. [9] described a cloud-based Emergency Medical System (EMS); this grants doctors to get to patient data easily and quickly, anywhere and anytime.
- Hsieh and Hsu [6] build a cloud telemedicine service, which allows collaboration in health care between onsite and off-site cardiologists.
- The work in [12] proposed a service based on cloud computing, which enables cloud applications to communicate with each other. The architecture proposed consists of a cloud logic mapping layer, a cloud broker layer and an event-driven layer for communications.
- A driven approach for developing a cloud solution is also described in [4]; the main objectives of the solution are the segregation of applications from the infrastructure, the possibility of computer-assisted modeling and the exact collaboration to make updates.

No one can deny that the maintenance and management of the traditional architecture and technology become more expensive and cumbersome, more than that, it is too hard for both to keep pace with the evolution of IT technology.

Therefore, a suitable architecture and framework need to be developed in order to deal with these limitations.

Cloud computing and Event-Driven architecture are the best way to implement HIS. With this combination, hospitals will improve the performance and quality of services with less efforts and costs.

3 Enabling Technologies

All the enterprises including hospitals and institutes are looking to adopt a new technology that optimizes the time and costs, increases the performance and improves the analysis of the data. This new solution should be flexible and scalable to deal with the evolution of IT in the future.

In this section, we will describe with more detail all the terms mentioned before.

3.1 Cloud Computing

Cloud computing is a new technology allowing access to a set of quasi-infinite IT resources, including hardware and software assets via a broad network access [2, 13].

Cloud computing should have five essential characteristics, namely a broad network access, on demand self-service, measured services, resource pooling and rapid elasticity.

Cloud services offered to users could be either in the form of software as a service (SaaS), users then pay a subscription according to the degree of use of the software, or in the form of a platform as a service (PaaS), users create their own applications with development platforms provided by the cloud provider, or finally as infrastructure as a service (IaaS).

Three Cloud deployment models (public, private and hybrid) are communally used. Firstly, a public cloud service provider produces computing resources that are publicly accessible through the Internet, such as applications, networks and storage. Public cloud services can be free or pay-per-use. Secondly, the private cloud is for a specific organization and does not allow sharing resources with other groups of peoples. Finally, a hybrid cloud usually consists of several sub-infrastructures (private, public). In this infrastructure, organizations provide and manage some IT resources in their own data centers, while providing and managing other IT resources externally.

In our work, we decide to use Cloud Computing technology, because the central advantages of the cloud are performance and low cost. All set of resources are cheaper. The Cloud users like healthcare institutes and hospitals can easily optimize time and cost of the deployment over a Cloud without hiring an IT personal to control the internal infrastructure. As a result, the user can focus on critical tasks without having to incur additional personnel and IT training costs.

3.2 Event-Driven Architecture

Nowadays, several emerging companies are using business models based on the on-demand access (which matches the cloud model) [8]. In addition, the access and the treatment of data in HIS is complex. The Event-Driven architecture (EDA) [5] is the ideal solution for using this information in real time [17], also making changes without any complexity [10].

We can define EDA as a design pattern made around production, detection and response to timely events [14, 19]. It is a design paradigm for dynamic, asynchronous, process-oriented context normalization, with minimal coupling, which makes it a better choice for HIS architecture on the cloud [3].

Event-Driven Architecture benefits:

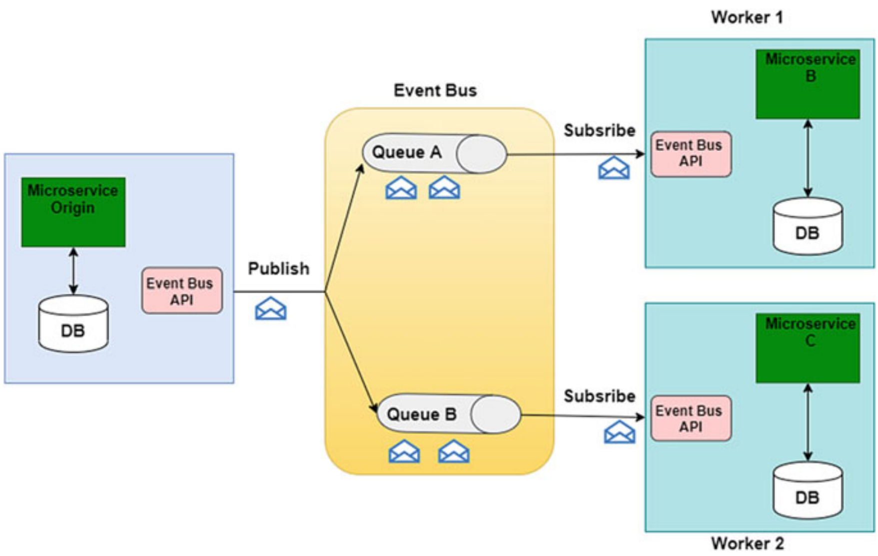


Fig. 1 Event-driven architecture model

- Event-Driven architectures are asynchronous that do not block any tasks. This give resources the possibility to cross freely to the next task once their unit of work is complete, without worrying about what happened before or what will happen in the future.
- Services do not require knowledge or dependencies of other services. When using events, the service runs independently without knowing other services, including its implementation details and transmission protocol.
- Services under the event model can be independently updated, tested and deployed more easily.
- Since the services are divided into an event-driven architecture, and as a microservices execute one task at the time, the management and the update of the service become easy.
- An event-driven solution with queues can "replay" past events to recover lost work. This is useful for preventing data loss when consumers need to recover.

Figure 1 describes the main architecture of a typical Event-Driven system.

4 The Technical Architecture

The technical architecture that we define will provide an HIS in a collaborative way where information systems will try to link devices, disparate systems and stakeholders.

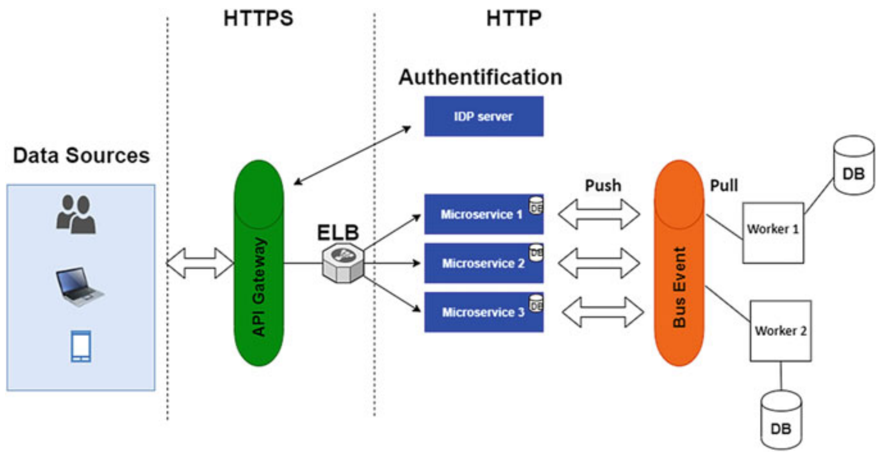


Fig. 2 Simple example of event-driven architecture model

This will bring different services together and create a healthcare solution in the cloud, to reach performance and minimize time and cost of the global solution.

Figure 2 explains with more details how event-driven architecture works in depth.

- HTTPS/HTTP protocols can be secured with a bunch of protocols and tools developed over the last decades, and HTTP is also a standardized communication option.
- The API gateway is located between the client application and the microservice. It works like a reverse proxy. It can also give other functions like authentication, SSL and caching.
- Elastic Load Balancing (ELB) assigns incoming service traffic to the right microservices.

From a functional point of view, the technical architecture that we propose will allow health establishments to reach the following long-term objectives:

- Provide the best possible service through accurate and timely sources of information; Give the best service through the exact information sources;
- Latency Minimal in connected services;
- Minimize manual executions in a goal to reduce human errors;
- Give a high availability of medical information solution;
- Receive an information by a sharing solution.

The technical architecture will respect the following points:

- Interoperability, Integrity and Data accessibility;
- Transformation and data storage;
- Mediation and data federation;
- Events generation;
- Complex health event processing;
- Monitoring.

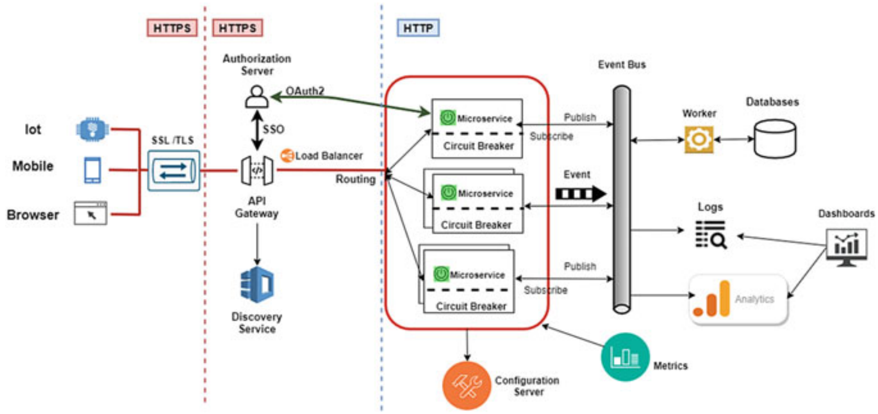


Fig. 3 The technical architecture

Figure 3 summarizes the overall architecture we presented previously; it presents each component with its information flow and gives a technical approach in order to better implement the solution in the cloud.

5 Methodology and Results

Our test consists of making an appointment in a health institute; after the validation of all the information, users receive a confirmation email. To achieve this goal, we will use docker and AWS technology.

In our case, we will use Amazon Web Services because they had achieved the certificate that enables them to store and manage the healthcare systems. Below are the steps for the implementation. First, we create and configure the infrastructure needed for this solution.

For this tested platform, we configured a security group that enable only SSH protocol (Port 22) for remote connection and HTTP protocol (Port 8080) to expose our website to the Internet. More details are presented in Figs. 4 and 5.

After the configuration and the deployment of the infrastructure was completed, we move on to configure Docker. Eventually, we used the public Ip address of our EC2 instance. We proceed then to Docker and microservices configuration.

With this configuration, it easy to make the implementation, upgrade, and the management of hospital information system. More than that, we can optimize the cost and time of the implementation.

The platform is developed in AWS and has given good satisfaction, which demonstrates the feasibility of our HIS solution.

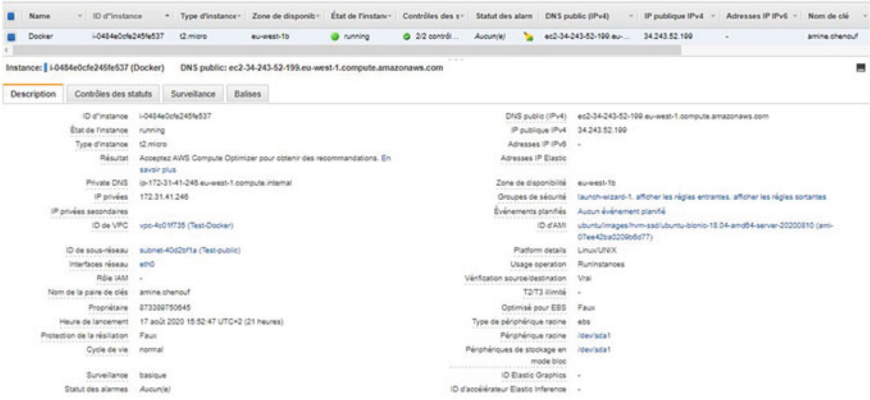


Fig. 4 Instance EC2 description

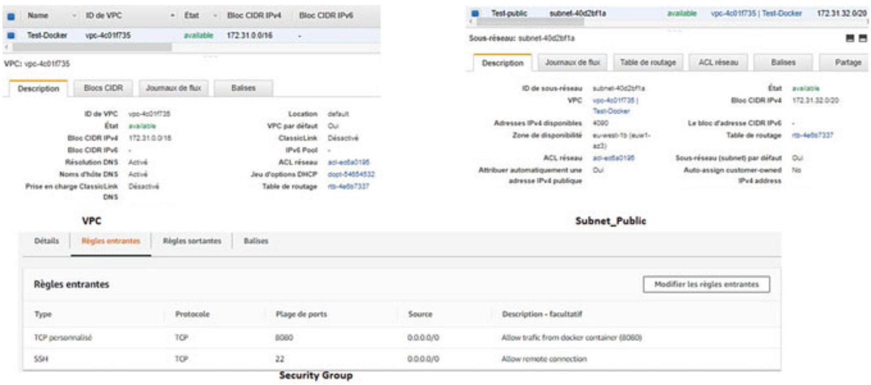


Fig. 5 Network configuration

6 Conclusion

This article presents in depth a technical architecture for Hospital Information Systems, based on Event-Driven architecture and combined with cloud technology as IT service provider. A proof of concept, consisting of a patient appointments platform, based on these principals (EDA, cloud) is presented and demonstrated. The results obtained demonstrated the feasibility of the proposed architecture.

As a future work, we aim at using the combination of an infrastructure as a code, automation scripts and orchestration technologies for the building, the implementation and the configuration of the overall HIS platform without any human interaction.

References

1. Aarab Z, Saidi R, Rahmani MD (2016) Event-driven modeling for context-aware information systems. In: 2016 IEEE/ACS 13th international conference of computer systems and applications (AICCSA), pp 1–8
2. Bahrami M, Singhal M (2015) A dynamic cloud computing platform for ehealth systems. In: 2015 17th international conference on E-health networking. Application Services (HealthCom), pp 435–438
3. Bo D, Kun D, Xiaoyi Z (2008) A high performance enterprise service bus platform for complex event processing. In: 2008 seventh international conference on grid and cooperative computing, pp 577–582
4. Charlton S (2011) Model driven design and operations for the cloud
5. Fitterer R, de Witte B (2009) Enabling pervasive healthcare by means of event-driven service-oriented architectures the case of bed management in mid-sized to large-sized hospitals. In: 2009 3rd international conference on pervasive computing technologies for healthcare, pp 1–4
6. Hsieh J, Hsu M (2012) A cloud computing based 12-lead ECG telemedicine service
7. Jemal H, Kechaou Z, Ben Ayed M, Alimi AM (2015) Cloud computing and mobile devices based system for healthcare application. In: 2015 IEEE international symposium on technology and society (ISTAS), pp 1–5
8. Klarl H, Preitsameter M (2006) Securing service-oriented and event-driven architectures results of an evaluation of enterprise security frameworks. In: 2006 IEEE services computing workshops, pp 89–89
9. Koufi V, Malamateniou F, Prentza A, Vassilacopoulos G (2017) Real-time process analytics in emergency healthcare, pp 147–150
10. Lan L, Li F, Wang B, Zhang L, Shi R (2014) An event-driven service-oriented architecture for the internet of things. In: 2014 Asia-Pacific services computing conference, pp 68–73
11. Takemura T, Kume N, Okamoto K, Kuroda T, Yoshihara H (2012) Development of an environment for information reuse on a hospital information system. In: The 6th international conference on soft computing and intelligent systems, and the 13th international symposium on advanced intelligence systems, pp 1384–1387
12. Livogiannis T, Tahvildari L (2011) A reference model for developing cloud applications
13. Tawalbeh LA, Habeeb S (2018) An integrated cloud based healthcare system. In: 2018 Fifth international conference on Internet of Things: systems, management and security, pp 268–273
14. Tran H, Zdun U (2013) Event actors based approach for supporting analysis and verification of event-driven architectures. In: 2013 17th IEEE international enterprise distributed object computing conference, pp 217–226
15. Waldrop M (2007) Scientific In: 2007 data center in a box—a shipping container stuffed with servers could usher in the era of cloud computing, vol 297, pp 90–93
16. Wang JW, Sun W, Ye X, Chen ST (2005) Construction of citizen healthy information system in xiamen. In: 2015 medical information, vol 18, pp 517–520
17. Wu J, Chen P, Wu N, Chao Y (2011) Event-driven open service platform based remote healthcare system. In: 13th international conference on advanced communication technology (ICACT2011), pp 1559–1562
18. Yu HQ, Zhao X, Zhen X, Dong F, Liu E, Clapworthy G (2014) Healthcare-event driven semantic knowledge extraction with hybrid data repository. In: Fourth edition of the international conference on the innovative computing technology (INTECH 2014), pp 13–18
19. Yuan ED, Watton M (2012) An event-driven service oriented architecture for space command and control decision making. In: 2012 IEEE aerospace conference, pp 1–9

VR Real-Time Monitoring System for Meteorological Observation Devices Integrated with 3D GIS

Qinqiang Zhou, Guiye Huang, Jianyong Li, and Binghuai Chen

Abstract Monitoring mode is single, fault monitoring alarm positioning is not fine and intuitive, and the display mode is monotonous and so on, which are the main disadvantages in most current integrated meteorological observation network monitoring system. Three-dimensional visualization online monitoring system is developed and applied for meteorological observation networks. The virtual modeling of various meteorological detection devices and their operating environment are carried out, utilizing the three-dimensional GIS and virtual reality techniques. The driving mapping rules of real-time observation data and device operation status between the actual scene and the virtual model are established, device fault classification and automatic positioning mechanism are created, then the virtual model component status of all kinds of meteorological detection device is real-time updated with the change of detection data and operation status, by which the simulation and reappearance of monitoring on-line, fault remote diagnosis and guiding maintenance are realized. As is shown in meteorological observation monitoring and support operation business, the level of automation and intelligence for the integrated meteorological observation network monitoring system has been greatly improved, and it has been an efficient monitoring guarantee means for the unattended operation of automatic observation.

Keywords Meteorological device · Virtual reality · Holographic virtual scene · 3D GIS · Fault diagnosis and location · Real-time monitoring

1 Introduction

With the rapid development of the Internet of things, computer and communication technology, the automation and informatization level of meteorological detection device operation and monitoring have been continuously improved. Feizhong et al. [1–4] have researched and developed a national comprehensive meteorological observation operation monitoring system, by which the plane fusion monitoring of

---

Q. Zhou (✉) · G. Huang · J. Li · B. Chen
Guandong Meteorological Data Center, Guangzhou 510080, China
e-mail: 41409858@qq.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_61

all elements, types and quantities of main observation device such as weather radar and automatic weather station and other major observation device has been realized. Wei et al. [5] designed the operation monitoring system for L-band radiosonde radar network, and Dongming et al. [6, 7] carried out the research and development of provincial meteorological detection device operation monitoring business support system in combination with personalized requirements. Scientific and information-based technical support is provided by the above-mentioned systems to give full play to the construction and operation benefits of the integrated meteorological observation station networks. However, there are still some problems in these monitoring systems, such as manual filling, large amount of work by the staff on duty, single display form of two-dimensional report interface, imprecise and non-intuitive positioning of fault monitoring and alarm, and more and more prominent operation and maintenance pressure with an increase in the number and type of devices.

In recent years, the application of new technologies based on 3D GIS, virtual reality (VR) and digital twins in the field of workshop 3D visualization monitoring is relatively mature [8–13], and much attention has also been attracted for its application in the field of meteorology. Many domestic meteorological departments have also carried out relevant research and practice. Shen et al. [14–17] systematically summarized the research status of virtual reality technology in weather phenomenon (cloud, snow, rain) simulation, weather system (flow field, temperature field) simulation and numerical forecast product simulation, as well as its application in meteorological maintenance support, film and television publicity, meteorological science popularization, meteorological disaster scene, etc.; Mei Honghui et al. [18–20] proposed the design and implementation of a three-dimensional visualization system for Doppler weather radar, lightning and other meteorological data. Tiesong et al. [21–24] made in-depth research on the application of 3D GIS platform in the field of meteorological visualization; Yang et al. [25–27] made preliminary exploration and application on the whole process of meteorological information visualization monitoring and automatic weather station visual monitoring. However, the research on 3D visualization monitoring of large-scale ground integrated meteorological observation station networks is relatively less, especially the research and development of operation monitoring 3D visualization system driven by real-time data is still blank, the operation scenario and reality of comprehensive large-scale ground meteorological station networks, the transparency and intelligence of the whole process of monitoring, such as visual comprehensive display of real-time running state and real-time data, accurate and visualized positioning of fault stratification are all not realized. We can only know the fault of observation device, but can't know the specific component failure or abnormality, so the intelligent and visual visualization effect is poor.

Virtual reality (VR) technology and three-dimensional geographic information (GIS) technology are used innovatively for the new generation 3D visual operation monitoring system, based on the intensive, efficient, fast, convenient, sharing concepts, by which nearly 3000 meteorological observation devices and transmission network operation status, as well as the real-time data of various automatic weather stations are integrated, and the 3D virtual devices are updated in real time.

The key business modules of operation monitoring, such as proposed operation scenario, 3D modeling and cutting of device, visualized automatic positioning and synchronous real-time alarm of fault level, and spatiotemporal consistency inspection of data, etc., are developed, by which the operation status and real-time/historical observation data of detection device can be shown intelligently and comprehensively, so as to effectively improve the support capacity and efficiency of meteorological device, promoting the centralized and unified operation and monitoring information and intelligent level of comprehensive meteorological observation station networks.

2 Functional Architecture

The 3D visual operation monitoring system is mainly composed of three parts: multi-source collection of operation and data transmission status for various detection device, real-time driving mapping of state information and real-time observation data processing to 3D virtual model, and visualization display for operation status and real-time observation data with integration of 3D GIS, as shown in Fig. 1.

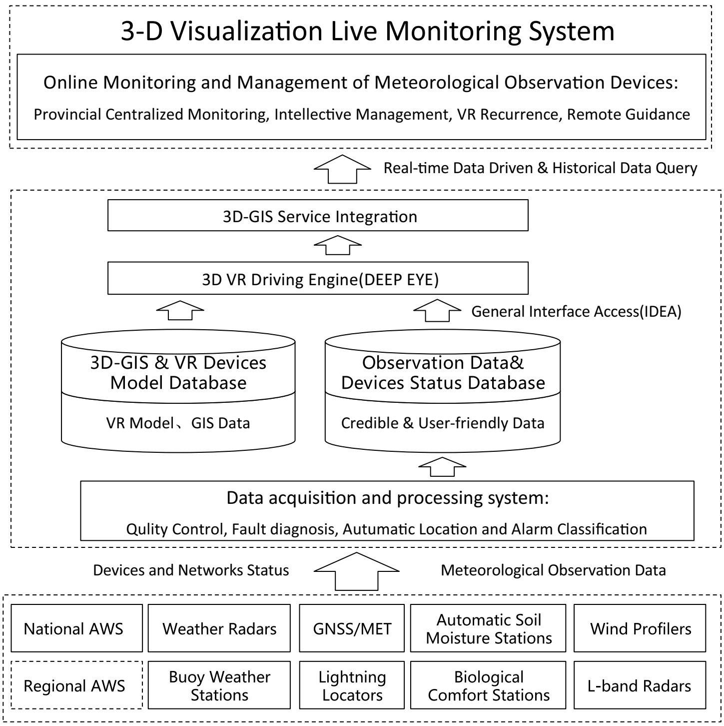


Fig. 1 Main function architecture for VR real-time online monitoring system for meteorological observation devices integrated with 3D GIS

Operation and transmission status data collection. The operation and transmission status information of the original 15 categories of detection device such as weather radar is collected and optimized based on the unified data structure design by using different methods such as TCP/UDP/FTP/Web page grabbing etc. The access of different devices and the analysis of different protocol data are mainly solved.

3D virtual scene modeling. Based on the comprehensive meteorological observation station network information, various detection device structures and business environment and other basic data, the three-dimensional virtual modeling of operation scene and structural components for various devices are carried out, and the three-dimensional GIS system is integrated to construct the three-dimensional visual virtual scene holographic model of the whole network devices.

Real-time data-driven mapping of 3D virtual model. After being analyzed and processed, the operation status, network transmission status and the observed data of various automatic weather stations are coupled with the 3D model-driven engine to update the virtual model status and the live data display in real time.

Auxiliary analysis of maintenance. The authenticity of extreme value of the same factor within the given space–time range is effectively checked and analyzed, by comparative analysis of time and space quality control of various kinds of automatic weather station live data. Fault query statistical and analysis based on the region and time interval can help to improve the support efficiency of the comprehensive meteorological observation devices.

3 Construction of Holographic Virtual Scene Mode

The holographic virtual scene model of integrated meteorological observation stations is constructed based on the actual operation scene and composition structure of various observation device, and unity 3D is used to build the virtual scene model, by which the independent operation scene library and component model library are generated for each type of device. Combined with the basic attribute information of various types of device, station location, elements and other basic attribute information of the observation stations, the mapping driving rules of imported live data and operation status are coupled with the 3D GIS being integrated, as is shown in Fig. 2.

3.1 Material Collection

The integrated meteorological observation stations include 15 kinds of main observation device, such as national automatic weather station, weather radar, wind profile, L-band sounding radar, GNSS/met, automatic soil moisture station, biological comfort instrument, traffic weather station, ship weather station, oil platform weather station, buoy weather station, lightning locator, regional automatic weather station, island weather station, etc. The virtual scene modeling materials mainly include station

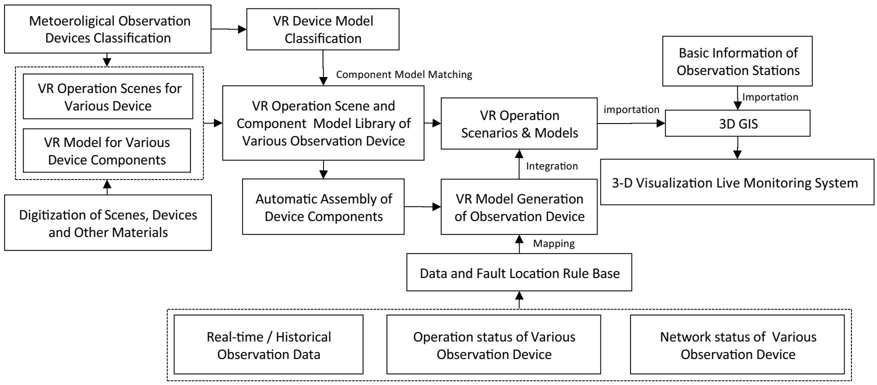


Fig. 2 Construction process of holographic VR scenes for various devices

scene, installation device drawings and photos, such as station construction drawings, built drawings, layout plans, device layout, device structure drawings and supporting manuals, station introduction text or video data, etc. All kinds of materials are digitalized completely one by one, which is used to construct 3D virtual scene of site facilities, buildings and observation device models.

3.2 *Scene Model Construction*

The holographic virtual scene model of all kinds of devices is divided into three levels: operation scene, observation device and structural component. According to the collected materials, the model component library is constructed, and the three-dimensional model of device is quickly assembled, taking weather radar as an example, as shown in Fig. 3. 15 types of device are modeled, including 12 weather radar stations and 16 wind profiler stations. The other 13 types of device are mainly installed in the ground observation field, so only one virtual model is built for each type of device.

3.3 *Basic Data Implantation*

The basic data implantation includes three-dimensional GIS geographic information data of Guangdong Province and basic information of various observation device, including category, station location (longitude and latitude), station name, introduction, contact information, etc., so as to bind the fast, accurate and realistic 3D modeling of station buildings, devices and equipment.

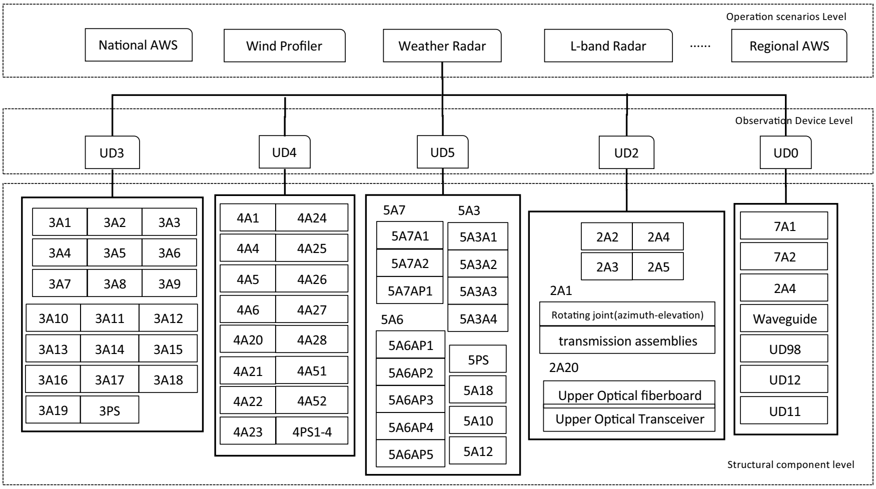


Fig. 3 Three-level structure of operation VR model for each observation device, taking weather radar as an example

In order to ensure the location consistency of various observation device scene models in Google 3D GIS geographic information system, the coordinates of actual operation scene of various devices on Google map are obtained. The virtual scene model of device is implanted into 3D 0.60 m high precision satellite remote sensing image geographic information system. The digital elevation model (DEM) data production and three-dimensional scene generation are completed.

4 Data Acquisition and Processing

The data source of 3D visualization operation monitoring system mainly includes real-time operation status of device or network and real-time observation data of various automatic weather stations. Data acquisition and processing are the important bases for driving real-time intelligent automatic positioning and 3D visualization display of device status.

4.1 Status Information

The operation status of 15 kinds of detection device is from state files separately, which are analyzed and saved to the database in real time. In order to show the various running states of the device and locate the components automatically, it is necessary to preprocess all kinds of state information, including fault judgment rules, hierarchical alarm rules, fault component location rules and so on.

Fault diagnosis. The fault judgment of 15 kinds of detection device is mainly divided into two categories. Weather radar, wind profiler, L-band sounding radar, GNSS/met and lightning locator are all defined by the state documents specified in the functional specification requirements. The fault classification of other automatic weather stations mainly includes power supply, collector, communication and sensor status of each element. Taking this as an example, the fault judgment rules are described in detail, as shown in Tables 1 and 2.

Automatic location of faulty components. According to the device fault judgment rules, the components or component relationships corresponding to each fault alarm of various detection devices are set, so that the components of virtual model driven by 3D engine can make real-time state change and fault prompt. Taking TWP3 wind profiler as an example, the fault component location rules are illustrated. According to the data file upload situation and status monitoring file of wind profile, the fault description and component location are shown in Table 3.

Classified alarm. If all data interruption, fault, module or component alarm, and other abnormal phenomena are given alarm prompt for 15 categories of device, the fault alarm prompt of various detection device will be more disordered, which is not conducive to the selection and priority support of the device with serious fault. Therefore, the fault alarm level is set according to the type, quantity and actual support business demand for each type of device fault alarm in order to carry out efficient and flexible management for all kinds of device monitoring and support according to the fault priority. All kinds of detection devices set the fault alarm level, respectively. The system will automatically block the display of fault alarm higher than the set level in the virtual scene model (the smaller the level number, the higher

Table 1 Judgment basis for fault or abnormality of DZZ1-2 automatic weather station

Code	Fault description	Judgment basis
1	Data Interruption	No data update in current real-time data table over 10 min
2	AC Power failure	AC for data collector cannot be detected by AC detection circuit
3	Low DC voltage	DC voltage detected by data collector is less than 9 V
4	High DC voltage	DC voltage detected by data collector is higher than 14 V
5	Abnormal state of meteorological element sensors	As is shown in Table 2 (temperature, humidity, air pressure, wind direction, wind speed, rainfall, grass temperature, eight-layer ground temperature, evaporation and visibility)

Table 2 Determination of sensor working state of DZZ1-2 automatic weather station

Code	Working state	Description
0	Normal operation	/
2	Fault or not detected	The loop current is abnormal or 0 for the temperature sensor
		The humidity sensor is not connected or the voltage output is 0
		The air pressure sensor is not connected or has no output
		The reed switch of rain sensor is closed for a long time
		Wind direction sensor abnormal. All 0 and all 1 of 7-bit gray code check if any of them is not 0 or 1
		The wind speed is 0 m/s for a long time
3	High sampling jump	Jump range is in accordance with the requirements of the functional specification
4	Low sampling jump	Jump range is in accordance with the requirements of the functional specification
5	Exceeding the upper limit for Sample data	Air temperature: <75 °C and >80 °C Air pressure: <400 hpa, >1100 hpa Relative humidity: <0%, >100%
6	Exceeding the lower limit for Sample data	1 min precipitation: <0 mm, >10 mm 0–5 min precipitation: <0 mm, >50 mm Average wind direction in 2 min and 10 min: <0° and >360° respectively 2 min and 10 min average wind speed: <0 m/s, >75 m/s Instantaneous wind speed: <0 m/s, >150 m/s Grass and ground surface temperature: <-90 °C, >90 °C 5 cm ground temperature: < −80 °C, >80 °C 10 cm ground temperature: < −70 °C, > 70 °C 15 cm ground temperature: < −60 °C, > 60 °C 20 cm ground temperature: < −50 °C, > 50 °C 40 cm ground temperature: < −45 °C, > 45 °C Ground temperature (80 cm, 160 cm and 320 cm): < −40 °C, >40 °C Visibility: <0 km, >70 km Evaporation: <0 mm, >100 mm
9	No check	Unable to judge the current working state
N	Null	Sensor is powered off or not configured

Table 3 Fault location and level of TWP3 wind profiler components

Fault description	Component location	Levels
All or part of products including ROBS, HOBS, OOBS, ORIGINAL, STATUS, RAD are not uploaded	Complete machine	1
Beam control unit failure	Transmitter beam control unit and antenna	1
Low power supply voltage of beam control unit		3
Abnormal standing wave value of antenna		3
Low reflected power value of antenna		2
Abnormal beam pointing state		2
Low transmitter power supply voltage (TX_5V < 4.0) or (TX_13V < 7.0) or (TX_28V < 23.0) or (TX_36V < 25.0) or (TX_50V < 35.0)	Transmitter power: 5 V/13 V/28 V/36 V/50 V	3
Over temperature alarm for Transmitter	Transmitter cabinet	3
The output power of transmitter is less than 1.5KW	Solid-state transmitter module	2
Transmitter duty ratio <8%		1
Abnormal reflected power value of transmitter		2
Abnormal status of RF switch	Transmitter RF switch assembly	2
Abnormal working pulse width of transmitter (high mode 0.8 μs, low mode 0.4 μs)	Solid state transmitter module	3
Abnormal standing wave state of transmitter	Transmitter cabinet	3
Transmitter input status abnormal		2
Transmitter output status abnormal		2
Low voltage of frequency synthesizer		2
Low power supply voltage of receiving channel		2
Abnormal excitation signal output value	Receiver cabinet	3
Abnormal output state of LO signal		3
Low digital IF internal DC power supply		3
No output or low voltage for UPS	UPS	2

the fault level), and only record the fault alarm details in the background for query, statistics and analysis. TWP3 wind profiler is taken as an example to illustrate the classification rules of fault alarm level, as shown in Table 3. Other detection device is the same and can be optimized and adjusted according to the actual monitoring and support business.

Level 1: There is no data uploading in more than 30 min, six kinds of product files including ROBS, HOBS, OOBS, ORI, STA and RAD are all missing within 30 min.

Level 2: If the peak transmission power is less than 2.0 kW, the service availability will be affected; or if the number of ROBS, HOBS, OOBS, ORI, STA and RAD files are less than 6, the data rate will be affected.
Level 3: Wind profile radar components alarm do not cause radar shutdown, or do not affect the normal operation of the profiler.
Level 4: Other alarms of wind profiler alarm do not affect normal business operation, but need to be paid attention to or maintained regularly.

5 Read-Time Data Mapping Driven

The organic integration of virtual reality technology, spatial GIS technology and simulation technology is realized in DEEP EYE 3D driving engine, which is the key hub to drive the virtual scene model update and change of meteorological observation device in real time according to the real data and running state changes. The data real-time driving mapping relationship between various devices and corresponding virtual models is shown in Fig. 4.

There are two kinds of data information obtained through data acquisition and processing. One is the real-time observation data of various automatic weather stations, and the other is the state information used to drive the 3D virtual model of device to update and change. The observation data display of each element on each VR sensor is updated in real time with the data upload frequency.

The queue structure is used to store the running status information in order to make the 3D virtual model state reflect the real running state in real time. The corresponding driver instruction queue is created on each type of 3D virtual model, in which the driving instructions generated after processing such as fault judgment, fault classification and automatic positioning are stored. According to the first in first out principle of queue, DEEP EYE 3D driving engine receives the driving instructions in sequence, automatically locates the detection device components in real time, updates the component status to red (fault) or yellow (alarm), and displays

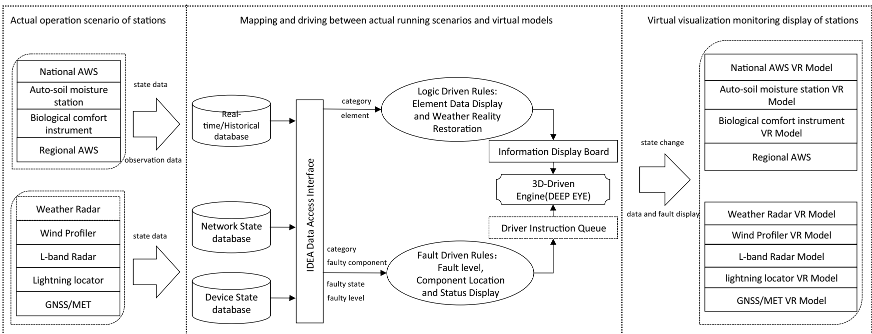


Fig. 4 Real-time data-driven mapping between devices and VR models

Fig. 5 3D visualization monitoring of virtual scene model of meteorological detection devices

the device operation status information, so as to realize the real-time 3D visual monitoring of the operation status of meteorological detection devices, as is shown in Fig. 5.

6 Conclusion

A VR visualization monitoring system with 3D-GIS for the comprehensive meteorological observation stations was initiated, by which the monitoring efficiency of meteorological devices has been effectively improved, unattended business of national ground weather stations has been effectively guaranteed, and the business availability, data rate and other assessment indicators have been improved by 1%. However, it still needs to be improved in some aspects, which will be the key research content in the future to meet the needs of a more comprehensive and refined meteorological device monitoring business.

(1) It is necessary to further sort out the transmission business process of all kinds of meteorological observation data stream, strengthen the visual monitoring of data flow and monitor the transmission process nodes of various meteorological observation data visually and accurately.

(2) Based on stream computing technology, the quality control analysis and monitoring of all kinds of original online data are further improved in the process of collection and transmission of various kinds of observation data.

References

1. Feizhong P, Lianchun S, Kejin W et al (2011) Design and practice of integrated meteorological observation operation monitoring system in China. Meteorology 37(2):213–218
2. Haihe L, Zhaolin M, Chunhui Z et al (2011) Integrated meteorological observation operation monitoring system. Meteorology 37(10):1292–1300
3. Feng L, Shiguang Q, Wei Z et al (2014) Integrated meteorological observation operation monitoring service and system upgrade design. Meteorol Sci Technol 42(4):539–544
4. Yan L, Feng L, Zhiqiang Z et al (2013) Construction of operation monitoring system for regional automatic weather stations in China. Meteorol Sci Technol 41(2):231–235
5. Wei L, Chunhui Z, Zhaolin M et al (2010) Design of operation monitoring system for L-band meteorological observation network. J Appl Meteorol 21(1):115–120
6. Dongming Z, Zhicheng W, Zhangwei W (2019) Design and development of provincial meteorological device operation monitoring system. Zhejiang Meteorol 41(1):35–38
7. Qinqiang Z, Yuanhong L, Jianyong L et al (2011) Key technologies for real-time monitoring of automatic weather station detection network. Meteorol Sci Technol 39(4):477–482
8. Cheng Z, Kaiting S, Jiang L et al, 3D visual monitoring system for workshop based on digital twin. Comput Integr Manuf Syst. https://kns.cnki.net/kcms/detail/11.5946.TP.20200817.0917.008.html.2020.08.17
9. Qiu F, Liu Z, Wu Y et al (2020) 3D visual monitoring system for discrete manufacturing machine shop. Mech Des Manuf (7):205–213
10. Jiawen L (2019) Research on key technologies of real-time 3D visualization monitoring of digital workshop. Nanjing University of Aeronautics and Astronautics
11. Xiangtao H, Hongliang L (2018) Research and implementation of 3D monitoring system for virtual workshop. Intell Manuf (7):31–37
12. Yanan J, Yanhui Z, Zhichao H et al (2020) Application of a 3D visualization technology in power monitoring system. Electr Technol 4:105–107
13. Hongyan N, Xubing C, Kai Y (2017) Research on 3D visualization online monitoring system of oil and gas station. Petrochem Technol 9:79–81
14. Shen Y, Xu Z, Li Y (2019) Application of virtual reality technology in meteorological field. Meteorol Sci Technol 47(3):531–536
15. Liping Z (2012) Research on key technologies of meteorological disaster scene based on virtual reality. Nanjing University of information technology
16. Chunhua W, Zhipeng L (2020) VR_ Application of AR technology in scene management of meteorological observation station. Electr Technol Softw Eng 10:34–35
17. Wenguang D, Yan L, Wenqiang D et al (2019) Application of VR technology in ground meteorological observation device support skill training. Prog Meteorol Sci Technol 9(6):49–52
18. Binghui Y (2013) Design and implementation of meteorological data visualization platform based on geographic information. Zhejiang University
19. Tingting B, Shengming J, Xiaoru Y (2020) Design and implementation of 3D visualization analysis platform for Doppler weather radar. Meteorol Sci Technol 48(4):490–495
20. Chuanlin Z, Xueliang W, Hongfei F et al (2017) Design and implementation of 3D visualization statistical analysis system for lightning data. Meteorol Sci Technol 45(1):59–63
21. Tiesong Z, Xiaoyun W, Wei L et al (2013) S architecture and open source WebGIS platform for meteorological observation station network visualization system. Meteorol Sci Technol 41(1):57–61
22. Mingyue L, Qilin Z, Wenqiang G et al (2011) Design and implementation of lightning data visualization map component based on GIS. Meteorol Sci Technol 39(6):823–827
23. Jun Y, Xiaohong X et al (2013) County level meteorological service platform based on 3D GIS. Meteorol Sci Technol 41(1):78–82
24. Xuequan Z, Wei W, Peng W et al (2015) Research and implementation of key technologies of 3D meteorological GIS platform. Meteorol Sci Technol 43(2):226–231

25. Yang L, Hu P, Wang Y (2019) Design and implementation of Shanghai meteorological information whole process visualization monitoring system. Prog Meteorol Sci Technol 9(3):29–35
26. Yucai C, Hongyan X, Xinyuan J et al (2012) Design of visual monitoring system for automatic weather station. Meteorol, Hydrol Oceanogr Instrum 1:39–44
27. Rui Y, Yaoxin S, Wenjie F (2014) Design and implementation of virtual training system for automatic weather station. Proceedings of national meteorological observation technology exchange conference: atmospheric detection and instrumentation Committee of China Meteorological Society

Tibaηsim: Information Access for Low-Resource Environments

Francis Dittoh, Hans Akkermans, Victor de Boer, Anna Bon, Wendelien Tuyp, and André Baart

Abstract In Africa, and other places around the world, there are numerous people who do not have access to information from the World Wide Web or other digital sources. This is not an issue of infrastructure only, but also of cultural and social factors, including low literacy. Most rural communities in the Northern Region of Ghana fall in this category, where information in written form and/or English is not accessible due to the inability of majority of community members to read and write. This paper presents Tibaηsim (originally "RadioNet"); a case-study of an appropriate ICT4D methodology in the development of an information delivery system hosted in low-resource areas, with empirical data from context analysis from the rural communities and other stakeholders. The paper also presents an evaluation of the system and the methodology, by way of User Evaluation and System Monitoring. The paper also shows how contextual issues are catered for through the methodology used. Tibaηsim focuses on available technologies and appropriate information formats by providing a system that relies on GSM and FM Radio, in the local language(s) of the community. Tibaηsim was deployed in 5 rural communities, reaching a total of almost 1000 people, providing them primarily farming-related information.

F. Dittoh (✉)
University for Development Studies, Tamale, Ghana
e-mail: fdittoh@uds.edu.gh

F. Dittoh · H. Akkermans · V. de Boer · A. Bon · W. Tuyp
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: Hans.Akkermans@xs4all.nl

V. de Boer
e-mail: v.de.boer@vu.nl

A. Bon
e-mail: a.bon@vu.nl

W. Tuyp
e-mail: w.tuijp@vu.nl

A. Baart
Utrecht, The Netherlands
e-mail: andre@andrebaart.nl

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_62

Keywords ict4d · Low-resource · Information access · Unconnected · Digital divide · Voice technologies · Mobile · Radio · User-centered design · Rural · Sub-saharan africa

1 Introduction

The digital divide is a problem that has plagued the world since the inception of the World Wide Web [1]. Only about 24.4% of Africans are connected to the internet [2]. The reasons for this low level of access, and subsequently to relevant information are; Lack of Infrastructure, Low Incomes and Affordability, User Capabilities; basic literacy and digital literacy, Lack of Incentives for Access; lack of awareness, relevant content, and cultural or social acceptance [3].

This paper presents Tibaηsim (a Dagbanli[1] word that translates to "Our Knowledge"); a case-study of an appropriate ICT4D methodology (see Fig. 1) in the development of an information delivery system hosted in low-resource areas with empirical data from context analysis from the rural communities and other stakeholders, which guided the design process and produced an information delivery system for rural communities. The paper further presents an evaluation of the system and the methodology, by way of User Evaluation; an analysis of Usability, Learnability, Availability. The paper shows how the aforementioned issues are catered for through the iterative and user-centered methodology used.

1.1 ICT Challenges in the Rural Context

In Ghana, voice-telephony, radio, and television are existing systems that function adequately, where the former two are more utilized in rural areas. However, there is an urban-rural divide, which means even further deterioration of services in the rural areas [4]. This is clearly seen in the areas of electricity and Internet connectivity which are often unreliable in rural communities [5]. Ghana has a literacy rate of 76.67% and a youth literacy rate of 90.6%. However, the rural north, which happens to be a major food source for the country, retains the lowest literacy rates [6]. Information available online is heavily biased toward text and thereby discriminates against those who cannot read and write [5]. Another barrier to connecting the unconnected has been that they do not find the information to be locally relevant enough to warrant the effort to gain access to it [3].

In the attempt to circumvent the issues of lack of infrastructure and unaffordability, it is important to center innovation on "technology-in-use" as opposed to "technology-as-invention" [7]. The extreme increase of mobile telephony in rural Africa [8] has led to the emergence of ICT4D projects, research and solutions that

[1] https://en.wikipedia.org/wiki/Dagbani_language.

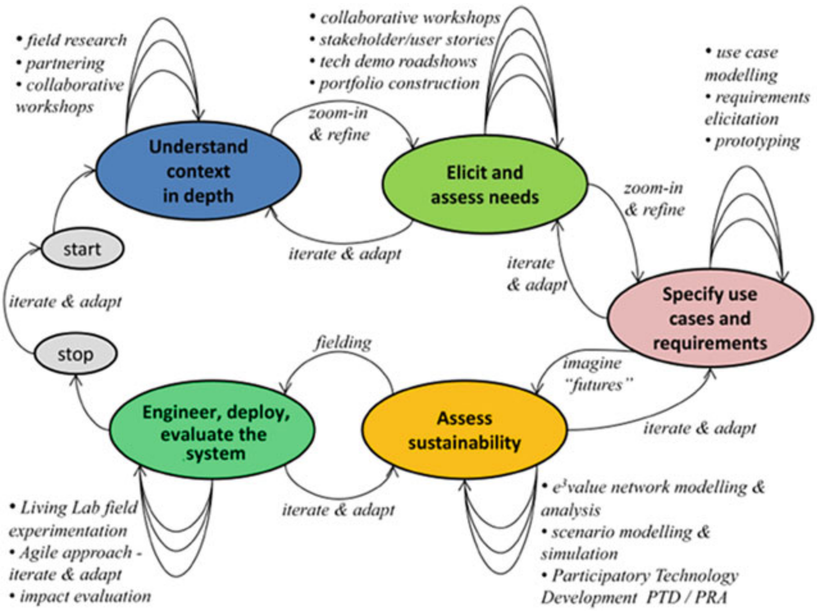


Fig. 1 Methodology

aim to utilize these existing technologies, namely, SMS [9, 10], Mobile Web [11], Voice Technologies [12], and Radio [12, 13] to aid various aspects of development. Some of these have failed completely [10] due to their reliance on technologies that are circumvented by low literacy (namely, SMS) and for the same reasons, some have not been able to scale up effectively in the rural areas [9] since, based on empirical evidence, SMS is hardly used by folk in rural areas in Ghana [6].

2 Methodology

This study used a collaborative, adaptive, and iterative methodology (see Fig. 1) to identify and tackle the issues from a socio-technical standpoint [5]. The study also forms a validation of the methodology by providing a practical application to an information system design and implementation.

The research focused on East Gonja District of the newly formed Savannah Region of Ghana. Five (5) communities were selected; typically small communities with about 20–30 households with roughly 200–250 people per community.

The evaluation of the issues pertaining to the digital divide, workshops with multidisciplinary groups, and interviews with stakeholders enabled us to *understand the context in depth* (see Fig. 1). Coupled with the understanding of the operational needs for the various stakeholders, led to aid *elicit and assess the needs* (see Fig. 1) of the proposed system.

Through an iterative process, stakeholders and end-users provided insight into the needs of the system. The key idea is to create an information access system for local communities by providing an FM broadcast system. This will enable rural communities to get access to relevant voice-based information in their own language(s). This will involve the design of a low-cost system, built on the Raspberry Pi with Solar-Power Banks that stores recorded voice fragments over GSM calls and broadcasts it over FM on short-range.

2.1 Key Requirements

Following an understanding of the context and the needs analysis, as well as considering the technologies available, the following are the key requirements of the system.

2.2 Non-functional Requirements

- Maintainability—NGOs/Companies and Institutions should have the ability to maintain the system and update information with ease.
- Availability—Users should have the ability to reach the system at all times. As such, barring mobile network issues, the system should be hosted on a platform that will be available at all times.
- Scalability—The system must have the ability to scale to different locations and for any number of subscribers and the system should be easily replicable.
- Reliability—The system must ensure that information is as accurate as possible.
- Usability—The targeted user group creates a requirement of simplicity in the user interface.
- Cost-Effectiveness—The whole system must work together to be financially sustainable.

2.3 Functional Requirements

The functional requirements are shown in a MoSCow format as follows Table 1.

Table 1 MoSCoW

Must have	Interactive voice response
	Local language(s)
	Regular information updates
	Short-range radio (FM) Transmission of information
	Monitoring
Should Have	Uninterrupted power supply
Could have	Community-level subscription-based service
	Wide-range radio (FM) Transmission of information

3 Implementation

3.1 Development

Based on the Context Analysis and Needs Assessment, Use-Case and Requirement Analysis, and the Feasibility and Sustainability Assessment, the various components of the system were *engineered, deployed and evaluated* (see Fig. 1) in multiple iterations with adaptations in a Agile approach coupled with an adapted living lab approach [5].

Tibaηsim was built on the Raspberry Pi 3B+,[2] tiny credit card size computer that was designed in the UK by the Raspberry Pi Foundation. The Raspberry Pi 3B+ runs a Quad Core 1.4GHz Broadcom BCM2837 64bit CPU, with 1GB RAM (2GB and 4GB RAM versions available) and requires a power supply of 2.5A. It also has a 40-pin extended GPIO, of which, vital to this project, GPIO Pin 4 transmits its FM signals. Tibaηsim also required a Huawei 3G Modem that is chan_dongle capable [3] and a GSM Sim Card. A power bank was also added to the package to ensure uninterrupted power supply.

The Kasadaka platform was used for the system's audio input module. Kasadaka enables the development and hosting of voice-based information services, targeted at low-resource environments. The VSDK allows for the rapid development of voice service (prototypes) in a web-based development environment [14]. This enables Information Providers to call into the system to leave messages, and at the backend, on confirmation, the audio fragment is stored in Kasadaka's file storage.

The Tibaηsim FM Broadcast System was built using Pi-FM-RDS which generates an FM modulation. It can include monophonic or stereophonic audio. The transmitter uses the Raspberry Pi's PWM generator to produce VHF signals [15]. The Broadcast System of Tibaηsim is the main user interface for the End-User. The system reads the audio files from the file storage system of the Voice Input Module and broadcasts them on a loop on FM at 107.9MHz.

[2] https://www.raspberrypi.org/.

[3] http://asterisk-service.com/en_US/page/chan-dongle-modems.

4 Evaluation

The System Usability Scale (SUS) is designed to obtain subjective feedback on overall usability and user satisfaction [16]. Apart from the 10-item SUS Score, using factor analysis, the SUS is able to provide additional information via two sub-scales: an 8-item "Usability" and 2-item "Learnability" scale [17, 18]. The SUS involved the administering of the test to all 106 participants after their use of the system. It is important to note that translations (in the right context) were necessary to effectively administer the SUS as it has been found that there is a possibility of a lack of comprehension in some of the SUS questionnaire items for even non-Native English speakers [17].

For reliability analyses of the data, the SUS showed good internal consistency (alpha = 0.743) [19] with an SUS Score of 80.52 (an Adjective Rating of "Excellent"), Learnability value of 71.93, and Usability of 82.67. The high scores for SUS, Learnability, and Usability, indicates a high quality of end-user's experience and the ability to utilize it with little to no prior training.

The Context Analysis and Needs Assessment of the methodology used necessitated the solicitation of required content from the end-users. Being involved in the development process, end-users as well as stakeholders were able to clearly specify the type of information that would be relevant to them. Empirical data was also gathered on the type of content end-users would find relevant. Notable among information considered relevant in the research area are *Crop Farming Practices*, *Market Prices*, *Information on seed*, *Climate Information*, and *Placing ads for sales of produce*. Of note is information from the World Wide Web, and example being climate information; this requires more technical work to enable automated updates to the system and was implemented as a separate integrable system dubbed *Mr. Meteo* [20]

Cost-effectiveness is of utmost importance and considered a Non-Functional Requirement (see Sect. 2.2). As such, design choices were made for low-cost hardware and open-source software (see Sect. 3.1). The system is also designed to enable rapid and little to no cost setup. For purposes of sustainability, it is imperative that running costs and affordability are catered for and as such, design choices are made to facilitate low-cost maintenance and affordability to end-users.

5 Discussion

Development and/or deployment of ICT systems for low-resource areas comes with a myriad of challenges [5], some general and others pertaining to the particular system. This is important to any study, being that these are issues that are not always readily apparent.

Development of systems for low-resource environments with the methodology used requires physical visits (possibly several) to communities. Some communities are remote, with bad road access, making the required multiple visits tedious. This, on the contrary, is the very reason why information access for such communities is vital. In this project, all communities were remote, with bad road access, but visits were necessary for information elicitation and deployment. However subsequently, it will be possible to deploy with the help of Agricultural Extension Agents who, as part of their existing jobs, visit these communities from time to time. Information dissemination through ICTs rely on telecommunications networks, but unfortunately, many rural communities lack sufficient access. In some rural communities in Ghana, there is little to no 4G, sparse 3G, and patchy GSM reception. This needs to be taken into account to build systems that do not require constant internet access, high bandwidth, constant uptime, etc. The details will depend largely on the particular community. The next iteration of development will use an outdoor, waterproof enclosure coupled with solar-power supply in addition to GSM boosters to augment the reception to the device.

In many developing countries, hardware needed to build ICT systems may not be readily available. For Tibaηsim, the required hardware (Raspberry Pi 3B+ and Huawei 3G Modems) had to be imported into the country due to its unavailability in Ghana. This caused some delays with the initial pilot.

Finally, a project specific issue worth mentioning; the FM broadcast range of the Raspberry Pi 3B+ with a simple wire antennae is about a 120 m. As much as this covers enough of the communities, the ideal would be a wider range to cover entire communities and cater for larger communities. The next iteration will consider boosting the range of the FM transmission. This however would require liaising with the National Communications Authority[4] in Ghana once transmissions begin to exceed official community ranges and/or becomes a commercial product.

6 Conclusion

Information access remains a problem in many parts of the world [1]. Developing countries, like Ghana, have numerous rural communities that have little access to up-to-date and relevant information [2]. Telecommunication infrastructure is lacking in these communities and limited, mostly, to GSM reception and there are other major issues like unaffordability of Internet-capable device and Internet and voice services, low levels of literacy, and unavailability of relevant content which further compounds the issue [3]. Available literature points us toward the use of "technology-in-use" [7]; in the case of Ghana, mobile telephony and FM radio, concentrating on voice technologies [12] and local languages.

This paper has presented a case-study of an appropriate ICT4D methodology [5] in the development of an ICT4D system which is hosted and used in low-resource environments. Empirical data from rural communities and other stakeholders was used, iteratively, to design a low-cost, voice-based information system for rural communities meeting the contextual requirements and operational goals of the various

[4] https://www.nca.org.gh/.

stakeholders. Furthermore, a System Usability Score (SUS) [16] test was administered to 106 community members from five (5) communities where the system was deployed and we presented an analysis of the SUS, Usability and Learnability scores which indicate that the requirements of the end-users were met and the system is easily used and learned.

Information is a very important commodity for development [21]. As such, bridging the digital divide is vital, not only to developing countries, but to the world, as the successes of attaining the Sustainable Development Goals (SDGs)[5] will benefit the planet as a whole. The process of systems development for low-resource regions is therefore critical to the successfully attaining SDGs. This process is not trivial and requires an appropriate methodology that is collaborative, user-centered, and iterative. This aids to fix issues that exist for the stakeholders and end-users as opposed to fixing perceived problems. The Tibaηsim project shows this development process for low-resource areas practically and hopes to serve as a guide for future projects.

The authors thank Christophe Guéret for the original Kasadaka idea and the University for Development Studies ICT for Food and Water in Ghana team led by Saa Dittoh. We also acknowledge the contributions of Kay Kutah Wilhelm (Savanna Agricultural Research Institute), The Internet Society Ghana Chapter, and The Internet Society for funding the pilot project. The authors also thank Chris van Aart, Amadou Tangara, Chris Reij, Stephane Boyera, Gossa Lô, Julien Ouedraogo, Matthieu Ouedraogo, Mary Allen, Nana Baah Gyan, Aman Grewal, Yacouba Sawadogo, Stefan Schlobach, and many others for their contribution to this continuing research project.

References

1. Worldwide digital population as of January 2020, Published by J. Clement, Feb 3, 2020
2. The State of the Internet Around the World (2019) https://www.statista.com/chart/17322/global-internet-access/
3. Schmida S, Bernard J, Zakaras T, Lovegrove C, Swingle C (2017) Connecting the next four billion: strengthening the global response for universal internet access. USAID, Dial, SSG Advisors, Seattle
4. Estache A, Vagliasindi M (2007) Infrastructure for accelerated growth in Ghana: investment needs, policies and institutions. World Bank
5. Bon A, Akkermans H, Gordijn J (2016) Developing ICT services in a low-resource development context. Complex Syst Inform Modeli Q 84–109
6. Dittoh F, van Aart C, de Boer V (2013) Voice-based marketing for agricultural products: a case study in rural Northern Ghana. In: Sixth international conference on information and communications technologies and development. New York
7. Heeks R (2009) The ICT4D 2.0 Manifesto: where next for ICTs and international development?. https://doi.org/10.13140/RG.2.2.21400.14086
8. International Telecommunication Union Statistics. http://www.itu.int/ITU-D/ict/statistics/
9. ESOKO. http://www.esoko.com
10. Google Trader. http://www.google.com.gh/local/trader
11. Boyera S (2004) The mobile web to bridge the digital divide?. W3C

[5] https://sustainabledevelopment.un.org/.

12. de Boer V et al (2012) Radio marché: distributed voice and web-interfaced market information systems under rural conditions. CAISE
13. Topham J (2016) "Making Waves" radio based learning brings education to young girls in the DRC. https://warchild.ca
14. Baart A et al (2018) Affordable voice services to bridge the digital divide: presenting the Kasadaka platform. In: Selected Papers from the WEBIST 2018 international conference. Springer LNBIP Book Series. Berlin, Germany, p 2019
15. FM-RDS transmitter using the Raspberry Pi. https://github.com/ChristopheJacquet/PiFmRds
16. Brooke J (1996) SUS-A quick and dirty usability scale. Usability Eval indus 189(194):4–7
17. Finstad K (2006) The system usability scale and non-native English speakers. J Usability Stud 1(4):185–8
18. Sauro J (2011) A practical guide to the system usability scale: background, benchmarks & best practices. Measuring Usability LLC., Denver, CO
19. Gliem JA, Gliem RR (2003) Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest research-to-practice conference in adult, continuing, and community education
20. Dittoh F, de Boer V, Bon A, Tuyp W, Baart A (2020) Mr. Meteo: providing climate information for the unconnected. In: 12th ACM conference on web science (WebSci '20 Companion). Southampton, United Kingdom
21. Allen Beth (1990) Information as an economic commodity. The Am Econ Rev 80(2):268–273

Bent QR Code Image Rectification Method Based on Image-To-Image Translation Network

Kazumoto Tanaka

Abstract This paper describes an image rectification method for bent quick response (QR) codes attached on cylindrical surfaces using a deep neural network-based image generator. The input to the proposed generator is the bent QR code image and the output is the rectified image such that the QR code reader can recognize the code. The generator is trained with an image-to-image translation framework using a large number of pairs of QR code images and their bent images. A weighted loss function between generated QR code image and its corresponding QR code image without distortion is used during training to focus on learning the rectification of the peripheral part of the image, which is affected significantly by curvature. In the study, pseudo QR codes were used for training. In an experiment where a bent QR code image is read using the proposed method, it is demonstrated that a QR code with a curvature degree larger than those in previous studies can be read.

Keywords Bent QR Code · Image rectification · Weighted loss function · Image-to-image translation · Generative adversarial network · Rectified image generator

1 Introduction

Quick response (QR) code is widely used in industry and daily life to present various information. The code is printed directly on a surface where information should be presented, or a sticker with the code printed on it is attached to the surface. Because the code reading may fail when the surface is curved, alignment patterns for distortion correction have been added to version 2 of the QR code [1]. However, this distortion correction capability is limited, and it is difficult to read the code on a surface with a large curvature degree. Most other methods proposed previously for bent QR code image rectification are image transformations based on geometric feature matching

K. Tanaka (✉)
Kindai University, 1, Umenobe, Takaya, Higashi-Hiroshima, Hiroshima 739-2116, Japan
e-mail: kazumoto@hiro.kindai.ac.jp

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_63

Fig. 1 A QR code attached on a cylinder. Because the left and right peripheral parts of the QR code are distorted significantly, it is difficult to read the code using the conventional method

[2–4]. Because these methods involved a simple linear transformation for correction, it is difficult to apply them to a significantly bent QR code image (Fig. 1).

This study proposes a deep neural network (DNN) that inputs an image of bent QR code attached to a cylindrical surface and outputs the rectified image. We have developed a DNN-based image generator with the framework of an image-to-image (I2I) translation network (pix2pix) [5], which is a generative adversarial network. The I2I translation requires paired examples for training the network. In this study, a large number of pairs of pseudo QR code images and their bent images were generated by computer graphics and used to train the network. During training, a weighted loss function between generated image and its corresponding image without distortion is used to focus on learning the rectification of the peripheral part of the image, which is affected significantly by curvature. Experimental results showed the generator can read a QR code with a curvature degree larger than those in previous studies.

This paper is organized as follows. The generator is described in Sect. 2. The experimental results and the discussion are presented in Sects. 3 and 4, respectively. Section 5 provides the conclusion.

2 Rectified Image Generator for Bent QR Code

2.1 Training Architecture for Generator

Let *QRi*, *QRi_b*, and *QRi_g* be the QR code image, bent QR code image of which a circumscribed rectangle is formed in a square shape, and generated QR code image, respectively. These images are squares of the same size (Fig. 2). The generator performs an image transformation on *QRi_b* to obtain *QRi_g* as a rectified image. The network architecture is the same as that of pix2pix [5].

The pix2pix generator is trained using target and reference images and attempts to convert the input image into the form of the reference image. In this study, *QRi_b* is the

Fig. 2 Left: Bent QR code image. Middle: *QRi_b*. Right: *QRi*. Circumscribed rectangle of *QRi_b* formed in a square shape

target image, and *QRi* is the reference image. Figure 3 shows the training architecture of the generator. The generator produces outputs for *QRi_b* in the dataset, whereas the discriminator attempts to predict whether the input image is a QR code image or one that is produced by the generator. To train the generator, two loss functions must be reduced. One is *MSELoss_g*, which is the mean squared error between each element in the prediction result map produced by the discriminator and true value (1.0) map. The other is *L1Loss*, which is the mean absolute error between each element in *QRi_g* and its corresponding *QRi*, but the absolute error is weighted using Eq. (1) to focus on learning the rectification of the peripheral part of the image, which is significantly

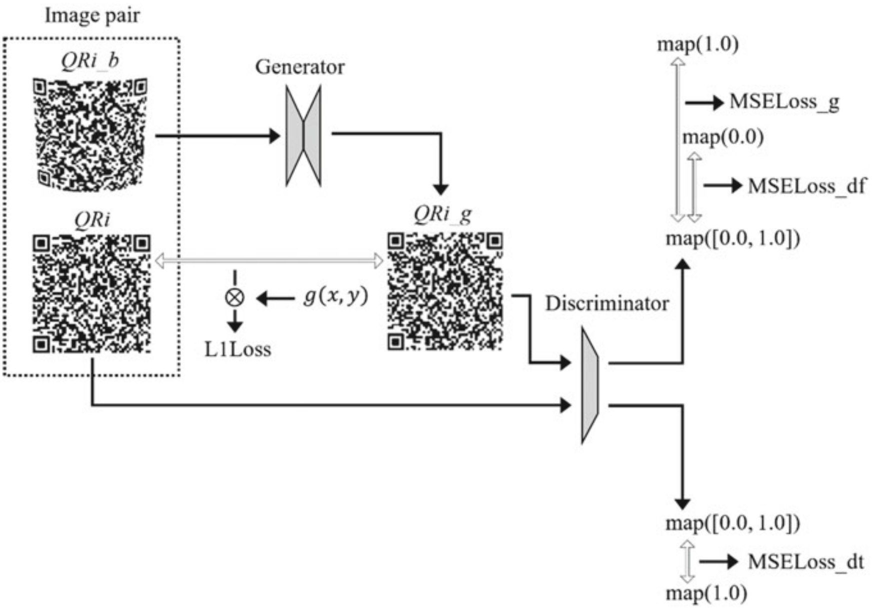


Fig. 3 Simplified view of generator training architecture based on I2I translation network. *QRi* and its corresponding *QRi_b* are stored in a dataset as an image pair

affected by curvature.

$$g(x, y) = \alpha\left(\beta - \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-\left(\frac{x}{w-1}-0.5\right)^2}{2\sigma^2}}\right), 0 \leq y \leq h - 1 \tag{1}$$

where x and y are the coordinates of *QRi_b*; w and h denote the width and height of *QRi_b*, respectively; α, β and σ are positive numbers.

Two loss functions were reduced to train the discriminator. One is *MSELoss_dt*, which is the mean squared error between the prediction result map for *QRi* and the true value (1.0) map. The other is *MSELoss_df*, which is the mean squared error between the prediction result map for *QRi_g* and the false value (0.0) map.

2.2 Datasets

A dataset for image pair *QRi* and *QRi_b* was prepared for the training. The image size was set to 256 × 256, and the images were created using OpenGL. In this study, a pseudo QR code was utilized for the training instead of the real QR code. The pseudo QR code was created such that each cell is painted black or white randomly and three finder patterns are set at the corners (Fig. 4); subsequently, 2000 images of pseudo QR codes were created. Each image was attached to a cylinder in a virtual space and filmed virtually (Fig. 5). The angle between the normal vector at the center of the code and the optical axis of the virtual camera was changed randomly within 10° for each filming. Furthermore, the radius of the cylinder was changed randomly within the range $0.4 \leq L/(\pi R) \leq 0.8$, where L and R denote the length of the sides of the pseudo QR code and the radius of the cylinder, respectively. A virtually filmed image was fit into a square measuring 256 × 256, Gaussian filtered, and then adaptively binarized to obtain *QRi_b*. A dataset for training was created using 2000 pairs of *QRi_b* and the pseudo QR code image (i.e., *QRi*). Similarly, a dataset for testing using 100 images was created.

Fig. 4 Example of pseudo QR code

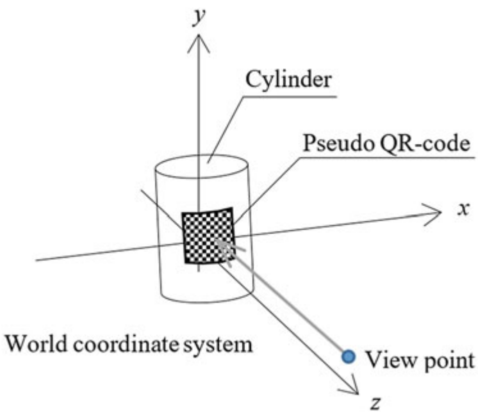


Fig. 5 *Left* Pseudo QR code attached on the virtual cylinder. *Right* Example of pseudo QR code image captured virtually from view point

2.3 Training and Testing

The generator and discriminator were trained with a batch size of "1" for 1000 epochs. The training time on a system with an NVIDIA GeForce GTX 1080 Ti GPU was approximately 24 h. The generator was tested using 100 images. Some of the test results are shown in Fig. 6.

As shown in the figure, the rectification was successful when $L/(\pi R)$ was 0.6 or less. However, beyond 0.6, the number of cases with failed restoration of the position and shape of black and white patterns increased gradually.

3 Experiment

To obtain bent QR code images of $L\big/(\pi R) = 0.4, 0.5, \ldots, 0.8$ for the experiment, the images were captured while changing the size of the real QR code attached to one cylinder. Five types of QR codes of models 2 and 8 were prepared, and 25 images were captured in total. The code area extracted from each image was fit into a square measuring 256×256 and then binarized adaptively. The binarized image was rectified by the generator, and the rectified image was verified to determine whether the reading through a QR code reading application was successful. Figure 7 shows an example of the resulting images. The recognition rate results are shown in Table 1.

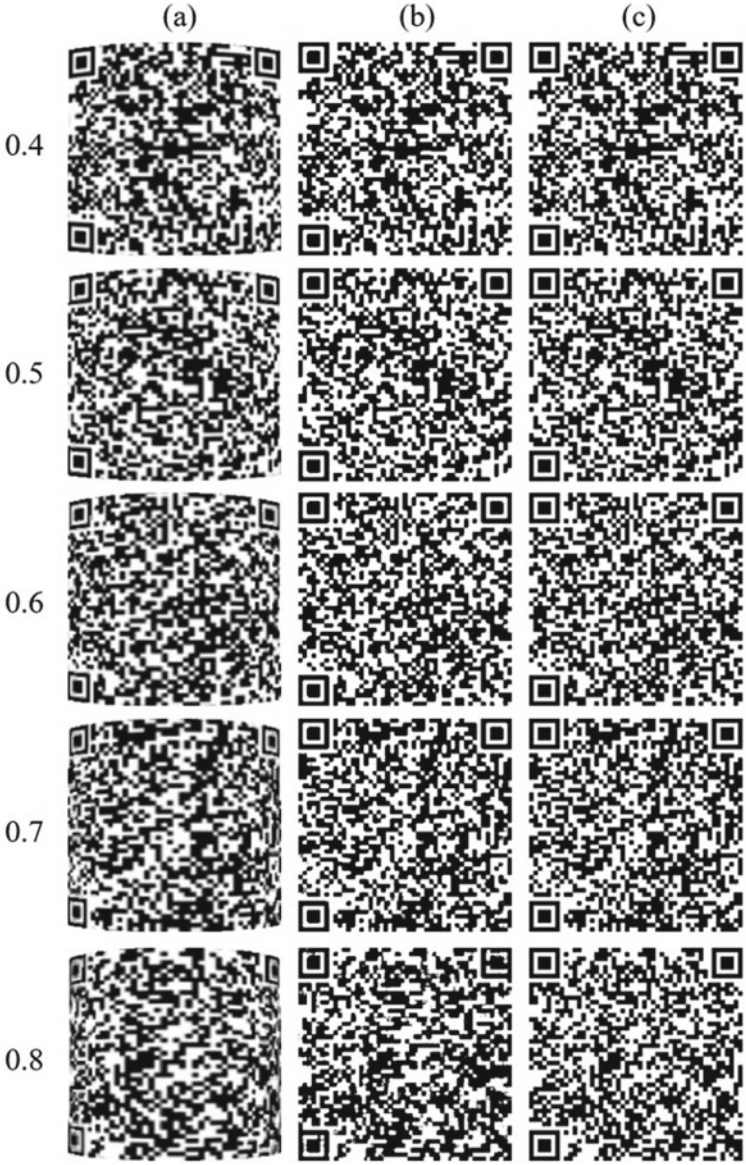


Fig. 6 Sample results. **a** *QRi_b* from dataset for testing. Curvature degree increases from top to bottom. The number on left side shows $L/(\pi R)$. **b** *QRi_g* generated from (**a**). **c** Ground truth (*QRi* paired with (**a**))

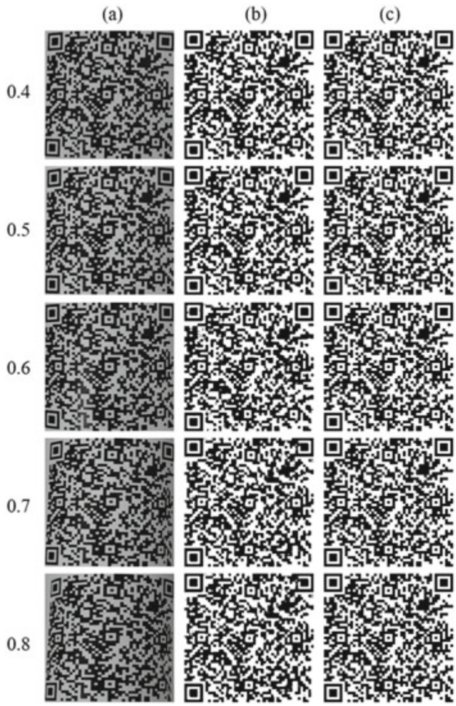


Fig. 7 Sample of experimental results. **a** *QRi_b* of real QR code. The number on left side shows $L/(\pi R)$. **b** *QRi_g* generated from (**a**). **c** Real QR code corresponding to (**a**)

Table 1 Experiment results

$L/(\pi R)$	Recognition rate (%)
0.4	100.0
0.5	100.0
0.6	60.0
0.7	0.0
0.8	0.0

4 Discussion

In the case of $L/(\pi R) = 0.4$ and 0.5, the recognition was successful; however, when it was 0.6, the recognition rate was 60%, and the recognition failure was due to rectification failure, as shown in Fig. 7. Meanwhile, in the test step, no rectification failure occurred in the case of $L/(\pi R) = 0.6$. This discrepancy occurred because during the actual filming, the aberration of the camera lens affected the distortion of the peripheral part of the image; however, images that did not account for this distortion were used in the training.

Next, we consider the recognition rate of a method reported previously [4]. The experiments in [4] showed that when the curvature rate described in [4] was converted to $L/(\pi R)$, the recognition was successful up to 0.43; however, cases involving curvature rates higher than that were not reported. Meanwhile, our method was successful up to 0.5, and the success rate of 0.6 was 60%. Therefore, the method successfully read the QR code with a curvature degree larger than that shown in the previous study.

Finally, the practical application of the proposed method is described. Generally, a GPU is required to calculate the DNN. However, current smartphones and mobile terminals that are often used to read QR codes do not have GPUs. Therefore, in this study, we assumed a method of sending a bent QR code image from a terminal to a GPU server and receiving the reading result.

5 Conclusion

We proposed an image rectification method for bent QR codes. The method is based on a rectified image generator that was trained using an I2I translation framework. The weighted loss function between generated QR code image and its corresponding QR code image without distortion was used in the training to focus on learning the correction of the peripheral part of the image, which was significantly affected by curvature. An experiment for reading bent QR code images using the proposed method showed that a QR code with a curvature degree larger than that shown in previous studies can be read. In future studies, we will primarily focus on improving the recognition rate in the case of $0.6 \leq L/(\pi R)$.

Acknowledgements This work was partially supported by JSPS KAKENHI Grant Number 18K02876.

References

1. Tiwari S (2016) An introduction to QR code technology. In: 2016 international conference on information technology. Bhubaneswar, pp 39–44
2. Li X, Shi Z, Guo D, He S (2013) Reconstruct algorithm of 2D barcode for reading the QR code on cylindrical surface. In: International conference on anti-counterfeiting, security and identification. Shanghai, pp 1–5
3. Xuan W, Peng C, Liuping F, Jianle Z, Peijun H (2017) Research on correcting algorithm of QR code image's distortion. In: IEEE international conference on communication technology. Chengdu, pp 1752–1756
4. Li K, Meng F, Huang Z, Wang Q (2019) A correction algorithm of QR code on cylindrical surface. J Phys: Conf Ser 1237(2)
5. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: International conference on computer vision and pattern recognition. California, pp 22–29

A New Concept of ICT on Eduinformatics in Higher Education

Kunihiko Takamatsu, Ikuhiro Noda, Bannaka Kenya, Tomoe Nakagawa, Yasuhiro Kozaki, Kenichiro Mitsunari, Masato Omori, Ryohei Adachi, and Yasuo Nakata

Abstract In this article, we propose a new concept of information and communication technology on eduinformatics in higher education. Eduinformatics is a novel field of education that combines both education and informatics. Eduinformatics not only deals with students' data but also provides new analysis methods and concepts to handle data in education, similar to bioinformatics. In September 2020, a subcommittee for universities from the Ministry of Education, Culture, Sports, Science, and Technology in Japan shared a perspective of the connection between education and research in universities in the future. In this paper, we assess our research question on how to connect education and research in universities to enhance the quality of higher education. To investigate this question, we introduce two practical examples from Kobe Tokiwa University in Japan, and we present results of recent eduinformatics. From these examples, we find that Information and Communication Technology (ICT) is the connection between education and research.

K. Takamatsu (✉) · I. Noda · B. Kenya · T. Nakagawa · K. Mitsunari · R. Adachi · Y. Nakata
Center for the Promotion of Excellence in Research and Development of Higher Education, Kobe Tokiwa University, Kobe 6530838, Japan

K. Takamatsu · I. Noda · B. Kenya · Y. Nakata
Organization for the Advancement of Higher Education, Kobe Tokiwa University, Kobe 6530838, Japan

K. Takamatsu
Life Science Center, Kobe Tokiwa University, Kobe 6530838, Japan

K. Takamatsu · Y. Nakata
Faculty of Health Sciences, Kobe Tokiwa University, Kobe 6530838, Japan

I. Noda · T. Nakagawa
Academic Development Division, Kobe Tokiwa University, Kobe 6530838, Japan

K. Mitsunari · M. Omori
Faculty of Education, Kobe Tokiwa University, Kobe 6530838, Japan

B. Kenya · R. Adachi
Department of Oral Heal, Kobe Tokiwa College, Kobe 6530838, Japan

Y. Kozaki
Faculty of Education, Osaka Kyoiku University, Osaka 582-8582, Japan

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_64

Keywords Eduinformatics · Information and communication technology · Higher education

1 Introduction

1.1 Eduinformatics

Approximately two decades ago, the human genome consortium, which consisted of many researchers from across the globe, had sequenced a rough draft of the human genome [1]. The human genome includes approximately 3 Giga base pairs. Many informaticians attended the human genome project using both computer and informatics, since the data were too extensive to handle manually. This phenomenon created a novel interdisciplinary field, known as bioinformatics, that is a combination of both the biology and informatics fields.

Since 2008, research on higher education is being actively pursued, and the establishment of Institutional Research (IR) promoted universities by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) in Japan [2]. This situation is similar to the era when bioinformatics was started.

Many educational researches are required to be "evidence-based." This type of research is difficult using conventional pedagogical techniques alone. If researchers can use informatics adequately, for example, data science, machine learning, statistics, and so on, higher education can also develop dramatically. For example, biology has developed dramatically with the advent of bioinformatics.

Eduinformatics is a novel field of education that combines both informatics and education [3]. It is a portmanteau of "education" and "informatics." Eduinformatics not only deals with students' data but also provides novel analysis methods and concepts to deal with data in education, similar to bioinformatics. In this paper, we review eduinformatics [4] (Fig. 1).

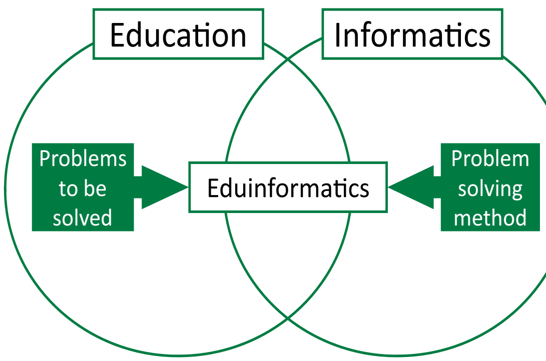


Fig. 1 The concept of Eduinformatics (from [4])

1.2 Connection Between Education and Research in University

Since the early 1990s, many universities in Japan have reformed to enhance the quality of higher education [5]. University reform was carried by universities themselves and was also promoted by the MEXT [5]. Moreover, in 2012, MEXT presented strict guidelines to universities for reform [6].

In September 2020, a subcommittee for universities in Japan's MEXT presented a perspective of the connection between education and research in universities for the future [7, 8]. The proposal stated that faculties of universities have the following features (excerpt).

- Faculties in universities are more interested in research than education.
- Many university faculty members find it difficult to balance education and research.

To alleviate these problems, the proposal suggested effective organizational and university management.

1.3 Research Question

This paper proposes to answer the following research question: how to connect education and research in universities to improve the quality of higher education. To investigate this question, first, we introduce practical examples from Kobe Tokiwa University in Japan, which show results of recent eduinformatics. Using these examples, we find that Information and Communication Technology (ICT) is the key to connecting education and research. Finally, we propose a new concept of ICT on eduinformatics in higher education.

2 Two Practical Examples

2.1 Center for the Promotion of Excellence in Research and Development of Higher Education in Kobe Tokiwa University

Kobe Tokiwa University belongs to Tamada educational institution, [9] which started in 1908. While Kobe Tokiwa College was established in 1967 [10] and Kobe Tokiwa University in 2008, some departments of Kobe Tokiwa College were merged with Kobe Tokiwa University. Currently, Kobe Tokiwa University includes two faculties and five departments. One is the faculty of education, another is the faculty of health science. The faculty of education has a department of child education, while the faculty of health science has a department of nursing, medical technology, and radiology.

Kobe Tokiwa University and Kobe Tokiwa College are both located in the same compound. Kobe Tokiwa College has two departments, the department of oral health and the department of correspondence course of nursing. Approximately 1,500 students are enrolled at Kobe Tokiwa University. The size of Kobe Tokiwa University is small as defined by MEXT [11]. The feature of Kobe Tokiwa University is that all departments have the same purpose, which is to obtain the national certification of medical technologists, nurses, public health nurses, school nurses, radiological technologists, nursery teachers, kindergarten teachers, elementary school teachers, and dental hygienists.

This means that most students determine their courses by the rules established by the MEXT or Ministry of Health, Labor, and Welfare in Japan, for special national certification. Therefore, it is difficult for faculty to focus on both education and research. Most faculties use their time and effort to educate students than focus on their research. To improve this situation, the Center for the Promotion of Excellence in Research and Development of Higher Education (CPERD) was established.

The CPERD in Kobe Tokiwa University was constructed in 2009 [12, 13]. Professor Ryohei Adachi, the first president of the CPERD, led the promotion of not only education but also research at Kobe Tokiwa University. The second president of the CPERD, Professor Yasuo Nakata, started open forums to discuss each faculty member's research projects. Moreover, Kobe Tokiwa University's staff can also attend these forums. Some staff write papers and present their research at the international conference about IR or Enrollment Management Institutional Research (EMIR), similar to researchers. This shows that it is important to work with both faculty and staff on the same problems.

The CPERD also plays a role in statistical mathematics analysis at Kobe Tokiwa University. Some members of CPERD are specialists in statistic and bioinformatics. They help other faculty members to conduct research, using many analysis methods such as statistics and informatics. This phenomenon shows that this partnership is good and effective in moving the research forward with the help of these specialists.

2.2 Examples of Research in Eduinformatics

In 2016, a complex network analysis and a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis were developed [14]. The complex network field analyzes features of huge and complex networks. In 1998, Watts et al. revealed "small-world" networks [15]. In the 1960s, SWOT analysis was developed as part of the Harvard Business Policy [16]. Using this new method, we attempted to find the larger problems on what is preventing university reform at Kobe Tokiwa University in Japan, in order to improve the quality of higher education [14]. We visualized the network data using Cytoscape [17] (Fig. 2). In the figure, the size of nodes is depended on the number

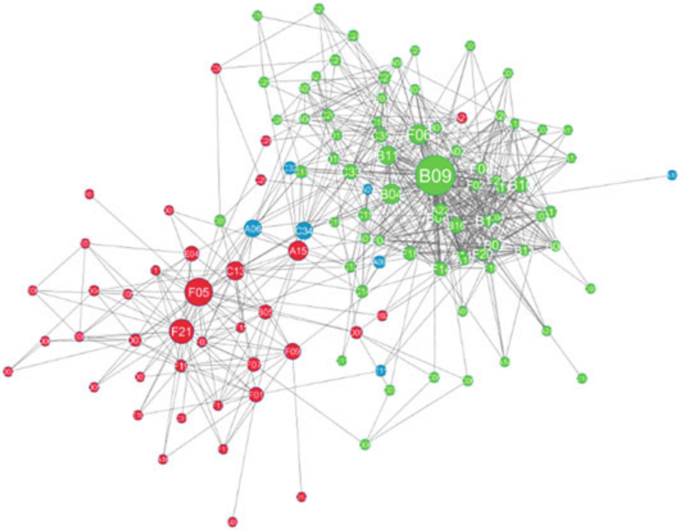


Fig. 2 Visualization of networks of SWOT analysis [14]

of edges that are connected to the node, and in the figure, the biggest node is B09, which denotes "faculty and staff did not have improvement awareness."

Competency-based education has become a topical subject in improving the quality of higher education [18]. We developed a common evaluation indicator called "Tokiwa competencies" [19]. As of 2017, the syllabi of the University display the relationships between the 19 Tokiwa competencies and courses. In early 2017, we developed and published new visualizations for curricula using syllabi, by a combination of multidimensional scaling methods, cosine similarity, and scatter plot [20] (Fig. 3).

3 Results and Discussion

Using the two practical examples discussed in Sect. 2, we attempt to answer our research question, that is, how to connect education and research in university to improve the quality of higher education.

First, we recognize that to analyze data in eduinformatics, we need to prepare the data. For example, when we obtain students' answers from paper questionnaires, we need to type these data into a computer. To obtain significant statistical results, we must use extensive data, and this is time consuming. However, when we use ICT to collect the answers, for example, Google Forms, it is easy to obtain a digital summary. This is easy since most students in Japanese universities now have smartphones.

We can use text mining methods, although, it is more difficult to separate Japanese words than English words, since words in the Japanese language are not separated by

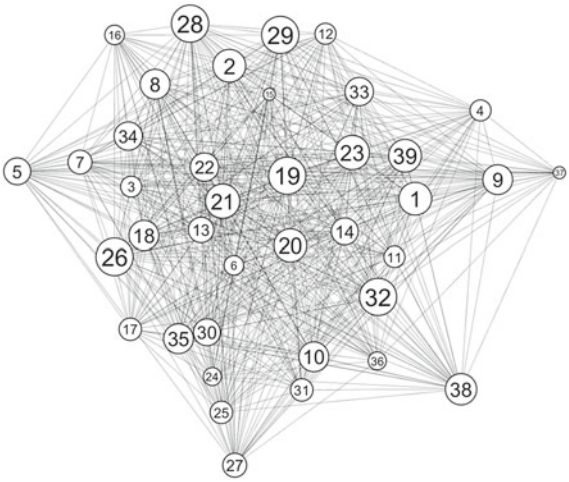


Fig. 3 A visualized network by Tokiwa competencies from courses

space. However, we can use morphological analyzers such as MeCab [21] to obtain accurate results. When we obtain answers to a descriptive type questionnaire using ICT, we can use text mining analysis and visualize by networks.

We, therefore, propose a new concept of ICT on eduinformatics in higher education (Fig. 4). As described in Sect. 2.1, faculty members focus on education than

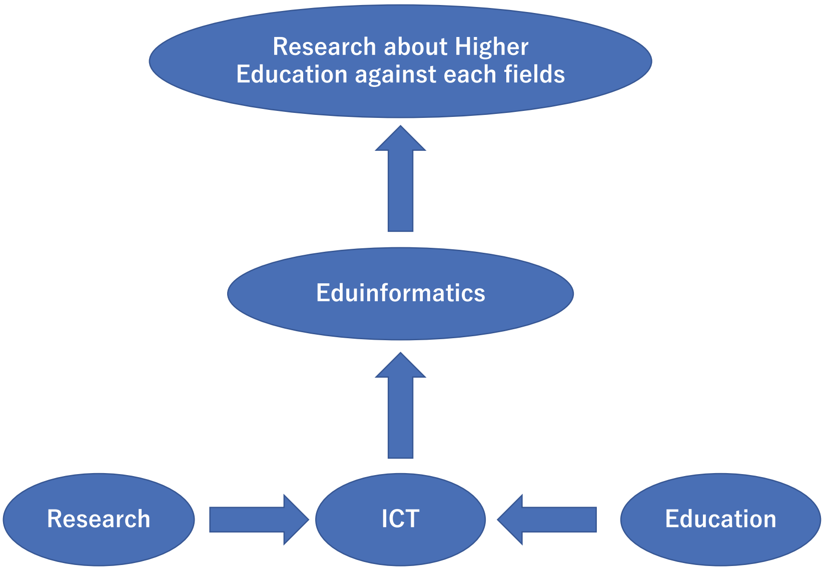


Fig. 4 New concept of ICT on Eduinformatics in Higher Education

research since they are required to conduct many lectures in the university. Therefore, in such a case, faculty members must improve the quality of education. Educational faculty can use ICT to combine research and education to obtain data and analyze them using eduinformatics. This type of research is usually called educational research in higher education. It is an effective method in enabling educational faculty members to perform research activities.

4 Conclusion

We have obtained answers to our research question. Our research question is how to connect education and research in universities to improve the quality of higher education. One of the answers is ICT. When educational faculties use ICT to obtain students' data, they can easily prepare and analyze these data to enhance the quality of higher education.

References

1. Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature. https://doi.org/10.1038/35057062
2. MEXT (2008) Construction of education for university. http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo4/houkoku/080410.htm. Accessed 1 Nov 2020
3. Takamatsu K, Murakami K, Kirimura T et al (2018) "Eduinformatics": a new education field promotion. Bull kobe Tokiwa Univ 11:27–44. https://doi.org/10.20608/00000958
4. Takamatsu K, Kozaki Y, Murakami K et al (2019) Review of recent eduinformatics research. IEEE/IIAI Int Congress on Appl Inf Technol 2019:27–32
5. Yamamoto S (2012) Higher education reforms in Japan: changing relationship between government and universities. In: State and market in higher education reforms. Brill Sense, pp 201–211
6. MEXT (2012) Implementation plan of university reform. https://www.mext.go.jp/b_menu/houdou/24/06/__icsFiles/afieldfile/2012/06/05/1312798_01_3.pdf. Accessed 5 Oct 2020
7. MEXT (2020) Perspective of education with education and research. https://www.mext.go.jp/content/20200917-mxt_koutou01-000009971_06.pdf. Accessed 5 Oct 2020
8. MEXT (2018) Grand design of higher education for 2040 (report). https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/1411360.htm. Accessed 5 Oct 2020
9. Kobe Tokiwa University. http://www.kobe-tokiwa.ac.jp/univ/. Accessed 5 Nov 2020
10. The history of Tamada educational institution. http://www.kobe-tokiwa.ac.jp/univ/guide/history.html. Accessed 5 Oct 2020
11. MEXT (2018) About Private university ordinary expenses subsidy in 2019 or later. https://www.mext.go.jp/a_menu/koutou/shinkou/07021403/002/002/__icsFiles/afieldfile/2018/09/19/1409177.pdf. Accessed 25 Oct 2020
12. Kobe Tokiwa University (2020) Education and research support in Kobe Tokiwa University. http://www.kobe-tokiwa.ac.jp/univ/education/ktu.html. Accessed 20 Oct 2020
13. Kobe Tokiwa University (2020) About Tokiwa University. http://www.kobe-tokiwa.ac.jp/univ/guide/history.html. Accessed 20 Oct 2020
14. Takamatsu K, Kirimura T, Bannaka K et al (2016) SWOT analysis and complex network analysis to enhance governance in universities by collaboration between faculty and staff. In:

2016 5th international institute of applied informatics (IIAI) international congress on advanced applied informatics (IIAI-AAI). Institute of Electrical and Electronics Engineers (IEEE). pp 1188–1189

15. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440–442
16. Rangkuti F (2004) Teknik Membedah kasus Bisnis. Anal SWOT
17. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303
18. Nodine TR (2016) How did we get here? A brief history of competency-based higher education in the United States. J Competency-Based Educ. https://doi.org/10.1002/cbe2.1004
19. Krimura T, Takamatsu K, Bannaka K et al (2017) Innovate the management of teaching and learning at our own university through collaboration between academic faculty and administrative staff. Bull Kobe Tokiwa Univ 10:23–32
20. Takamatsu K, Murakami K, Kirimura T et al (2017) A new way of visualizing curricula using competencies: Cosine similarity, multidimensional scaling methods, and scatter plotting. In: 2017 6th international institute of applied informatics (IIAI) international congress on advanced applied informatics (IIAI-AAI). Institute of Electrical and Electronics Engineers (IEEE), pp 192–197
21. Kudo T (2005) MeCab: yet another part-of-speech and morphological analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html. Accessed 12 Apr 2016

Towards an Ontological Approach for the Integration of Information on Operation and Maintenance in BIM for Road Infrastructure

Sara Ait-Lamallam, Imane Sebari, Reda Yaagoubi, and Omar Doukari

Abstract Building information modeling (BIM) is a new process to manage the full lifecycle of a construction project. For road infrastructure, the operation and maintenance (O&M) phase has major importance as it secures its sustainability. However, this phase is not satisfactorily integrated into the BIM process. This limitation emerges from two predominant factors: First, the current industry foundation classes (IFC) standard does not yet support objects related to infrastructure. Second, the information applied to the O&M phase, represented in the IFC, does not allow complete management of this phase. This research paper aims to propose an ontological approach, called IFCInfra4OM, to integrate O&M information in BIM for road infrastructure. In addition, it features the importance and requirements of O&M in road infrastructure and proposes the ontological approach to integrate these requirements within the IFC schema. For this purpose, the methodological diagram is presented. Its steps outline how the analysis of existing international standards is necessary for better integration of BIM into the road infrastructure domain. The proposed methodology develops the criteria to be considered for such an objective. Finally, the scientific contribution of the research is presented in the light of the prevailing field literature. Furthermore, future works for building IFCInfra4OM are submitted.

Keywords Building information modeling · Industry foundation classes · Operation and maintenance · Road infrastructure · Ontological approach · International standards

S. Ait-Lamallam (✉) · I. Sebari · R. Yaagoubi
School of Geomatics and Surveying Engineering, Hassan II Institute of Agriculture and Veterinary Medicine, Rabat, Morocco
e-mail: s.aitlamallam@iav.ac.ma

O. Doukari
CESI Engineers School of Nanterre, Paris, France

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_65

1 Introduction

Building information modeling (BIM) is a collaborative work process around a 3D digital twin of a building [1]. It grants the intervention of several stakeholders on the same structure model to handle it throughout its lifecycle [2]. This process appeared in architecture, engineering and construction (AEC), it has brought a fundamental change in the way of considering the act of designing and building [3].

Sustaining a continuous flow of information exchange during the whole building lifecycle is an advantage of BIM. It enables information availability and accessibility for the O&M phase with no additional efforts [4]. The O&M stage is the most expensive among all phases of a building lifecycle [5]. Indeed, according to [6], the costs of O&M are estimated up to 80% for roads, railways and bridges. These costs could be reduced by integrating BIM in O&M operations.

Nowadays, scientists and industries are working jointly to adopt this new paradigm of structuring O&M information and project management for infrastructure [7, 8]. In the same vein, this paper introduces an approach to integrate O&M information in BIM for road infrastructure. Section 2 provides a general overview of current BIM processes for infrastructure with more emphasis on road infrastructure. Section 3 highlights the importance of the O&M phase for road projects. Section 4 details the proposed methodology to develop an ontology that integrates the O&M information within IFC schema. The last section concludes this paper and gives insights into future work.

2 Literature Review

2.1 BIM for Infrastructure

After a notable relapse in the AEC sector, BIM was introduced to rehabilitate and renovate the field of construction. Nowadays, BIM standards are developed by Building Smart International (BSI), which is an international organization that improves data exchange in the construction industry. BSI works closely with Open Geospatial Consortium (OGC) and International Organization for Standardization (ISO) [9]. This collaboration allows developing open standards for BIM applications and software such as the Industry Foundation Classes (IFC) [10] and ISO 16,739–1 [11].

After its considerable success in the building industry, BIM has proven its efficiency for infrastructure as well. BIM is being increasingly adopted in infrastructure projects for: (1) general infrastructure [12, 13], (2) road infrastructures [14, 15], (3) bridges [16, 17], (4) standards for transportation infrastructure [2] and (5) infrastructure management [18]. In the following, more details are presented for BIM standards and applications in road projects.

BIM Standards for Infrastructure. The IFC is an open data standard that enables a trouble-free exchange of information about a structure (building or infrastructure). It is regulated by the Standard for the Exchange of Product Model Data (STEP) and is specified in EXPRESS (ISO 10,303–11), which is a standard modeling language [19]. The current versions of open BIM standards are IFC2.3 for buildings and IFC4 for infrastructure facilities [20]. IFC development has included the infrastructure sector through the BSI IFC Infra initiative since 2011 [9]. The version IFC4 represents an extension of the 2.3 version [21]. It describes the alignments that represent a common basis for all types of linear infrastructure.

Two other research projects are in progress: (1) IFCBridge [7] that describes specific objects for bridge structures, and (2) IFCRoad for objects related to roads [8]. Despite the integration of BIM in the infrastructure sector, there is no official IFC version that describes specific objects of different types of infrastructure neither in terms of geometry nor in terms of semantics [22].

BIM for Roads Infrastructure. Integrating BIM into road infrastructure projects is still problematic. On one hand, the development of IFC standards, which determine the data structure for these types of facilities, is still in progress. On the other hand, the use of BIM in scientific research works, outside the context of standardization, is often applied in the design and construction phases and remains minimal concerning the O&M one [22].

The constraints of BIM integration within the O&M stage are generated for different reasons. First, there are limitations created by information relating to the management of the O&M phase, which is included in the IFC standard. Second, restrictions are caused by IFC sophistication degree for infrastructure. The model proposed for IFCRoad is oriented toward the structuring of data into classes of objects, essentially, for the conception and construction phase. While managing an infrastructure lifecycle, the BIM standard needs to include more semantics related to object properties that refer to all its phases.

2.2 O&M Phase Requirements for Infrastructure

The aim of O&M is to secure an optimal functioning of any given facility until the end of its service life. If integrated into BIM, O&M information will be available and accessible to all stakeholders of a facility project. This integration will decrease the time and cost of the O&M phase considerably. Several efforts have been deployed to use BIM for managing the O&M in infrastructure projects.

Importance of O&M in Infrastructure Projects. Transport infrastructure projects, such as roads, are characterized by their high complexity [12]. This complexity is due to different conditions in which they exist, such as traffic, climate and geological and topographical variability. They are also extending over a vast territory, which raises the issue of crossing several zones characterized by various geo-referencing systems [23]. Besides, they can interfere with other types of infrastructure along their path, such as bridges, railways and tunnels. These particularities make their management more complex in comparison with building projects.

The O&M phase for the above-mentioned infrastructures is exceptionally long. According to [24], the stakeholders are involved for a minimum period of 20 years compared with the design and construction phases, which could take only a few years [12]. O&M is considered as the costliest phase in an infrastructure life cycle [6].

The ultimate purpose of the O&M phase is to ensure that road infrastructure is optimally functional until reaching the end of its lifecycle. Therefore, a properly handled O&M phase of a road infrastructure is key to the sustainability of the project. BIM provides great opportunities to achieve such a purpose by: (1) integrating O&M information in early phases of the lifecycle for new roads and (2) implementing this information in the BIM As-Built model for existing ones. For both cases, an already existing structure to contain this information should be found in the BIM standard schema that allows fully O&M phase management.

Integrating O&M in BIM. Studies conducted in the scientific and industrial fields, to integrate BIM and O&M information, can be classified as follows: (1) Mapping BIM to GIS-based facility management systems [25, 26]. This approach uses mapping rules between BIM and GIS standards based on the semantic similarities between them, (2) Using classification, for data transfer, to link a classified BIM model with an asset management system, [5](3) Using BIM models to visualize the O&M content of a database like [27, 28]. It is an approach that links a separate database with a BIM model to enhance infrastructure management, (4) Developing an O&M taxonomy linked to objects of a BIM model [29]. This approach aims to attach information directly to BIM objects.

In the standardization context, several international efforts have been made to define the information requirements that a facility management system must fulfill for the operation phase. These initiatives are: (1) The ISO 55,000 standard [30] that specifies a requirement checklist of good management practices relevant to all facility types, (2) the PAS 1192–3 standard published by the British Standards Institution [31]. It specifies how an Asset Information Model should be created and how it should be used and maintained through the lifecycle of buildings and infrastructure assets to reach a level 2 BIM modeling, (3) the EurOTL [32] standard presents a base framework that allows creating libraries of object types, related to transportation infrastructure, to be integrated with road project management systems, (4) the BSI technical report on the requirements of infrastructure managers in a BIM process [33]. This latest determines the minimum information necessary for the management of infrastructure in its investment phase.

For the infrastructure sector, transferring useful knowledge for the management of O&M phase remains the least effective compared with other phases [34]. This problem is caused by several reasons. The limitation of standards' objectives is one of these reasons. For instance: (1) OKSTRA [35], a norm designed for transport infrastructure exchange for both semantic and geometric information, between softwares [36], (2) COINS [37], which is an information model that allows information exchange in the construction phase and (3) Inframodel, a standard based on a classification system, allows the exchange of several infrastructure layers [38]. Another reason for the problematic integration of O&M in BIM is the management systems of assets that contain redundant information. Additionally, the absence of involvement of O&M contractors in the early stages of the projects' lifecycle [12] reduces the effectiveness of this phase management.

In the context of road projects, the current IFC version for infrastructure (IFC4) contains some relevant information for the O&M phase such as [22]:"Date of Information", "Function", "Structure Material", "Manufacturer", "Risk", "Criticality" and "Inspection Frequency". However, these attributes do not allow full management of O&M for road infrastructures [33].

In short, there are worldwide scientific efforts to integrate the O&M phase in different structures' lifecycle. These works aim to determine the useful information for O&M to be integrated with BIM for infrastructure. However, the existing solutions are still unable to fully manage such a phase. To overcome this shortcoming, this paper proposes an ontological approach to enrich IFC BIM standard with O&M information for road infrastructure. This ontological approach aims to add meaningful O&M information to the IFC schema with respect to road objects and their relations.

2.3 Ontology and Domain of Knowledge

In information science, ontology is used for semantic description, schemas interoperability and knowledge sharing via machine-understandable rule sets [39]. An ontology defines concepts and their relationships in a clear and unique manner to allow their understanding in the same way between systems and machines (allowing their re-use, adaptation, extension, improvement or validation). With the increasing importance of knowledge interchange, several solutions have adopted ontologies as their conceptual keystone [40].

In this paper, an ontological approach is proposed to enrich the IFC schema. It allows building a reference data model with specific concepts and connections unlike any other software system. This allows a direct understanding of information by the end-user based on this model without the need for interpretation. It also enables its recycling in several systems. In addition, an ontological approach enables the transition from knowledge to information, which is a basic principle in standard development processes [41]. Additionally, ontology is a concept-oriented approach that always provides a stable perception. In contrast, an engineering approach is a context and goal-oriented that tolerates several interpretations of the same concept [41].

3 Research Scope and Methods

The ontological approach proposed in this paper concerns two knowledge domains: (1) Operation and Maintenance and (2) the BIM for road infrastructure. Its scope is related to the integration of information from the first domain within the concepts of the second as illustrated in Fig. 1.

The O&M domain is a vast knowledge domain where diversity is interpreted by several classes of information and several structure types. Information can be grouped into three levels as illustrated in Fig. 2. Each group contains information relevant to a project stakeholder role. All these information groups need to be integrated with O&M phase management system to enable an efficient system performance [42].

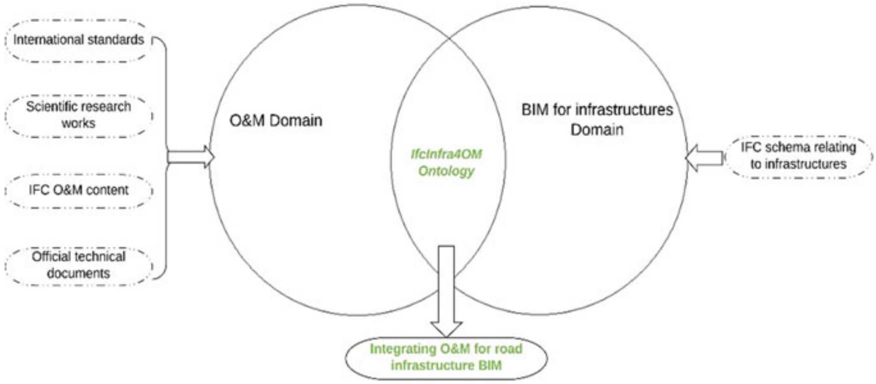


Fig. 1 IFCInfra4OM domains and scope

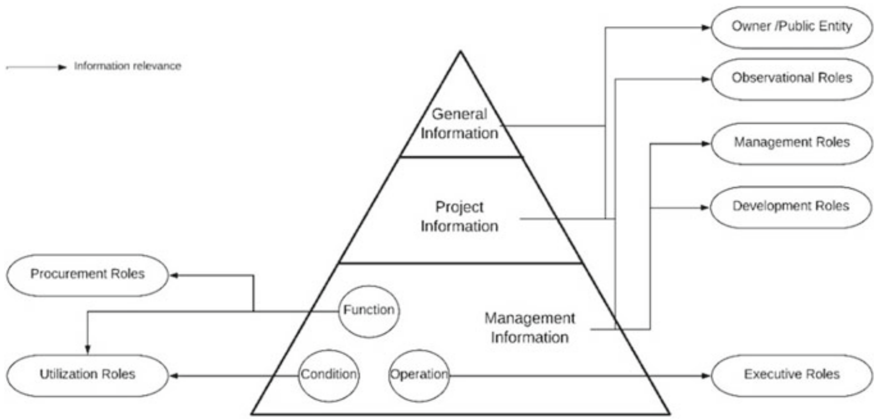


Fig. 2 Information levels in O&M domain and its relevance for stakeholders' roles

General information is about policies and organizations, while project information concern contracts documents, technical documents and asset status and records.

Management information is more diverse:

- Functional information: are information about functions and behaviors relating to the assets. They are grouped into four main categories: information that describes the value and performance of the asset, the risk and guarantees, the requirements for energy exploitation and appropriate level of use of the asset [22].
- Condition information: relative to technical conditions in which the asset or facility is located.
- Operation information: information relevant to an operation.

The BIM, nowadays, is a focal point for transportation infrastructure. Meanwhile, BIM is still a work in progress for roads. The integration of O&M information must take into account the candidate standard schema IFCRoad proposed by the BSI. This will guarantee the interoperability between the IFCInfra4OM model and the future Roads BIM standard.

3.1 Methodology to Build the IFCInfra4OM Ontology

Figure 3 illustrates the methodology adopted to establish IFCInfra4OM ontology. It follows three major steps:

- In the literature review, background materials are: (1) International standards; (2) research works and (3) technical documents (published by specialized organizations). One of the most important shortfalls in these resources is that they contain information non-specific to road infrastructure. On one hand, and as explained in Section 2, information concerns general facilities and might be non-specific

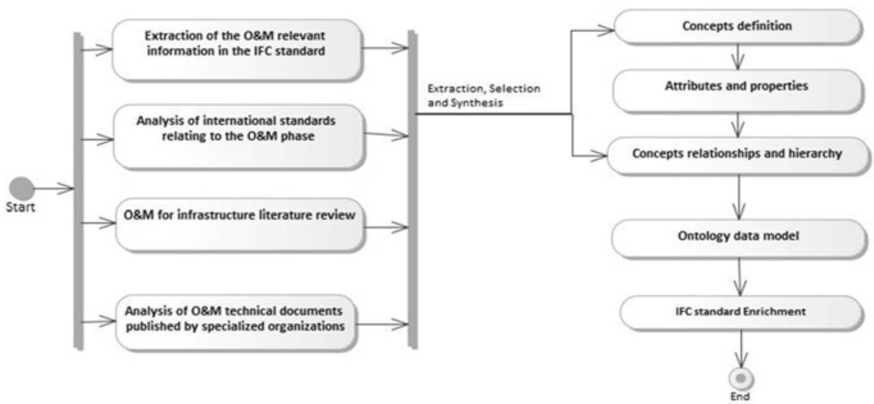


Fig. 3 Methodology proposed for building IFCInfra4OM ontology

to roads. Besides, information can concern the investment phase of a facility and non-specific to O&M stage, which is a part of this phase. On the other hand, the documents are of two genres: universal (displaying and adhering to global standards) and local (specific to a particular region). All these data lead to decide on a level of development (LODt) to adopt while building the ontology. LODt is defined as “A frequently mentioned concept to describe information richness of BIM objects” [43]. It represents the semantic level to which information is detailed. To decide on a LODt, three criteria are discussed in the next step.

- A thorough analysis (extraction, selection, synthesis) of research material is provided in Sect. 2. This step aims to capture relevant information about O&M phase of road infrastructure following the three levels illustrated in Fig. 2. This analysis adheres to the following criteria:

  1. *Universality*: the enrichment of the IFC standard is carried out through universal concepts and not local or region-related ones. For this purpose, holistic concepts were defined.
  2. *Road infrastructure oriented*: Selecting relevant information for this research framework.
  3. *Interoperability*: First, the ontology meets, with no redundancy, O&M information (cited in Section 3), objects entities, and abstract entities already existing in the current official IFC schema and that may relate to O&M phase. Then, it is coherent with the new object entities proposed in the IFCRoad schema, such as IFCRoad entity [8].

- The third step is ontology building. It aims to define concepts, properties, attributes and relationships between concepts. It is conducted according to a defined LODt.

4 Discussion

It is intended, through this research paper, to contribute in making the management of the O&M phase more efficient and sustainable for roads. For this purpose, it is proposed to build an ontology and integrate it with the IFC4 official standard with respect to the future IFCRoad candidate schema. The research contribution is highlighted in Fig. 4. This figure shows several approaches that are used to manage the O&M phase in the context of BIM. They were also detailed in Sect. 2.2 and are represented in Fig. 4 as the black process. These approaches enable a mapping between a BIM model and an existing system (an asset management system, a database or software) to retrieve information, relevant to the O&M phase, that is not supported by the BIM model. This process, although it is the most used, is intricate due to access problems and incompatibilities with the mapped systems. It also needs recurrent adaptations over time due to software specifications change.

The green process is what this research contribution suggests. It shows the IFC4 enrichment workflow and aims to manage the roads O&M phase using only a BIM

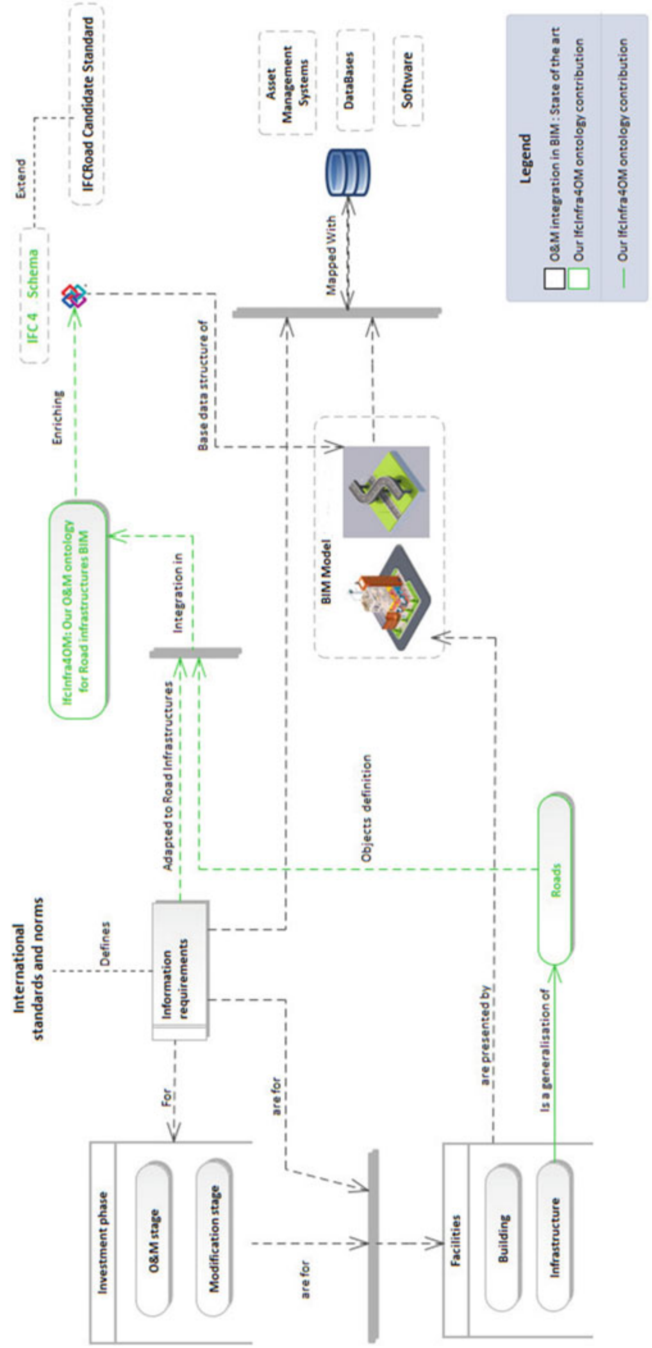


Fig. 4 IFCInfra4OM contribution in the light of the state of the art

model. First, this will make O&M information available directly in the BIM data structure. Second, it will enable to retrieve O&M information from the BIM model to a mapped system which will make it a direct database. Additionally, hierarchy defined in BIM standard for object connection will permit complex analyses for better decision-making. Finally, the respect of the future IFCRoad candidate schema will permit a fast integration of O&M information in the coming version of the IFC towards a whole lifecycle management for roads.

5 Conclusion

Building information modeling integration into the road infrastructure field of knowledge is a novelty. It presents several gaps in O&M use throughout the lifecycle of roads. This paper puts forward the importance of BIM for facilities and describes the O&M phase in the lifecycle of road infrastructure. The ontological approach, proposed in this paper, is a new perspective for a better understanding of the links between O&M and BIM for infrastructure, and a good strategy to enable recycling of concepts. It is also a convenient approach to fit the existing BIM standard for future research work. This research paper is a preparatory work for a data model of the IFCInfra4OM ontology that will be submitted. To validate the methodology employed in this research, a case study will be used to enrich the IFC according to IFCInfra4OM prototype.

References

1. Akob Z, Abang Hipni MZ, Rosly MR (2019) Leveraging on building information modelling (BIM) for infrastructure project: Pan Borneo Highway Sarawak Phase 1. In: IOP conference series: materials science and engineering. Institute of Physics Publishing, p 012060. https://doi.org/10.1088/1757-899X/512/1/012060
2. Amann J, Flurl M, Jubierre JR, Borrmann A (2014) An approach to describe arbitrary transition curves in an IFC-based alignment product data model. In: Computing in civil and building engineering. American Society of Civil Engineers. Reston, VA, pp 933–941. https://doi.org/10.1061/9780784413616.116
3. Tolmer CE, Castaing C, Diab Y, Morand D (2017) Adapting LOD definition to meet BIM uses requirements and data modeling for linear infrastructures projects: using system and requirement engineering. Vis Eng 5:21. https://doi.org/10.1186/s40327-017-0059-9
4. Borrmann A, König M, Koch C, Beetz J (2018) Building information modeling: technology foundations and industry practice. Springer International Publishing.https://doi.org/10.1007/978-3-319-92862-3
5. Heaton J, Parlikad AK, Schooling J (2019) Design and development of BIM models to support operations and maintenance. Comput Ind 111:172–186. https://doi.org/10.1016/j.compind.2019.08.001
6. Pelzeter A (2007) Building optimisation with life cycle costs – the influence of calculation methods. J Facil Manag 5:115–128. https://doi.org/10.1108/14725960710751861
7. Borrmann A, Muhic S, Hyvärinen J, Chipman T, Jaud S, Castaing C, Dumoulin C, Liebich T, Mol L (2019) The IFC-Bridge project—extending the IFC standard to enable high-quality

exchange of bridge information models. In: Proceedings of the 2019 European conference on computing in construction. University College Dublin, pp 377–386. https://doi.org/10.35490/ec3.2019.193
8. Plume J, Erismann C (2019) Delivery plan for infra extensions to IFC. 1–13. https://forums.buildingsmart.org/. Accessed 26 Nov 2020
9. BuildingSmart International: Annual Report 2017. https://buildingsmart-1xbd3ajdayi.netdna-ssl.com/wp-content/uploads/2018/05/2017-bSI-annual-report_final_comp.pdf Accessed 26 Nov 2020
10. BuildingSmart International: Industry Foundation Classes (IFC). https://www.buildingsmart.org/standards/bsi-standards/industry-foundation-classes/. Accessed 26 Nov 2020
11. International Organisation for Standardization: ISO 16739–1:2018, Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries—Part 1: Data schema. https://www.iso.org/standard/70303.html. Accessed 26 Nov 2020
12. Kordestani Ghalenoeei N, Saghatforoush E, Athari Nikooravan H, Preece C (2018) Evaluating solutions to facilitate the presence of operation and maintenance contractors in the pre-occupancy phases: a case study of road infrastructure projects. Int J Constr Manag 1–13. https://doi.org/10.1080/15623599.2018.1512027
13. Ng ST, Xu FJ, Yang Y, Lu M (2017) A master data management solution to unlock the value of big infrastructure data for smart, sustainable and resilient city planning. Procedia Eng 196:939–947. https://doi.org/10.1016/j.proeng.2017.08.034
14. Chong HY, Lopez R, Wang J, Wang X, Zhao Z (2016) Comparative analysis on the adoption and use of BIM in road infrastructure projects. J Manag Eng 32:05016021. https://doi.org/10.1061/(asce)me.1943-5479.0000460
15. Sankaran B, France-Mensah J, O'Brien W (2016) Data integration challenges for CIM in large infrastructure: a case study of Dallas horseshoe project. Int Conf Comput Civ Build Eng (ICCCBE 2016). Osaka Univ. Tokyo, Japan, pp 1015–1022
16. Heikkilä R (2012) National guidelines for bridge information modeling and automation. In: 2012 proceedings of the 29th international symposium of automation and robotics in construction, ISARC 2012. International association for automation and robotics in construction I.A.A.R.C). https://doi.org/10.22260/ISARC2012/0002
17. Park T, Kang T, Lee Y, Seo K (2014) Project cost estimation of national road in preliminary feasibility stage using BIM/GIS platform. In: Computing in civil and building engineering—proceedings of the 2014 international conference on computing in civil and building engineering. American Society of Civil Engineers, Reston, VA, pp 423–430. https://doi.org/10.1061/9780784413616.053
18. Marzouk M, Abdel Aty A (2012) Maintaining subway infrastructure using BIM. In: Construction research congress 2012. American Society of Civil Engineers, Reston, VA, pp 2320–2328. https://doi.org/10.1061/9780784412329.233
19. Chen G, Luo Y (2017) A BIM and ontology-based intelligent application framework. In: Proceedings of 2016 IEEE advanced information management, communicates, electronic and automation control conference, IMCEC 2016. Institute of Electrical and Electronics Engineers Inc., pp 494–497. https://doi.org/10.1109/IMCEC.2016.7867261
20. BuildingSmart International: Industry Foundation Classes Version 4.1.0.0. https://standards.buildingsmart.org/IFC/RELEASE/IFC4_1/FINAL/HTML/. Accessed 26 Nov 2020
21. Borrmann A, Amann J, Chipman T, Hyvärinen J, Liebich T, Muhic S, Mol L, Plume J, Scarponcini P, IFC infra overall architecture project documentation and guidelines. https://www.buildingsmart.org/wp-content/uploads/2017/07/08_bSI_Overall. Accessed 26 Nov 2020
22. Floros GS, Boyes G, Owens D, Ellul C (2019) Developing IFC for infrastructure: a case study of three highway entitieS. ISPRS Ann. Photogramm. Remote Sens Spat Inf Sci IV-4/W8, 59–66. https://doi.org/10.5194/isprs-annals-IV-4-W8-59-2019
23. Markič Š, Donaubauer A, Borrmann A (2018) Enabling geodetic coordinate reference systems in building information modeling for infrastructure. In: Proceeding of the 17th international conference on computing in civil and building engineering. Tampere, Finland

24. Lin FC (2005) The integration between design and maintenance of building automation: a decision support approach. Fac Built Environ Eng Queensl Univ Technol. https://eprints.qut.edu.au/4478/1/4478.pdf. Accessed 15 Dec 2020
25. Deng Y, Cheng JCP, Anumba C (2016) Mapping between BIM and 3D GIS in different levels of detail using schema mediation and instance comparison. Autom Constr 67:1–21. https://doi.org/10.1016/j.autcon.2016.03.006
26. Wetzel EM, Thabet WY (2015) The use of a BIM-based framework to support safe facility management processes. Autom Constr 60:12–24. https://doi.org/10.1016/j.autcon.2015.09.004
27. Aziz Z, Riaz Z, Arslan M (2017) Leveraging BIM and big data to deliver well maintained highways. Facilities 35:818–832. https://doi.org/10.1108/F-02-2016-0021
28. Yin X, Liu H, Chen Y, Wang Y, Al-Hussein M (2020) A BIM-based framework for operation and maintenance of utility tunnels. Tunn Undergr Sp Technol 97:103252.https://doi.org/10.1016/j.tust.2019.103252
29. Motamedi A, Hammad A, Asen Y (2014) Knowledge-assisted BIM-based visual analytics for failure root cause detection in facilities management. Autom Constr 43:73–83. https://doi.org/10.1016/j.autcon.2014.03.012
30. International Organisation for Standardization: ISO 55000:2014, Asset management—Overview, principles and terminology. https://www.iso.org/standard/55088.html. Accessed 26 Nov 2020
31. British Standard Institution (2014) PAS 1192–3:2014 Specification for information management for the operational phase of assets using building information modelling. BSI Stand. Publ. 1:1–44. https://bugvaorg.files.wordpress.com/2018/09/bsi_pas_1192-3_2014.pdf. Accessed 26 Nov 2020
32. EUROTL Framework overview. https://www.roadotl.eu/static/eurotl-ontologies/index.html. Accessed 15 Dec 2020
33. Jackson P (2018) Infrastructure asset managers BIM requirements delivering the information "Asset Managers" need and can trust using openBIMTM. buildingSMART. https://www.buildingsmart.org/wp-content/uploads/2018/01/18-01-09-AM-TR1010.pdf. Accessed 26 Nov 2020
34. O'Keeffe A, Alsem D, Corbally R, van Lanen R (2017) INTERLINK D.2 & D.3 Combined WPA & WPB report. Technical report, conference of European directors of roads. https://www.cedr.eu/wpfb-file/interlink-d-2-d-3-wpa-wpb-report-final-issue-pdf/. Accessed 15 Dec 2020
35. OKSTRA®. http://www.okstra.de/. Accessed 26 Nov 2020
36. Schultze C, Buhmann E (2008) Developing the OKSTRA ® Standard for the needs of landscape planning in context of implementation for mitigation and landscape envelope planning of road projects
37. BIM Locket: COINS. http://www.coinsweb.nl/Introduction_COINS_2.htm. Accessed 26 Nov 2020
38. BuildingSmart International: Inframodel Data Exchange. https://buildingsmart.fi/en/infrabim-en/inframodel-data-exhange. Accessed 26 Nov 2020
39. Gu H, Li H, Yan L, Liu Z, Blaschke T, Soergel U (2017) An object-based semantic classification method for high resolution remote sensing imagery using ontology. Remote Sens 9:329. https://doi.org/10.3390/rs9040329
40. Wardhana H, Ashari A, Kartika A (2018) Review of ontology evolution process. Int J Comput Appl 179:26–33. https://doi.org/10.5120/ijca2018916537
41. Katranuschkov P, Gehre A, Scherer RJ (2003) An ontology framework to access IFC model data. J Inform Technol Constr 8:413–437. http://www.itcon.org/2003/29. Accessed 15 Dec 2020
42. Xu X, Ma L, Ding L (2014) A framework for BIM-enabled life-cycle information management of construction project. Int J Adv Robot Syst 11:126. https://doi.org/10.5772/58445
43. Volk R, Stengel J, Schultmann F (2014) Building information modeling (BIM) for existing buildings — literature review and future needs. Autom Constr 38:109–127. https://doi.org/10.1016/j.autcon.2013.10.023
44. Kim J-U, Kim Y-J, Ok H, Yang S-H (2015) A study on the status of infrastructure BIM and BIM library development. In: 2015 international conference on computational science and computational intelligence (CSCI). IEEE, pp 857–858. https://doi.org/10.1109/CSCI.2015.52

Group's Influence Value in Logistic Regression Model and Gradient Boosting Model

Quang-Vinh Dang, My-Linh Tran, Minh-Hung Dang, Thi-Minh-Trang Tran, Huu-Nghia Nguyen, Thi-Minh-Hue Cai, and Thanh-Duyen Phan

Abstract Measuring the influence of particular data point is an important task in designing a data hub. In recent years, several research works have addressed the problem. In this paper we analyze these approaches and show that they are too complicated to apply in practice. We propose a new lightweight approach to approximate the influence of data points. We evaluate our proposal on the popular Home Credit dataset to show the effective.

Keywords Data valuation · Influence function · Gradient boosting

Q.-V. Dang (✉)
Industrial University of Ho Chi Minh city, Ho Chi Minh City, Vietnam
e-mail: dangquangvinh@iuh.edu.vn

M.-L. Tran
Quy Nhon University, Binh Dinh, Vietnam
e-mail: tranmylinh@qnu.edu.vn

M.-H. Dang · T.-M.-T. Tran · H.-N. Nguyen · T.-M.-H. Cai · T.-D. Phan
TranData - FPT Software, Quy Nhon, Binh Dinh, Vietnam
e-mail: hungdm16@fsoft.com.vn

T.-M.-T. Tran
e-mail: trangttm3@fsoft.com.vn

H.-N. Nguyen
e-mail: nghianh7@fsoft.com.vn

T.-M.-H. Cai
e-mail: huectm@fsoft.com.vn

T.-D. Phan
e-mail: duyenpt7@fsoft.com.vn

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_66

1 Introduction

In recent years, the rise of data hub where multiple organizations can share and reuse the data of each other raise several crucial questions regarding security, trust, privacy, and particularly data evaluation that we address in this paper. On the other hand, the call for federated learning wherein the training process is distributed over multiple sides requires an effective evaluation metric to prevent free riders and malicious participants. On the other hand, the quality of the data from different entities may vary widely. Therefore, a key question often asked by stakeholders of a ML system is how to determine the influence or the value of groups data points are provided by data contributors. If we fail to do so, the system will be tricked easily by contributing noisy or malicious data. A natural way of tackling the data valuation problem is to calculate the value of each point in the group and then the value of each group is the sum of the component points.

The influence function is a method that calculates the value of each data point. It does the task by calculating the impact if we remove a single data point from the training set. Indeed the naive way is to train everything again from scratch, but it is not the idea of the influence score. There have many studies and applications to influence function for individual points; but there is little research on influence function for group points, especially for tree model. Influence boosting [6] will be used for ensemble-based models, such as Random Forest (RF) or Gradient Boosted Decision Trees (GDBT) to determine the value of group data points.

In this paper, we first show how to calculate influence function of group data points for logistic regression. Then we applied influence boosting for tree model and compared the performance with logistic regression model. We use the popular Home Credit dataset released on Kaggle website [2] to evaluate our approach.

Our contributions are: (i) we show that the state-of-the-art influence measuring methods are not suitable to be used in practice as they require a huge computational power; and (ii) we present a novel method to approximate the influence scores of a group of data points instead of a single data point; and (iii) we evaluate our proposal in a real-world dataset.

2 Measuring Group Effects

2.1 Related Works

Data evaluation is a relatively new problem rising recently. Particularly, we focus on two papers. The first one is about the influence function of group for parameter model [4] and the other is about influence function of training point for gradient boosted decision trees model [6].

2.1.1 Review Influence Function of Group for Parameter Model

The influence score for groups has been introduced recently in [4]. The study provides an approach to calculate the influence function of a subset of training data.

Consider learning a parameter model (for instance, a logistic regression model or SVM). With input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Our task is to find a parameter set $\theta \in \Theta$ that maps from input to output. We are provided with n training points notated as $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and we define a loss function $L(x, y, \theta)$ that is double-differentiable and convex in Θ.

The parameters are selected to satisfy the following condition:

$$\hat{\theta}(\mathbf{1}) = argmin_{\theta\in\Theta}\left[\sum_{i=1}^{n} L(x_i, y_i; \theta)\right] + \frac{\lambda}{2}\|\theta\|_2^2 \tag{1}$$

that minimize the L_2-regularized empirical risk. In the $\hat{\theta}(\mathbf{1})$, $\mathbf{1}$ is the all-one vector to reflects the fact that at the beginning all the training points are assigned the same uniform sample weights.

Our objective is to evaluate the impact on the model of different training data subsets. We define the sample weight vector $w \in \{0, 1\}^n$ as $w_i = \mathbb{I}((xi, yi) \in W)$ and consider the modified parameters

$$\hat{\theta}(\mathbf{1} - w) = argmin_{\theta\in\Theta}\left[\sum_{i=1}^{n}(1 - w_i)L(x_i, y_i; \theta)\right] + \frac{\lambda}{2}\|\theta\|_2^2 \tag{2}$$

corresponding to the retrained model after excluding W.

The *actual effect* $\mathcal{I}_f^* : [0, 1]^n \to \mathbb{R}$ of the subset w is

$$\mathcal{I}_f^*(w) = f(\hat{\theta}(\mathbf{1} - w)) - f(\hat{\theta}(1)) \tag{3}$$

where the evaluation function $f : \Theta \to \mathbb{R}$ measures the quantity of interest.

The problem with computing the actual effect $\mathcal{I}_f^*(w)$ is that it can be prohibitively costly to retrain the model to compute $\hat{\theta}(\mathbf{1} - w)$ for each w subset. Influence functions provide $\mathcal{I}_f^*(w)$ with a relatively powerful first-order approximation that prevents retraining as following.

Consider the function:

$$\begin{aligned} q_w &: [0, 1] \to \mathbb{R} \\ & tw \longmapsto q_w(tw) = f(\hat{\theta}(\mathbf{1} - tw)) \end{aligned}$$

So that the *actual effect*:

$$\mathcal{I}_f^*(w) = q_w(1) - q_w(0)$$

Using first-order Taylor's approximation :

$$q_w^{'}(0) \approx q_w(1) - q_w(0)$$

So we define *predict effect* $\mathcal{I}_f(w)$:

$$\mathcal{I}_f(w) \overset{\text{def}}{=} q_w^{'}(0) = \nabla_\theta f(\hat{\theta}\,(\mathbf{1}))^\top \left[\frac{d}{dt}\hat{\theta}(\mathbf{1} - tw)\Big|_{t=0} \right] \\ = \nabla_\theta f(\hat{\theta}\,(\mathbf{1}))^\top\, H_{\lambda,\mathbf{1}}^{-1}\mathbf{g_1}(w) \tag{4}$$

where $\mathbf{g_1}(w) = \sum_{i=1}^{n} w_i \nabla_\theta \ell(x_i, y_i; \hat{\theta}(\mathbf{1}))$, $H_\mathbf{1} = \sum_{i=1}^{n} \nabla_\theta^2 \ell(x_i, y_i; \hat{\theta}(\mathbf{1}))$, and $H_{\lambda,\mathbf{1}} = H_\mathbf{1} + \lambda I$.

In the work of [3] the authors showed that the influence of groups of data points can be estimated via the influence score of single data point, though it is not a complete match. Therefore, we can use (4) to calculate the influence score for groups of data points for logistic regression model.

2.1.2 Review Influence Boosting for Gradient Boosted Decision Trees

Similar to Koh et al., 2019 [5], with Gradient Boosted Decision Trees (GBDT) model, Boris Sharchilev et al., 2018 [6] calculated influence actual for training points and influence function formula.

With space input X_{train} and space output Y_{train}, we have a GBDT model with standard training ensemble $F(x) := \sum_{t=1}^{T} f^t(A^{t-1})(x)$, where A^{t-1} is intermediate prediction on x and $f^t()$ is leaf value of x in t^{th} tree. To calculate influence score for individual points $x_{train} \in X_{train}$, we need to retrain model without x_{train}. Since retrain multiple times increases computation cost, we assume that the tree structure does not change with each retrain; so we only need to recalculate for each tree (for each retrain) in algorithm LeafRefit [6].

In GBDT model, we define the *actual effect*:

$$\mathcal{I}_{grad}^*(x_{train}, x_{test}) := L(y_{test}, F(x_{test})) - L(y_{test}, \hat{F}_{\setminus x_{train}}(x_{test})) \tag{5}$$

where $\hat{F}_{\setminus x_{train}}$ is the model retrain without x_{train}.

We have the *predict effect* (has been showed [6]):

$$\mathcal{I}_{grad}(x_{train}, x_{test}) := \frac{\delta L(y_{test}, F(x_{test}))}{\delta w_i(x_{train})} \tag{6}$$

Constructing the *LeafInflence* algorithm to calculate the derivative of the leaf value is similar to the algorithm to calculate the leaf value [6].

Algorithm *LeafInfluence*

Input: training point x_{i_0}, leaf formula type *formula*

Output: new leaf values $\{\hat{f}_l^t\}_{t=1,l=1}^{T,L}$, $M(\mathbf{A}_{ij}^o)$

for t =1 **to** T **do**:

$$\frac{\delta f_l^t(A^{t-1})}{\delta w_{i_0}} = -\frac{I_l^t(i)(f_{G;l}^t + g_i^t) + \sum_{j \in I_l^t} w_j h_j^t J(A^{t-1})_{ij}}{H_{G;l}^t}, l = 1...L$$

or

$$\frac{\delta f_l^t(A^{t-1})}{\delta w_{i_0}} = -\frac{I_l^t(i)(h_i^t f_{H;l}^t + g_i^t) + \sum_{j \in I_l^t} w_j (k_j^t f_{H;l}^t + h_j^t) J(A^{t-1})_{ij}}{H_{G;l}^t}, l = 1...L$$

$$M(\mathbf{A}_{ij}^t) = M(\mathbf{A}^{t-1})_{ij} + \frac{\partial f_{P(x_j)_t}^t}{\partial w_j}, i = 1...n, j = 1...n$$

end for

return $\{\frac{\delta f_l^t(A^{t-1})}{\delta w_i}\}_{t=1,l=1}^{T,L}$

According to their conclusion, through the experimental results, the authors suggested that the calculation can be used in practice. Therefore, we use this algorithm to calculate the influence score of single data points for GBDT model and expand to calculate the influence function of groups data points.

2.2 *Calculating the Influence Function of Group*

We are given a training data X_{train}. The dataset is divided into m groups $X_1, \ldots, X_m$.

2.2.1 Retrain Tree Model

We train model with gradient boosting algorithm on training data X_{train}. With a test point (x_{test}, y_{test}), we have a predict $F(x_{test})$. Then we retrain model on training data $X_{train} \setminus X_i$ for $i = 1, \ldots, m$; and we have predict $F_i(x_{test})$.

We define the influence function of a train group with a test point:

$$\mathcal{I}_i^*(X_i, x_{test}) = L(y_{test}, F(x_{test})) - L(y_{test}, F_i(x_{test})), i = 1, \ldots m \quad (7)$$

where $L(y, F(x))$ is the loss function in gradient boosting model.

2.2.2 Sum of Influence Training Points

In GBDT model, the influence function on x_{train} is $\mathcal{I}_{grad}(x_{train}, x_{test})$ (6), with $x_{train} \in X_{train}$. We define the influence function of a train group with a test point:

$$\mathcal{I}_i^{**}(X_i, x_{test}) = \sum \mathcal{I}_{grad}(x_j, x_{test}) \quad (8)$$

where $i = 1, \ldots m, x_j \in X_{train} \setminus X_i$.

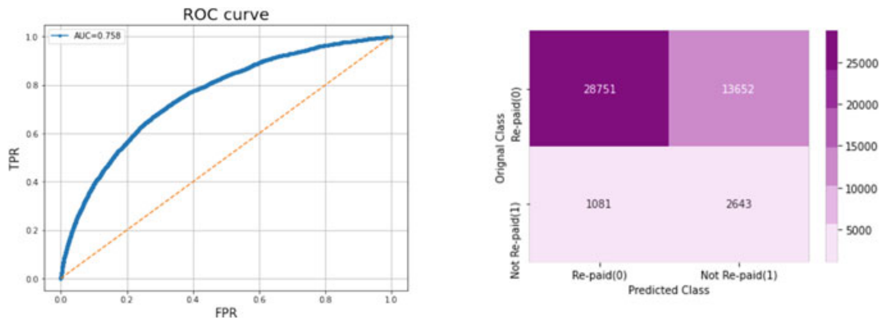


Fig. 1 AUC score Logistic model for Home Credit data

3 Experiment

For our experiments with Logistic regression and GBDT, we implement methods to calculate influence function show in [3, 6]. The dataset used for our experiments is [2] Home Credit.

The dataset contains about 200,000 record with more than 500 features in each data point. The data we used contains about 200 features. We pick 20 subsets from the dataset with number data points in each subset random from 5,000 to 15,000 data points.

3.1 Influence Value for Logistic Regression

First we train model logistic regression to predict loan risk. We set parameter are: learning rate (0.01), because the dataset is imbalanced so we set class weight (0.08 : 0.92). The result of model logistic after training is 0.78 (Fig. 1).

Using the influence function [3] we get influence score for each subset and plot in Fig. 2.

3.2 Influence value for Gradient Boosting model

For our experiments with GBDT, we use CatBoost [1] an open-source implementation of GBDT by Yandex (Fig. 3).

We train the CatBoost tree with 100 epochs, depth 5 at each tree. The learning rate is 0.15 and we set class weight (0.05 : 0.95). Then we apply influence boosting to calculate influence score for each subset and plot in Fig. 4.

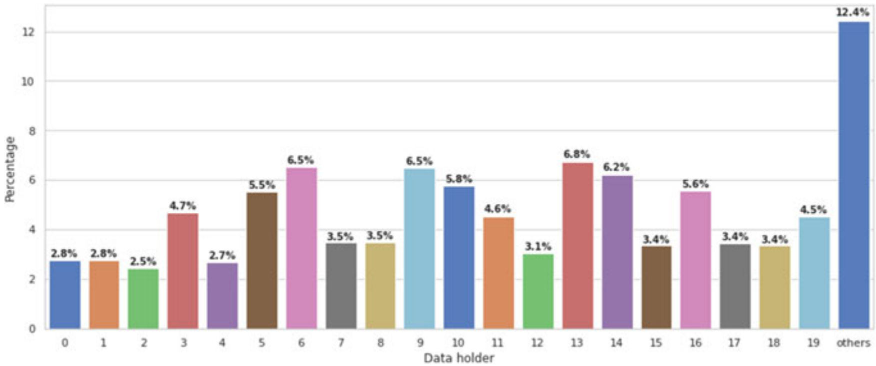


Fig. 2 Influence value for each group (Influence function)

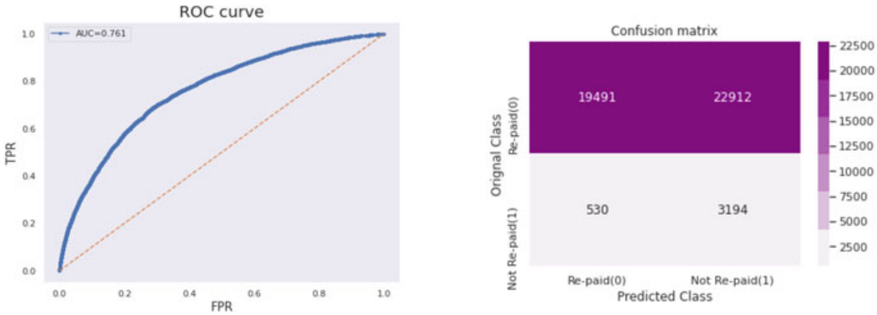


Fig. 3 AUC score CatBoost model for Home Credit data

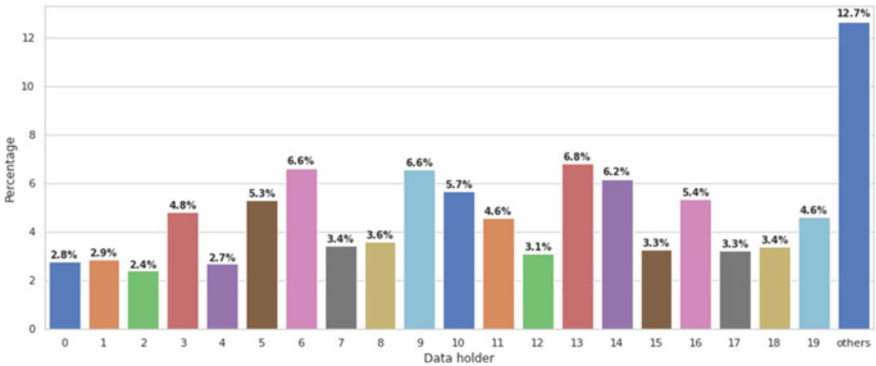


Fig. 4 Influence value for each group (Influence boosting)

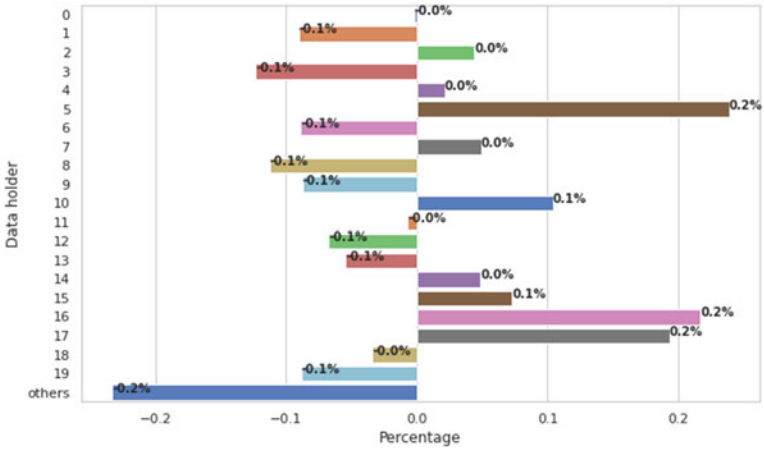
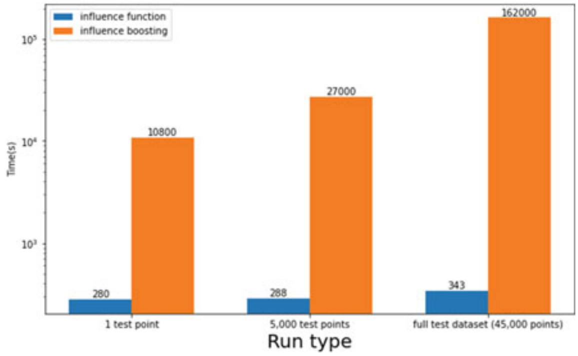


Fig. 5 Different percent between influence function and influence boosting

Fig. 6 Running time between GDBT and logistic regression



We can see that the influence scores are computed by two methods that are correlated with each other. We conclude that we can approximate the influence scores by a much lighter method (Fig. 5).

3.3 Running Time Comparison

We display the running time of two algorithms in Fig. 6. We can clearly observe that the logistic regression algorithm takes much less time than the GDBT while retaining the high accuracy influence scores.

4 Conclusions

In this paper we present and analyze the problem of data valuation in data hub and federated learning context. We study a few state-of-the-art methods and show their inefficiency in dealing with big data in practice. We propose a lightweight proxy to approximate the influence score of data points. We evaluate our proposal using a real-world dataset. In the future, we plan to work on the dynamic problems when data points are added and removed frequently.

Acknowledgements My-Linh Tran, VINIF.2020.ThS.QN.02 was funded by Vingroup Joint Stock Company and supported by the Domestic Master/Ph.D. Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VinBigdata).

References

1. Dorogush AV, Ershov V, Gulin A (2018) Catboost: gradient boosting with categorical features support. arXiv:1810.11363
2. Kaggle: Home credit default risk. https://www.kaggle.com/c/home-credit-default-risk/data. Accessed 31 Sept 2020
3. Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. arXiv:1703.04730
4. Koh PW, Steinhardt J, Liang P (2018) Stronger data poisoning attacks break data sanitization defenses. arXiv:1811.00741
5. Koh PWW, Ang KS, Teo H, Liang PS (2019) On the accuracy of influence functions for measuring group effects. In: Advances in neural information processing systems, pp 5254–5264
6. Sharchilev B, Ustinovsky Y, Serdyukov P, de Rijke M (2018) Finding influential training samples for gradient boosted decision trees. arXiv:1802.06640

Keyword Extraction Algorithm Based on Pre-training and Multi-task Training

Lingqi Guo, Haifeng Sun, Qi Qi, and Jingyu Wang

Abstract The generalization ability of the supervised model is relatively weak in keyword extraction technology. For enhancing the robustness of the model, a keyword extraction method is proposed inspired by the pre-training model. After pre-training with plenty of corpus and fine-tuning with specific datasets, the proposed method performs more robust in keyword extraction tasks. In addition, multi-task training is added in the fine-tuning stage to improve the accuracy of the model. Plenty of comparative experiments show that the proposed method is very significant in improving the robustness and accuracy of the model.

Keywords Keyword extraction · Pre-training · Multi-task training · Model robustness · Transfer learning

1 Introduction

Keyword extraction technology refers to the technology of extracting representative words from text. It is the foundational work of text research such as text retrieval, abstract generation and text classification.

In order to extract keywords accurately and effectively, the researchers put forward two effective ways. One is the unsupervised method, which extracts keywords by their statistical features, such as word frequency, whether the candidate words are capitalized and so on. The representatives of such methods are TF-IDF [1], YAKE [2]. Some researchers, inspired by PageRank [3] algorithm, build a word graph model to extract keywords, such as TextRank [4]. The other one is a supervised method. After manually labeling keywords, the machine learning model is trained using the word frequency and location information of the words. The representative of this kind of algorithm is Kea proposed by Ian et al. [5].

L. Guo (✉) · H. Sun · Q. Qi · J. Wang
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
e-mail: guolingqi@ebupt.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_67

However, statistical features still have their limitations. With the rapid development and wide application of deep learning technology in recent years, researchers pay more attention to extract keywords by neural network and have achieved good results. Marco et al. [6] proposed to use GloVe to encode text, then train BiLSTM (Bi-Directional Long Short-Term Memory) model to extract keywords. Previous keyword extraction mainly used models to extract words that already in the original text but could not cope with the situation that keywords did not in the original text. In order to solve this problem, Rui Meng et al. [7] proposed to use Copy mechanism and Seq2Seq model to generate keywords, and this work also achieved good results.

After summarizing the previous work, it is found that the accuracy and recall rate of the unsupervised model are not high, but the generalization ability is relatively strong; the accuracy of the supervised model is high, but the generalization ability is weak. The main reason for this situation is that the unsupervised learning method has learned the statistical features of texts, which are generally of high migration, but it is difficult to learn the complex semantic relations between words. However, the supervised learning model can learn the relations between texts, but these features are difficult to migrate in cross-domain situations. Inspired by the pre-training model in natural language processing in recent years, this paper applies a large number of training data in the open domain for pre-training and proposes a method combining both pre-training and multi-task to better complete the keyword extraction task.

In recent years, more pre-training language models have been proposed, and their good performance and better robustness in natural language processing prove that two-stage training method combining pre-training and fine-tuning training is effective. The two-stage training method applies open domain corpus in the pre-training stage, which makes the model more extensive. In the fine-tuning stage, the datasets of specific fields or specific tasks are applied, which makes the model more suitable for downstream tasks. Marco et al. also used two-step training to extract keywords, which proved the effectiveness of this method.

In multi-task training, because the goal of auxiliary task is similar to that of keyword extraction task, but the difficulty of auxiliary task is lower, so the auxiliary task can help keyword extraction task to remove noise and make results correct in actual training process.

This paper uses the encoder of transformer [8], which performs well in the field of natural language processing, as the main component of the method proposed in this paper. In order to improve the generalization ability of the model, a large number of papers in open domain are used to pre-train the model. Then, in order to better adapt to the field of downstream tasks, the pre-training model is fine-tuned by datasets in specific fields. For improving the precision of the model and keeping the consistency with the pre-training tasks, the multi-task training method is adopted in the fine-tuning stage. After the test set is predicted, some rules are added to sort out the predicted results.

After comparing with other mainstream models, it is proved that the model proposed in this paper has greatly and effectively improved the robustness and accuracy of key phrases extraction, compared with the mainstream model.

The rest of the paper is organized as follows. Section 2 introduces the related work, while the details of data sets and methodology are presented in Sect. 3. Section 4 discusses the experimental results on different performance measures. Conclusions and future work are provided in Sect. 5.

2 Related Work

2.1 Keywords Extraction

Keywords refer to words that are more scientific and efficient to reflect the content of the article, and keyword extraction technology is a basic but important technology in the field of natural language processing. In 1972, Jones and others proposed TF-IDF algorithm, which was used by researchers to extract keywords and achieved good results. Nowadays, the technology of keyword extraction has made remarkable progress.

There are two main ways to extract keywords. One is the unsupervised learning method, which mainly uses the statistical features of words to get keywords through weight calculation. The typical algorithm in this aspect is YAKE algorithm proposed by Campos et al., which makes use of the statistical characteristics of word frequency, the position of the first occurrence of the word, whether the candidate words are capitalized, and then uses the weight to find the ranking of the candidate list to get keywords. Some researchers have used graphs to extract keywords. The typical work is TextRank, PositionRank [9] and other algorithms. Influenced by the pre-training model, SUN et al. [10] proposed an unsupervised method to extract keywords by using word vectors and sentence vectors.

Another main way to extract keywords is supervised learning method. This method often requires a lot of manual labels, but its accuracy is usually higher than that of unsupervised methods. KEA algorithm proposed by Ian et al. trains machine learning models such as Naive Bayesian Classifier to extract keywords through statistical features. With the development of deep learning, people pay more attention to how to extract keywords accurately by using deep learning model. In 2018, Marco et al. used BiLSTM to extract keywords. This method uses word vectors as features to train neural networks, which has good robustness. Zhang et al. [11] proposed a new LSTM structure, which uses attention mechanism to extract keywords through two-stage training. R. Meng et al. took the keyword extraction task as generation task, and generated keywords according to text content by using Seq2Seq model and Copy mechanism. The previous unsupervised methods performed poorly in accuracy, while the supervised methods performed well in specific fields but poor in migration. The method proposed in this paper uses a large amount of open domain data for pre-training, which makes the model both accurate and migratory.

2.2 Pre-training Language Model

As tagged data are difficult to obtain, researchers began to study how to use plenty of unlabelled data to improve the ability of language models. The work of Devlin et al. [12] proves that the pre-training language model is very effective. The main method of pre-training model is to pre-train the model with a large number of unlabeled data samples, and then fine-tune the model with specific datasets so that the model can perform better in downstream tasks. In 2018, Devlin et al. built a bi-directional pre-training language model BERT by using the transformer model, and made great progress in the field of natural language processing. After that, GPT-2, a pre-training model put forward by

Radford et al. [13], achieved good results in generative tasks. In 2019, more work to improve the BERT model was put forward. For example, Zhang et al. [14] integrated knowledge into the BERT and proposed a new pre-training model ERNIE. In addition, ALBERT [15] based on BERT has a faster encoding speed. RoBERTa, a pre-training model trained by Vaswani et al. [16] with more data, turns up. In the same year, Yang et al. proposed a new autoregressive language model XLNet [17] based on autoregressive and self-coding mode.

Inspired by the pre-training model, in order to make full use of the data to improve the generalization ability of the model, the method proposed in this paper is to pre-train the keyword extraction model first, and on the basis of pre-training, fine-tune the training to make the model more suitable for specific datasets.

3 Methods

The main content of this chapter is the details and training methods of the model proposed in this paper. The model structure is mainly described in Sect. 3.1, and more details of model training method are described in Sect. 3.2.

3.1 Model Architecture

The main structure of the keyword extraction model proposed in this paper is the encoder structure in transformer. The method of pre-training and fine-tuning is used to alleviate the situation of poor cross-domain performance of supervised keyword extraction algorithm. The structure of the model is shown in Fig. 1. The model mainly includes text coding layer, encoder, sentence classification layer and keyword extraction layer.

Assuming that the input text sequence is $\left\{w^1, w^2, \ldots w^n\right\}$, the text sequence is coded as vector sequence $\left\{v^1, v^2, \ldots v^n\right\}$ through the text coding layer.

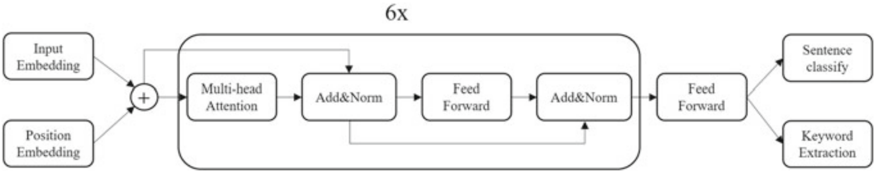


Fig. 1 Model structure

According to Vaswani's work, in the multi-head attention layer of encoder, three different weight matrices are initialized for each $head_i$, which are Value vector (V_i), Key vector (K_i) and Query vector (Q_i), and d_{K_i} represents the dimension of Key vector. Then the calculation formula of self-attention is:

$$head_i = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_{K_i}}}\right)V_i \tag{1}$$

The formula for calculating multi-head attention is as follows, where W_0 is the weight matrix and $MultiHead$ is the multi-head attention.

$$MultiHead = Concat(head_1, \ldots, head_i)W_0 \tag{2}$$

The sentence classification layer is softmax layer, which is used to classify the encoder results and calculate the distribution probability. If the input text contains keywords, the label is 1, otherwise it is 0. The purpose of setting this layer is to better judge the prediction result quality of keyword extraction task. The expression of the i-th sample result y_{CLS}^i I output by the sentence classification layer is:

$$y_{CLS}^i = Softmax(W_C H_C + b_C) \tag{3}$$

where H_C is the output by the full connection layer, W_C is the weight matrix, and b_C is the bias vector. The loss function expression of training sentence classification task is:

$$L_{CLS} = -\sum_i y_{CLS}^{i'} \log\left(y_{CLS}^i\right) \tag{4}$$

where $y_{CLS}^{i'}$ represents the correct label of the ith sample, and y_{CLS}^i represents the posterior probability of the i-th sample predicted by the sentence classification layer.

The keyword extraction layer is also softmax layer, which is used to find the probability distribution of keywords in the sequence. The value y_{KE}^i of the i-th word tag predicted by the keyword extraction layer is expressed as:

$$y_{KE}^i = Softmax(W_K H_K + b_K) \tag{5}$$

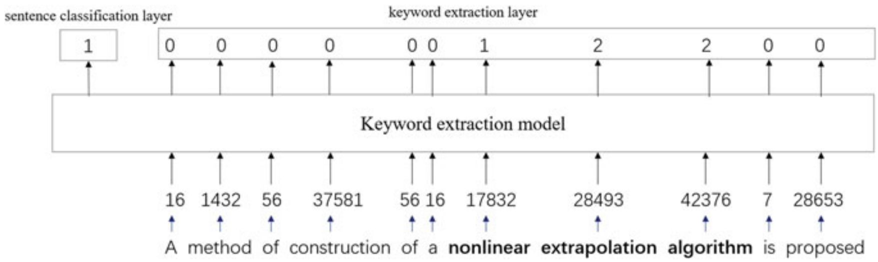


Fig. 2 The input and output diagram

where H_K is the output of the fully connected layer, W_K is the weight matrix, and b_K is the bias vector. The loss function expression of training keyword extraction task is:

$$L_{KE} = -\sum_i y_{KE}^{i\prime} \log\left(y_{KE}^i\right) \tag{6}$$

where $\boldsymbol{y_{KE}^{i\prime}}$ represents the correct label of the ith word in the text, and $\boldsymbol{y_{KE}^i}$ represents the posterior probability of the ith word predicted by the keyword extraction layer.

The input and output diagram of the model is shown in Fig. 2. The input text is "a method of construction of a nonlinear extrapolation algorithm is proposed", which is converted into an index sequence, and then encoded into a vector sequence. The annotated keyword of the input text is "Nonlinear Extrapolation Algorithm", so the output label at the sentence classification layer is 1. In the output sequence of the keyword extraction layer, the output label at the beginning of the keyword is 1, the output label at the following keyword parts is 2, and the output labels at other positions are 0.

3.2 Pre-training and Fine-Tuning

For enhancing the generalization ability of the model and making full use of the data, this paper uses two stages of pre-training and fine-tuning. In the pre-training stage, a large number of academic papers are used as data. After the paper text is divided into n sentences, the i-th sample is $\left(s^i, l^i\right)$, where s^i represents the ith sentence, and l^i represents whether there are keywords in s^i. The value of l^i is as follows:

$$l^i = \begin{cases} 1 & If\ \ s^i\ contains\ \ eywords \\ 0 & If\ s^i\ does\ not\ contain\ keywords \end{cases} \tag{7}$$

The output of the model in the pre-training stage is only the output of the sentence classification layer, so the loss function $L_{pre-train}$ in the pre-training stage is:

$$L_{pre-train} = L_{CLS} \tag{8}$$

where L_{CLS} stands for the loss calculation function of sentence classification task.

After the pre-training, the parameters of the model are saved, and fine-tuning training is carried out for different datasets. In order to improve the precision of fine-tuning training, the method proposed in this paper adopts multi-task method to carry out fine-tuning training. Then the i-th data in the training set can be expressed as $\left(s^i, l_c^i, l_k^i\right)$, where s_i represents the text of the i-th data, l_c^i represents whether the sentence contains keywords or not, and l_k^i is a label sequence with the same length as s^i, which words in s^i are keywords. Let the label k^p of l_k^i represent whether the p-th word in s^i is a keyword, then the value of k^p is as follows:

$$k^p = \begin{cases} 0 \; If \;\; the \, p-th \;\; word \;\; is \;\; not \;\; a \, keyword \\ 1 \; If \;\; the \;\; p-th \, word \, is \, the \, starting \;\, position \, of \;\, the \, keyword \\ 2 \; If \;\; the \;\; p-th \;\; word \, is \, a \, keyword \end{cases} \tag{9}$$

The two tasks in the fine-tuning stage are sentence classification and keyword extraction, so the loss function $L_{fin-tuning}$ in the fine-tuning stage is:

$$L_{fin-tuning} = WL_{CLS} + L_{KE} \tag{10}$$

W represents the weight, L_{CLS} represents the loss function of sentence classification task, and L_{KE} represents the loss function of keyword extraction task.

When the model makes prediction, some rules are applied to sort out the results:

1. Delete keyword prediction results with length greater than 4 words;
2. Delete results with special punctuation marks (such as results with # and @);
3. If the sentence classification layer predicts that the result of a sample is 0, the keyword result of the sample is deleted. For example, input the text “I like apples”. If the output result of the keyword extraction layer is sequence (0, 0, 1), and the output result of the sentence classification layer is 0, the keyword result “apples” will be discarded.
4. Sort the extracted keywords according to the probability weighted average output by the keyword extraction layer;
5. Use the editing distance, the similar results in the keyword candidate list are deleted, and the keyword results with the highest prediction probability are retained.

4 Experiments

This chapter mainly describes the details of the experiment and proves the effectiveness of the method proposed in this paper through comparative experiments.

Section 4.1 mainly talks about the source of experimental data, experimental operation steps and model evaluation methods; Section 4.2 talks about the comparison experiment of model details settings; Section 4.3 talks about the comparison between the method proposed in this paper and other mainstream methods.

4.1 Experimental Data

The experimental data of this paper are mainly abstract parts of academic papers of various disciplines. In the pre-training stage, 100,000 pieces of training set data, 10,000 pieces of dev set data and 10,000 pieces of test set data were obtained from abstracts of academic papers in various fields, such as physics, chemistry and machinery. As shown in Table 1, in the pre-training task, the proportion of negative samples and positive samples in training set, dev set and test set is close to 1:1. Firs, the rules are used to judge whether the sentence contains keywords, and then the sentence samples are labeled with or without keywords according to Eq. (6). There are three public datasets used in the fine-tuning experiment: two keyword extraction datasets in the field of computer science and one keyword dataset in the open field. The details of the three datasets used are as follows:

1. **Krapivin** [18]: This dataset contains 2304 papers published by ACM. The key phrases in this dataset are marked by the author. In our experiments, the first 400 papers of this dataset are selected as test sets, 50 as dev sets and the rest as training sets for fine-tuning training.
2. **NUS** [19]: This dataset contains 211 articles, with keywords marked by volunteers and article authors. There is no official standard for dividing training sets and test sets. In this paper, 150 articles are used as training sets, 20 as dev sets, and the rest as test sets.
3. **Semeval2010** [20]: This dataset contains 284 academic papers in various fields collected by ACM, which is the official dataset of Semeval2010 Task5 Competition. Among them, 100 articles are used as test sets, and the rest are used as training sets and dev sets.

The specific experimental operation steps of the method proposed in this paper are as follows:

1. Remove non-English texts and texts with a long length;
2. Use NLTK [21] tool to divide the article into sentences, and then split each sentence into word sequences;

Table 1 Experimental data analysis of pre-training

	Total samples	Positive samples	Negative samples
Train set	100,000	56,386	43,614
Dev set	10,000	4728	5272
Test set	10,000	5645	4355

3. Use rules to generate tags of sentence classification tasks and keyword extraction tasks, and cleaning the datasets used in pre-training for the second time, so as to keep the integrity of the text and the accuracy of the tags as much as possible;
4. Conduct pre-training, save model parameters after pre-training, and then carry out fine-tuning of the pre-trained model on the text dataset in the computer field;
5. After the fine-tuning, predict the results of the test set, and then use the rules to sort out the predicted results. See Sect. 3.2 for specific rules;
6. After sorting out the prediction results, evaluate the performance of the method and calculate the F1@5 score of the test set.

In the stage of method evaluation, in order to be consistent with the mainstream evaluation methods, the text of the model prediction test set is used to list the top five results of the model prediction probability from high to low and the correct label to calculate the F1@5 score.

4.2 Ablation Experiment

To ensure the validity of the experimental results, the deep learning framework used in this paper is Tensorflow [22]. Adam optimizer is used in both pre-training and fine-tuning training, and the initial learning rate is set to 0.0001. In the fine-tuning stage, if the loss decreases less than 0.01 after five rounds of training, the learning rate becomes 0.2 times the original one. The maximum sequence length is 256, the encoder layer number is 6, and the 8-head attention mechanism is adopted. Therefore, the following comparative experiments are set up in this paper:

1. The transformer model is used to extract keywords, without pre-training stage;
2. The encoder model of transformer is used to train the keyword extraction task directly, without adding sentence classification task and pre-training;
3. The encoder model is trained separately in each dataset, and two tasks of sentence classification and keyword extraction are added in the training stage;
4. After pre-training the model, only the keyword extraction task is added in the fine-tuning training of the model;
5. After pre-training the model, two tasks of sentence classification and keyword extraction are added in the fine-tuning stage of the model.

As shown in Table 2, pre-training was added in Experiment 4 and Experiment 5, and the results were greatly improved compared with Experiment 1 and Experiment 3. After adding multi-tasks, it is found that the performance of the training model in Experiment 5 is obviously better than that in Experiment 4, and the improvement is close to 3%. It is concluded that the performance of the model has been significantly improved after multi-task training.

In the multi-task training stage, this paper selects the sentence classification task, which is similar to the keyword extraction task, and the sentence classification task is

Table 2 The results of control experiment

Experiment	Pre-train	Sentence classification	Keywords extraction	Krapivin	NUS	Semeval
1			√	0.211	0.214	0.208
2			√	0.208	0.215	0.204
3		√	√	0.230	0.240	0.233
4	√		√	0.283	0.325	0.275
5	√	√	√	**0.312**	**0.354**	**0.310**

Input: The standardization of grids based on **web services** has resulted in the need for scalable **web service discovery** mechanisms to be deployed in grids.

Keywords: web services;discovery;
Results of Experiment 4: web services;web service discovery mechanisms;
Results of Experiment 5: web services;discovery;

Fig. 3 Example of predicted results

simpler than the keyword extraction task, so it can help the model to extract keywords more accurately in the learning stage.

As shown in Fig. 3, the input text is: "the standardization of grids based on web services has resumed in the need for scalable web service discovery mechanisms to be deployed in grids". The labeled keywords are "web services" and "discovery". The prediction results of the model obtained in Experiment 4 are "web services" and "web service discovery mechanisms", and the prediction results of the model obtained in Experiment 5 are the same as the labeled results. According to the above results, the model can extract keywords more accurately after multi-task training.

After comparative experiments, it is proved that the method proposed in this paper has been greatly improved after adding the pre-training task. At the same time, after adding multi-task training, it has a better performance in keyword extraction task. To conclude, it proves that the model training method used in this paper is effective.

4.3 Comparative Experiment with Mainstream Model

In order to verify the robustness and accuracy of the proposed method, in this section, the method proposed in this paper is compared with other mainstream methods in the field of keyword extraction.

In this paper, four mainstream and well-performing keyword extraction models are selected: TF-IDF and YAKE algorithms are selected in unsupervised algorithms; KEA and CopyRNN are selected in the supervised algorithm. YAKE and TF-IDF algorithms belong to unsupervised keyword extraction methods based on statistical

Table 3 Results on Krapinvin, NUS and Semeval

Model	Type	Krapivin	NUS	Semeval
TF-IDE	Unsupervised	0.113	0.139	0.120
YAKE	Unsupervised	0.215	0.159	0.151
KEA	Supervised	0.096	0.068	0.027
CopyRNN	Supervised	0.305	0.342	0.291
The proposed mode;	Supervised	**0.312**	**0.354**	**0.310**

features, KEA belongs to supervised keyword extraction method based on statistical features, and CopyRNN model belongs to supervised keyword generation method.

Compared with the other mainstream methods, the experimental results are shown in Table 3, and the model proposed in this paper has the best results in the current dataset. Compared with TF-IDF, YAKE and KEA models, the method proposed in this paper has been greatly improved. But compared with CopyRNN, our method cannot deal well with the situation that keywords are not in the text. In addition, when analyzing the experimental results, we find that CopyRNN's generalization ability is slightly weak, and it is difficult to accurately generate keywords in the open domain. In the computer domain dataset, the model proposed in this paper makes smaller improvement, while in the open domain dataset, the model proposed in this paper makes greater improvement.

Through the above experimental analysis of mainstream models, it can be concluded that although the model proposed in this paper is a supervised training method, it has better generalization ability than mainstream supervised models and can accurately predict keywords in cross-domain datasets or open domain datasets.

5 Conclusion and Future Work

The existing keyword extraction algorithm cannot take into account both the generalization and accuracy of the model. In this paper, a two-step training method is proposed to extract keywords. First, a large number of open domain text data are used to pre-train the model, and then specific domain datasets are used to fine-tune the model. After experimental analysis, after adding pre-training and multi-task, the accuracy and generalization ability of the model can be significantly improved, so that the model can extract keywords more accurately and have better migration.

Compared with the current mainstream keyword extraction model, it is found that the method proposed in this paper has a good performance. However, the method proposed in this paper also has some shortcomings. Compared with the keyword generation model, the method proposed in this paper cannot extract keywords from the text with missing keywords. In the future work, we will focus on keyword extraction algorithms so that the model can accurately extract keywords with less data.

References

1. Jones KS (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc
2. Campos R, Mangaravite V, Pasquali A et al (2018) YAKE! collection-independent automatic keyword extractor. In: European conference on information retrieval. Springer, Cham, pp 806–810
3. Page L, Brin S, Motwani R et al (1999) The pagerank citation ranking: bringing order to the web. Stanford InfoLab
4. Mihalcea R, Textrank TP (2004) Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 404–411
5. Witten IH, Paynter GW, Frank E et al (2005) Kea: practical automated keyphrase extraction. In: Design and usability of digital libraries: case studies in the Asia Pacific. IGI global, pp 129–152
6. Basaldella M, Antolli E, Serra G et al (2018) Bidirectional lstm recurrent neural network for keyphrase extraction. In: Italian research conference on digital libraries. Springer, Cham, pp 180–187
7. Meng R, Zhao S, Han S et al (2017) Deep keyphrase generation. arXiv:1704.06879
8. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
9. Florescu C, Positionrank CC (2017) An unsupervised approach to keyphrase extraction from scholarly documents. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp 1105–1115
10. Sun Y, Qiu H, Zheng Y et al (2020) IFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. IEEE Access 8:10896–10906
11. Zhang Y, Tuo M, Yin Q et al (2020) Keywords extraction with deep neural network model. Neurocomputing 383:113–121
12. Devlin J, Chang M W, Lee K et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
13. Radford A, Wu J, Child R et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
14. Zhang Z, Han X, Liu Z et al (2019) ERNIE: Enhanced language representation with informative entities. arXiv:1905.07129
15. Lan Z, Chen M, Goodman S et al (2019) Albert: a lite bert for self-supervised learning of language representations. arXiv:1909.11942
16. Liu Y, Ott M, Goyal N et al (2019) Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692
17. Yang Z, Dai Z, Yang Y et al (2019) Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems, pp 5754–5764
18. Krapivin M, Autaeu A, Marchese M (2009) Large dataset for keyphrases extraction. University of Trento
19. Nguyen TD, Kan MY (2007) Keyphrase extraction in scientific publications. International conference on Asian digital libraries. Springer, Berlin, Heidelberg, pp 317–326
20. Kim SN, Medelyan O, Kan MY et al (2010) Semeval-2010 task 5: automatic keyphrase extraction from scientific articles. In: Proceedings of the 5th international workshop on semantic evaluation, pp 21–26
21. Loper E, Bird S (2016) NLTK: the natural language toolkit. cs/0205028
22. Abadi M, Barham P, Chen J et al (2016) Tensorflow: a system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16), pp 265–283

Tough Times, Extraordinary Care: A Critical Assessment of Chatbot-Based Digital Mental Healthcare Solutions for Older Persons to Fight Against Pandemics Like COVID-19

Guang Lu, Martin Kubli, Richard Moist, Xiaoxiao Zhang, Nan Li, Ingo Gächter, Thomas Wozniak, and Matthes Fleck

Abstract Pandemics like COVID-19 confine and isolate people due to the necessary physical distancing, lockdown or quarantine countermeasure, which can in particular deteriorate the social environment for older people. Chatbots are a significant part of COVID-19 countermeasures for providing digital mental health services to older people in order to 'keep them alive' during the fight against this pandemic. This study performs a critical assessment of the state of the art and the research gaps regarding the topic and proposes high-level, strategic solutions to help researchers and practitioners speed up the corresponding project planning and mitigate the potential development risks. The design of mental health chatbots is grounded in the fields of psychology and culture and utilizes the most advanced information and communication technologies including conversational interfaces, Artificial Intelligence, Machine Learning and Natural Language Processing. A strong interdisciplinary and intercultural collaboration is essential to better understand the mental hardships of older people during such times and take effective measures to alleviate their suffering. Research in this field provides long-term, sustainable value to our society.

Keywords COVID-19 · Pandemic · Chatbot · Psychology · Mental health · Culture · Older people · Digital technology · Conversational interfaces · Artificial Intelligence · Machine learning · Natural Language Processing

---

G. Lu (✉) · R. Moist · I. Gächter · T. Wozniak · M. Fleck
Institute of Communication and Marketing, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland
e-mail: guang.lu@hslu.ch

M. Kubli
aiaibot, Zurich, Switzerland

X. Zhang
School of Law, Humanities and Sociology, Wuhan University of Technology, Wuhan, China

N. Li
Roche Diagnostics, Rotkreuz, Switzerland

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_68

1 Introduction

1.1 Motivation of This Study

The outbreak and the rapid spread of the COVID-19 pandemic have caused untold distress and suffering for vulnerable groups of people, such as older persons all over the world. The United Nations' policy brief 'The Impact of COVID-19 on Older Persons' [1] highlighted that the fatality rate for people over 80 years old is disproportionately high and was five times the global average as of the end of April 2020. Older people affected by the pandemic are facing a wide range of potential crises, spanning from health issues, isolation and neglect, to social and economic ramifications. To mitigate the worrisome impacts of the COVID-19 pandemic on older persons, the United Nations have called for ambitious countermeasures, both in the short- and long term. One key aspect is mental health care for older people, which can be significantly enhanced with digital solutions and *Artificial Intelligence* (AI) techniques [2–4].

AI-enabled digital mental health care will have great market potential and impact on the healthcare industry but is still in its infancy [5]. So far, a representative research activity has been focused on modelling the spread and prevalence of COVID-19 cases in different geographical regions [6], which is complementary to medical research. These efforts provide policymakers with the basis to formulate and deploy effective measures, which rely on regulations, solidarity and innovative applications of big data technologies [7]. In addition, an approach that focuses more on the well-being of individuals has been demonstrated [8]. Digital techniques, including AI-enabled chatbots, were employed to screen citizens who are facing potential infection risks. Indeed, utilizing digital solutions for diagnosis, surveillance and prevention of COVID-19 has effectively postponed the dissemination of the virus and enhanced the early detection and treatment of the disease [9]. However, current research mostly ignores mental health care for frail, older people, who are most affected by the pandemic and the unprecedented countermeasures and isolation [10]. Although some endeavours exist in general to integrate AI with digital mental health-care [11], the scale of the COVID-19 crisis strongly requires better use of AI and chatbots in order to assess mental health and deliver care for this specific demographic group.

Hence, this paper aims to provide a critical assessment of the state of the art and the research gaps within the field. Potential solutions are proposed on a strategic level regarding implementing chatbot-based digital mental healthcare systems for older persons to fight against pandemics like COVID-19. The authors are aware that research into the direction of digital mental health must consider and adhere to ethical standards and procedures [12]; however, this aspect is out of the scope of this present study due to its psychological and technological focus.

1.2 Chatbots as Part of COVID-19 Countermeasures

Chatbots respond directly to the input they receive, allowing for a very interactive way of taking surveys [13]. They have been found beneficial in information dissemination, symptom monitoring and support behaviour change and mental health in times of the corona crisis [9]. Thus, researchers and practitioners propose to leverage this approach to assess the mental health of people through carefully selected questions and sentiment analyses. An ideal chatbot system will leverage an intuitive, engaging user interface and use advanced *Natural Language Processing* (NLP) techniques. Furthermore, it will be able to expand and adapt its functions continuously to the specific psychological needs of the users. With users' direct feedback, best practices and frameworks can be established and tested. Overall, the development of a chatbot-based mental healthcare system for older people requires seamlessly integrating the existing chatbot technology, advanced data processing and domain knowledge in psychology. Experience gained from the above subjects is also important beyond the current COVID-19 pandemic as isolation and mental health of older people will continue to be a social issue in the coming decades [5].

2 Recent Research and Development of Chatbot-Based Digital Mental Healthcare Solutions

2.1 Discussion on Mental Health, Chatbot and AI

The following scientific studies lay a solid base for developing pandemic-oriented digital mental healthcare solutions. Vaidyam et al. [14] performed a literature review on chatbots and their applications in the fields of mental health and psychiatry. It was reported that there is a considerable potential to apply chatbot technologies in psychiatric treatment. The early feedbacks from the users of the chatbots were also positive regarding leveraging those digital means for maintaining mental health. Miner et al. [15] focused on the fundamental issues of safety, trust and oversight problems that frequently occur while incorporating conversational AI in psychotherapy. It was argued that clinician-AI collaboration is key to the success of an AI-driven psychotherapy. Similarly, Gamble [16] investigated the literature, and discussed the promise of the chatbot technology in assisting people with mental health issues. The author strongly called for industry regulations in order to supervise solving the efficacy, privacy, safety and security challenges accompanying AI and digital mental health.

2.2 *Chatbot Prototype for Digital Mental Health Care*

Several chatbot prototypes have been developed and tested for digital mental health care during the corona pandemic. Denecke et al. [2] designed a mobile-based application called SERMO, which has an integrated chatbot performing *Cognitive Behaviour Therapy* (CBT) [17] on people who look for mental support. The CBT was conducted based on a natural language evaluation of the interactions between users and the chatbot. The outcomes of the user experience questionnaire demonstrated good efficiency, perspicuityand attractiveness for the software. Holt-Quick et al. [3] proposed a novel conversational agent by the name of Headstrong, which is theoretically rooted in CBT. Headstrong deployed activity programs in order to counter the negative impacts of COVID-19 using positive psychology on youth in New Zealand. The initial field tests showed the feasibility of the existing chatbot architecture in promoting mental resilience. Ouerhani et al. [4] implemented a chatbot by the name of COVID-Chatbot that is able to detect the mental stress of people caused by lockdown and quarantine during the pandemic. Advanced NLP techniques were adopted in four functional modules of the chatbot, i.e. data collection, information retrieval, depression detection and action generation. The performance of the chatbot was highlighted in a test of binary sentiment classification of users' emotional states, characterized by an F1 score of approximately 80%.

3 Chatbot-Based Digital Mental Health Care for Older Persons

Using chatbot-based digital solutions to help mitigate or eliminate mental health issues during pandemics like COVID-19 is still inadequate and underfunded for older persons. (1) It is crucial to identify any differences in the struggles and mental health issues between senior citizens in different regions or countries and their responses to digitally delivered mental health care. (2) A prototype for a smart conversational agent for older people, which adheres to ethical standards [12], is still strongly called for, with which conversational interactions can be lively and easily scaled and the mental state of users can be assessed via emotion detection techniques and even influenced using CBT. (3) An overarching framework for guiding best practices in the field is in particular lacking on how to effectively interact with older people with digital channels, such as chatbots. Based on these gaps, this section targets the existing *research questions* (RQs) in the corresponding domains and proposes some strategic solutions. Additional field observations and envisaged project design methods conceived among the authors are also shared as a possible inspiration to researchers and practitioners who are planning projects for the topic.

3.1 RQs on Psychological Level and Solutions

RQ 1: How can one in practice best capture the mental states and psychological needs of distressed, older people, based on the new experience in surveying, digital interaction, and social service to this demographic group?

In order to tackle RQ 1, both quantitative (survey) and qualitative (interview) approaches must be adopted for older persons, aiming to collect valuable data with respect to their mental states and changes during COVID-19. Here, for example, one can target people who are at least 60 years old. In addition, the group of the older persons can be differentiated at least into two sub-groups, i.e. the ones who are living in the community (including those living alone) and the ones who are living in a care home. This is pivotal due to the fundamentally distinct living environments that are encountered by those people, extenuating any potentially affecting factors that might lead to unwanted biases or incomplete perspectives. On one hand, a certain amount of participants needs to be recruited within a given timeframe from each of the sub-groups in the selected region, allowing for a coherent and exhaustive analysis of the surveyed respondents' mental health issues. These surveys should normally consist of carefully designed questionnaires, including self-evaluation, general health, generalized anxiety and depression and insomnia severity evaluation [18]. Besides this, it is essential to investigate the psychological needs of older people being cared for via appropriate oral communication, chatting content, and chatting style. Equally important, one must also test their knowledge in the field of digital mental health and understand their major attitudes towards a chatbot-based mental healthcare system. On the other hand, a sampled set of representatives from each sub-group must be interviewed either face-to-face or via online tools in order to collect added, beneficial data to supplement the survey-based studies. This is necessary as the interviewees' facial expressions and body language can be observed, which leads to a more comprehensive understanding of their psychological needs. With these approaches, researchers seek to not only identify the sources of, e.g. anxiety and loneliness but also put forth advisable responses to stimulate positive emotions and to ease or reverse negative ones. This is key to the framework of digital mental health care for older persons.

RQ 2: What are the major differences in older people's psychological reactions to the pandemic-related hardships between regions or countries of different languages, and what are the resulting implications to the applications of chatbots and digital mental health solutions in these culturally different groups?

Statistically, a pandemic of the COVID-19 scale occurs every one hundred years and can affect the course of nations, their cultures and human behaviour [19]. The potential differences in culture may affect older people's psychological responses to pandemic-caused hardships. Hence, it is suggested that future research perform comparable, survey- and interview-based studies in regions having different languages and potentially different cultural environments, aiming to disclose the latent interdependencies between culture and the mental health of older persons during COVID-19. In effect, one must recruit a similar amount of participants for surveys and perform equivalent interviews within each of the mentioned sub-groups. Researchers are suggested to use methods of text analytics measuring and differentiating carefully older persons' answers as well as psychological stances to the designed questionnaires and interviews in order to entail insights into the desired topics [20, 21]. The strength of mental responses of older people to COVID-19 must be quantified and correlated to the ratings calculated for representative culture values. Moreover, older people's language habits and demands on chatbots' linguistic abilities, communication tone and length of statements must be studied carefully. These approaches allow a balanced investigation of how to develop the functions of chatbots in order to meet the psychological needs of older persons who have distinctive cultural backgrounds.

3.2 RQs on Technological Level and Solutions

RQ 3: How can one assess effectively the human conversational data with chatbots through analysing, e.g. sentiment, communication style and engagement intensity, so that the system will have the most predictive power of the mental health needs of a user?

Arguably, the current research on chatbots and mental health focuses mostly on rule-based architectures instead of *Machine Learning* (ML)-driven intelligent engines [22]. This fact is attributed to the reliability and the security of the rule-based models as well as the only very lately emerging technological breakthroughs in NLP [13]. However, the additional 'survey' data obtained from users' discourses while using the chatbot will permeate and perpetuate into the training of an ML-assisted conversational agent during the second phase of the development. To ensure a valid assessment and usage of the conversational data, one must develop four core functionalities for the chatbot, i.e. conversational interface, emotion detection, suggestion of mindfulness activities for adjusting emotions and prediction of thoughts and feelings for personalized recommendations [2]. In these circumstances, users will be able to either frame their own sentences in a conversation or select easily the answers proposed by the chatbot. Human emotion patterns will be identified directly from their language interactions with the chatbot using both lexicon- [23] and ML-based [4] methods. The lexicon-based method can be treated as the starting point for the development, leaving the ML-enabled approaches being combined subsequently, also incorporating the users' communication style and engagement intensity features. To provide constructive suggestions of mindfulness activities to older persons, it requires comparing and understanding their fundamental cognitive and mental states in the COVID-19 crisis before and after using chatbots. The user experience can be enhanced further significantly if the chatbot can comprehend and predict the users' intentions and feelings via continuously analysing the human–machine conversations, and provide tailored, personalized CBT as guidance [24].

RQ 4: What are the key factors that can increase the acceptance and effectiveness of chatbots and digital mental health solutions for older people?

The solution for this RQ aims to mark a significant and novel contribution to the design aspects of mental health chatbots in order to achieve maximal acceptance and effectiveness. Generally, the quality of health chatbots can be measured in four dimensions, i.e. bot response and conversation, user-bot interaction, bot development and user experience [24]. It is suggested to leverage the user experience surveys in order to render transparent a variety of crucial and specific quality metrics for older people, such that the performance of the chatbot can be inspected and improved continuously. Here, the focal points on the bot response and conversation are the chatbot's personality, response flexibility, flow and length of conversation and dialogue structure [24]. To ensure an optimal user-bot interaction, the chatbot must take the distinctive nature of the older persons cautiously into consideration, and layout pre-determined but tailored answers elaborately for simplicity in interactions. In addition, both the rigid syntax and advanced NLP techniques [13] must be adapted and integrated into the semi-automated chatbot dialogue functions. For older people's convenience in usage, text interface can be complemented with clicking buttons and graphical appearances such as images, gifs or short videos. Overall, one must adopt a solution-oriented approach in development obeying a lean process in order to ensure a spiralling validation process for accomplishing a user-friendly and interactive chatbot interface. This will also have general implications for digital mental health solutions.

4 Conclusion

Developing chatbot-based digital mental healthcare solutions for older persons during pandemics like COVID-19 strongly calls for interdisciplinary and intercultural collaboration. Future research must unite the strengths of all relevant domains. The strategic solutions proposed in this paper could be utilized as the starting point for planning corresponding projects. More valid practices are required in order to analyse the efficacy of assessing the mental health issues of distressed, older persons with chatbots during the COVID-19 pandemic. In addition, the key elements for successful implementation and high adoption rate for the chatbots must be unveiled. Dedicated digital mental healthcare services are also believed to be applicable beyond the COVID-19 pandemic. Finally, the authors advise respective research projects to proactively incorporate ethical considerations already in the early stages of their research endeavours [25].

References

1. United Nations: Policy brief: the impact of COVID-19 on older persons (The Version of May 2020). https://www.paho.org/en/documents/policy-brief-impact-covid-19-older-persons (2020)
2. Denecke K, Vaaheesan S, Arulnathan A (2020) A Mental Health Chatbot for Regulating Emotions (SERMO)—concept and usability test. IEEE Trans Emerg Top Comput 14:1

3. Holt-Quick C, Warren J, Stasiak K, Williams R, Christie G, Hetrick S, Hopkins S, Cargo T, Merry S (2020) A chatbot architecture for promoting youth resilience. https://arxiv.org/abs/2005.07355
4. Ouerhani N, Maalel A, Ghézela HB, Chouri S (2020) Smart ubiquitous chatbot for COVID-19 assistance with deep learning sentiment analysis model during and after quarantine. https://www.researchsquare.com/article/rs-33343/v1
5. Varnai P, Farla K, Glasgow D, Grange S, Romeo S, Simmonds P (2018) AAL market and investment report 2018—A study prepared for the AAL programme (Active and assisted living) by technopolis group. https://www.innosuisse.ch/inno/en/home/go-global/aal.html
6. Wu K, Darcet D, Wang Q, Sornette D (2020) Generalized logistic growth modeling of the COVID-19 outbreak: comparing the dynamics in the 29 provinces in China and in the rest of the world. Nonlinear Dyn 101:1561–1581
7. Hua JL, Shaw R (2020) Corona Virus (COVID-19) "Infodemic" and emerging issues through a data lens: the case of China. Int J Environ Res Public Health 17:2309
8. Martin A, Nateqi J, Gruarin S, Munsch N, Abdarahmane I, Zobel M, Knapp B (2020) An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. Sci Rep 10:19012
9. Tayarani N, Mohammad H (2020) Applications of artificial intelligence in battling against Covid-19: a literature review. Chaos, Solitons & Fractals (In Press)
10. Vigo D, Psych L, Patten S, Pajer K, Krausz M, Taylor S, Rush B, Raviola G, Saxena S, Thornicroft G, Yatham LN (2020) Mental health of communities during the COVID-19 pandemic. Can J Psychiatry 65(10):681–687
11. D'Alfonso S (2020) AI in mental health. Curr Opin Psychol 36:112–117
12. Kelly BD, Drogin E, McSherry B, Donnelly M (2020) Mental health, mental capacity, ethics, and the law in the context of COVID-19 (Coronavirus). Int J Law Psychiatry 73:101632
13. Gao JF, Galley M, Li LH (2019) Neural approaches to conversational AI: question answering, task-oriented dialogues and social chatbots. Now Foundations and Trends. https://arxiv.org/pdf/1809.08267.pdf
14. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB (2019) Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 64(7):456–464
15. Miner AS, Shah N, Bullock KD, Arnow BA, Bailenson J, Hancock J (2019) Key considerations for incorporating conversational AI in psychotherapy. Front Psychiatry 10:746
16. Gamble A (2020) Artificial intelligence and mobile apps for mental healthcare: a social informatics perspective. Aslib J Inf Manag 72(4):509–523
17. Rothbaum BO, Meadows EA, Resick P, Foy DW (2000) Cognitive-behavioral therapy. In: Foa EB, Keane TM, Friedman MJ (eds) Effective treatments for PTSD: practice guidelines from the international society for traumatic stress studies. Guilford Press, pp 320–325
18. Wang M-Y, Liu X-L, Guo H, Fan H-Z, Jiang R-H, Tan S-P (2020) Mental health of middle-aged and elderly population during outbreak of Corona-virus disease 2019. Chin J Mult Organ Dis Elderly 19:241–245
19. Van Bavel JJ, Baicker K, Boggio PS, Capraro V, Cichocka A, Cikara M, Crockett MJ, Crum AJ, Douglas KM, Druckman JN, Drury J, Dube O, Ellemers N, Finkel EJ, Fowler JH, Gelfand M, Han SH, Haslam SA, Jetten J, Kitayama S, Mobbs D, Napper LE, Packer DJ, Pennycook G, Peters E, Petty RE, Rand DG, Reicher SD, Schnall S, Shariff A, Skitka LJ, Smith SS, Sunstein CR, Tabri N, Tucker JA, van der Linden S, van Lang P, Weeden KA, Wohl MJA, Zaki J, Zion SR, Willer R (2020) Using Social and behavioural science to support COVID-19 pandemic response. Nat Human Behav 4:460–471
20. Kjell ONE, Kjell K, Garcia D, Sikström S (2019) Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. Psychol Methods 24(1):92–115
21. Ahmad F, Abbasi A, Li JJ, Dobolyi DG, Netemeyer RG, Clifford GD, Chen H (2020) A deep learning architecture for psychometric natural language processing. ACM Trans Inf Syst 38(1):6

22. Abd-alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M (2019) An overview of the features of chatbots in mental health: a scoping review. Int J Med Inf 132:103978
23. Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development of psychometric properties of LIWC2015. https://repositories.lib.utexas.edu/handle/2152/31333
24. Fadhil A, Schiavo G (2019) Designing for health chatbots. https://arxiv.org/abs/1902.09022
25. Fleck M, Maltseva K (2018) Towards an ethical management framework for artificial intelligence. Academy of Management Global Proceedings (Surrey)

Towards Author Profiling from Modern Standard Arabic Texts: A Review

Asmaa Mansour Khoudja, Mourad Loukam, and Fatma Zohra Belkredim

Abstract One of the most popular topics that many researchers showed interest in recently is the possibility of extracting the personal/demographic characteristics of the authors from their texts such as gender, age, political affiliation, or native language. This is known as Author Profiling (AP). There is a growing interest in AP these last years, in view of its unlimited applications like crime investigations, security, or marketing analysis. In this paper, we review the state of the art about the main author profiling problems, as well as the most used techniques and features, focusing mainly on the Standard Arabic language.

Keywords Author profiling · Text mining · Modern Standard Arabic Language · Artificial intelligence · Machine learning

1 Introduction

In the last previous decades, there has been a significant interest in the enormous increase of information on the web, particularly on social media (Facebook, Twitter, blogs, etc.). But mostly, texts are anonymous and users don't reveal their real identity. This makes the identification of the authors of these anonymous texts a hard task to do. At the same time, extracting relevant information from such texts became crucial nowadays. Thus, this task is considered as one of the most popular topics in the research community, in sight of its unlimited applications like crime investigations, security or marketing analysis, and historic/literary text authorship verification.

A. Mansour Khoudja (✉) · M. Loukam · F. Z. Belkredim
Hassiba Benbouali University of Chlef, Chlef, Algeria
e-mail: a.mansourkhoudja@univ-chlef.dz

M. Loukam
e-mail: m.loukam@univ-chlef.dz

F. Z. Belkredim
e-mail: f.belkredim@univ-chlef.dz

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_69

For instance, Author Profiling helps in solving crime through forensic techniques used to analyze the characteristics of writing styles from documents, signatures, and threatening anonymous letters to identify the person who broke the law or committed the crime. In another perspective, such as marketing, companies would be interested to know the age category of their customers (young or old), in order to adopt new strategies or products to each category. The goal is to increase their profit. AP can even be used to detect a potential violent committer who has the intention of committing targeted violence such as terrorists or sexual predators from social media. AP is also beneficial in plagiarism, where author profiling and plagiarism detection are considered together because these topics are often related to a fraudulent behavior [1].

Many works have been done in AP for many languages, mainly English. For the Modern Standard Arabic language, AP may be considered as a new field where many studies must be initiated.

The Arabic language is the natal tongue of more than 423 million Arabic people without neglecting the fact that it's also the language of approximately 1.9 billion Muslims worldwide which means that the Arabic language is among the five most spoken languages in the world. Unfortunately, there still exists a huge gap when it comes to Arabic language processing in the field of AP, and a considerable quantity of data in Arabic is unstructured, underexploited, and has to be analyzed. However, the lack of resources and the numerous levels of linguistic representation (phonology, orthography, etc.) in Arabic makes it a very challenging but also very appealing language to be studied.

2 Author Profiling

Author profiling is a discipline in which people's writing is analyzed in an attempt to predict the demographic and/or personal characteristics of the author [1] such as his age group, gender, political affiliation, his mother language, and psyche characteristics.

2.1 Author's Gender

In 2003, [2] explored gender identification. In particular, a huge difference was found between male and female when using pronouns and some of the noun modifiers. Where female authors were more involved using these features while male authors were more informational. Also, in 2006, [3] tried to distinguish between male and female authors where they found that texts written by men were predominantly informational, containing, for example, more determiners and adjectives, while texts written by women were more involved, with higher rates of use of pronouns, negation markers, and the present tense. As of (2013–2020), author profiling became a part of a shared task organized at PAN[1] which is a contest for plagiarism analysis and authorship identification. It has become a major contest for the text mining community. The highlight of this competition has been mostly on identifying gender and age. In PAN13, [4] proposed the approach that was based on an emphasis on age and gender discrimination through learning the author's profile. It consisted of four steps, which are: calculating words' occurrences, selecting classes, and creating **ARFF**[2] files. In PAN14, [5] addressed the task with the idea of second-order attributes (building document vectors of the targeted profiles). The proposed idea deepened the analysis incorporating information among texts in the same profile, focusing on sub-profiles. For PAN 2018, [6] proposed a Fusion between text and image using Neural Network, they called it TIFNN, to identify gender. Where they applied deep learning to solve the task using RNN for text classification along with word embeddings and on the other hand, for image classification, they used CNN (ImageNet-based). They tried to combine the two approaches using the fusion component.

2.2 Author's Gender in Arabic

In PAN14, [5] addressed the task with the idea of second-order attributes (building document vectors of the targeted profiles). The proposed idea deepened the analysis incorporating information among texts in the same profile, focusing on sub-profiles.

For PAN18, [6] proposed a Fusion between text and image using Neural Network, they called it TIFNN, to identify gender. Where they applied deep learning to solve the task using RNN for text classification along with word embeddings and on the other hand, for image classification, they used CNN. They tried to combine the two approaches using the fusion component and they obtained overall the best result (81.98%).

2.3 Author's Gender in Arabic

In fact, Arabic is a challenging language because of the numerous levels of linguistic representation (phonology, orthography, morphology, and syntax). In fact, gender is morphologically marked when using Arabic texts. In [7] it was said that one of the related structures with regard to the distinction between male and female authors are foretelling phrases containing a first-person subject. For instance, in the Arabic equivalent of "I am happy", happy is morphologically distinguished as either feminine or masculine. An additional example would be 'منذ صغري و أنا شغوف بالقراءة' *(Since my childhood, I am passionate about reading)*. From these examples, it is clear in

[1] PAN is for plagiarism analysis and authorship identification: https://pan.webis.de/events.html.

[2] An ARFF is an attribute relation file format which is a text file (ASCII) that describes a group of instances that shares a set of attributes. They were built by the University of Waikato.

the adjective that the author's gender is masculine. Moreover, from this example 'حصلت على الباكالوريا من الثانوية الجديدة للبنات'*(I got my baccalaureate degree from the new secondary Girls school.),* we can guess that the author's gender is feminine because she studied at a girls' school. Extracting these features helps to detect gender differences very accurately.

In 2016, [8] aimed to detect the gender of authors' tweets in the Arabic language with the BOW approach accompanied by adding the name of the authors of the tweets and the total number of words for each tweet. In another study, [9] explored the Gender Identification problem for Arabic text classification. One of their aims was to check whether female authors write more emotionally than male authors tend to do. The final outcome showed that the dataset gives no evidence that this is true.

Recently, [10] have focused on gender and educational level classification in Arabic social media.

2.4 Author's Age

The writing style of an author steadily rises or falls as the person grows older. It has been observed that stylistic changes occur with age for every writer. For example, when the author's age increases, the word length increases with it as attested by [11].

Based on the content of the text, it is assumed that 13–17-year-old users write about what is related to youth and teenagers, school subjects, etc. and 23–27-year-old users write more about parties, college, life goals, and travel. While 33–47-year-old users post more about politics, religion, family life, and marriage life [3].

2.5 Author's Age in Arabic

In the Arabic language, the age of the author can be obvious from the text, for example, 'بعد أيام سأحتفل بعيد ميلادي الثلاثين'*In a few days, I'm going to celebrate my thirteenth birthday*). In this case, the age of the author is clear in the text therefore, it is easy to extract it (content-based). As well as in 'منذ عشرين سنة خلت كتبت مقالة في جريدة الأهرام'*(Twenty years ago, I wrote an article in Al Ahram).* Where the author's age is not specified but, obviously, the author is a middle or an elderly aged person (may be over 40 years old). Some other hints, for example, tweets about exams, schools, and university breaks, can indicate that the users were high school or university students [12]. The age classification is quite a complex and difficult task which is mainly based on the text's content.

2.6 Author's Political Affiliation and Ideology

Applying AP to different languages is an essential contribution in facing online terrorism. That's why the political affiliation prediction has gained an enormous interest from a number of researchers. There have been many contributions to political text classification such as [13] where the authors aimed to find members of a certain USA English forum who often used cursing or words of insults and supported the violence used towards a specific category of the society. The idea was to detect messages expressing anger and hostility. In a different study, [14] examined the classification of gender, age, and political affiliation from Swedish politicians' speeches. Different feature selections were performed. The classification of the political affiliation turned out to be the highest accuracy rate after including all forms in the feature sets. On the other hand, the feature sets that were restricted to verbs only or function words resulted in the highest scores for gender prediction and the lowest scores for political affiliation classification.

2.7 Author's Political Affiliation and Ideology in Arabic

While terrorist organizations like Al-Qaeda still exist, analyzing Arabic became an important task to stand against the spread of online terrorism. [13] worked on extracting from online messages linguistic features in order to elaborate stylistic features for terrorist's patterns of communication. They took an existing framework and made some modifications on it to analyze online authors and tried it on a web forum messages for both Arabic and English, and those messages were related to some known extremist groups. Koppel et al. [15] produced a system to classify texts from the Arabic language into organizational affiliation and ideology. They built a corpus of over 100 documents from several Islamic organizations and chose the 1,000 most common words as features. Tests showed that the method achieved a near-perfect accuracy.

2.8 Author's Native Language/Dialects Identification

The identification of native Language is the process of detecting the mother language of an author who writes in another language. Moreover, dialect identification or language variety is determined when we distinguish between similar languages or when the language itself differs in speaking. For example, Brazilian Portuguese versus Portuguese, Arabic (gulf, Egyptian, Maghrebi, etc.), or American and British English. For instance, the fragment of the text below *"Language Variety Identification analyses the behaviour of …*" can be easily attested as British English rather than American, because in American English behaviour is written without "u". Whereas, in the following text "*Native Linguage Identification analyzzes the behavior of …*", "Linguage" and "analyzes" may be spelt the way they are because of the first language interference from the Italian words "*lin*guaggio" and "anali*zz*a"[16]. Basile et al. [17] tried to create a single model for all language varieties that they were given in PAN17. They used word-unigrams and character-grams as features and other additional features that included POS tags, Twitter handle, and geographic entities, that worsened, rather than improved the performance. For Modeling, a simple single model was built that classifies the gender and the language variety of every user at the same time. Kodiyan et al. [18] solved the same problem with a deep learning approach without extracting features manually, which had proven recently to be very successful in natural language processing. The model that they built contained a bidirectional Recurrent Neural Network used with a Gated Recurrent Unit (GRU) with an Attention Mechanism. The model obtained an average accuracy for every language with 75,31% in gender and 85,22% in language variety classification.

2.9 Author's Native Language/Dialects Identification in Arabic

When it comes to Arabic, native language and dialects identification catch the attention of researchers due to its crucial importance in detecting threatening messages, improving online security gaps, and detecting potential terrorist attacks. However, since there is a variety of Arabic dialects, this task becomes more challenging as it requires representative annotated datasets to be available for each dialect. This is why only a few studies have been conducted on this problem, [12] created a large manually annotated Arabic dataset from various social media sources covering 16 Arabic countries and 11 dialectal regions. Where the main aim was to design a resource for developing AP tools.

3 Features

When it comes to author profiling, finding the best textual feature sets is the main aim to better perform text classification. Campbell and Pennebaker [19] stated that there are two types of features that have been considered to be the most important features in text classification: content-based features and style-based features. Since authors write in different styles and topics and as the years go by, style and topics of authors change as they get older. Therefore, these two features play an essential role in author profiling.

3.1 Style-Based Features

Style-based features or style markers are features of writing style that makes the task of author profiling easier and more advanced [13]. Four main categories of features can be highlighted: lexical, syntactic, structural, and semantic.

Lexical features are divided into two groups word or character-based features. The lexical features based on words have properties like the number of the total words, the number of words for every sentence, the distribution of the word length, and the abundance of the vocabulary. While character-based lexical features contain character types (letters, numbers …), the total number of letters for every sentence or for every word, etc. [13]. For instance, as mentioned before [7] worked on developing a system using text attribution tools. This system had been developed with the aim of extracting the authors' profiles from Arabic and English emails. For each document, a feature vector was calculated (Number of words/number of sentences) and other features such as the frequency of punctuation, frequency of characters, or word length. Also [20] addressed the task of author profiling at PAN competition in 2016, using stylistic and lexical features extraction (n-grams, punctuation features).

Syntactic features indicate the patterns used to compose statements. They contain characteristics like function words. They are very useful to differentiate between authors. For example, it might seem that using the word thus or hence would cause no noticeable difference but actually it can cause a huge stylistic difference [13]. Argamon et al. [2] identified some simple classes of lexical features and some syntactic features (less than 50). Where a huge difference was found between authors' gender when using pronouns and some of the noun modifiers. Where female authors were more involved using these features. On the other hand, male authors were more informational.

Structural features contain text's organization and layout properties, such as the number of paragraphs, numbered points, spacing, greetings, and signatures [13]. Structural features were used in the work of [21] where they described an approach for author profiling task using 311 and 476 features in which structural features were among them in the form of number of conversations, paragraphs, sentences, and words per sentences, number of special characters, etc.

Semantic features demand further analysis in linguistic and contain characteristics like synonyms, nouns' number, the aspect of verbs, and the tense of verbs [22]. For example, [23] presented a computational model able to detect sarcasm in the social network Twitter using seven groups of features, among them synonyms (common vs. rare synonyms use) and ambiguity. Some of these features were designed to predict variance and surprisingness.

Morphological features, spelling errors in an unedited text, or grammatical errors can be seen as other types of features [24]. There exist other style-based features: frequencies, POS (Parts of Speech), Twitter-specific elements, and readability measures. Where we can see these style-based features used in the work of [17] in which they tried to build one model for gender, language classification, and language varieties using word-unigrams and character-grams as features and other additional features that included POS tags, Twitter handle, and geographic entities. But in this case, these additional features worsened, rather than improved the performance.

3.2 Content-Based Features

Content-based features seek to classify the text's content. Among many techniques that could be used in the content-based approach to classify the content of a text would be n-grams (splitting a sentence to understand more closely the context and topics of the given sentence). Due to the existence of dissimilarity in topics between different age groups and gender, using n-grams, with different values of n, it can be possible to calculate the different frequencies of topics mentioned by the different author groups [25]. For example, in blogs, topics such as football and computers are used more by male authors while words like shopping and husband are used by female authors. Schler et al. [3] used features that are based both on style and content. Where, for the content-based, the features were simple words based on content, along with special words extracted from LIWC [26] categories, and they concluded that females use "involvedness" in their writings, while male tend to use "information" which are terms used by [27]. Also, men were concerned more about technology and politics. And teenagers wrote more about friends and mood swings. In the end, they concluded that the same features that discriminate between gender in writing style also discriminate between age in writing style. And they figured out that features based on content were slightly better than features based on style in age prediction; however, in gender prediction, combining both features proved to be more useful.

Other content-based features have been also explored. Among them, we can cite: sentiment words, dictionary words, term vectors, informal words, and bag of words.

4 Conclusion

This paper reviewed the state of the art of Author Profiling particularly in the standard Arabic language. The main aim was to show what had been done in this field before and highlight what should be done in the upcoming studies to fill in the huge existing gaps especially for the Arabic language in the field of AP.

References

1. Oakes MP (2019) Author profiling and related applications, no February 2019
2. Argamon S, Koppel M, Fine J, Shimoni AR (2003) Gender, genre, and writing style in formal written texts. Text Talk 23(3):321–346

3. SchlerJ, Koppel M, Argamon S, Pennebaker J (2006) Effects of age and gender on blogging. In Proceedings of 2006 AAAI spring symposium on computational approaches for analyzing weblogs
4. MechtiS, Jaoua M, Belguith LH, Faiz R (2013) Author profiling using style-based features. Notebook papers CLEF2
5. López-MonroyAP, Montes-y-Gómez M, Escalante HJ, Pineda LV (2014) Using intra-profile information for author profiling. CLEF (Working Notes), pp 1116–1120
6. Takahashi T, Tahara T, Nagatani K, Miura Y, Taniguchi T, Ohkuma T (2018) Text and image synergy with feature cross technique for gender identification. Working notes papers CLEF
7. Estival D, Gaustad T, Pham SB, Radford W, Hutchinson B (2007) Tat: an author profiling tool with application to arabic emails In: Proceedings of the Australasian language technology workshop 2007, pp 21–30
8. Alsukhni E (2016) Investigating the Use of machine learning algorithms in detecting gender of the Arabic Tweet author, vol 7, no 7, p. 319–328
9. AlsmearatK, Shehab M, Al-Ayyoub M, Al-Shalabi R, Kanaan G (2016) Emotion analysis of Arabic articles and its impact on identifying the author's gender. In: Proceedings of IEEE/ACS international conference computer system application AICCSA, vol 2016, July, 2016
10. Al-Ghadir AI, Azmi AM (2019) A study of arabic social media users—posting behavior and author's gender prediction. Cognit Comput 11(1):71–86
11. Can F, Patton JM (2004) Change of writing style with time. Comput Hum 38(1):61–82
12. ZaghouaniW, Charfi A (2018) Guidelines and annotation framework for Arabic author profiling. arXiv1808.07678
13. Abbasi A, Chen H (2005) Applying authorship analysis to extremist-group Web forum messages. IEEE Intell Syst 20(5):67–75
14. Dahllöf M (2012) Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability. Lit Linguist Comput 27(2):139–153
15. KoppelM, Akiva N, Alshech E, Bar K (2009) Automatically classifying documents by ideological and organizational affiliation. In: 2009 IEEE international conference on intelligence and security informatics, pp 176–178
16. Franco-Salvador M, Kondrak G, Rosso P (2017) Bridging the Native Language and Language Variety Identification Tasks. Procedia Comput Sci 112:1554–1561
17. BasileA, Dwyer G, Medvedeva M, Rawee J, Haagsma H, Nissim M (2017) N-gram: new groningen author-profiling model. arXiv1707.03764
18. KodiyanD, Hardegger F, Neuhaus S, Cieliebak M (1866) Author Profiling with bidirectional rnns using attention with grus: notebook for PAN at CLEF 2017. In: CLEF 2017 conference and labs of the evaluation forum, Dublin, Ireland, 11–14 September 2017, vol 1866
19. Campbell RS, Pennebaker JW (2003) The secret life of pronouns: flexibility in writing style and physical health. Psychol Sci 14(1):60–65
20. ModaresiP, Liebeck M, Conrad S (2016) Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. In: CLEF (Working Notes), pp 970–977
21. Meina M et al. (2013) Ensemble-based classification for author profiling using various features. Notebook papers CLEF
22. Stamatatos E (2009) A survey of modern authorship attribution methods, vol 60, no 3, pp 538–556
23. BarbieriF, Saggion H, Ronzano F (2014) Modelling sarcasm in Twitter, a novel approach, pp 50–58
24. ArgamonS, Koppel M, Pennebaker JW, Schler J (2006) Automatically profiling the author of an anonymous text, no Sebastiani 2002
25. HolbE (2019) Using author profiling to determine the age group of an author, no June
26. Pennebaker JW, Francis ME, Booth RJ (2001) Linguistic inquiry and word count: LIWC 2001. Mahw Lawrence Erlbaum Assoc 71(2001):2001
27. Biber D (1995) Dimensions of register variation: a cross-linguistic comparison. Cambridge University Press

Business Value from ICT Investments: A Configurational Exploratory Perspective

Rodrigo Bogarin

Abstract For the last 30 years, there have been numerous attempts to explain the impact on organizations from investments in ICT. However, the results have not been conclusive. Factors considered in isolation, as well as the direct impact in financial performance, are replaced by an indirect impact and a systemic view in the business value generation. From a configurational perspective that considers IT Governance, Absorptive and Organizational Agility capabilities, a different definition of Business Value from IT, and an exploratory approach based on case comparison this study aims to open a new venue to identify a recipe of how business value follows to ICT investments. By understanding how this value is generated organizations can also execute complementary activities to get additional benefits. This study included medium and large private organizations in Costa Rica. As an important finding, it is important to mention that benefits from ICT investments are mainly based on internal factors under the organization's control and closely related to strategy whether this is to be a follower or a trailblazer.

Keywords ICT business value · IT governance · Organizational agility · Absorptive capability

1 Introduction

The IT Business Value (ITBV) has been and continues to be a topic of interest on the research in the areas of Information Systems and Organizational Strategy [1, 16, 52]. This interest is not only from scholars but also from managers that on one hand are aware of the need to invest more in ICT but at the same time are doubtful if this one is their best investment option [50]. It is expected that the level of investment increases as the Fourth Industrial Revolution settles, referring to the intensive use of ICT to support economical activities as coined in 2016 by Klaus Schwab from the World Economic Forum. At the same time, the COVID-19 has also made managers

R. Bogarin (✉)
Costa Rica Institute of Technology, Cartago 30101, Costa Rica
e-mail: rbogarin@itcr.ac.cr

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_70

re-evaluate the strategic role ICT has performed in organizations and it is most likely to be emphasized in the upcoming years as they move faster in Digital Transformation processes.

Even though there has been a lot of research being conducted on this topic with a peak number of publications in 2000 [51] there are still some unanswered questions and the topic is only partially explained [2] letting arguments of the nineties that claimed there is no value generated from IT still with some support [30, 52, 55, 60].

Meanwhile, studies have shown some contradictory results with respect to the impact of productivity of IT [13, 27, 29, 30, 47] scholars agreed on the need for evolution of methods and techniques taking on this topic once thought as simple it is now found to be more complex than expected [14, 30, 55]

When trying to understand the benefits from ICT according to [42] there are five principles to consider: (1) value of ICT is not inherent to them, (2) benefits from ICT are perceived when people are able to perform their jobs in a different way, (3) benefits are to be recognized and called out by managers and end users instead of IT managers, (4) all IT projects produce results but not all of them are benefits, and (5) benefits should be actively managed.

Due to the lack of conclusive results of the impact of ICT investments, difficulties when modeling benefits [22, 52], difficulties when modeling and interpreting the interaction of three or more factors [19], and a general acceptance for new perspectives [30], this study proposes a different approach based on identifying configurations of causal conditions in the generation of ITBV instead of the traditional identification of independent variables and their net effect. Initially, we consider configurations based on IT Governance, Organizational Agility, and Absorptive capability and then explore additional conditions needed or sufficient to generate business value featured as benefits of different types got from the ICT investments.

This paper is divided into five sections. Section 2 reviews the theoretical lens applied to this topic since the nineties and defines the initial conditions in this study. Section 3 describes the method and technique used here. Section 4 presents the results which are discussed in Sect. 5. Finally, we discussed conclusion, limitations, and future work in Sect. 6.

2 Literature Review

The impact of ICT has been of interest to scholars. According to [15] more than four hundred publications are related to this topic. In the last few years, studies have taken a different approach [48]. Based on the dominant perspectives in these studies is too possible to identify two periods of approximately fifteen years each.

2.1 1990–2005 ICT and Direct Impact on Financial Performance

Studies in this period are mainly characterized by the financial performance as a main dependent variable, by the theories used (mainly RBV), and the use of regression and case techniques. It was common to find the dependent variable to be the financial impact and was commonly measured using Return on Equity (ROE) or Return on Assets (ROA) [52]. Not being able to explain the real relationship between IT and productivity allowed to coin the Productivity Paradox [9] which is claimed to be resolved. Authors not only identify a positive relationship between IT and productivity but also found that IT investments push for complementary activities including training, decision support systems, and process redesign. Studies also recognize that by itself IT investments are not enough to generate an impact on the financial performance [38].

A classic study [7] established the relationship between the ICT capability and the organizational performance and intangible assets. These results are coincident with the ones got by when this study was reproduced with a few procedural changes [51].

From the point of view of investment evaluation, there are several proposals including methods multicriteria [8], the Value–Advantage–Risk framework [33], and the real-option approach [4, 18]. After a study that starts from the premise that ICT investments are different from other capital investments, it is found that no significant differences in the methods but instead on the financial analysis and the difficulty to monetize intangible benefits [3]. As described in [5] there are five failure factors associated to ICT investment including (1) who oversees the business case, (2) lack of rigor when evaluating and accepting, (3) lack of rigor when identifying costs and benefits, (4) risk identification, and (5) lack of updates and reviews to documents. Lack of other complementary investments for example related to organizational change can also be added to the previous list as noted by [54].

A meta-analysis [31] conducted late in this period summarized not only the interest of scholars for this topic but also tried to explain the variability in the results according to sample size, techniques used, choice of dependent variable, industry, and information sources making clear that consensus is far from being reached.

2.2 2005-Actual ICT and Complex Relation to Business Value

After 2005, there seems to be more awareness of the importance of ICT for organizations. In a study [22], CIOs point out the relevance of the alignment of IT and business strategies. Concepts from the first period like Productivity Paradox continue to be questioned as well as the use or abuse of the RBV in this context [53]. A new meta-analysis [34] shows an expansion of the scope of business value to include the development of new organizational capabilities and alignment with organizational strategy as new ways to measure the impact of the ICT. Additionally, ICT capability is identified as an important component for innovation and efficiency in the organization [68].

Theoretical lens is expanded to include not only the RBV but also the Contingency Theory and Dynamic Capabilities as suggested in [32, 64, 65]. Same happens with acceptance of the concept of synergies of ICT with other factors not only internal but also external as suggested in [11, 12, 40, 49], and that the results of these interactions are highly contextual and very difficult to be generalized [21].

Dependent variable of financial performance is questioned as not able to capture the impacts of ICT since it is normally based on financial measures of historical information [41] and some of these impacts are not only difficult to clearly identify but also not necessarily perceived immediately. Other studies also include alternative ways to measure ICT impact [40, 56].

The last 5 years suggest a new interest in the strategical and transformational benefits of new technologies under the umbrella of Digital Transformation processes which assume that definitions, concepts, and methods are yet to be described [28, 39, 48].

2.3 Configurational Perspective

The concept of "fit" in strategic information technology management research [6] suggests six different perspectives: moderation, mediation, matching as bivariate relationships meanwhile considers covariation, profile deviation, and gestalts as multivariate alternatives and considered the latter to provide richer and fuller explanations than the former. Also, considered gestalts to be better suited to theory building and characterized them as "feasible sets of internally consistent and equally effective configurations" [62, p. 432].

A configurational perspective offers an alternative to the dominant perspectives in the ICT research: variance and process. While the variance perspective looks to establish relations between predictors and outcomes, it assumes each predictor has an independent effect on the outcome also fails to deal with situations where an outcome is a result of several values of the predictor. On the other hand, process perspective has limitations to recognize holistic or systemic where there are several predictors interacting [24].

The objective of Configuration Theory is to identify patterns and combinations of causal conditions and describe how its synergetic effects produce an outcome [19]. Also, part of this theory is the concept of asymmetrical causality where elements influencing the presence of an outcome are not necessarily related to its absence. Hypotheses are set in terms of sufficiency, necessity, both, and none which use comparative configural methods to test. When comparing these methods with regression analysis there are three main differences: math theories, hypothesis types, and concepts of causality [61].

Assuming, ICT investment level will continue to increase as companies go into Digital Transformation processes, we consider an initial configuration based on Agility, Absorptive Capability, and IT Governance, taking into account those acting together and consequences of it as the Alignment Paradox [56].

IT Governance. As part of the Corporate Governance, IT Governance refers to process, structures, and mechanisms that allow business and IT managers to support IT and business alignment and business value creation through IT-based investments [26]. Its importance and benefits have been widely discussed [63, 66, 68].

Absorptive Capacity. Defined as "a set of process and routines that allow firms to acquire, assimilate, transform, and exploit knowledge to produce a dynamic capacity" [35, p.382]. It has appeared in different studies with different conceptions and measures [10, 17, 70].

Organizational Agility. Refers to firm's ability to respond to changes with innovative and fast answers. Extend both the flexibility usually associate to IT-based process and the strategic one associated with unstructured changes [36]. Alignment Paradox has been mentioned when discussing IT and firm performance only to conclude that organizational agility is a mediator between the two [33, 56, 59].

Business Value of Information Technology. It has not traditionally been the consensus of what really is the business value of Information Technology [57]. However, in the last few years, there is a tendency to characterize the business value in terms of benefits informational and transactional. Later, extended to include strategic ones and finally to include transformational benefits [25].

3 Method

This study followed an exploratory sequential approach divided into two phases. The first one, based on a comparative analysis of 30 private-owned medium and large firms from Costa Rica and included the application of fuzzy set Qualitative Comparative Analysis (*fsQCA*) [44] considering Organizational Agility, Absorptive Capability, and IT Governance as causal conditions and IT Business Value as the outcome. Second one, conducted additional *fsQCA* analysis considering a new causal condition: Management Commitments [44] and triangulation and complementary analysis from additional interviews.

As described in [45] *fsQCA* is a method that sets itself as a bridge between the traditional qualitative and quantitative approaches being especially applicable in situations where samples are between 12 and 50. But more important fsQCA as a variant of QCA uses Boolean logic, theoretical sets, and fuzzy logic to establish relationships of sufficiency, necessity, both, or none between causal conditions and an outcome. Cases are evaluated for their membership degree to the different theoretical sets (causal conditions) and configurations (combinations of causal conditions created from set operations) are derived by applying simplification rules. A solution is then identified that shows over the threshold metrics of *consistency* and *coverage*. See [46] for a step-by-step example.

Comparative methods differ from regression analysis in three important features [61]: the math theories they are based on (Boolean vs. Linear Algebra), the type of hypothesis they are meant to (Implication vs. Covariation), and the concept of Causal Complexity (Conjunction vs. Interaction) and although very heated discussions have and continue to happen about the merits of each one, both approaches have limits and possibilities.

In this study, *fsQCA* offered not only the possibility for a new perspective to the IT Business Value but allowed an iterative process to theory building. Sources for each case included managers from both IT and business and the use of instruments based on [25, 35–37, 43, 67]. Meanwhile, generalization of results is based on the concept of analytical generalization as described in [69].

4 Results

Analysis using *fsQCA* suggests that before considering all causal conditions at once, you should consider the relationship between each causal condition and the outcome. For this analysis IT Governance showed values of $consistency\, X \leq Y(0.79)\; y\; X \geq Y(0.87)$ is interpreted as *necessity*. For Agility, a consistency of $X \leq Y(0.93)\; y\; X \geq Y(0.8)$ that is interpreted as *sufficiency* and for Absorptive Capability $consistency\, of\; X \leq Y(0.89)\; y\; X \geq Y(0.77)$ which also shows as *sufficiency*.

Next steps include filter configurations that did not show any cases or did not meet the defined threshold of minimum cases of 2 or minimum consistency of 0.89 (remainders). Finally, the researcher suggested how each condition should be taken into account when building the intermediate solution. This solution refers to a simpler than the complex that takes into account remainders based on researcher suggestion and not all of them as it will be in the parsimonious solution.

The complex solution using [46] is shown in Fig. 1. The Venn diagram using QCA v3.10 for R by A. Dusa is shown in Fig. 2.

Second iteration of fsQCA included a new causal condition Management Commitments [43] with a new complex solution shown in Fig. 3.

A Venn diagram for the new solution is in Fig. 4.

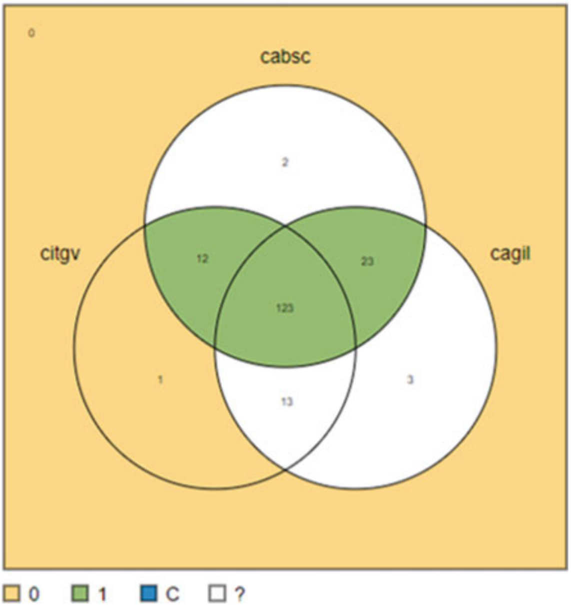
Even though the new causal condition inclusion took solution coverage to a higher level, the effect was the opposite in the solution coverage which decreased from 0.77 to 0.65, so we decided to continue looking for additional causal condition from in-depth interviews. From those and after an analysis guided by [23], a new causal condition: Company Culture to ICT with four dimensions: Organizational Commitment, Leadership, Organizational Change Management and Communication was identified.

```
Model: citbv = f(cagil, cabsc, citgv)
Algorithm: Quine-McCluskey

--- COMPLEX SOLUTION ---
frequency cutoff: 2
consistency cutoff: 0.877016
                      raw         unique
                    coverage     coverage    consistency
                   ----------   ----------   ----------
cagil*cabsc        0.688017     0.0929753    0.973684
cabsc*citgv        0.680785     0.0857438    0.927516
solution coverage: 0.77376
solution consistency: 0.920713
```

Fig. 1 Complex solution for ITBV

Fig. 2 Venn diagram for ITBV




```
Model: citbv = f(ccomp, cagil, cabsc, citgv)
Algorithm: Quine-McCluskey

--- COMPLEX SOLUTION ---
frequency cutoff: 2
consistency cutoff: 0.966746
                         raw        unique
                       coverage    coverage    consistency
                      ----------  ----------   ----------
ccomp*cabsc*citgv     0.657025    0.657025     0.964367
solution coverage: 0.657025
solution consistency: 0.964367
```

Fig. 3 Complex solution for ITBV with new causal condition MC

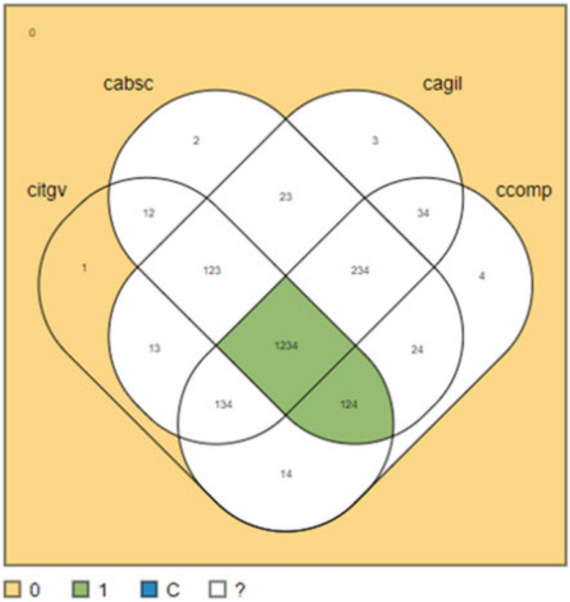


Fig. 4 Venn diagram for ITBV with new causal condition

5 Discussion

Figure 1 shows a compound solution with two terms with Absorptive Capability as the common one. This is interpreted by saying that there are two ways benefits from ICT investments can be obtained. First from the strategic and careful alignment of IT applications built from the organizational knowledge and the second one when this knowledge is embedded in fast and innovative responses to changes no matter where they came from. Only two of the three causal conditions are present in the final solution which might show some support for the Alignment Paradox [33, 58].

The second iteration included a new causal condition: Management Commitments that included management commitment to Strategical Choices, Platform Development, Working Intelligently, and Action-Oriented Evaluation [44]. When considered this causal condition jointly with the previous three conditions solution showed as a one-term condition formed by the conjunction of Management Commitment, Absorptive Capability and IT Governance and excluded Organizational Agility which on one hand discouraged the Alignment Paradox effect but on the other hand contributed to a decrease on the coverage of the solution. A set coincidence analysis between Organizational Agility and Management Commitments is 0.73 which explains most of the decrease in the solution coverage.

If instead of considering all four causal conditions only Management Commitments and Organizational Agility are considered the coverage will increase to 0.87 and the consistency to .95 which would be suggesting that new configurations can be based on either of these two causal conditions.

Both iterations suggested that there is room for refining the configurations looking for an extended coverage of the ITBV. From in-deep interviews with experts, a new causal condition (Company Culture for ICT) was identified, and it is characterized by a technology-savvy leadership, a commitment to IT beyond management, an organizational culture that welcomes change and promotes company-wide communications. This new causal condition would most probably be acting in conjunction with the original ones, as seems to represent a superset of Management Commitments.

It is not until 2020 that ICT investments were not only challenged but also valued as a survival table during the pandemic. Firms realized, some the hard way, that even after had invested millions of dollars in ICT those investments have not really made it to the core of the business, showing a low level of Absorptive Capability. As always, for others pandemic opened opportunities that depending on their strategy and its ability to respond (Organizational Agility) drove to important growth.

6 Conclusions, Limitations, and Future Work

Nowadays, with events like the pandemic or the upcoming 4IR, firms need to make sure that their ICT investments are really paying off. This study proposed a new one based on a configurational perspective and identified initially three causal conditions and how they jointly relate to ITBV. Looking to expand the coverage of the explanation, new conditions were identified that show Commitments (from management and firm wide) as required. Identified causal conditions are mostly internal and associated to capabilities that can be increased by management decisions and complementary activities which leads to think that benefits obtained from ICT are basically under firm control. Finally, management understands that some benefits will not be perceived in purely financial terms and not always in the immediate future. There are practical implications from this study that shed some light onsome organizational capabilities important for technology-driven initiatives. For scholars, opened a new venue by proposing a whole new perspective and expand the scope of application of the QCA method previously mainly applied to other social sciences.

Every study has limitations and this one is no exception. No doubt most important limitation would be related to results generalization which can be addressed in several ways: by replicating the study with a large-N looking for a statistical generalization, or by extending this work as part of a mixed methods setup.

This study can be expanded in several ways. By considering new causal conditions for additional coverage, by testing the implication hypothesis while looking for a statistical generalization [20], by replicating the study in other countries, by conducting longitudinal studies that better reflect medium to long term benefits from ICT investments, and the role ICT will have in a post-pandemic world.

References

1. Ashrafi R, Mueller J (2015) Delineating IT resources and capabilities to obtain competitive advantage and improve firm performance. Inf Syst Manag 32(1):15–38
2. Baker J, Song J, Jones D (2008) Refining the IT business value model: evidence from a longitudinal investigation of healthcare firms. In: ICIS 2008 proceedings
3. Ballantine J, Stray S (1999) Information systems and other capital investments: evaluation practices compared. LogisTIC Inf Manag 12(1/2):78–93
4. Benaroch M (2000) Justifying electronic banking network expansion using real options analysis. MIS Q 24:197
5. Bennington P, Baccarini D (2004) Project benefits management in IT projects: an Australian perspective. Project Manag J
6. Bergeron F, Raymond L, Rivard S (2001) Fit in strategic information technology management research: an empirical comparison of perspectives. Omega 29(2):125–142
7. Bharadwaj A (2000) A resource-based perspective on information technology capability and firm performance: an empirical investigation. Manag Inf Syst Q 24(1):169–1969
8. Borenstein D, Betencourt P (2005) A multi-criteria model for the justification of IT investments. Infor 43(1):1–21
9. Brynjolfsson E (1993) The productivity paradox of information technology. Commun ACM 36(12):77
10. Camison C, Fores B (2010) Knowledge absorptive capacity: new insights for its conceptualization and measurement. J Bus Res 63(7):707–715
11. Cao G, Duan Y (2013) Systemic capabilities as emergent properties: the source of IT business value. In: 7th annual IEEE international systems conference, proceedings, SysCon 2013
12. Cao G, Wiengarten F, Humphreys P (2011) Towards a contingency resource-based view of IT business value. Syst Pract Action Res 24(1):85–106
13. Carr NG (2003) IT does not matter. Harvard Business School June(1):12
14. Chae H, Koh C, Prybutok V (2014) Information technology capability and firm performance: contradictory findings and their possible causes. MI 38(1):305–326
15. Daulatkar S, Sangle P (2017) Causality in information technology business value: a review. Bus Process Manag J 21(3):482–516
16. Dehning B, Richardson VJ, Urbaczewski A, Wells JD (2004) Reexamining the value relevance of e-commerce initiatives. J Manag Inf Syst 21(1):55–82
17. Flatten T, Engelen A, Zahra S, Brettel M (2011) A measure of absorptive capacity: scale development and validation. Eur Manag J 29(2):98–116
18. Fichman R (2005) Beyond valuation: "options thinking" in IT project management. Calif Manag Rev 47(2):74–96
19. Fiss P (2007) A set-theoretical approach to organizational configurations. Acad Manag Rev 32(4):1180–1198
20. Fiss P, Sharapov D, Cronqvist L (2013) Opposites Attract? Opportunities and challenges for integrating large-N QCA and econometric analysis. Polit Res Q 66(1):191–198
21. Gelhard C, von Delft S, Gudergan S (2016) Heterogeneity in dynamic capability configurations: equifinality and strategic performance. J Bus Res 69(11):5272–5279
22. Gerow J, Thatcher J, Grover V (2015) Six types of IT-business strategic alignment: an investigation of the constructs and their measurement. Eur J Inf Syst 24(5):465–491
23. Goulding C (2002) Grounded theory: a practical guide for management business and market researchers. SAGE Publications, USA
24. Gregor S (2006) The nature of theory in information systems of theory in information systems. MIS Q 30(3):611–642
25. Gregor S, Martin M, Fernandez W, Stern S, Vitale M (2006) The transformational dimension in the realization of business value from information technology. J Strat Inf Syst 15(2006):249–270

26. Herz T, Hamel F, Uebernickel F, Brenner W (2011) IT governance mechanisms in multi-sourcing: a business group perspective. In: Proceedings of the annual Hawaii international conference on system sciences, pp 5033–5042
27. Hitt L, Brynjolfsson E, Walsham G (1994) The three faces of IT value: theory and evidence. In: Proceedings of the 15th international conference on information systems (ICIS), pp 263–277
28. Henriette E, Feki M, Boughzala I (2015) The shape of digital transformation: a systematic literature review. In: Ninth mediterranean conference on information systems, Greece, pp 431–443
29. Im K, Dow K, Grover V (2001) A reexamination of IT investment and the market value of the firm: an event study methodology. Inf Syst Res 12(1):103–117
30. Khallaf A, Omran MO, Zakaria T (2017) Explaining the inconsistent results of the impact of information technology investments on firm performance: a longitudinal analysis. J Account Organ Chang 13(3):359–380
31. Kohli R, Devaraj S (2003) A meta-analysis of structural variable in firm-level empirical research. Inf Syst Res 14(2):127–145
32. Li T, Chan Y (2019) Dynamic information technology capability: concept definition and framework. J Strat Inf Syst 28(4)
33. Liang H, Wang N, Xue Y, Ge S (2017) Unraveling de alignment paradox: how does business-IT alignment shape organizational agility? INFORMS 28(4):863–879
34. Lim J, Dehning B, Richardson V, Smith R (2011) A meta-analysis of the effects of IT investment on firm financial performance. J Inf Syst 25(2):145–169
35. Limaj E, Bernroider E, Choudrie J (2016) The impact of social information system governance, utilization, and capabilities on absorptive capacity and innovation: a case of Austrian SMEs. Inf Manag 53(3):380–397
36. Lu Y, Ramamurthy K (2011) Understanding the link between information and organization agility: an empirical examination. MIS Q 35(4):931–954
37. Mirani R, Lederer A (1998) An instrument for assessing the organizational benefits of IS projects. Decis Sci 29(4):803–838
38. Melville N, Kraemer K, Gurbaxani V (2004) Review: information technology and organizational performance: an integrative model of it business value. MIS Q 28(2):283–322
39. Morakanyane R, Grace A, O´Reilly P (2017) Conceptualizing digital transformation in bussiness organization: a systematic review of literature. In: BLED 2017 proceedings 21
40. Nevo S, Wade M (2011) Firm-level benefits of IT-enabled resources: a conceptual extension and an empirical assessment. J Strat Inf Syst 20(4):403–418
41. Ong S, Chen P (2014) The effects of IT: from performance to value. Ind Manag Data Syst 114(1):70–85
42. Peppard J, Ward J, Daniel E (2007) Managing the realization of business benefits from IT investments. MIS Q Exec 6(1):1–11
43. Peteraf M, Barney J (2003) Unraveling the resource-based tangle. Manag Decis Econ 24(4):309–323
44. Quaadgras A, Weill P, Ross J (2014) Management commitments that maximize impact from IT. J Inf Technol 29(2014):114–127
45. Ragin C (2008) Redesigning social inquiry: fuzzy sets and beyond. University of Chicago Press, USA
46. Ragin C, Sean D (2017) Fuzzy-set/qualitative comparative analysis 3.0. Irvine, California: Department of Sociology, University of California
47. Rai A, Patnayakuni R, Patanayakuni N (1997) Technology investment and business performance. Commun ACM 40(7):89–97
48. Reiss J, Amorim M, Melao N, Matos P (2018) Digital transformation: a literature review and guidelines for future research. In: WorldCIST'18, AISC 745, pp 411–421
49. Rivard S, Raymond L, Verreault D (2006) Resource-based view and competitive strategy: an integrated model of the contribution of information technology to firm performance. J Strat Inf Syst 15(1):29–50

50. Roztocki N, Weistroffer H (2009) Event studies in information systems research : an updated review. In: Proceedings of the Americas conference on information systems, (August), pp 1–11
51. Santhanam R, Hartono E (2003) Issues in linking information technology capability to firm performance. MIS Q 27(1):125–153
52. Schryen G (2013) Revisiting IS business value research: what we already know, what we still need to know, and how we can get there. Eur J Inf Syst 22(2):139–169
53. Seddon P (2014) Implications for strategic IS research of the resource-based theory of the firm: a reflection. J Strat Inf Syst 23(4):257–269
54. Sherer S, Kohli R, Baron H (2003) Complementary investment in change management and IT investment payoff. Inf Syst Front 5(3):321–333
55. Stoel MD, Muhanna WA (2009) IT Capabilities and firm performance: a contingency analysis of the role of industry and IT capability type. Inf Manag 46(3):181–189
56. Tallon P (2003) The alignment paradox. CIO Insight 1(40):75–76
57. Tallon P (2014) Do you see what I see? The search for consensus among executives' perceptions of IT business value. Eur J Inf Syst 23(3):306–325
58. Tallon P, Pinsonneault A (2011) Competing perspectives on the link between strategic information technology alignment and organizational agility: insights from a mediation model. MIS Q 35(2):463–486
59. Tallon P, Queiroz M, Coltman T, Sharma R (2019) Information technology and the search for organizational agility: A systematic review with future research possibilities. J Strat Inf Syst 28(2019):218–237
60. Tambe P, Hitt LM (2012) The productivity of information technology investments: new evidence from IT labor date. Inf Syst Res 23(3):599–617
61. Thiem A, Baumgartner M, Bol D (2016) Still lost in translation! A correction of three misunderstandings between configurational comparativists and regressional analysts. Comp Pol Stud 49(6):742–774
62. Venkatraman N (1989) The concept of fit in strategy research: toward verbal and statistical correspondence. Acad Manag Rev 14(3):423–444
63. Weill P, Ross J (2005) A matrixed approach to designing IT governance. MIT Sloan Manag Rev 46(2):26–34
64. Wiengarten F, Humphreys P, Cao G, Mchugh M (2013) Exploring the important role of organizational factors in IT business value: taking a contingency perspective on the resource-based view. Int J Manag Rev 15(1)
65. Wilden R, Devinney T, Dowling R (2016) The architecture of dynamic capability research identify the building block of configurational approach. Ann Acad Manag 10(1):997–1076
66. Wilkin C, Couchman P, Sohal A, Zutshi A (2016) Exploring differences between smaller and large organizations corporate governance of information technology. Int J Account Inf Syst 22(1):16–25
67. Wu S, Straub D, Liang T (2015) How information technology governance mechanisms and strategic alignment influence organizational performance: insights from a matched survey of business and IT managers. MIS Q 39(2):497–518
68. Xue L, Ray G (2012) Efficiency or innovation: how do industry environments moderate the effects of firm's IT assets portfolios? MIS Q 36(2):509–528
69. Yin R (2008) Case study research and applications: design and methods, 6th edn. Sage Publications, USA
70. Zahra S, George G (2002) Absorptive capacity: a review, reconceptualization and extension. Acad Manag Rev 27(2):185–203

FasteNet: A Fast Railway Fastener Detector

Jun Jet Tai, Mauro Sebastián Innocente, and Owais Mehmood

Abstract In this work, a novel high-speed railway fastener detector is introduced. This fully convolutional network, dubbed FasteNet, foregoes the notion of bounding boxes and performs detection directly on a predicted saliency map. FasteNet uses transposed convolutions and skips connections, the effective receptive field of the network is 1.5× larger than the average size of a fastener, enabling the network to make predictions with high confidence, without sacrificing output resolution. In addition, due to the saliency map approach, the network is able to vote for the presence of a fastener up to 30 times per fastener, boosting prediction accuracy. FasteNet is capable of running at 110 FPS on an Nvidia GTX 1080, while taking in inputs of 1600 × 512 with an average of 14 fasteners per image. Our source is open here: https://github.com/jjshoots/DL_FasteNet.git.

Keywords Object detection · Convolutional neural networks · Railway fastener detection

1 Introduction

Traditionally, the inspection of railway fasteners has been done by visual inspection [1]. However, modern inspection methods are taking over this task. Examples include the Viola–Jones algorithm [2] and symmetry-based pyramid histogram of oriented gradients (PHOG) algorithm [3]. Gibert et al. used support vector machines (SVMs)

J. J. Tai (✉) · M. S. Innocente
Autonomous Vehicles & Artificial Intelligence Laboratory (AVAILab), Coventry University, 7th Floor Friars House, Manor House Drive, Coventry CV12TE, UK
e-mail: taij@uni.coventry.ac.uk

M. S. Innocente
e-mail: Mauro.S.Innocente@coventry.ac.uk

O. Mehmood
Omnicom Balfour Beatty, Clifton Park Ave, Rawcliffe, York YO305PB, UK
e-mail: owais.mehmood@balfourbeatty.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_71

to perform fastener and defect detection [4]. The work was also extended to fully convolutional neural networks [5]. Comparisons show that deep learning methods work well even with limited training data while being easier to design. More crucially, deep learning models tend to have lower inference time and good generalizability. Recent work by Song et al. [6] utilized a ResNet18 [7] backbone network on top of Faster-RCNN [8]. Their implementation achieved a recall and precision of more than 99% but was capped at 35 FPS on an Nvidia Titan X for 210 fasteners detected per second. Wang et al. proposed using the popular YOLOv2 and YOLOv3 architectures to perform fastener component detection [9]. Their implementation on an Nvidia GTX 1080Ti worked on close-up images of railway fasteners and performed with a mean average precision (mAP) of 93% at 35 FPS. These works show that deep convolutional neural networks (DCNNs) can be applied to railway fastener detection, like other computer vision tasks before it. However, the task of fastener, defect, and component detection is unique. Generally, the scale and rotation of the fasteners do not change, there is also no overlap between fasteners. This is true even in the case of sleeper detection. This is in contrast to general object detection tasks that may be subject to scale variance, occlusion, overlap, and other visual interruptions. For this reason, many existing network architectures are inelegant and can be improved. This work introduces several novel ideas:

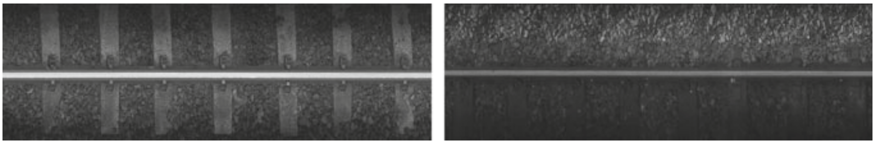
- Bounding boxes prevent the network from making more than one prediction per fastener without output suppression. Even so, predicting more than 10 bounding boxes at every output spatial location is excessive. FasteNet instead predicts a saliency map. The bounding boxes are then derived from this output using a first principles method, boosting accuracy by allowing the network to predict up to 30 times per fastener without requiring more than a one channel output feature map.
- There is no need to train the network to work on a large range of fastener scales. Due to this, FasteNet has an effective receptive field (ERF) only around 50% larger than necessary.
- FasteNet has different regimes during training and inference. During inference, outputs are thresholded and ceiled to increase prediction accuracy by eliminating low confidence predictions. During training, this step is removed to allow the network to fully train for every output location.

The rest of this paper is arranged as follows; in Sect. 2, our proposed system architecture is presented. In Sect. 3, the novelties that improve system performance are introduced. Finally, the paper is concluded in Sect. 4.

2 Methodology

2.1 Problem Description

FasteNet's purpose is to count fasteners in an image; not novel, but serves as a proof of concept for the ideas introduced here. A set of 997 images are generated by a New Measurement Train, some examples are shown in Fig. 1. The fasteners in



(a) (b)

Fig. 1 Example of railway images in the dataset where **a** all fasteners are visible, **b** the image is very badly illuminated

the images include fasteners that are occluded, damaged, or missing. Furthermore, the fasteners are exposed under various lighting conditions due to time of day; no artificial illumination is used. The images also contain fasteners of various types—PR clip, E clip, Fast clip, C clip, and J clip. The images are solely black and white, making detection harder by removing color information.

The problem is treated as a localization one. As long as the network prediction area overlaps with a ground truth bounding box, the prediction is considered a true positive (TP), otherwise it is perceived as a false positive (FP). The network must output the number of fasteners it sees in the image. This is to enforce the idea that the algorithm must clearly distinguish one fastener from another fastener, as opposed to simply outputting a saliency map. This variation of object detection strays away from the general bounding box detection one [8, 10, 11]. As long as the prediction area lies over a real fastener is sufficient.

2.2 FasteNet Design

Figure 2 provides a top level overview of FasteNet's architecture, as well as training and inference regimes. The input to the network is a black and white image defined here as $w_i \in \mathbb{R}^{1\times m\times n}$ where $m \leftrightarrow 64|m$ and $n \leftrightarrow 64|n$. It outputs a saliency map of $x_i \in \mathbb{R}^{1\times \frac{m}{8}\times \frac{n}{8}}$. The backbone network of FasteNet has just enough layers to provide it with a roughly 64×64 ERF. This is a slightly hand-wavy method of calculating the ERF; in reality, the ERF more closely represents a Gaussian distribution on the true receptive field [12]. This ERF is roughly $1.5\times$ the size of a fastener in the image. However, using only convolutional and pooling layers to achieve this ERF would translate to a very poor output resolution of 1/64th the input resolution. Many implementations of object detection work well with low output resolutions by using multiple bounding boxes per spatial element and then suppressing repeat predictions [10, 13]. This approach relies on very deep networks to provide the output layers with enough semantic value and then uses bounding boxes to handle localization precision loss.

Our approach takes a different look at the problem. FasteNet uses transposed convolutions to increase the output spatial resolution to 200×64 (1/8th the input size) and then predicts an objectness score at each output location. To reduce information

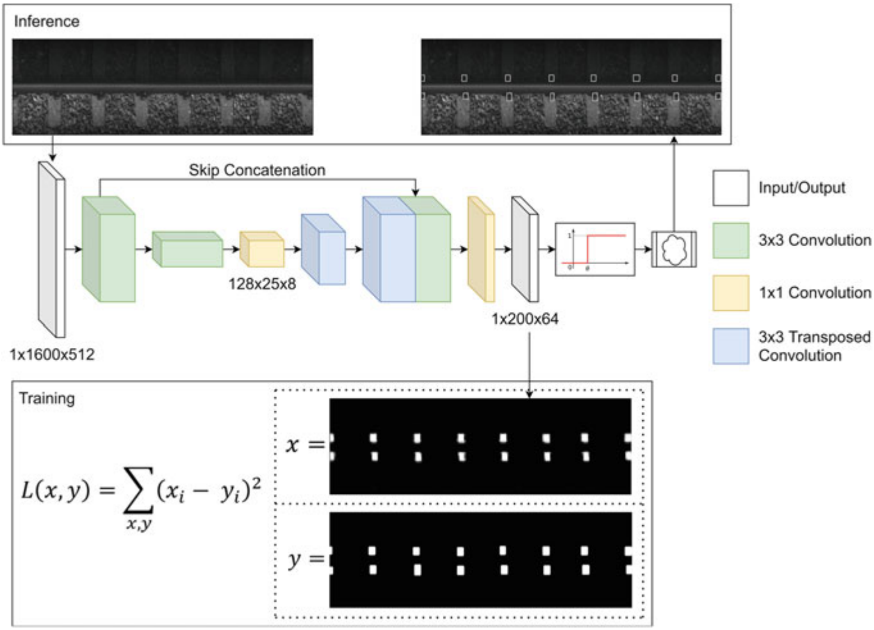


Fig. 2 Top level view of FasteNet. During training, the network trains against segmentation masks. During inference, the network outputs are thresholded before a contour finding algorithm is applied to find individual fasteners

loss and alleviate the vanishing gradient problem [14], feature maps from an earlier layer are concatenated to layers before the output layer. 1×1 convolutions are then used to produce the output saliency map. This allows up to 30 predictions per fastener. Other tidbits that boost FasteNet's performance include batch normalization to combat internal covariate shift during training and to provide regularization. Leaky ReLU units with a negative gradient of -0.1 are used to prevent dead neurons. The overall network architecture is shown in Table 1, note that the spatial size depends on the input size; any input size divisible by 64 can be computed by FasteNet—our implementation uses an input size of 1600×512 *during inference*.

During training, FasteNet trains against semantic segmentation masks. An example is shown in Fig. 2. During inference, FasteNet thresholds all output pixels, $x_i \leftarrow \delta_{[x>\theta]}$, where δ is the Kronecker delta, and then uses a boundary walking algorithm by Suzuki et al. [15] to identify each cluster at the output. The cluster bounds are used to derive bounding boxes. Semantically, given a chain of points $c_i \in \mathbb{R}^{n\times 2}$ denoting the contour for a cluster, the top left $\mathbf{b}_1$ and bottom right $\mathbf{b}_2$ corners of the bounding boxes are defined as $\mathbf{b}_1 = \{\min_{i\in n} \mathbf{c}_{i_x}, \max_{j\in n} \mathbf{c}_{j_y}\}$ and $\mathbf{b}_2 = \{\max_{i\in n} \mathbf{c}_{i_x}, \min_{j\in n} \mathbf{c}_{j_y}\}$. Other algorithms for drawing bounding boxes can also be used. In an extreme case, bounding boxes can be foregone exclusively, and the shape of the contour itself can be used to identify instances of fasteners.

2.3 Training and Evaluation

The network is trained with 1,000 256×256 random crops per image for 700 out of 997 available images. The mean squared error (MSE) objective is used, $L(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x},\mathbf{y}} (x_i - y_i)^2$, where $\mathbf{x}$ and $\mathbf{y}$ represent the network prediction and ground truth mask, respectively. The Adam optimizer [16] with a learning rate of 1e-6 and weight decay of 1e-2 is set as the optimizer. The network converges after the first 20 epochs with a precision of 96% and a recall of 89% when validated on the remaining 297 images. Hard negative mining was then used to train the network on difficult training images (more than 1 FP or FN). The network was then explicitly trained on these examples for 2 epochs, before continuing on the whole 700 images for another 2 epochs. This was repeated 3 times. In the end, a precision of 99% and a recall of 90% was accomplished on the validation dataset.

3 Performance Analysis

The model is validated on the remaining 297 images from the dataset. Evaluation with metrics done by other authors [4, 5, 17, 18] is not possible as the datasets are not public and the detection classes are different. Instead, FasteNet is compared to two other internal models. This is to validate our hypothesis that a network designed in this manner has better performance than a brute force DCNN implementation. These two networks are denoted here as VanillaNet and LargeNet, shown in Tables 3 and 4 with a comparison in Table 2.

VanillaNet does not utilize transposed convolutions and consists of 11 layers. LargeNet uses transposed convolutions with a larger ERF of roughly 128×128. Note that because of the output scaling ratio of this network, the inputs are cropped to 1536×512. These two additional networks are trained on the same training regime that was used on FasteNet in Sect. 2.3. Special considerations are made to ensure that each network has roughly the same number of parameters to reduce variance in a representational capacity.

By varying the threshold θ, the rate of predictions can be varied to form the precision–recall curves for each network shown here in Fig. 3. The performance of FasteNet is on top. LargeNet and VanillaNet implementations may perform as well or better than FasteNet when given more training time due to deeper network architectures and hence theoretically more representational capacity. However, this sub-percentage performance increase is compensated by the loss in some computational speed, which is especially the case in VanillaNet. At the end of the training, FasteNet is able to output bounding boxes around each fastener as shown in Fig. 4.

Table 1 Network architecture of FasteNet. The outputs from layer 3 are concatenated channel wise to outputs of layer 10 to form the input to layer 11

Layer #	Operation	Input size	Output size	Kernel size	Pad size	Pool size	Activation
1	Convolution	$1@\ 1600 \times 512$	$32@\ 800 \times 256$	3×3	1×1	2×2	Leaky ReLU
2	Convolution	$32@\ 800 \times 256$	$32@\ 400 \times 128$	3×3	1×1	2×2	Leaky ReLU
3	Convolution	$32@\ 400 \times 128$	$64@\ 200 \times 64$	3×3	1×1	2×2	Leaky ReLU
4	Convolution	$64@\ 200 \times 64$	$64@\ 100 \times 32$	3×3	1×1	2×2	Leaky ReLU
5	Convolution	$64@\ 100 \times 32$	$128@\ 50 \times 16$	3×3	1×1	2×2	Leaky ReLU
6	Convolution	$128@\ 50 \times 16$	$128@\ 25 \times 8$	3×3	1×1	2×2	Leaky ReLU
7	Convolution	$128@\ 25 \times 8$	$64@\ 25 \times 8$	1×1	–	–	Leaky ReLU
8	Transposed convolution	$64@\ 25 \times 8$	$64@\ 50 \times 16$	4×4	1×1	–	Leaky ReLU
9	Transposed convolution	$64@\ 50 \times 16$	$64@\ 100 \times 32$	4×4	1×1	–	Leaky ReLU
10	Transposed convolution	$64@\ 100 \times 32$	$64@\ 200 \times 64$	4×4	1×1	–	Leaky ReLU
11	Convolution	$128@\ 200 \times 64$	$1@\ 200 \times 64$	1×1	–	–	Sigmoid

Table 2 Summary of FasteNet in comparison with two other networks built for an evaluation of FasteNet's design

Network	Parameter count	FLOPS	FPS
FasteNet	494k	4.00G	110
VanillaNet	500k	25.93G	25
LargeNet	514k	6.38G	80

4 Conclusion

In this work, a methodology for designing DCNN models for railway maintenance tasks was introduced. Formally, bounding box regression is replaced with contour finding on thresholded saliency maps. To the best of our knowledge, this approach is first. This concept takes advantage of the unique nature of railway track images when compared to generic object detection challenges. FasteNet performs fastener detections on images of 1600×512 at a speed of 110 FPS on an Nvidia GTX 1080 with up to 99.1% precision using a very lightweight network.

As it stands, there are several limitations to FasteNet—susceptibility to perspective and size warping being a major one, limiting its use case to non-airborne mounted camera systems. FasteNet may also be overfitting to the dataset, given that the dataset is considered small, it may not fully capture the distribution of such an application in real life.

In the future, several things will be looked at. Firstly, there is hope to make the dataset public and contain more samples. Second, detecting fasteners alone is not particularly useful. FasteNet will aim to be applicable to fastener, sleeper, rail, and ballast defects by having one extra channel per class of detection. The goal is to leverage the principles of FasteNet to develop similar models for railway inspection and/or road inspection. FasteNet can also be adapted to perform multitask classification as in the case of [5]. Additionally, an expansion of the dataset would allow the network to generalize better. Data augmentation methods such as a range of affine transforms, lux manipulation, self adversarial training, or simply gathering more data can be used.

Table 3 Network architecture of VanillaNet

Layer #	Operation	Input size	Output size	Kernel size	Pad size	Pool size	Activation
1	Convolution	1@1600 × 512	32@800 × 256	3 × 3	1 × 1	2 × 2	Leaky ReLU
2	Convolution	32@800 × 256	64@800 × 256	3 × 3	1 × 1	–	Leaky ReLU
3	Convolution	64@800 × 256	32@800 × 256	3 × 3	1 × 1	–	Leaky ReLU
4	Convolution	32@800 × 256	64@400 × 128	3 × 3	1 × 1	2 × 2	Leaky ReLU
5	Convolution	64@400 × 128	128@400 × 128	3 × 3	1 × 1	–	Leaky ReLU
6	Convolution	128@400 × 128	64@400 × 128	3 × 3	1 × 1	–	Leaky ReLU
7	Convolution	64@400 × 128	128@200 × 64	3 × 3	1 × 1	2 × 2	Leaky ReLU
8	Convolution	128@200 × 64	64@200 × 64	3 × 3	1 × 1	–	Leaky ReLU
9	Convolution	64@200 × 64	128@200 × 64	3 × 3	1 × 1	–	Leaky ReLU
10	Convolution	128@200 × 64	64@200 × 64	3 × 3	1 × 1	–	Leaky ReLU
11	Convolution	64@200 × 64	1@200 × 64	1 × 1	1 × 1	–	Sigmoid

Table 4 Network architecture of LargeNet. The outputs from layer 3 are concatenated channel wise to outputs of layer 12 to form the input to layer 13

Layer #	Operation	Input size	Output size	Kernel size	Pad size	Pool size	Activation
1	Convolution	1@1536 × 512	32@768 × 256	3 × 3	1 × 1	2 × 2	Leaky ReLU
2	Convolution	32@768 × 256	64@384 × 128	3 × 3	1 × 1	2 × 2	Leaky ReLU
3	Convolution	64@384 × 128	64@192 × 64	3 × 3	1 × 1	2 × 2	Leaky ReLU
4	Convolution	64@192 × 64	64@96 × 32	3 × 3	1 × 1	2 × 2	Leaky ReLU
5	Convolution	64@96 × 32	64@48 × 16	3 × 3	1 × 1	2 × 2	Leaky ReLU
6	Convolution	64@48 × 16	64@24 × 8	3 × 3	1 × 1	2 × 2	Leaky ReLU
7	Convolution	64@24 × 8	128@12 × 4	3 × 3	1 × 1	2 × 2	Leaky ReLU
8	Convolution	128@24 × 8	64@12 × 4	1 × 1	–	–	Leaky ReLU
9	Transposed convolution	64@12 × 4	64@24 × 8	4 × 4	1 × 1	–	Leaky ReLU
10	Transposed convolution	64@24 × 8	64@48 × 16	4 × 4	1 × 1	–	Leaky ReLU
11	Transposed convolution	64@48 × 16	64@92 × 32	4 × 4	1 × 1	–	Leaky ReLU
12	Transposed convolution	64@92 × 32	64@192 × 64	4 × 4	1 × 1	–	Leaky ReLU
13	Convolution	128@192 × 64	1@192 × 64	1 × 1	–	–	Sigmoid

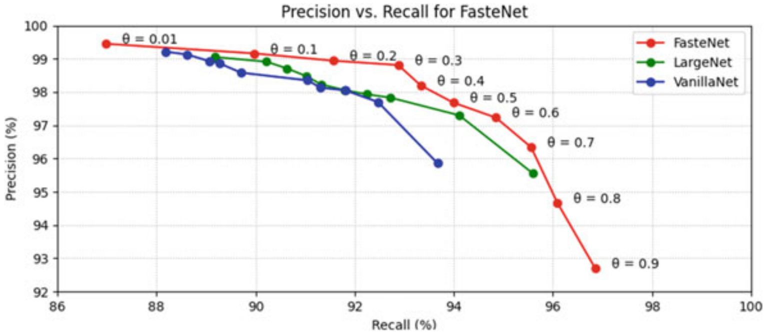


Fig. 3 Precision–Recall Curve of FasteNet, VanillaNet, and LargeNet when performed on the unseen 297 images of the dataset with different threshold values

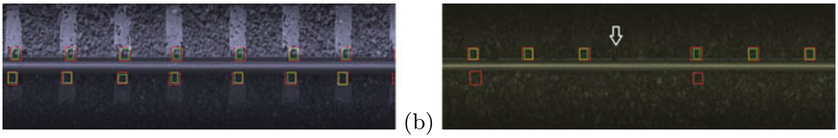


Fig. 4 Green boxes are FasteNet predictions, red boxes are ground truth annotations. **a** 'False' Positives, the network predicts the presence of the fasteners even if small and not annotated in the ground truth. **b** False Negatives resulting from fastener occlusion and poor lighting conditions. Even so, the network hedges against detecting defect fasteners, identified with the white arrow here

References

1. Marquis B, LeBlanc J, Tajaddini A (2014) Vehicle track interaction safety standards. In: ASME/IEEE joint rail conference, vol 45356. American Society of Mechanical Engineers, p V001T02A016
2. Xia Y, Xie F, Jiang Z (2010) Broken railway fastener detection based on adaboost algorithm. In: 2010 international conference on optoelectronics and image processing, vol 1. IEEE, pp 313–316
3. Liu J, Li B, Xiong Y, He B, Li L (2015) Integrating the symmetry image and improved sparse representation for railway fastener classification and defect recognition. In: Mathematical problems in engineering
4. Gibert X, Patel VM, Chellappa R (2015) Robust fastener detection for autonomous visual railway track inspection. In: 2015 IEEE winter conference on applications of computer vision. IEEE, pp 694–701
5. Gibert X, Patel VM, Chellappa R (2016) Deep multitask learning for railway track inspection. IEEE Trans Intell Trans Syst 18(1):153–164
6. Song Q, Guo Y, Yang L, Jiang J, Liu C, H Mu (2019) High-speed railway fastener detection and localization system. arXiv:1907.01141
7. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
8. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
9. Wang T, Yang F, Tsui K-L (2020) Real-time detection of railway track component via one-stage deep learning networks. Sensors 20(15):4325

10. Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC (2017) Dssd: deconvolutional single shot detector. arXiv:1701.06659
11. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv:1804.02767
12. Araujo A, Norris W, Sim J (2019) Computing receptive fields of convolutional neural networks. Distill 4(11):e21
13. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
14. Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertain Fuzziness Knowl-Based Syst 6(02):107–116
15. Suzuki S et al (1985) Topological structural analysis of digitized binary images by border following. Comput Vision Graph Image process 30(1):32–46
16. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
17. Liu J, Huang Y, Zou Q, Tian M, Wang S, Zhao X, Dai P, Ren S (2019) Learning visual similarity for inspecting defective railway fasteners. IEEE Sens J 19(16):6844–6857
18. Wei X, Yang Z, Liu Y, Wei D, Jia L, Li Y (2019) Railway track fastener defect detection based on image processing and deep learning techniques: a comparative study. Eng Appl Artif Intell 80:66–81

CSP Machine in the Standard C++ Library Context: Implementation and Sample Applications

Milen Loukantchevsky

Abstract The C.A.R. Hoare's theory of Communicating Sequential Processes (*CSP*) is a strong basis for analysis and synthesis of concurrent computer systems. Its manifestation as abstraction, a parallel systems specification language and a parallel programming language is unique. If the absence of a protected address space is ignored, the *std::thread* class could be regarded near close to *CSP* process semantics. Unfortunately, in the standard *C++* library there is not type even approximately close to *CSP* channel. Hence, the main goal set: proposal of suitable implementation of *CSP* semantics of inter-process communications through the *C++ 11* standard library means in seek of higher degree of structuring of communications of objects of type *std::thread*. As a result, a compact *C++* template library named *csp* is developed. It encapsulates the means of messaging and synchronization between objects of type *std::thread*. The *csp* namespace defined includes *csp::chan*, *csp::mux*, *csp::sink* and *csp::forks* classes. The *csp::chan* class is strong implementation of *CSP* channel semantics, with two methods *csp::chan::send()* and *csp::chan::recv()* corresponding to *CSP* communications commands <*!*> for output and <*?*> for input. The *CSP* alternative command for nondeterministic selection is implemented by the method *csp::mux::recv()* using multiple wait on *std::condition_variable* and randomized choice between true guards. By practical reasons, through the additional classes, *csp::sink* and *csp::fork* supported the communication schemes <*n:1/n*> and <*1:n*> too. The code of the developed library and examples of its use are placed in the *GitHub* under *MIT* license.

Keywords CSP · C++ 11 · Concurrency · Message passing · Nondeterminism

1 Introduction

Leading methodological principle of science is to direct investigations to the essential contradiction of the subject domain. One such contradiction in the area of

M. Loukantchevsky (✉)
University of Ruse, 8 Studentska Str, POB 7017, Ruse, Bulgaria
e-mail: mil@ieee.org

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_72

computer systems is between explicit parallelism of current computer architectures and sequential thinking [17].

The C.A.R. Hoare's theory of Communicating Sequential Processes (*CSP*) is a strong basis for analysis and synthesis of concurrent computer systems [1, 12, 17, 20]. It's unique in its simultaneous manifestation as an abstract model, a parallel systems specification language and a parallel programming language. Therefore, *CSP* is one of the potential means of overcoming the above mentioned contradiction.

From engineering point of view, *CSP* defines an abstract parallel machine—*CSP* machine (Fig. 1). This machine consists of two subsystems: a set of computing nodes and a communication medium. Computer nodes provide local protected execution environment for executing one or more concurrent asynchronous processes. The communication medium, in turn, consists of a set of channels. The processes grouped by pairs could communicate only by message passing over a communication channel. Thus, concurrent programming paradigm has shifted from traditional "processes and shared objects" [19] to "processes and channels" [23].

Over its long history, *CSP* has found numerous of successful implementations: from *transputers* and its *OCCAM* language, through *CSP* distributed frameworks like [3] to modern *Go* and *Python* languages [8, 21, 23] and chip multiprocessor platform *xCORE/XC* [18, 22]. Despite that, the constructive potential of this parallel computational model is far from being expended.

If the absence of a protected address space is ignored, the *std::thread* class could be regarded near close to *CSP* process semantics. But in the *C*++ standard library,

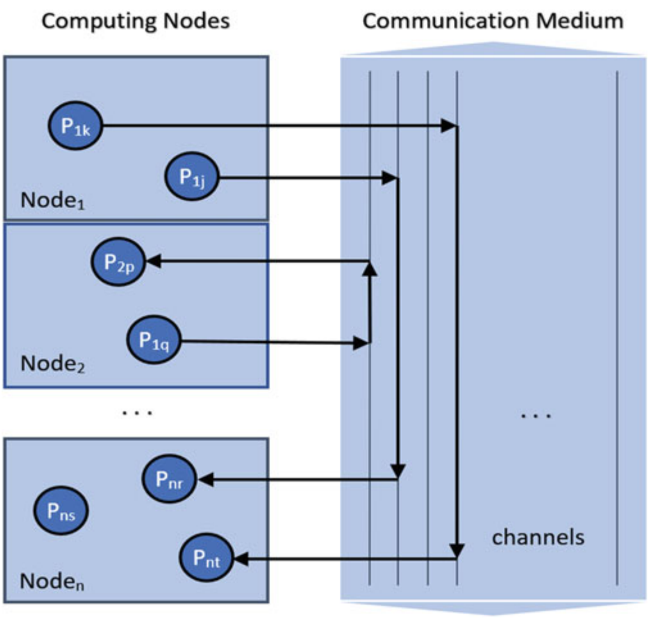


Fig. 1 Generalized *CSP* machine

Fig. 2 Communication scheme <*1:1*>

there is not type even approximately close to *CSP* channel nor *CSP* nondeterministic selection [13, 23].

Hence, the main goal set[1]: to propose a suitable implementation of the *CSP* semantics of inter-process communications through the *C++ 11* standard library means in seek of higher degree of structuring of communications of objects of type *std::thread* and hiding of low level synchronization mechanisms.

The choice of the C++ 11 standard, rather than any of the following, is based on next prerequisites:

- *C++ 11* standard is widely accepted and supported.
- Of all the standards that follow it (*C++ 14* and *C++ 17*), only *C++ 20* has more "dramatic" changes of interest in the area of concurrency (*std::jthread*, *std::semaphores*), and *C++ 20* is currently in its final approval process [6].

Undoubtedly, after *C++ 20* standard adoption there could be appropriate revision of *csp* library proposed.

2 CSP Template Library Classes

2.1 General View

The *C++* library developed, adheres to modern *C++* principles of template libraries. Named *csp*, it encapsulates the means of messaging and synchronization between objects of type *std::thread* as well as hiding of low level mechanisms. The *csp* namespace defined includes *csp::chan*, *csp::mux*, *csp::sink*, and *csp::forks* classes.

The communication channel is the basic building block of the *CSP* machine's communication medium, as seen in Fig. 1. Hence, there is a need for separate class *csp::chan* providing point-to-point communication scheme (Fig. 2).

[1] For the author, this is in a sense an evolution of his previous works, such as *fiberOS/CSP* [15, 17].

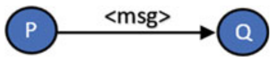


Fig. 3 Pair of communicating processes

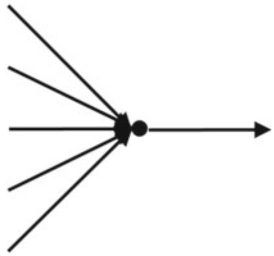


Fig. 4 Communication scheme *<n:1>*

It follows the *CSP* semantics of inter-process communications that suggests the simplest possible kind of channels: unidirectional, point-to-point *<1:1>*, unbuffered, with direct naming. Such channels do not actually imply flow control, but a special kind of bilateral synchronization known as "*rendezvous*".

Let us assume that we have a parallel system *S* consisting of two processes—producer *P* and consumer *Q* described by the next *CSP* equation

$$S = \{P||Q\}, P = \{Q!\,msg\}, Q = \{P?\,msg\} \tag{1}$$

Both processes *P* and *Q* form a communication pair (Fig. 3). They interact over *point-to-point* (*p2p*, *<1:1>*) communication channel through pair of communication commands: *<!>* and *<?>*, output (send) and input (receive), respectively.

To communication scheme *<1:1>* provided with the *csp::chan* class it is necessary to provide one more of type *<n:1>* (Fig. 4).

In parallel systems, as well as in concurrent ones, *nondeterminism* is an intrinsic feature and result of random choice of transitions. Dijkstra's *guarded command* is a mean to control and effectively use of the nondeterminism in parallel systems [7]. Guarded command further evolved by C.A.R. Hoare into *CSP alternative command*

$$\{\mathrm{G}_1 \rightarrow \mathrm{P}_1 \Box \mathrm{G}_2 \rightarrow \mathrm{P}_2 \Box \cdots \Box \mathrm{G}_\mathrm{n} \rightarrow \mathrm{P}_\mathrm{n}\} \tag{2}$$

where the *i*th alternative $\mathrm{G}_\mathrm{i} \rightarrow \mathrm{P}_\mathrm{i}$ consists of a guard G_i and subordinated process P_i. The necessary condition for the subordinated process P_i to start execution is its guard G_i to evaluates *true*. The guards have the same priority and they all are evaluated simultaneously. As a consequence, there could be more than one guard evaluated true. In such a case, the choice between their subordinated processes is random.

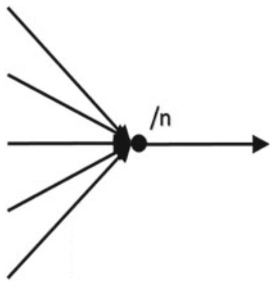


Fig. 5 Communication scheme *<n:1/n>*

The alternative command, in fact, represents generalization of deterministic *<if-else>* operator. But only if all guards are mutually exclusive, we will have *<if-else>* deterministic behavior, otherwise, the behavior observed will be nondeterministic [17].

Special interest in terms of process interactions is the case when the guards are input (receive) commands

$$\{P_1?x \rightarrow SKIP \Box P_2?x \rightarrow SKIP \Box \cdots \Box P_\mathrm{n}?x \rightarrow SKIP\} \tag{3}$$

provided by the class *csp::mux* and its member function *csp::mux::recv().*

The communication schemes *<1:1>* and *<n:1>* are in principle sufficient to perform any possible type of inter-process communications. But forpractical reasons, the additional classes *csp::sink* and *csp::fork* are included in the *csp namespace* as well.

The *csp::sink* class supports a communication scheme *<n:1/n>* . This is a kind of input fork, in which, multiple waits are performed at the input and messages from *n* sources are received simultaneously as many as available (Fig. 5).

While *csp::mux::recv()* implements *CSP* alternative command (Eq. 3), *csp::sink::recv()* corresponds to the next

$$\{P_1?x_1 \rightarrow SKIP || P_2?x_2 \rightarrow SKIP || \cdots || P_\mathrm{n}?x_n \rightarrow SKIP\} \tag{4}$$

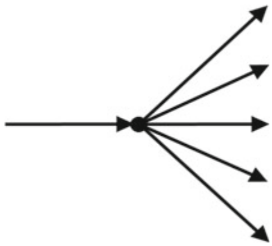
where available inputs are executed in parallel and command ends after the last input received.

Conversely to *csp::sink*, the *csp::fork* class performs multiple waits by output thus supporting the communication scheme *<1:n>* with multicast send (Fig. 6).

As opposed to *csp::sink::recv()* and Eq. 4, *csp::sink::out()* corresponds to the Eq. 5, where input (receive) commands of Eq. 4 are replaced by output (send) commands

$$\{Q_1!x_1 \rightarrow SKIP || Q_2!x_2 \rightarrow SKIP || \cdots || Q_\mathrm{n}!x_n \rightarrow SKIP\} \tag{5}$$

Fig. 6 Communication scheme <*1:n*>



Here, all outputs are executed in parallel and command ends after the last output committed.

To comply with the volume restrictions set here, we will limit ourselves to presenting only the class *csp::mux*. For those interested, concise introduction to the topic could be found in [16], and the developed *csp* library itself as well as testbed applications are placed in the *GitHub* repository [14].

2.2 Multiplexor Class

As already said, it is necessary to maintain <*n:1*> scheme (Fig. 4), and that in more general nondeterministic form given by Eq. 3. In first versions of the *csp* library, this was done by a separate function *csp::alt()* later transformed into the method *csp::mux::recv()* being part of the multiplexor class *csp::mux*. Its descriptor includes conditional variable [4] to synchronize on and privately owned mutex.

```
template <class T_ID, class T_DATA>
class mux
{
public:
  using CHAN = csp::chan<T_ID, T_DATA>;

  // Dispatching Modes
  enum class Dispatching {by_place, uniform};
  //  by_place  - deterministic mode
  //  uniform   - default (nondeterministic) mode
  . . .
protected:
  std::mutex mu_cv;  // mutex of the cond. variable
  std::condition_variable cv; // conditional variable
  std::vector<std::shared_ptr<CHAN>> in;  // inputs
};
```

The multiplexed input channels C_1, C_2, …, C_n of Fig. 7 are contained in the data member *in* of type *std::vector <std::shared_ptr <CHAN>>* .

When started *csp::mux::recv()* is blocked on std::mux conditional variable. It is marked at Fig. 7 by $G[n]$ as association with array of n input guards. When at least one of these guards signals, *csp::mux::recv()* unblocks and builds the list L with all guards evaluated to true. The index k of the selected guard contained in L is determined by the uniform distributed random function $\hat{g}$. We have here, in fact, a form of Data Level Parallelism (*DLP*) sequencer which select one of the input streams and output it to the exit of the circuit.

As in the < *1:1* > communication scheme, so here is used predicative form of *std::condition_variable::wait()* implemented as *lambda* [2, 10, 11]. This corresponds to multiple wait over *std::condition_variable*, which is resolved according to *CSP* alternative command nondeterministic semantics. In addition to the default nondeterministic dispatching of input messages, is provided "*by place*" dispatching too.

```
//  Alternative Command n-channel
//  Dispatching:
//    uniform  - default (nondeterministic) mode
//    by_place - deterministic mode
int recv(T_DATA& dst, Dispatching dispatch =
Dispatching::uniform)
{
  int id = -1;

  //  WaitOnForMultipleEvents
  std::unique_lock<std::mutex> lk(mu_cv);
  cv.wait(lk,
      [this, &id, &dispatch]()
      { //  The Predicate, implemented as lambda
        //  checked before std::wait(lock)
        if(dispatch == Dispatching::by_place)
        {
          // Dispatching by guard place
          for(auto j = 0; j < this->in.size(); j++)
          {
            if(this->in[j]->get_sent_status() == true)
            {
              id = j;
              return true;
            }
          }
        }
```

```
        else if(dispatch == Dispatching::uniform)
        {
          // Nondeterministic uniform dispatching
          std::vector<int> vGuardsInTrue;
          for(auto j = 0; j < this->in.size(); j++)
          {
            if(this->in[j]->get_sent_status() == true)
            {
              vGuardsInTrue.push_back(j);
            }
          }

          if(vGuardsInTrue.size() > 0)
          {
            int k = 0;  // index of selected true guard
            if(vGuardsInTrue.size() > 1)
            {   // more than 1 guards are true
              // select one of them randomly
              std::uniform_int_distribution<int>
                distribution(0, vGuardsInTrue.size() - 1);
              k = distribution(random_engine);
            }

            id = vGuardsInTrue[k];
            return true;
          }
        }

        id = -1;
        return false;
      });

  in[id]->recv(dst);

  return id;

– }
```

Interest represents that section of lambda, marked in the above listing with gray background, corresponding to $k = \widehat{g}(L)$. The vector *vGuardsInTrue* stands there for *L*. The specified section could be used for verification of nondeterministic properties of the primitive according to *CSP* alternative command semantics.

At the end of the *csp::mux::recv()* primitive, is returned the identifier *id* () of the channel selected to receive from which is useful to trace the source of the message received.

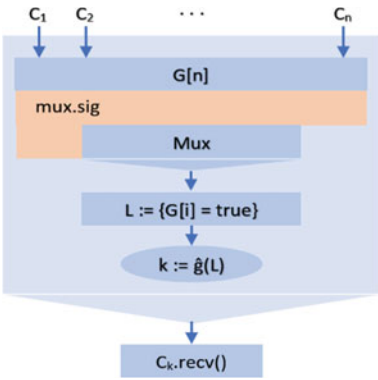


Fig. 7 Generalized operating structure of *csp::mux::recv()*

The random engine itself used in *csp::mux::recv()* is defined as static and randomized at the beginning of the *csp* namespace following the standard pattern.

```
namespace csp
{
  //  Nondeterministic Uniform Dispatching Setup
  static auto seed = std::chrono::system_clock::
                          now().time_since_epoch().count();
  static std::default_random_engine random_engine(seed);
  . . .
}
```

Useful feature of *csp::mux* class is the possibility of limiting the number of input channels *in.size()* to one. This is potential prerequisite for generalization because of unification of the type of inputs when used as parameters—channels could be replaced by multiplexors.

Close by functionality example to here described *csp::mux::recv()* is the *CSP* alternative command implementation in the exokernel *fiberOS/CSP* [15, 17]. There are separate 2-channel and *n*-channel *ALT()* primitives. But these primitives are tightly bounded to exokernel *fiberOS/CSP* and cannot be used separately.

Another example of *CSP* alternative command implementation is th <*select*> operator within the platform *xCORE/XC* produced by *XMOS* [18, 22]. Moreover, this is done at microarchitectural level. Unfortunately, such support is not common in general purpose computer architectures yet. Moreover, the author's studies revealed that *xCORE/XC* implementation is not true nondeterministic.

3 Conclusion

Following the practice of the standard *C++* template, a compact library named *csp* is developed. It encapsulates the means of messaging and synchronization between objects of type *std::thread* and hiding of low-level synchronization mechanisms.

The *csp* library evolution begins from a point-to-point < *1:1* > communication over *CSP* proto channel and goes through several versions: *0.1*, *0.2*, …, *0.8RC*, *0.9RC* and *0.91RC* till now. The last one supports not only the basic communication schemes < *1:1* > and < *n:1* > through *std::chan* and *std::mux* classes, but also the communication schemes < *n:1/n* > and < *1:n* > through additional classes *csp::sink* and *csp::fork*.

Numerous of testbeds were developed. On the one hand they are used in the development of the *csp* library. And on the other, as examples of application of the library in solving typical tasks, different communication patterns separately and in combinations (*p2p-mux*, *sink-mux*, *fork-mux*, etc.), support of *co-procedures* with decentralized and centralized management, *active objects*, etc. Some of them are presented here.

The overall development was performed by the *Embarcadero C++ Builder®* development environment and its *Clang-enhanced C++* compiler [5, 9]. However, both the *csp* library itself and its testbeds include only standard constructions and hence are portable at source level for use with other *C++ 11* compatible environments. Finally, the code of the developed library and examples of its use are posted onto the author's *GitHub* repository [14] under *MIT* license.

References

1. Abdallah A, Jones C, Sanders J (2005) Communicating sequential processes: the first 25 years. Springer, Berlin
2. Bartlomiej F (2020) C++ lambda story. https://leanpub.com/cpplambda. Accessed 26 Sept 2020
3. Charousset D, Hiesgen R, Schmidt T (2014) CAF—the C++ actor framework for scalable and resource-efficient applications. In: AGERE! 2014, October 20, 2014, Portland, OR, USA, ACM 978-1-4503-2189-1
4. C++ Reference: The condition_variable class. https://en.cppreference.com/w/cpp/thread/condition_variable. Accessed 2020/11/02
5. C++ Support in Clang. http://clang.llvm.org/cxx_status.html#cxx17. Accessed 01 Nov 2020
6. Current ISO C++ Status. https://isocpp.org/std/status. Accessed 27 Nov 2020
7. Dijkstra E (1975) Guarded Commands, nondeterminancy and formal derivation of programs. Commun ACM 18(8):453–457
8. Donovan A, Kernighan B (2016) The Go Programming language. Pearson Education Inc., New York
9. Embarcadero RAD Studio Docwiki. http://docwiki.embarcadero.com/RADStudio/Sydney/en. Accessed 01 Nov 2020
10. Gottschling P (2016) Discovering modern C++: an intensive course for scientists, engineers, and programmers. Pearson Education Inc., Boston
11. Gregoire M (2014) Professional C++, 3rd edn. Wiley, Indianapolis
12. Hoare CAR (1978) Communicating sequential processes. Commun ACM 21(8):666–677

13. Josuttis N (2012) The C++ standard library: a tutorial and reference, 2nd edn. Pearson Education Inc., Boston
14. Loukantchevsky M (2020) CSP Library GitHub Repository. https://github.com/milphaser/CSP. Accessed 13 Dec 2020
15. Loukantchevsky M (2020) fiberOS/CSP GitHub Repository. https://github.com/milphaser/fOS. Accessed 13 Dec 2020
16. Loukantchevsky M (2020) Implementation of the CSP semantics of inter-process communications using the C++11 standard library. In: Reports awarded with "Best Paper" crystal prize—59th science Conference of Ruse University, Ruse. ISBN 978-954-712-826-2
17. Loukantchevsky M, Kostadinov N, Avakyan H (2019) A testbed of nondeterminism in educational context. In: Proceedings of the 20th international conference on computer systems and technologies, New York, NY, USA, ACM, pp 304–307. ISBN 978-1-4503-7149-0. https://doi.org/10.1145/3345252.3345260
18. May D (2009) The XMOS XS1 Architecture. XMOS Ltd.
19. Schneider F, Andrews G (2020) Concepts of concurrent programming. https://www.cs.cornell.edu/fbs/publications/concepts_espirit.pdf. Accessed 26 Nov 2020
20. Schneider S (2000) Concurrent and real-time systems: the CSP approach. Wiley, Chichester
21. Wang Y (2020) Concurrency with Python: CSP and Coroutines. https://bytes.yingw787.com/posts/2019/02/09/concurrency_with_python_csp_and_coroutines/. Accessed 27 Nov 2020
22. Watt D (2009) Programming XC on XMOS Devices. XMOS Ltd.
23. Williams A (2012) C++ concurrency in action: practical multithreading. Manning Publications Co, Shelter Island

Spatio-Temporal Causal Relations at Urban Road Networks; Granger Causality Based Networks as an Insight to Urban Traffic Dynamics

Glykeria Myrovali, Theodoros Karakasidis, Georgia Ayfantopoulou, and Maria Morfoulaki

Abstract Being spatio-temporal variations of urban road traffic a critical information for understanding and predicting accurately traffic, the current paper focuses on urban road traffic dynamics understanding by introducing the notion of causality. Using 15-min aggregated travel time series from taxi GPS data, causal networks are developed. The results reveal potentials arising from mining causality beyond correlation notion among urban road paths as well as the contribution of causal networks to a decision support system for traffic management. The knowledge on causal relations and the characteristic time lags on the 'transfer' of the information (traffic) among the road paths is a key knowledge for traffic management, since it gives the possibility to proactively intervene in the affected road paths and to inform users for alternative routes. Being high the extendibility and transferability potentials of the proposed approach, exploitation in other transport-related problems appears promising.

Keywords Granger Causality · Correlation · Floating Taxi Data · Travel Time Time Series · Causal Networks · Traffic Management · Transport Planning · Decision Making · Traffic Information Provision Systems · Smart Cities

---

G. Myrovali (✉) · G. Ayfantopoulou · M. Morfoulaki
Hellenic Institute of Transport/Centre for Research and Technology Hellas (CERTH/HIT), 6th km Charilaou-Thermi Road, 57001 Thessaloniki, Greece
e-mail: myrovali@certh.gr

G. Ayfantopoulou
e-mail: gea@certh.gr

M. Morfoulaki
e-mail: marmor@certh.gr

T. Karakasidis
Physics Department of Civil Engineering, University of Thessaly, 35100 Lamia, Greece
e-mail: thkarak@uth.gr

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_73

1 Introduction

The gradual (structured or even unstructured) development of cities has increased the complexity of urban functions, part of which is the urban transport network and urban traffic. Urban transportation is characterized by inherent complexity that increases with the increase of city's size (increasing number of subsystems consisting the overall system). Furthermore, the structure of the road network (topological structure), socio-economic factors, weather effects, traffic management policies, the trip purpose, the selected mode, trips' distribution, travel time and cost (of all alternatives) and interaction among individual decisions are interrelated issues adding in system's complexity ([1]-[8]).

Urban traffic is undoubtedly a non-static phenomenon, presenting variations over both space (variations per road path) and time:

- Short term variations caused by delays at intersections and turning maneuvers linked to parking space search or to the specific needs of the road users leading to shock-wave phenomena and traffic propagation
- Daily variations—peak and off-peak hours
- Weekly variations—different day profiles, i.e. typical weekdays, Saturday and Sunday profile
- Long term variations or variations over the years that are caused by sociodemographic and other global trends [9].

Another category of variations could be the 'shock' variations arising from global threats as the unprecedented situation with corona-virus in 2020—the last type of variations is a 'new entry' dimension that suddenly (but still unknown in what level) affects almost all the above types of time variations.

A series of research papers refer to the spatiotemporal variation of traffic flow; multivariate ARIMA models, Spatiotemporal ARIMA (STARIMA), ARIMA with exogenous variables (ARIMAX) and dynamic multivariate models as Vector Autoregressive Moving Average (VARMA) has been widely used. The evolution of the above models was needs—driven; to explore the autocorrelation in time and space of the traffic variables ([10, 11]). The correct identification and calculation of autocorrelation among observations is the basis of an optimal statistical representation of spatio-temporal relations [12], while it also consists of a necessary input for machine learning approaches (ML) [13] since the absence of such a dimension could result in an erroneous representation of data variation or even in overestimation of ML models performance [14]. The interested reader is referred to paper [15], where an analytic review of models for predicting the spatio-temporal variation of traffic is made.

Reviewing of models seeking to incorporate the spatiotemporal information of traffic reveals a common principal assumption; that the spatio-temporal autocorrelation at the datasets can be sufficiently described based on globally fixed measures as, for example, that the extent of autocorrelation among observations is fixed in space and/or time [16]. However, it seems that the assumptions of a static spatio-temporal catchment area fails to describe dynamic systems as of urban road traffic.

The current situation is generated upon the previous traffic state while it affects also (two directional) the traffic state of a close road path or even a path that is not in its close neighbourhood ([17]-[19]). Relevant questions arising from the above are: which are the causal relations among road paths per traffic state, in which degree and how does this effect change over the different traffic states (i.e. from an uncongested to a congested state or under an urgent event)?

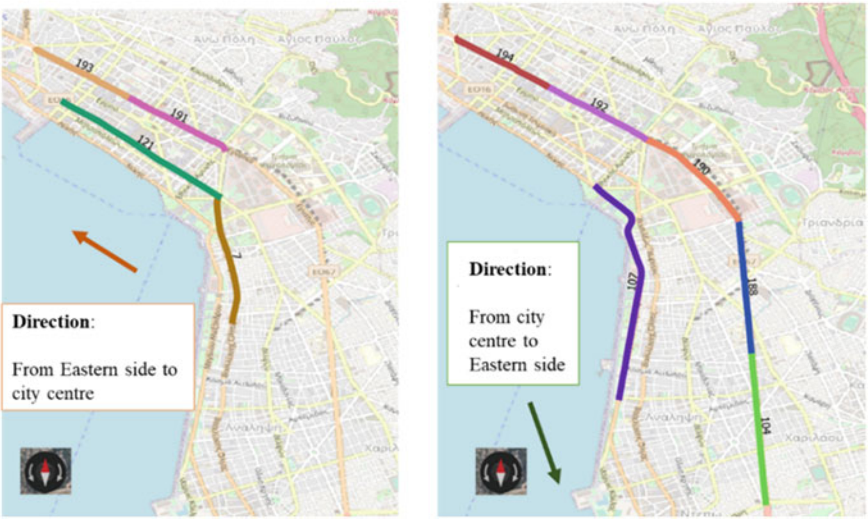
Spatial autocorrelation ([20–22]), correlation coefficient and cross correlation ([23–25]) are the most commonly used indicators for describing the spatio-temporal evolution of traffic. The current paper, going a step forward while assessing the need to identify causality in the relations among roads paths, introduces the measure of Granger causality in the above analysis for avoiding reflecting just the 'mere' or spurious correlations (results based just on the correlation analysis). Furthermore, based on graph theory and visualization powers, the paper proposes also the transformation of the spatio-temporal information to networks for getting a quick and reliable insight in traffic dynamics.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of the study area and of the traffic data used in the analysis, Sect. 3 describes the methodological process applied. Section 4 presents the results of the analysis, and finally, Sect. 5 summarizes the findings, implications of the study, and directions for future research.

2 Study Area and Data Description

2.1 The Study Area; the Road Network of the City of Thessaloniki

The study area is part of the road network of the wider area of Thessaloniki, the second largest Greek city, with a population of over 800.000 inhabitants (2011). Car dependence during the previous years in Thessaloniki was significantly high (private car modal share of over 60%) while around 20% of the trips were conducted by Public Transport and 3% with taxi [26], numbers that have however started changing; recent data on travel mode choice reveal a decrease in the share of trips by private car and an increase in more sustainable modes and ways of transport (walking, bikes, e-scooters and public transport). Today, the city is considered a smart transport laboratory; the stable cooperation among core stakeholders in the framework of Thessaloniki Smart Mobility Living Lab (https://smartmlab.imet.gr/) and the exploitation of new technologies have brought a new era in traffic planning and management. Floating car data from approximately 1200 taxis, multiple traditional sensors, cameras, radars and point-to-point Bluetooth detectors create the bedrock of innovation for the city giving the opportunity to city transport practitioners to make efficient decisions for the transport system [27].



	path7	path104	path107	path121	path188	path190	path191	path192	path193	path194
Matlab coding	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Path length (km)	1.3	1.6	2.3	1.5	1.4	1.1	0.95	0.95	0.85	0.85
Direction	west	east	east	west	east	east	west	east	west	east
Number of lanes	Bus lane + 2 operational		Bus lane + 4 operational	Bus lane + 2 operational						
Subpart of greater axis*	V. Olgas	K. Karamanli	M. Alexandrou	Tsimiski	K. Karamanli	Egnatia	Egnatia	Egnatia	Egnatia	Egnatia
Daily Average Flow / Passenger Car Unitσ (PCU)	35.000	35.000	45.000		35.000 – 40.000					

Fig. 1 Examined urban road paths, Thessaloniki (Gr)

The analysis presented in the current paper is made per road path and comparative results are reached (multivariate time series analysis). The examined paths along with key characteristics (with coding, direction, location) are presented in Fig. 1.

2.2 *The Traffic Datasets—Travel Time Timeseries Per Road Path (15 min Step)*

The traffic time series used for the current analysis are based on taxi GPS data—data being elaborated by the Hellenic Institute of Transport (CERTH/HIT) in the framework of the Thessaloniki Smart Mobility Living Lab. The dataset comes from a fleet of in total 1200 taxis circulating and generating pulses every 6–10 s, depending on their speed ([28]-[29]). The number of timestamps processed for the current analysis is 13.966.878 (total timestamps in the first semester of 2017).

Map matching of timestamps and travel time calculation per road path per quarter of day took place as described in [30]. Average travel times per quarter of weekdays per road path were calculated creating the reference time series for the current work (10 time series of 96 lags / quarters of day length).

3 Methods

Having as ultimate scope, the development of easily readable and interpretable graphs for understanding the road network dynamics and the relations among urban road paths, a three-step methodology was followed: i)cross correlation and ii) causalities (Granger causality) calculation formulated the spatio-temporal dimension of the urban road traffic system, which, in turn, formed the basic information for knowledge networks building (visualization of causal relations among road paths). The main software used is MATLAB (for exploiting 'Multivariate Granger Causality Toolbox' (MVGC) [31]).

3.1 Time Series Cross Correlation

In statistics, correlation defines the degree of similarity between two variables or time series (or else, variables' movement association) receiving values among -1 and 1; -1 for perfect negative correlation or inverse correlation coefficient showing fully alignment of variables' movement in the opposite direction, 1 for the perfect positive correlation and 0 when the variables present entirely different evolution.

When we are examining the movement of time series at different time lags than zero, the procedure is called cross correlation. For two time series, i.e. xt, yt that can be related to future or past time lags, the sample cross correlation function (CCF) can help us mine the lags of the x-variable that might be useful predictors of yt (therefore, it is possible that x leads y or x lags y). The cross—correlation function of the jointly stationary xt and yt (with lag h) time series is [32]:

$$\rho_{\mathrm{X}\Upsilon}(\tau) = \frac{1}{\mathrm{N}-\tau}\,\frac{\sum_{t=1}^{\mathrm{N}-\tau}(\boldsymbol{x}_t-\bar{\boldsymbol{x}})\left(\boldsymbol{y}_{t+\tau}-\bar{\boldsymbol{y}}\right)}{\sqrt{\sum_{t=1}^{\mathrm{N}-t}(\boldsymbol{x}_t-\bar{\boldsymbol{x}})^2}\sqrt{\sum_{t=1}^{\mathrm{N}-t}\left(\boldsymbol{y}_t-\bar{\boldsymbol{y}}\right)^2}} \tag{1}$$

where, N is the length of the time series, τ is the (each time) examined time lag, x-and(y)- the average values of the two time series.

In the current analysis, we calculated and examined the cross correlation function among each pair of time series (travel time time series in examined paths) in order to identify indications for relations and interrelated dynamics. However, there is always the issue of spurious correlation, in which the examined variables or time series are associated but not causally related—this could be the result of an unseen factor or even a coincidence [32].

3.2 Time Series Causality Analysis

Causality refers to the existence of causal relationship among variables. According to [33], a signal (time series) Yt causes Xt when we are in a position to better predict Xt when we use information not only contained in its past values, but also in past values of Yt. According to Granger (1969) [32] when MSE (Mean Square Error) of forecasting Xt using just itself is higher that when taking also into account Yt then we say that Yt causes Xt.

A principal analysis behind causality is vector autoregressive models—therefore, we develop two autoregressive models, one containing just Y and the other including also X, and the test that follows is whether the variance of the second model is significantly lower than this of the first. In that case, we say that causality exists [34].

At first step, we calculate the MSE of the prediction, k steps ahead, of Y based on previous values of the same variable. For this case, the vector model of Y is

$$Y_t = a_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + \ldots + A_p Y_{t-p} + e_t \tag{2}$$

This model is called restricted since it contains just the one of the two time series. The MSE of the model is given by the Eq. 3.

$$MSE_Y = \frac{1}{n-p} \sum_{t=p}^{n-1} {e^2}_{t+1} \tag{3}$$

where et is the prediction error, that is equal to the difference between the predicted value for t = k and the actual value.

The second step is the MSE calculation of Y predictions taking into account the effect of X. In this case, the vector autoregressive (VAR) model, that is called unrestricted, becomes

$$\begin{aligned} Y_t = a_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + \ldots + A_p Y_{t-p} \\ + B_1 X_{t-1} + B_2 X_{t-2} + \ldots + B_p X_{t-p} + e_t \end{aligned} \tag{4}$$

while, respectively, the MSE of the second model (MSE(Y/(X))) is calculated and the two values are compared. If.

$$MSE_{(Y/(X))} < MSE_Y \tag{5}$$

then we can say that X causes Y.

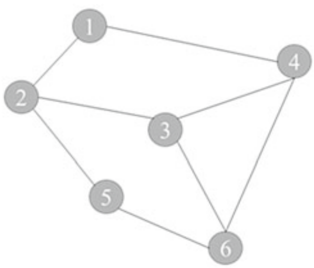
For testing the above, Granger causality index is calculated by Eq. 6:

$$F_{X \to Y} = log \frac{MSE_Y}{MSE_{(Y/(X))}} \tag{6}$$

If the value of Granger causality index is above zero (F_(X → Y) > 0) , then we can say that causality exists (from X to Y) while the larger the value of the index the higher the effect of X to Y. In case F_(X → Y) is closely to zero, we can say that there is no causality between the two variables.

3.3 Causal Graphs Development

The last decades, network construction from time series, that is mostly based on mixing graphs theory with time series analysis, has been widely used (basically in econometrics and neuroscience). It is also very encouraging for spatio-temporal phenomena as atmospheric and oceanic data. Ultimate goal is to capture the dynamics (relations-causalities-system identification) from the interpretation of network's topological structure (Fig. 2). The interested reader is directed to ([35]-[39]) for various methodologies that have been proposed over the years for the construction of such networks.



$$A = \begin{array}{c} \\ (1) \\ (2) \\ (3) \\ (4) \\ (5) \\ (6) \end{array} \begin{array}{c} \begin{array}{cccccc} (1) & (2) & (3) & (4) & (5) & (6) \end{array} \\ \left[\begin{array}{cccccc} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{array}\right] \end{array}$$

Fig. 2 Adjacency matrix example, *where in the current work, the nodes are the road paths and the links are representing the existence of significant causal relation among them (1 in the matrix shows the existence while the measure of significance is the combination of granger causality)*

In the current work, the measures of cross correlation at the first upper or lower point of the CCF was taken in combination with Granger causality determined the distance at the adjacency matrix for the networks building according to the following equation:

$$Gcausality = (Vcausality, Ecausality) = \begin{vmatrix} 0 & caus_{1,2} & \ldots & caus_{1,k} \\ caus_{2,1} & 0 & \ldots & caus_{2,k} \\ \ldots & \ldots & \ldots & \ldots \\ caus_{k,1} & caus_{k,2} & \ldots & 0 \end{vmatrix} \tag{7}$$

where, Vcausality represents the nodes of the network (10 road paths in our case), Ecausality refers to the links (relatively high causalities) and causi, j is the distance measure associated with the causality of the pair i-j. Graph theory applies for converting the causality matrices into networks [40].

4 Results

Average travel times for crossing the examined paths were calculated at 15-min scale based on the taxi GPS data of the 1st semester of 2017 that have been collected by CERTH/HIT in Thessaloniki. General characteristics (skewed data) and patterns (i.e. morning peak at around 07:00 and at midday, considerable less traffic after 21:00) are mined from the sequence plots that verify the existing knowledge (Fig. 3) for traffic daily performance in the city.

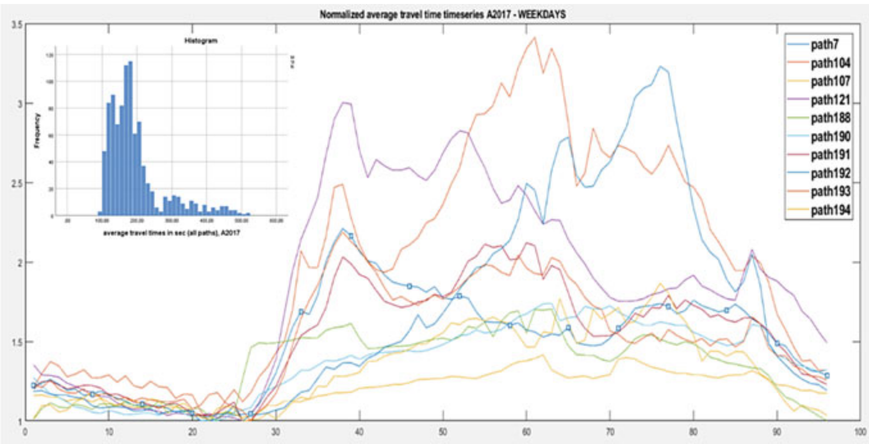


Fig. 3 Travel time data distribution (skewed data) and traffic patterns-average travel time time series per path, weekdays A2017 (normalized values)

path	7	104	107	121	188	190	191	192	193	194
Path's length (m)	1291	1559	2282	1544	1353	1115	945	941	854	852
N	11040	11040	11040	11040	11040	11040	11040	11040	11040	11040
Mean	146	283	192	324	160	139	184	193	185	143
Median	143	262	188	291	163	139	173	155	169	134
Mode	153	180	190	180	180	140	181	121	150	120
Std. Deviation	55	145	88	144	60	44	104	126	83	89
Skewness	6	4	20	1,8	16	10	12	5	9	12
Kurtosis	75	57	493	9	460	251	233	54	154	234
Minimum	60	35	97	32	43	58	56	53	55	30
Maximum	1243	3056	2599	2176	2223	1489	2591	2831	2029	2463

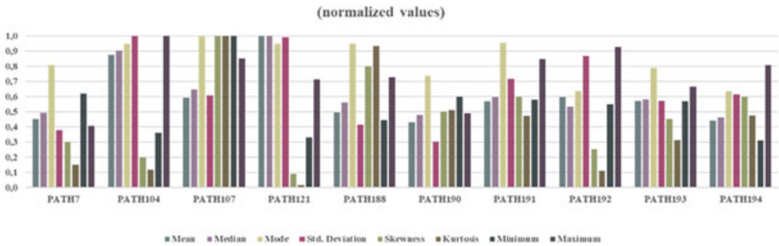


Fig. 4 Descriptive statistics per road path

Skewness and kurtosis of travel time data is obvious also from the table of the descriptive statistics per path (Fig. 4). An interesting finding is that, the skewness of the distribution varies among road paths while initial indications that are being further examined by the authors are drafted injecting prior knowledge of the reference area urban road traffic; travel time distribution's skewness varies in different road paths according to their operational characteristics and the level of congestion during the day. For example, road path 121, the central road axis of the city serving a wider area that attracts a large percentage of daily trips and where congestion phenomena appear in a frequent basis (v/c $\approx$ 0.98), presents the lower comparatively skewness. From the other side, road path 107 (v/c $\approx$ 0.55) that does not present severe congestion phenomena presents the higher skewness. Authors are further examining skewness and daily traffic variations in order to conclude to concrete results.

Skewness and kurtosis of the data can be affected by non-recurrent events, however, since the same reference period is taken into account for all paths, the current work is not further focusing on the influence of events on the descriptive statistics.

Sample cross correlation functions were then explored per pair of time series for visual identification of lagged periods and direction of correlation (Fig. 5).

The results of the cross correlation and granger causality analysis per pair of time series are presented in Fig. 6.

The interpretation of the cross correlation and causality matrices presented in Fig. 7 results among others in the following significant remarks:

- the highest causality is identified from path 121 to path 193, parallel central axis in the same direction—high cross correlation in 0 lags
- the parallel road paths 190 and 107 (same direction, to East) are highly causally related

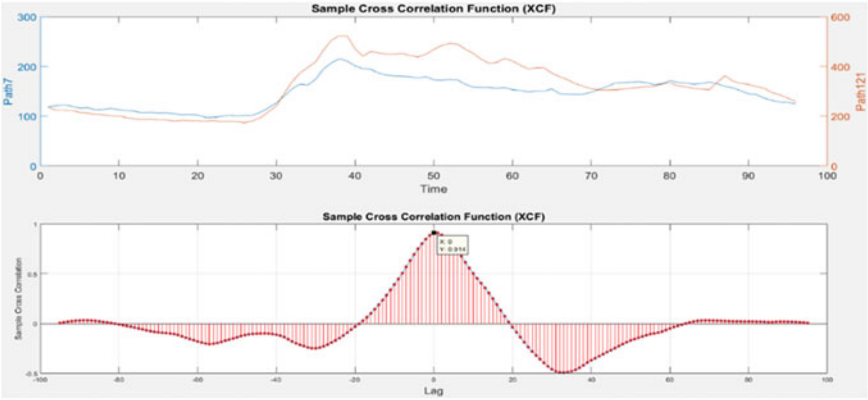


Fig. 5 Sample cross correlation functions path 121—path 7, weekdays A2017

Cross - correlation										
	PATH7	PATH104	PATH107	PATH121	PATH188	PATH190	PATH191	PATH192	PATH193	PATH194
PATH7	1	0,67	0,7	0,91	0,78	0,7	0,86	0,59	0,86	0,64
PATH104		1	0,89	0,74	0,84	0,92	0,91	0,85	0,84	0,89
PATH107			1	0,78	0,78	0,94	0,82	0,87	0,78	0,87
PATH121				1	0,81	0,74	0,92	0,75	0,95	0,75
PATH188					1	0,81	0,87	0,61	0,85	0,7
PATH190						1	0,86	0,87	0,78	0,86
PATH191							1	0,63	0,94	0,79
PATH192								1	0,5	0,9
PATH193									1	0,67
PATH194										1

Time lags (15')										
	PATH7	PATH104	PATH107	PATH121	PATH188	PATH190	PATH191	PATH192	PATH193	PATH194
PATH7		0	9	0	-2	2	1	23	0	3
PATH104			2	-1	0	1	0	4	0	0
PATH107				-9	3	1	-2	1	4	3
PATH121					-2	7	1	20	0	12
PATH188						5	2	2	3	1
PATH190							-2	1	-3	0
PATH191								0	0	0
PATH192									0	-1
PATH193										0
PATH194										

Granger Causality										
	PATH7	PATH104	PATH107	PATH121	PATH188	PATH190	PATH191	PATH192	PATH193	PATH194
PATH7		0,01	0,19	0,18	0,54	0,04	0,01	0	0,01	0,22
PATH104	0,03		0,36	0,37	0,48	0,68	0,02	0,05	0,08	0,07
PATH107	0,33	0,68		0,42	0,61	0,97	0,51	0,06	0,38	0,44
PATH121	0,4	0,01	0,01		0,55	0,01	0,02	0,03	0,03	0,01
PATH188	0,01	0	0,33	0		0,3	0,05	0	0,01	0,01
PATH190	0,18	0,39	0,44	0,32	0,83		0,38	0,08	0,31	0,32
PATH191	0,59	0,21	0,33	0,75	0,81	0,27		0,02	0,4	0,08
PATH192	0,04	0,5	0,53	0,43	0,41	0,34	0,09		0,18	0,53
PATH193	0,35	0,1	0,09	1	0,72	0,07	0,19	0		0,22
PATH194	0,3	0,45	0,16	0,4	0,7	0,37	0,38	0,31	0,5	

Fig. 6 (i) cross correlation, (ii) respective time lags and (iii) Granger causality results (normalized values)

Fig. 7 Causal network of the examined urban road network [© Graph Online]

- road path 188 seems to play a pivotal role for the examined network (high causal influence to the rest road paths)—it is estimated that its proximity to the city center and its connection to the urban peripheral highway road is of crucial importance for the whole network operation

Exploiting the graph theory, the causality matrix is converted to a causality network as depicted in Fig. 7.

The development of networks based on this type of spatio-temporal information for the urban road traffic is estimated to provide a quick understanding of the underlying dynamics and the potential effect of mobility interventions in a path to the whole network. For example, it is easily understandable from the network in Fig. 7, a network—wide trend visualization, that road path 188 is a very crucial road path for the network and variations in traffic in road path 188 are reflected in different time lags (information to be taken from cross-correlation time lagged) in almost all the rest road paths or that path 107 is a path highly reflecting traffic variations in many other paths (thickness of lines are in line with the causality).

5 Conclusions and Discussion

Travel time time series, from taxi GPS data in the city of Thessaloniki (GR) used in the analysis, seem to verify in general the traffic state understanding—the proof of reliability of taxi GPS data compared with the general traffic is not included in the scopes of the current paper, however, the authors have already tested them with experimental probe vehicles applying generalized linear models (gamma with log link function) for identifying their accuracy compared to general traffic (private cars).

The analysis conducted in the framework of the current work reveals encouraging results from the combined exploitation of multivariate time series and graph theory for the transport sector. It is expected that the outcome of this research could be a main building block of a wider Decision Support System in Traffic planning and management ('CCN Decision Support System'—correlation, causality network). The causality reasoning in correlation among road paths is a novelty to the best of the authors knowledge for the spatio-temporal analysis of traffic. Although the current work took into account same traffic variables per road path (travel time time series), Granger causality can be also found among different variables describing the transport system, i.e. causality among timeseries of travel times—flows—delays at different points/areas of the networks which shows a high transferability potential for supporting other applications (i.e. tailored volume delay functions). Similarly, the transfer of the components of the analysis at different levels, i.e. interurban road network is also possible. Exploitation potentials also exist at various areas; from a support mechanism to a navigation product to a support in fleet management (based on the knowledge of where traffic will increase in the near future—time lags). The current analysis was made on average travel time series from a large period which can give the grand picture of the urban transport system performance (recurrent urban traffic). The authors highly support that the whole analysis should be undertaken in frequent basis in order to take case-specific results since the complexity of urban transport system is high.

The next steps of the analysis include testing data aggregation levels in order to compare results; from general network trends and paths' relations (more aggregated level—i.e. hourly time intervals) to a more detailed check of sensitivity to traffic changes (day of week and time of day) [41] and checking the effect of events on the results.

References

1. Barthélemy M (2011) Spatial networks. Phys Rep 499(1–3):1–101. https://doi.org/10.1016/j.physrep.2010.11.002
2. Strano E, Nicosia V, Latora V, Porta S, Barthelemy M (2012) Elementary processes governing the evolution of road networks. Scie Rep 2(296). https://doi.org/10.1038/srep00296
3. van Ommeren J, Rietveld P, Nijkamp P (2000) Job mobility, residential mobility and commuting: A theoretical analysis using search theory. Ann Reg Sci 34(2):213–232. https://doi.org/10.1007/s001689900004
4. Polyzos S (2015) Urban development. Kritiki Publications, Athens, Greece, pp 541 [in Greek]
5. Polyzos S, Tsiotas D, Minetos D (2013) Determining the driving factors of commuting: an empirical analysis from Greece. J Eng Sci Technol Rev 6(3):46–55. https://doi.org/10.25103/jestr.063.09
6. Chowell G, Hyman JM, Eubank S, Castillo-Chavez C (2003) Scaling laws for the movement of people between locations in a large city. Phys Rev E. https://doi.org/10.1103/physreve.68.066102

7. Kwon J, Mauch M, Varaiya P (2006) Components of Congestion. Trans Res Record J Trans Res Board 1959(1):84–91. https://doi.org/10.1177/0361198106195900110
8. Wen T-H, Chin W-C-B, Lai P-C (2017) Understanding the topological characteristics and flow complexity of urban traffic congestion. Phys A Stat Mech Appl 473:166–177. https://doi.org/10.1016/j.physa.2017.01.035
9. Weijermars WAM, Netherlands Research School for Transport IL (2007) Analysis of urban traffic patterns using clustering. Netherlands TRAIL Research School
10. Getis A (1995) Cliff, A.D. and Ord, J.K. 1973: Spatial autocorrelation. London: Pion. Progress in Human Geography, 19(2), 245–249. https://doi.org/10.1177/030913259501900205.
11. Bartholomew DJ, Box GEP, Jenkins GM (1971) Time series analysis forecasting and control. Operat Res Q (1970–1977) 22(2):199. https://doi.org/10.2307/3008255
12. Hackney JK, Bernard M, Bindra S, Axhausen KW (2007) Predicting road system speeds using spatial structure variables and network characteristics. J Geogr Syst 9(4):397–417. https://doi.org/10.1007/s10109-007-0050-4
13. Angarita-Zapata JS, Masegosa AD, Triguero I (2019) A taxonomy of traffic forecasting regression problems from a supervised learning perspective. IEEE Access 7:68185–68205. https://doi.org/10.1109/access.2019.2917228
14. Patil GP (2009) Impacts and Wider Impacts on Statistics. Geogr Anal 41(4):430–435. https://doi.org/10.1111/j.1538-4632.2009.00772.x
15. Ermagun A, Levinson D (2018) Spatiotemporal traffic forecasting: review and proposed directions. Transp Rev 38(6):786–814. https://doi.org/10.1080/01441647.2018.1442887
16. Cheng T, Haworth J, Wang J (2011) Spatio-temporal autocorrelation of road network data. J Geogr Syst 14(4):389–413. https://doi.org/10.1007/s10109-011-0149-5
17. Chandra SR, Al-Deek H (2008) Cross-correlation analysis and multivariate prediction of spatial time series of freeway traffic speeds. Transp Res Record J Transp Res Board 2061(1):64–76. https://doi.org/10.3141/2061-08
18. Sun S, Zhang C, Zhang Y (2005) Traffic flow forecasting using a spatio-temporal Bayesian network predictor. In: Duch W, Kacprzyk J, Oja E, Zadrożny S (eds) Artificial neural networks: formal models and their applications—ICANN 2005. ICANN 2005. Lecture notes in computer science, vol 3697. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11550907_43.
19. Yang S, Shi S, Hu X, Wang M (2015) Spatiotemporal context awareness for urban traffic modeling and prediction: sparse representation based variable selection. PLoS ONE 10(10). https://doi.org/10.1371/journal.pone.0141223
20. Hallenbeck ME, Rice M, Smith B, Cornell-Martinez C, Wilkinson J (1997) Vehicle volume distributions by classification. FHWA,U.S. Department of Transportation, Washington, D.C.
21. Black WR, Thomas I (1998) Accidents on Belgium's motorways: a network autocorrelation analysis. J Transp Geogr 6(1):23–31. https://doi.org/10.1016/s0966-6923(97)00037-9
22. Flahaut B, Mouchart M, Martin ES, Thomas I (2003) The local spatial autocorrelation and the kernel method for identifying black zones. Accid Anal Prev 35(6):991–1004. https://doi.org/10.1016/s0001-4575(02)00107-0
23. Yue Y, Yeh AG-O (2008) Spatiotemporal traffic-flow dependency and short-term traffic forecasting. Environ Plann B Plann Des 35(5):762–771. https://doi.org/10.1068/b33090
24. Chandra SR, Al-Deek H (2009) Predictions of freeway traffic speeds and volumes using vector autoregressive models. J Intell Transp Syst 13(2):53–72. https://doi.org/10.1080/15472450902858368
25. Xie J, Choi Y-K (2017) Hybrid traffic prediction scheme for intelligent transportation systems based on historical and real-time data. Int J Distrib Sens Netw. https://doi.org/10.1177/1550147717745009
26. Mitsakis E, Stamos I, Grau JMS, Chrysochoou E, Iordanopoulos P, Aifadopoulou G (2013) Urban mobility indicators for Thessaloniki. J Traff Logist Eng 1(2):148–152. https://doi.org/10.12720/jtle.1.2.148-152
27. Aifadopoulou G, Salanova JM, Tzenos P, Stamos I, Mitsakis E (2019) Big and open data supporting sustainable mobility in smart cities—the case of Thessaloniki. In: Nathanail E, Karakikes I (eds) Data analytics: paving the way to sustainable urban mobility. CSUM 2018. Advances in intelligent systems and computing, vol 879. Springer, Cham

28. Stamos I, Salanova Grau JM, Mitsakis E, Aifadopoulou G (2016) Modeling effects of precipitation on vehicle speed: floating car data approach. Transp Res Record J Transp Res Board 2551(1):100–110. https://doi.org/10.3141/2551-12
29. Myrovali G, Karakasidis T, Charakopoulos A, Tzenos P, Morfoulaki M, Aifadopoulou G (2019) Exploiting the knowledge of dynamics, correlations and causalities in the performance of different road paths for enhancing urban transport management. In: Book: Decision support systems IX: main developments and future trends, 5th international conference on decision support system technology. https://doi.org/10.1007/978-3-030-18819-1_3
30. Salanova Grau JM, Mitsakis E, Tzenos P, Stamos I, Selmi L, Aifadopoulou G (2018) Multi-source data framework for road traffic state estimation. J Adv Transp 2018:1–9. https://doi.org/10.1155/2018/9078547
31. Barnett L, Seth AK (2014) The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. J Neurosci Methods 223:50–68. https://doi.org/10.1016/j.jneumeth.2013.10.018
32. Granger CWJ (1969) investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3):424. https://doi.org/10.2307/1912791
33. Wiener, N. (1956). Nonlinear prediction and dynamics. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability (Vol. 3, pp. 247–252). Berkeley, University of California Press.
34. Charakopoulos A (2015) Analyzing and identifying spatial dynamics using advanced time series analysis methods, PhD thesis, University of Thessaly
35. Zhang J, Small M (2006) Complex network from pseudoperiodic time series: topology versus dynamics. Phys Rev Lett 96(23). https://doi.org/10.1103/physrevlett.96.238701
36. Xu X, Zhang J, Small M (2008) Superfamily phenomena and motifs of networks induced from time series. Proc Natl Acad Sci 105(50):19601–19605. https://doi.org/10.1073/pnas.0806082105
37. Yang Y, Yang H (2008) Complex network-based time series analysis. Phys A 387(5–6):1381–1386. https://doi.org/10.1016/j.physa.2007.10.055
38. Gao, Z., & Jin, N. (2009). Flow-pattern identification and nonlinear dynamics of gas-liquid two-phase flow in complex networks. Physical Review E, 79(6). https://doi.org/10.1103/physreve.79.066303.
39. Gao Z-K, Jin N-D, Wang W-X, Lai Y-C (2010) Motif distributions in phase-space networks for characterizing experimental two-phase flow patterns with chaotic features. Phys Rev E 82(1). https://doi.org/10.1103/physreve.82.016210
40. Bondy JA, Murty USR (1976) Graph theory with applications, vol 290. Macmillan, London
41. Vlahogianni E, Karlaftis M (2011) Temporal aggregation in traffic data: implications for statistical characteristics and model choice. Transp Lett 3(1):37–49. https://doi.org/10.3328/tl.2011.03.01.37-49

The Role of e-Commerce in Organic Farming in Latvia

Denis Vasiliev

Abstract Competition between organic and convenience farms is acute. Large convenient farms harness economies of scale and enhanced access to the market through collaboration with major retail chains to outcompete smaller farms. E-commerce could potentially improve small farmer access to end consumers, without the need to establish collaboration with retail chains. Lack of e-commerce adaptation between organic farmers could signal potential competitive disadvantage of organic farmers in Latvia, vis a vis representatives of convenient agriculture. In order to evaluate internet usage among Latvian organic farmers for commercial purposes, a list of 3264 Latvian certified farms was examined. It was found that less than 10% of the farmers have websites and less than 15% are present on social media. Moreover, a large proportion of farmers that are present on the Internet do not use their platforms for purposes of e-commerce. Such a low commercial usage of digital communication tools among Latvian farmers signals possible link with overall decline in organic farming industry. Therefore, it is heavily recommended to promote e-commerce among Latvian organic farmers.

Keywords e-Agriculture · Sustainability · Communications

1 Introduction

A recent UN Report [1] suggests that the world has failed to meet biodiversity conservation targets. Given the fact that in Europe around 40% of land area is farmland, one of the key drivers of biodiversity loss in the region is agricultural intensification [2]. Biodiversity loss, in turn, may undermine human food security as a result of ecosystem service loss, including pollination, pest control, and soil fertility [3]. Thus, promotion of sustainable agricultural practices, including organic farming, is between the key goals of the EU Biodiversity Strategy [4]. Despite substantial policy effort as the EU level, such as greening measures in CAP and agri-environmental schemes [5],

D. Vasiliev (✉)
Turiba University, 68 Graudu Street, Riga 1058, Latvia
e-mail: turiba@turiba.lv

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_74

competition between organic and convenient farms remains fierce across the EU [6]. Access to the market is between the factors that puts smallholders at a competitive disadvantage vis a vis larger convenient farms [7], as due to economies of scale, the former are often unable to secure collaboration with national scale retail chains. The way to increase competitive advantage in such situation is to get the organic food directly to environmentally conscious consumers, potentially cutting out intermediaries from the supply chain [8]. In a modern world, this could be achieved through harnessing one of e-agriculture tools—e-commerce [9]. This strategy becomes particularly relevant in the context of current pandemic situation, whereby even large retail chains switch to e-business.

Several ongoing processes in Latvia indicate that organic smallholders suffer from strong competition with larger convenient farms. These include generally steadier increase in land area under commercial crops, farm consolidation, and land abandonment [10]. Furthermore, relatively few organic brands are represented at the shelves of Latvian major retail chains. At the same time, internet coverage in Latvia is very good, providing internet access to more than 80% of households [11]. Moreover, internet usage has sharply increased in the country over the last year [12], providing a tremendous opportunity for e-commerce. In this context, I decided to evaluate internet usage among Latvian organic farmers for commercial purposes. To the best of knowledge, no directly comparable research has been undertaken so far. This research, may have a significant contribution to development of Latvian organic farming industry and may benefit overall sustainable development of the country. Furthermore, positive effects of the research may be scaled up in future, in case internet usage among organic farmers for e-commerce purposes would be evaluated across Europe and maybe even across the World.

2 Materials and Methods

In present research data from Latvian Agricultural Data Centre [13] was used as a sampling frame. The source contains a full list of all certified organic farmers in Latvia, including both individual farmers and companies (thereafter a term “farms” will be used for organic agricultural companies). A total number of records on the database is 3264. A sample of 300 farmers was selected from the sampling frame using random sampling method. For this purpose, a random number table from a book “Reproduced from Million Random Digits” by RAND corporation was used [14]. A Google search was undertaken on each selected farm/farmer name. All farms/farmers from the sample were recorded into an Excel spreadsheet. The spreadsheet had the following columns: name of the farm/farmer, location, product category (i.e. growers, livestock producers, poultry and egg producers), website (present/absent)), purpose of the website (e-commerce, advertising, other), website address, Social media page (present/absent), Social media page address, social media platform, purpose of the Social media page (e-commerce, advertising, other). The websites and social media pages were examined in detail, in order to understand whether they are used for the purposes of e-commerce, advertising or for other purposes. Proportion of farms/farmers with websites and/or presence on Social media was determined. Also, proportion of farms/farmers with websites as well as proportion of sites used for e-commerce, advertising and other purposes was estimated. Proportions of corporate farms with websites was compared with proportion of private farms with web-sies. Similarly, number of farms and individual farmers present on Social media was compared. Most popular Social media platforms used by farms and individual farmers were identified. Number of farms with websites was sorted by category. Also, proportion of farms/farmers present on the Internet was estimated by location.

3 Results

Results of this study suggest that only 18% of Latvian organic farmers/farms have some sort of presence on the Internet. Only 8% of Latvian organic farms have websites. Individual farmers tend not to have personal websites. Moreover, just about 2/3 of these websites are used for e-commerce and the other 1/3 is used for attracting tourists, rather than for product sales. Twelve percent of farmers/farms are present on social media. Individual farmers prevail over corporate farms at social platforms, with 30 individual farmers and corporate 6 farms in present study sample represented at social media. Interestingly, from corporate farms only those that had websites were present on social media. Presence on social media between individual farmers was mainly for private purposes, rather than commercial ones, with only 5 individual farmers from this study sample explicitly offering their products for sale. In contrast, corporate farms that had social media accounts used them exclusively for commercial purposes. The following social media platforms were used by Latvian organic farmers/farms: Facebook, Draugiem.lv, Instagram and LinkedIn. LinkedIn was used only by corporate farmers, with all but one farm that was present on social media had an account on a platform. Only a half of corporate farms that were present on social media also had a Facebook account. Individual farmers were represented at Facebook, Draugiem.lv, Instagram, however only Facebook was used for commercial purposes. Interestingly, dairy farms had much weaker presence on the internet than farms growing fruits and/or vegetables (i.e. growers) representing less than 1/5 of farms that had some sort of internet presence. All dairy farms from the sample used websites for attracting tourists, rather than for e-commerce. Interestingly, larger proportion of farms/farmers that were located closer to large cities (e.g. Riga, Daugavpils, Ventspils) (75%) were found to be represented on the Internet (having a website or Social media page) in comparison to those located further from larger cities (more than 50 km away), with just 20% of the latter present on the Internet.

4 Discussion

Results of present study suggest that majority of Latvian organic farmers do not harness opportunities of the e-commerce. Even those farms that have websites or are present at social media often use these digital tools for the purposes other than sales of their products. Private farmers lag behind corporate farms in use of digital media for securing access to the market. Also, farms located further than 50 km from larger cities tend not to be present on the internet. Overall, adaptation of e-commerce among organic agriculturalists in Latvia negligible. The findings of present study clearly indicate that farmers within organic farming industry may be losing out to competitors i.e. large convenient farms, due to failure to ensure effective access to the market. Unlike large convenient farms that collaborate with large retail chains, organic farmers could potentially benefit from direct access to the customers through internet channels [15]. Moreover, internet presence could facilitate branding of organic products, thereby allowing the producers of organic food gain higher profits and build a loyal customer base [16]. Underusage of the digital tools among Latvian organic farmers is indeed surprising, given the high level of internet penetration and coverage in Latvia. More research is needed to reveal the main motives of Latvian organic farmers, that prevent them from developing e-commerce. Still, the most plausible explanation to reluctancy of e-commerce adaptation is lack of knowledge and skills in digital trade and marketing among Latvian organic farmers. Thus, organizing training for the representatives of organic farming sector in e-commerce and could enable the farmers to effectively use the digital tools and this way increase competitive advantage of the sector ultimately delivering substantial benefits to Latvian economy and the environment [17]. Such training may be easily promoted across Latvia, as majority of organic farmers are members of professional organizations, such as LBLA (Latvian Biological Farmer Association) [18].

Another possible reason for underuse of digital tools by organic farmers is the fact that maintaining and promoting e-commerce websites may appear to be a very time-consuming task. Given the fact that farmers are extremely busy for more 6 months during the year, apparent labor-intensity of digital marketing may deter the farmers from adaptation of the tools. Perhaps, the best possible solution here could be application of adaptive management approaches, very common in nature conservation [19]. Adaptive management is based on collaboration among multiple stakeholders [20]. Given the fact that success of organic farmers is important for nature conservation and sustainable development, environmental non-governmental organizations (ENGOs) may be very keen to establish collaboration with the farmers. Alliances between ENGOs and farmers in turn have a great potential to enhance effectiveness of farmer e-commerce as ENGOs generally have a strong focus on communications [21]. In other words, this would allow organic farmers to “outsource” promotion of their e-commerce websites to ENGOs. A particularly effective approach here could be application of affiliate marketing tools, whereby influential ENGOs that have access to broad audiences of environmentally conscious people could place links to organic farmer websites at their internet pages and social media [22].

A particularly important aspect in promotion of e-commerce among Latvian organic farmers is to ensure a wide uptake [23]. For this purpose, adaptive management approach could provide a solution. Indeed, ENGOs and local authorities could implement a range of communicational activities promoting e-commerce among the farmers. In addition, trade organizations, such as Latvian Chamber of Commerce and Industry [24] and Investment Agency and Development Agency of Latvia [25] could organize series of events aiming to promote more active internet usage in Latvian organic farming industry. Overall, there needs to be a common effort among a range of stakeholders in order to help increase competitive advantage of Latvian organic farmers through promotion of digital strategy adaptation.

In line with Food and Agriculture Organization of the United Nations (FAO) recommendations [26], there is a need for national scale e-agriculture strategies to be implemented across the world. It is therefore highly recommended to promote e-commerce among Latvian farmers in the context of broader e-agriculture. The approach would allow to improve not only access to the market for organic farmers, but also increase their revenues and competitive and vantage through increasing efficiency of their operations. It is also important to note, that the findings of present study may be relevant not only to Latvia. Indeed, more research on Internet usage among organic farmers across Europe and perhaps all over the world, world allow making more confident conclusions based on comparative studies between different countries (e.g. comparing tends in Western Europe and Eastern Europe). Also, further research, could help persuading multiple stakeholders at national and transnational scales across entire regions to promote e-commerce and perhaps e-agriculture, thereby significantly contributing to implementation of the UN Sustainable Development Goals [27].

5 Conclusions

The aim of this study was to evaluate internet usage among Latvian organic farmers for commercial purposes. It was found that majority of Latvian farmers do not use e-commerce tools. This can potentially limit their access to the market, thereby reducing competitive advantage over larger conventional farmers. It is therefore recommended to promote usage of digital commercial tools among Latvian organic farmers. This would require providing training and establishing collaborations with multiple stakeholders. Application of adaptive management approach could ensure wide farmer uptake of e-commerce in Latvia, strengthening the industry and contributing to sustainable development and nature conservation across the country. Research in this area would benefit organic farming industry and sustainable development in general if scaled up at European or even Global level. Larger scale comparative studies are needed in order to draw confident conclusions and propose policy recommendations.

References

1. Secretariat of the Convention on Biological Diversity. Global Biodiversity Outlook 5. Montreal (2020)
2. ECA Biodiversity in Farming. https://www.eca.europa.eu/Lists/ECADocuments/AP19_09/AP_BIODIVERSITY_EN.pdf. Last accessed 08 Nov 2020
3. Richardson RB (2010) Ecosystem services and food security: economic perspectives on environmental sustainability. Sustainability 2(11):3520–3548
4. EC (2011) Our life insurance, our natural capital: an EU biodiversity strategy to 2020. Brussels
5. Peer G et al (2019) A greener path for the EU Common Agricultural Policy. Science 36(6452):449–451
6. Agence Bio. Organic Farming and Market in the EU (2019)
7. Dolveac L (2016) An overview on the supply chain for European organic food market. Econ Sci 9(58–2):315–330
8. Melovic B et al (2020) Attracting green consumers as a basis for creating sustainable marketing strategy on the organic market—relevance for sustainable agriculture business development. Foods 9(11):1552–1577
9. Okediran OO, Ganiyu R (2019) E-agriculture reviewed: theories, concepts and trends. FUOYE J Eng Technol 4(1):125–130
10. Central Statistical Bureau of Latvia. Agriculture in Latvia 2018. https://www.csb.gov.lv/en/statistics/statistics-by-theme/agriculture/agricultural-prices/search-in-theme/321-agriculture-latvia-2018. Last accessed 05 Nov 2020
11. Ventarl Statistical Bureau of Latvia (2020) 82 % of households have access to Internet. https://www.csb.gov.lv/en/statistics/statistics-by-theme/science-ict/computers-internet/key-indicator/82-households-have-access-internet#:~:text=In%202018%2C%2081.6%20%25%20households%20had,increased%20by%208.7%20percentage%20points. Last accessed 07 Nov 2020
12. Zalmane D (2020) Sharp increase in internet use in Latvia. Latvian Public Broadcasting. https://eng.lsm.lv/article/economy/economy/sharp-increase-in-internet-use-in-latvia.a354202/. Last accessed 08 Nov 2020
13. Latvian Agricultural Data Centre. https://www.ldc.gov.lv/lv/statistika/biologiska_lauksaimnieciba/. Last accessed 01 Sept 2020
14. RAND Corporation: A Million Random Digits with 100,000 Normal Deviates**.** Free Press (2000)
15. Huo Y, Mu H (2017) Research on the development of E-commerce model of agricultural products. MATEC web conference on 100. 13th global congress on manufacturing and management (GCMM 2016)
16. Simmons GJ (2007) "i-Branding": developing the internet as a branding tool. Mark Intell Plann 25(6):544–562
17. Darch H, Lucas T (2002) Training as an e-commerce enabler. J Workplace Learn 14(4):148–155
18. LBLA Homepage (2020) http://www.lbla.lv/. Last accessed 27 Nov 2020
19. McCarthy MA, Possingham HP (2007) Active adaptive management for conservation. Conserv Biol 21(4):956–963
20. Allen CR et al (2011) Adaptive management for a turbulent future. J Environ Manage 92(5):1339–1345
21. Molina-Murillo SA, Smith TM (2005) How much is too much? Exploring life cycle assessment information in environmental marketing communications. 24(1/2):199–224
22. Abdelhadi MH et al (2020) Impact of affiliate marketing on customer loyalty. J Facul Tour Hotels-Univ Sadat City 4(1/1):50–71
23. Kuehne G et al (2017) Predicting farmer uptake of new agricultural practices: a tool for research, extension and policy. Agric Syst 156:115–125
24. Latvian Chamber of Commerce and Industry (2020) https://www.chamber.lv/en/node/36. Last accessed 20 Nov 2020

25. Investment and Development Agency of Latvia (2020) https://www.liaa.gov.lv/en. Last accessed 15 Nov 2020
26. FAO (2020) http://www.fao.org/in-action/e-agriculture-strategy-guide/en/. Last accessed 01 Dec 2020
27. The UN (2020) https://www.un.org/sustainabledevelopment/sustainable-development-goals/. Last accessed 12 Dec 2020

A Machine Learning Approach on Earthmoving Fleet Selection

Case Study: Burj Hammoud Landfill Project, Lebanon

Anthony Abdelmassih, Rafic Faddoul, and Fadi Geara

Abstract This research underlines the importance of simulation modeling by presenting a fleet selection optimization methodology featuring a detailed simulation of earthmoving projects and a genetic optimization algorithm. Two genetic algorithm techniques for fleet selection are proposed and compared in a load and carry application. The simulation model takes into account specific cost criteria that go into equipment selection such as the ownership and operating costs of each machine. The impact of different scenarios on the Profit & Loss function (fitness value) and machine selection is extensively detailed and analyzed. The research work underlines topics such as machine size class and maintenance and repair. The methodology is applied to a real case study that took place in Lebanon. The proposed approach is shown to yield a significant increase in the efficiency of the selected construction equipment fleets.

Keywords Simulation · Optimization · Genetic algorithm · Machine learning · Earthmoving operation · Equipment selection · Maintenance and repair

1 Introduction

The construction business is an ever-changing industry facing increasingly bigger challenges, complexity, and above all fierce competition. Whether it's a vertical construction as in the race to build the tallest building, or horizontal construction as when delivering nature-challenging airport, highways, dams, or bridges, the contractors should pay extra attention to minute aspects of their project in order to meet the

A. Abdelmassih (✉) · R. Faddoul · F. Geara
University St. Joseph (USJ), Beirut, Lebanon
e-mail: anthony.abdelmassih@net.usj.edu.lb

R. Faddoul
e-mail: rafic.faddoul@usj.edu.lb

F. Geara
e-mail: fadi.geara@usj.edu.lb

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_75

projects' timeline, budget, and quality requirements. Statistics have shown that 40% of construction projects miss their timeline, 50% go beyond the budget and 30% are never completed [1]. According to the UK Department of the Environment, Transport, and the Regions [2], only 45% of construction projects are being completed within budget. In order to be competitive, contractors have to cut their margins, increase the efficiency of their workforce and assets, and keep their costs at their lowest.

There are three major inputs in the building construction process: (i) construction equipment, (ii) labor, and (iii) materials [3].This paper focuses on the optimization of the selection and operation of the construction equipment fleet. Modeling and simulating construction operations, made possible through ICT, allow to simulate earthwork operations, to visualize the costs, and to make necessary analysis and improvements, aiming to optimize the fleet operations. The slow adoption of simulation in construction was partly due to the complexity of the construction process itself and consequently to the effort required to prepare a realistic model capable of capturing all the relevant parameters of the problem [4]. However, in the last decade, soaring hardware capabilities and software efficiency resulted in simulators coming a long way in making construction models faster, easier, and more user-friendly. Devising an optimal planning process for equipment selection and construction methods is required both prior, and during construction works. The optimization problem is a multi-objective constrained problem [1] in the sense that one must respect schedule constraints, minimize costs and environmental deleterious impacts, maximize work quality, etc.

This paper proposes a methodology combining simulation and genetic optimization for an optimal selection of construction equipment fleet. The following sections will include (i) literature review (ii) Simulation modeling of construction sites, (iii) genetic algorithm. The proposed methodology is illustrated in a numerical application where a load and carry operation is optimized.

2 Literature Review

Many researchers have tackled the simulation-based optimization approach in construction applications. Marzouk [5] proposes a simulation-based Genetic Algorithm (GA) using discrete-event simulation methodology and defining a mixture of qualitative and quantitative optimization variables. The optimization algorithm, using elitism, allows the selection of an optimal fleet combination. Marzouk defines three fleet scenarios where each scenario defines the types of the involved equipment from each category (haulers, loaders, spreaders, and compactors) and hence their respective specific characteristics. The algorithm does not allow chromosomes mixing different machine types for a given category.

Fu [6] simulates a load and carry operation and adopts the total cost of ownership (TCO) as the fitness to evaluate each of the fleet combinations yielded by the genetic algorithm. The aim of the genetic algorithm was to minimize the TCO while maintaining a minimum production level. While Fu methodology allows the optimization of equipment type within each category, it does not allow for more than one type for each equipment category in the resulting optimal fleet. Shi [7] applied a neural network approach to predict the results of future projects, based on previous ones. The output of the system consists of four parameters among which the production rate and utilization rate, while fleet characteristics (such as number of trucks, loading time, and capacity of the trucks) are used as input variables. Gehbauer et al. [8] based their research on bionics, studying the techniques and technologies of nature and learning from these. The research work investigates the application of ant colony in the construction industry, highlighting behavioral interactions, communication between the stakeholders, and describing ant-behavior algorithmic approaches. Similarly, within bionics, Zhang [9] proposed a two-staged ranking selection criteria based on particle swarm optimization (PSO) for the selection of the optimal equipment configuration in earthmoving operation.

The proposed methodology in this paper lies within two folds; simulation and genetic optimization. The simulation part of the proposed methodology allows detailed mathematical expression of the fitness while using a thorough breakdown of the costs that are generated by owning and operating a fleet of construction machines. The proposed methodology also takes into account global operation costs such as project delay costs. The proposed algorithm is stochastic, in the sense that relevant uncertainties related to cost, operation times, maintenance costs, repair costs etc., are taken into account. As for the genetic optimization part, two genetic algorithms are proposed in this paper. In contrast with Fu [6] and Marzouk [5], the proposed algorithm will not constrain the optimal solution by any predefined sub-population or fleet scenario. For instance, instead of having only one gene for the hauler type, the presented methodology will allow for chromosomes containing several genes for haulers, where each gene defines a specific machine type. Also, the proposed methodology will offer a wider selection of equipment types. The proposed model can select an unconstrained number of loading and hauling units from a pool of several loading and hauling types.

3 Methodology—Simulation Modeling and Genetic Algorithm

In this paper, we proposed a methodology for selecting an optimal fleet configuration for construction equipment using GA. The fitness of the GA is calculated using discrete event simulation. The operating and ownership costs including maintenance and repair costs are finely detailed in the fitness of each candidate solution. Two different genetic algorithm codes are evaluated. All input parameters for the model are based on real data collected from real-world construction sites and from manufacturers' specifications. That was possible thanks to the hands-on experience of one of the authors in the field.

3.1 Simulation Model

Simulation has been a hot research topic for the last few decades. Several simulation languages were introduced such as CYCLONE, Stroboscope [10], and AP2 Earth and Simphony [4]. The construction and earthmoving industry was in need for simulation-based optimization approaches due to many convoluted aspects specific to this industry making it intractable to resort to analytical closed form solutions. Construction projects, and in particular earthmoving works, are complex systems during which the uncertain availability of several resources must coincide in time and space to carry out particular tasks. Construction projects are affected by uncertainties such as weather conditions, work accidents, breakdowns of the machines, and availability of external resources that can destabilize the synchronization of site operations. Safety and environmental regulations and constraints add to the planning complexity of construction projects. Global constraints increase the interdependencies among decision variables. For example, budgetary constraints and/or reliability constraints introduce interdependency between decision variables for otherwise independent tasks [11, 12]. This renders construction operations even more complex and simulation-demanding where traditional and conventional optimization strategies become useless.

Three different simulation modeling approaches are usually employed; namely: (i) discrete event simulation (DES), (ii) agent-based modeling (ABM), and (iii) system dynamics (SD). Each modeling method serves a level of abstraction and details, and can more accurately map a specific scenario or problem. The proposed simulation is based on the DES modeling approach. DES is needed for maximum details and low abstraction levels where precisely defined operations are modeled. It consists of dividing a continuous process into discrete parts to simplify the analysis. Specifically, it is used to capture the interaction between resources (potentially random) and activities (potentially having random outcomes). A series of blocks is used to define the transitions within the load and carry applications. The hauling units are defined as agents, while the loading units are seized from the resource pool as resource pools.

3.2 Genetic Algorithm

While simulations can model specific scenarios or operations given a specified set of input parameters, GAs are needed for finding the optimal set of parameters given a specified "fitness" function. Parameters sets are denoted "Chromosomes" in the GA paradigm. As such, the proposed model can be considered as an unsupervised machine learning approach which learns from simulated datasets. Genetic algorithm, originally introduced by John Holland in the 1960s and inspired by natural selection, is an optimization method that allows the evolution from one population to a new population using crossover and mutation according to Mitchell [13]. What ensures that the new population has a greater probability of being a more fit population than the previous population are the selection methods, which favor the selection of the fittest parents. In the "roulette wheel" method, the chance of the chromosomes to reproduce is proportional to their fitness. "Elitism" allows the genetic algorithm to retain the best chromosomes i.e., the chromosomes with the highest fitness to the next generation avoiding that they will be lost if not selected to reproduce or if genetically modified in the processes of mutation and cross over. "Elitism" has also been adapted in the presented methodology to improve the genetic algorithm's performance and ensure the generations' fitness is continuously converging toward a more fit generations.

In our proposed methodology, chromosomes represent specific fleet compositions. The genes in each chromosome represent the number of a specific predefined machine type, whether a wheel loader, excavator, articulated hauler, etc. The earth-moving simulation model returns the profit and loss (P&L) used as the fitness required by the GA. The P&L is calculated based on the volume of moved earth, the operating and ownership cost of each machine. This paper presents two different GA models, referred to as "GA 1" and "GA 2." In "GA 1," a chromosome consists of a one-dimensional array of genes. Each gene of the chromosome is associated with a random integer within a pre-defined range representing the number of units of specified machine type. "GA 1" uses a single point crossover operator. The mutation operator of "GA 1" (having a probability of 15%) consists of randomly choosing a gene in the chromosome and assigning a random integer (within the specified range) to that gene. "GA 1" uses the "roulette wheel" selection procedure to choose the chromosomes that will reproduce to generate the population of the next generation."GA 2" uses a procedure for generating chromosomes based on Bean's [14] random keys approach. In Beans random key approach, chromosomes are represented as an indexed string, or vector, of randomly generated real numbers in the interval [0, 1]. The index of each value represents a particular equipment instance. A single point crossover operator is used. The mutation operator consists of assigning a uniform random number in the interval [0, 1] of a randomly selected gene. Solutions are thus represented by a permutation vector. To decode a chromosome into a solution, the values of the genes are sorted in ascending order (one might choose a different sorting as for example a descending order sorting), thus producing a permutation corresponding to the indices of the sorted elements. Equipment associated with the top-sorted indexes are selected until one reaches an index where at least one of the constraints is no longer satisfied. Through the dynamic of the genetic algorithm, indexes (i.e., equipment) that should be selected will evolve low key values. Relevant constraints for our optimization problem might be budgetary constraints, availability of the equipment, etc.

Hence, in contrast with the previous algorithm, where every gene in the chromosome defines the number of every construction machine type, in Bean's proposed model, every gene represents a single instance of a machine unit. A chromosome according to Bean's model will have more genes than "GA 1" i.e., one gene for each machine instance. Random keys approach allows genetic algorithms to have more efficient mutations and crossover of the chromosomes, and thus faster convergence toward the best fleet configuration and scenario. We will test such a claim in our numerical application.

4 Case Study—Fleet Optimization in a Landfill Application

The simulation model along with the genetic algorithm code designed and described previously in this paper has been tested over a project taking place in Burj Hammoud, Lebanon, located northeast of Beirut on the edge of the Eastern Mediterranean sea. This project worth more than 100 million USD consists of removing an old 3.5 million cubic meter, 47 m high waste dump. According to the World Bank [15], this waste dump has been accumulated throughout the Lebanese civil war from 1975 to 1990. After the civil war, Burj Hammoud landfill remained open until it was stabilized in 1997. The landfill which later became a "mountain" was still unresolved as it is sinking into the sea and open to polluting the air, until the award of the project was mentioned. The period to remove the waste dump is 24 months.

4.1 Site Layout and Problem Description

The main task in this project is to load the trucks with the sorted waste and haul it on an average of two kilometers off-road before being dumped into specific isolated cells and returning empty to the loading site. A very wide selection of machinery, mainly from the large size class, operate on this project ranging from wheel loaders and excavators to articulated haulers. Figure 1 is a bird-eye view from the site in Lebanon. Selecting the right fleet of equipment is crucial to ensure profitable finances and efficient site operation. The project being located on the Mediterranean Sea shore

Fig. 1 Site landscape

is constrained by tight and severe regulations related to environment and protecting the sea life, not only on the Lebanese coasts but also for the neighboring countries.

4.2 Control Variables

The simulation model has both fixed and variable parameters. Two types of fixed parameters can be defined; parameters that are related to the site and operation scenario, and parameters that are related to the ownership and operating costs of each machine (Table 1). Variable parameters are genetically modified and controlled using an evolutionary algorithm.

Among the first type of fixed parameters are the hauling distance, price earned per cubic meter of material moved, diesel cost, and the total volume of material to be transported. The second set of fixed parameters defines the ownership and operating aspects of running each of the available machine models and types.

Table 2 is a list of construction equipment, selected from Volvo Construction Equipment product range, being used in our numerical application along with their specific machine features and related costs, both ownership and operating costs. Calculations of the maintenance cost and repair cost are detailed in the next section, however, it is worth noting that the hourly maintenance cost is calculated based on 250 h service intervals. The average speed of the trucks used in the simulation is based on the maximum speed of each articulated hauler. Calculating the cycle time, fuel consumption, bucket capacities, and the haulers' speed is based on the Performance Manual [16] available from Volvo CE.

Table 1 List of fixed parameters

Fixed parameters			
Site		Machines	
Hauling distance	km	Speed	km\hr
Price per cbm	$	Cycle time	sec
Total volume	m3	Capacity (bucket/trailer)	m3
Project deadline	month	Fuel consumption	L/hr
Penalty per day	$	Repair (mean sigma)	$
Operator cost	$/hr	Maintenance cost	$/hr
Diesel cost	$/L	Ownership cost	$/hr

Table 2 List of Volvo Construction Equipment selected for the simulation

Articulated Haulers	Max speed (km/h)	Capacity Heaped with tailgate (m3)	Fuel consumption	Maintenance mean ($)	Repair cost mean ($)	Repair cost sigma ($)	Ownership cost depreciated over 7 years ($/hr)
A60H	54.9	35.1	50	7	1000	500	9.78
A40F	57	24.7	36	4.9	800	400	6.52
A25F	52.7	15.9	20	3.38	650	300	4.89
Wheel loaders and excavators	Bucket capacity (m3)	Cycle time (sec)	Fuel consumption (L/hr)	Maintenance cost ($)	Repair cost mean ($)	Repair cost sigma ($)	Repair cost sigma ($)
L350F	8.4	36	33	4.67	1200	500	7.34
L180H	4.4	36	20	3.08	600	200	4.24
EC750D	4.4	19.75	45	3.38	1000	500	8.97
EC380D	2.25	16.3	23	2.85	500	200	4.08

4.3 Preventive Maintenance and Repair

Preventive maintenance and repair are included in the simulation knowing that their related expenses are the most uncertain cost among both ownership and operating costs [1]. In our application, preventive maintenance cost is the cost of filters and lubricants. The maintenance costs are sourced from average market prices. Four different preventive maintenance scenarios are considered (based on the service interval time for maintenance, whether 250 h, 500 h, 1000 h, or 4000 h). An integer gene having four different possible values is included in the chromosome to define the maintenance scenario. The maintenance costs calculated in Table 2 based on the 250 h scenario will be proportionally adjusted to fit the maintenance service interval. As for the repair cost, it will change according to the selected preventive maintenance scenario. If a machine is subject to strict maintenance programs, it will become less susceptible to failures and breakdowns. The total repair cost will be lower on a well-serviced machine than on a poorly serviced one. We assume that the time to breakdown follows an exponential distribution. We assume that the repair cost will follow a normal distribution where each machine has a specific mean and standard deviation for the repair cost. These values are gathered from real service histories of machines from the same or similar models.

4.4 Fitness

The fitness being quantified and optimized is the Profit & Loss (P&L) resulting from considered fleet selections. The Profit & Loss allows to capture in one objective function several aspects of the project such as the TCO, the allocated budget for this job, the operator cost, the diesel costs, the emission costs incurred on the project, the project execution time, and the impact of the different maintenance programs. The fitness will be quantified in US Dollars. The purpose of the genetic algorithm is to find a fleet combination that maximizes the profit, i.e., gets the job completed using the lowest operating cost. The P&L will be

$$P\&L = TM \times \text{Pcbm} - \text{EqCost} - \text{PenCost} \tag{1}$$

$$EqCost = \sum_{i=1}^{n} EqOpHr_i \times (MaintCost_i + OpCost_i + FuelCost_i) + \text{RpCost}_i + \text{OwnCost}_i \tag{2}$$

where

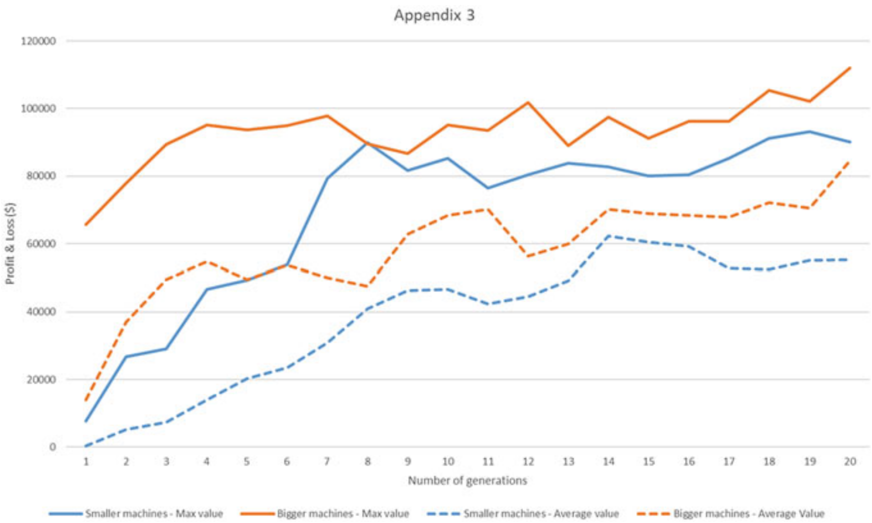
P&L is the profit and loss ($), TM is the total material transported (m^3), Pcbm is the price per cbm transported ($/$m^3$), EqCost is the cost of all equipment: ownership and operating cost ($), PenCost is the penalty cost if the project goes beyond deadline ($), i is the equipment number representing wheel loaders, excavators and articulated haulers, MaintCost is the maintenance cost of each machine per hour of machine operation ($/hr), OpCost is the cost of operator per hour of machine operation($/hr), FuelCost is the cost of fuel consumed per hour of machine operation ($/hr), RpCost is the cost of repair per hour ($), OwnCost is the ownership cost combining both a fixed cost and an hourly rate of ownership ($). The OwnCost imposes a fixed cost for every equipment mobilized to the project in addition to an hourly rate based on the machine operating hours. The fixed cost is derived from the initial price of the equipment normalized to this project, based on the total simulation time.

4.5 Main Findings

Figures 2, 3, and 4 depict the results obtained from running the proposed methodology. The x-axis shows the number of generations (20), each generation consists of 20 chromosomes (simulation), while the y-axis shows the fitness. Each chart presents two types of curves; the maximum value of a chromosome and the average value of all chromosomes in each generation. It can be seen that in the early generations, the fitness has low values. As the generations evolve, the fitness improves until the curve reaches a flat where the best fitness is reached. It should be mentioned that even though elitism is used in the optimization method and thus the fittest chromosomes are saved to the next generation, it is not possible to record an identical fitness value for two similar chromosomes, due to the stochastic nature of the simulation model. In the following sections, several scenarios have been tested and benchmarked against each other. Below is the list of simulated scenarios:

1. Machine size
2. Genetic algorithms methodologies
3. Maintenance and repair.

Machine size. To capitalize on Marzouk [5], Fig. 2 presents results for two new scenarios. In the first scenario, only the smaller size machines can be selected (L180H, EC380D, A40F, A25F), while in the second scenario, bigger size machines are selected (L350F, EC750D, A60H, A40F). Marzouk [5] adopted fleet scenarios where he pre-matched wheel loaders and haulers. The simulation run offering only bigger machines converged to a higher fitness compared to the runs offering only smaller machines. Bigger machine classes can replace several smaller machines, offer lower TCO per unit of material moved, and incur lower operator's costs. Marzouk [5] methodology, will not always offer improvements to the fitness and this will be shown in future research works. While contractors tend to unify and consolidate equipment

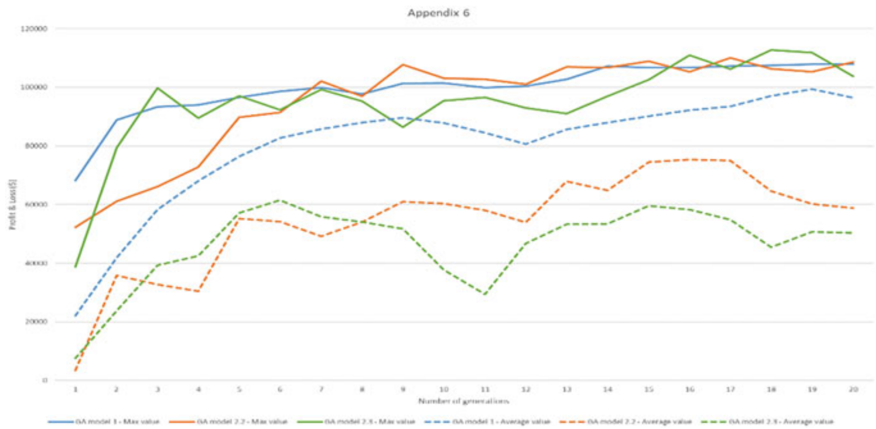


Scenarios	Fleet							
Machine size	L350F	L180H	EC750D	EC380D	A60H	A40F	A25F	Profit & Loss ($)
Smaller machines	X	3	X	2	X	8	13	93132
Bigger machines	2	X	2	X	11	2	X	112708

Fig. 2 Machine size

models, the aim of the presented methodology is to investigate machine selection variation with the widest machine offer. When parameters are modified, machine selection will adapt and the wider the available machine selection, the greater the versatility in finding the best fitness.

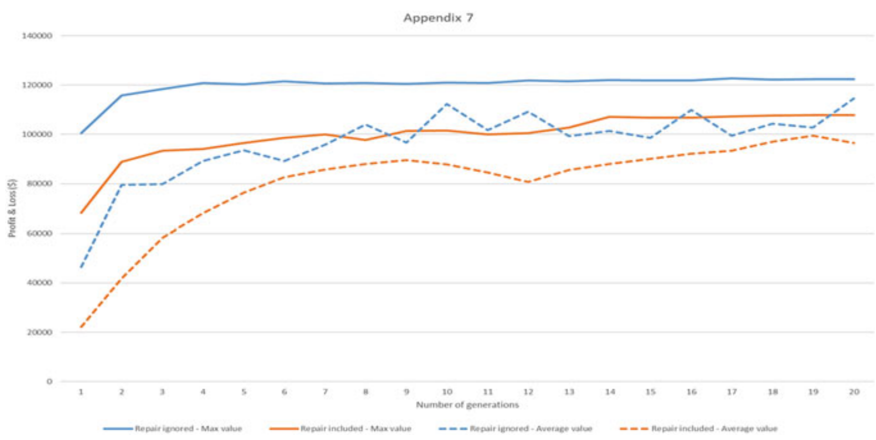
Genetic algorithms methodologies. As described in Sect. 3.2 of the paper, two different genetic algorithm methodologies are coded and tested “GA 1” and “GA 2.” The methods differ from one another in the way the genes encode the fleet information. “GA 2” is based on Bean’s [14] Random Key for Sequencing. The latter was coded in two different methods referred to in Fig. 3 below as “Model 2.2” and “Model 2.3.” While all methods converge toward global near-optimal solutions, it can be noted that Bean’s methodology resulted in improved optimums at earlier generations (as in in “Model 2.2”) compared to the method employed in “GA 1” referred to below as “Model 1.” It can also be noted that “GA 2” has a higher standard deviation compared to “GA 1” as a result of the methodology used in creating the chromosomes for each iteration. This is concluded based on a low average curve but high maximum value. This justifies a slower but more consistent convergence of “GA 1” toward the best fitness as opposed to a faster finding of the optimum, but a less consistent and “shakier” curve shape. It should be highlighted again that different fleet configurations can offer similar optimized fitness, adding more choices and freedom in fleet selection.



Scenarios	Fleet							
Genetic Algorithm	L350F	L180H	EC750D	EC380D	A60H	A40F	A25F	Profit & Loss ($)
Model 1	1	0	2	0	7	4	7	107846
Model 2.2	1	1	1	0	7	4	4	110014
Model 2.3	1	2	1	0	9	5	1	112724

Fig. 3 Genetic algorithms

Maintenance and Repair. Figure 4 presents the result of two scenarios: (i) the blue curve takes into account the preventive maintenance while ignoring the failures and repair costs; (ii) the red curve takes into consideration possible failures of the machines depending on the preventive maintenance program. As mentioned in Sect. 4.3, the proposed methodology designs four different maintenance programs (mP: 1, 2, 3, and 4), where mP:1 adopts the strictest preventive maintenance schedule, thus the highest maintenance cost per operating hour, while mP: 4 indicate almost no preventive maintenance and thus the lowest maintenance cost per hour. On the other hand, mP:1, in contrast with mP:4, defines a failure function that lowers the chance of machine facing unscheduled stops and breakdowns. The breakdown time in the presented model is considered an idle time for the machine. It can be seen, that in the scenario where failure and repair are simulated, the chromosome with the best fitness, converges toward mP:1. On the other hand, when failure and repair are ignored, the chromosome uses mP:4, meaning lowest maintenance cost since machine failure is not possible. In the case where failure and repair are included, the reason that mP converges to 1, can be assigned to the fact that the maintenance cost is lower than the repair cost (failure cost and interval), dictated by the assumptions presented in Table 2 (repair cost sigma and mean). In other words, the optimization algorithm adopts the highest maintenance cost, ensuring the lowest repair cost. Further development in ICT and data analytics could help predict better the repair cost and the time to failure of specific machines in a particular application, allowing more informed decisions when it comes to preventive maintenance.



Scenarios	Fleet								
M&R	L350F	L180H	EC750D	EC380D	A60H	A40F	A25F	mP	Profit & Loss ($)
Ignored	3	0	0	0	4	9	0	4	122706
Included	1	0	2	0	7	4	7	1	107846

Fig. 4 Maintenance and repair

5 Summary and Future Scope

This paper presents a fleet selection optimization methodology featuring a detailed simulation of earthmoving projects and a genetic optimization algorithm. The flexibility of the modeling allows for an easy adaptation of the approach to various earthmoving sites. Two different genetic algorithms are proposed and compared. The impact of different scenarios on the P & L function (fitness value) and the machine selection is extensively detailed and analyzed. This paper underlines topics such as machine size class, maintenance and repair. As mentioned in the last section, ICT and data analytics from real applications can support some of the assumptions in the simulation model, allowing more informed decisions, especially in maintenance and repair. A promising future research direction would be to investigate the impact of the adoption of future technologies on machine selection and machine size class, both loading and hauling units; future technologies such as automation and electrification. Moreover, safety is a major topic that should be addressed whether in automated or human-operated machines.

References

1. Peurifoy R, Ledbetter W, Schexnayder C (1996) Construction planning, equipment, and methods, 5th edn. The McGraw-Hill Companies Inc., New York
2. Jackson S (2002) Project cost overruns and risk management. school of construction management and engineering, The University of Reading
3. Nunnally S (2010) Construction methods and management, 8th edn. Prentice Hall, Upper Saddle River, N.J.
4. AbouRizk S, Hajjar D (1998) A framework for applying simulation in construction. Can J Civ Eng 25(3):604–617
5. Marzouk M, Moselhi O (2002) Simulation optimization for earthmoving opurations gsing aenetic Algorithms. Constr Manag Econ 20(6):535–543
6. Fu J (2012) Simulation-based optimization of earthmoving operations using genetic algorithm. Division of Traffic and Logistics. Royal Institute of Technology (KTH), Stockholm, Sweden
7. Shi JJ (1999) A neural network based system for predicting earthmoving production. Constr Manag Econ 17(4):P463-471
8. Gehbauer F, Stock P, Steffek P (2006) Swarm intelligence and ant colony approach—can they be applied in optimizing construction processes? Proceedings IGLC 14:15–30
9. Zhang H (2008) Multi-objective simulation-optimization for earthmoving operations. Autom Constr 18:P79-86
10. Martinez JC (1996) Stroboscope: State and resource based simulation of construction operations. PhD dissertation, MI: University of Michigan, Ann Arbor
11. Faddoul R, Soubra AH, Raphael W, Chateauneuf A (2013) Extension of dynamic programming models for management optimisation from single structure to multi-structures level. Struct Infrastruct Eng 9(5):432–447
12. Faddoul R, Raphael W, Chateauneuf A (2018) Maintenance optimization of series systems subject to reliability constraints. Reliab Eng Syst Saf 180:179–188
13. Mitchell M (1999) An introduction to genetic algorithms. A Bradford Book The MIT Press, Cambridge, pp 124–130

14. Bean J (1994) Genetic algorithms and random keys for sequencing and optimization. Operat Res Soc Am 6(2):154–160
15. World Bank (2011) Republic of Lebanon, Country Environmental Analysis. 62266, p 31
16. Volvo Performance Manual (2015), Global Marketing, Volvo Construction Equipment

Modified TPACK Framework for Teachers' Efficiency, Students' Performance and Students' Engagement

Daniel Lai, Sook Ling Lew, and Shih Yin Ooi

Abstract In the digital era, the environment of the classroom had gone through a drastic change where technology is slowly being introduced. Instead of conducting classes and lessons through the traditional or rather conventional way of teaching, the integration of educational technology in the classroom is being practiced recently. As teachers are the person who is being dedicated to deliver and transfer knowledge to students, their capabilities and abilities in conducting classes and lessons effectively are being spotlighted by society. Hence, for teachers to be well prepared in conducting classroom activities, a modified TPACK framework is proposed to assist teachers from seven aspects, which are technological knowledge, pedagogical knowledge and content knowledge, technological content knowledge, technological pedagogical knowledge, pedagogical content knowledge and last but not least technological pedagogical content knowledge since the thoroughness of teachers' knowledge on the stated seven aspects will be affecting how they handled classes and lessons effectively. Since the existing TPACK framework in assessing teachers' general knowledge about technology and how it interrelates with their pedagogical and content knowledge, the integration of the mobile interactive system with on-screen writing can further narrow down the scope focusing on how skillful and efficient are they to integrate mobile technology into classes and lessons. As a solution, this study found that the implementation of a mobile interactive system with an on-screen writing feature as the integration of educational technology in the classroom tailored for students who needed extra attention is effective in terms of preparing teaching plans and teaching materials. Therefore, this study will be resulting in different levels of teachers' efficiency, students' engagement and students' performance in classroom teaching and learning. This study intends to identify how TPACK framework elements interrelate with each other and the relationship between modified framework and teachers' efficiency, students' performance and students' engagement.

D. Lai · S. L. Lew (✉) · S. Y. Ooi
Multimedia University, Melaka, Malaysia
e-mail: sllew@mmu.edu.my

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_76

Keywords TPACK framework · Technology integration · Teaching and learning process · Interactivity and engagement · Mobile interactive system · Mobile technology · Educational technology

1 Introduction

1.1 Background

Due to the vast evolution of technology, many sectors are undergoing gradual changes including the education sector especially the integration of technology in the teaching and learning process. In the traditional classroom environment, there is minimal to no involvement of technology where the teaching and learning process is leaning toward the teacher-centered approach [1]. As such, the quality of learning is depending on the teacher's quality. The motivation of learning remains uncertain among students because the source of motivation is coming from the teacher himself or herself. Hence, the interaction between teachers and students remained minimal. Additionally, it has been proven that the relationship between students and their behavior towards learning via mobile technology is showing a positive attitude [2]. Moreover, it is essential to assess teachers' efficiency in conducting classes in conjunction with the presence of technology. Therefore, the TPACK framework is adopted and modified in this study to assess teachers and students whether teachers are efficient enough for conducting classes with the integration of technology and how students being beneficial through modified TPACK framework in terms of their academic performance as well as their engagement in classes conducted.

1.2 Problem Statement

In today's classroom, the conventional way of teaching is still practiced in most schools. This kind of classroom has existed for decades and can be found in all levels of education. Students who undergo the one-way teaching method, which is commonly practiced in the conventional classroom, are accepting that knowledge without further analysis and interpretation in the received content [3]. Moreover, the interaction between teachers and students will be very minimal, which leads to limited engagement between teachers and students. All in all, two major problems are being concerned in this study, which are:

1. Lack of engagement between teachers and students in the modern classroom.
2. The readiness of teachers' knowledge in facilitating classes with the integration of technology in the modern classroom.

1.3 Objectives

In this study, the TPACK framework is being adopted and served as an instrument to assess teachers from eight different aspects. These aspects include technological knowledge, pedagogical knowledge, content knowledge, technological pedagogical knowledge, technological content knowledge, pedagogical content knowledge and technological pedagogical and content knowledge. This study is intended to modify the adopted TPACK framework, which resulting in addressing teachers' efficiency, students' performance and students' engagement. Hence, the objectives of this study are the following:

1. Identify how the TPACK framework elements interrelate with each other.
2. Identify the relationship between modified framework and teachers' efficiency, students' performance and students' engagement.

2 Literature Review

2.1 Technology Integration for Education

In the twenty-first century, education, ICT, innovation and science technology are the main pillars of knowledge society [4]. Nowadays, students are considered digital natives because interactive technology has already been part of their daily lives [5]. In the past, schools are relying solely on teachers to deliver knowledge to students in the conventional method. At present, students are the generation that is exposed to technological devices and applications in which these technological related systems had changed their lifestyle gracefully. Nevertheless, the influences that can affect and limit the integration of technology in the classroom can be classified as external and internal influences [5].

From the aspect of external influences, poor infrastructure is considered as one of the major external influences. The most ideal way for students to become more effective in learning is to let them have more flexibility in terms of time and place to learn. Besides, the integration of technology into the classroom serves two principles, which are constructivist and sociocultural principles [5]. As for constructivists, learners are creating knowledge based on their interactions with the environment, building on existing knowledge and relevancy of the content is also taking into consideration or even instructional activities. As for sociocultural, the platform is provided by technology and instruments for engagement with one or more individuals immediately. However, teachers often trained within a short period and asked to explore the rest of it on their own. This causes lower efficiency as teachers have limited knowledge in selecting the more appropriate technology to support teaching and learning.

As for internal influences, without effective technological integration, it is unreasonable to claim it is possible to reach these goals without the appropriate online technologies. Despite the availability and accessibility of technology nowadays, teachers are those who neglect and hesitate to integrate technology into the classroom. Teachers tend to stick with the old-fashioned way of teaching and shifting to a new method and teaching environment can create frustration making them straying themselves away from the use of technological devices.

2.2 TPACK Framework

Nowadays, schools are mostly equipped with technological equipment in terms of hardware, software, online materials for both teaching and learning. Teachers can guide students into global technology and be competitive [6]. Additionally, online teaching and learning is focusing on two major sectors, which are pedagogy and educational technology. Lecturers need to be well equipped with the necessary knowledge and skills to integrate technology into their teaching process [7]. Hence, the TPACK model that was introduced by Mishra and Koehler [7] is ideal for identifying technology, pedagogical and content knowledge of teachers and transforming them to be more effective in teaching and learning [8]. TPACK model allows teachers to recognize and understand the practical way of delivering teaching materials and contents with the use of technology.

Based on the TPACK framework, the crucial elements of the TPACK framework can be categorized as technological, pedagogical and content knowledge, which represents the three vital interactions of the model. Teachers who seek to exploit TPACK for becoming effective teachers need to have content knowledge of their discipline, pedagogical knowledge to effectively transfer their ideas to learners, and the knowledge to employ appropriate educational technologies for teaching and learning [9]. However, this knowledge is treated differently and uniquely while the three keys knowledge of TPACK remains, which are technological, pedagogical and content knowledge.

3 Method

In this study, a systematic literature review (SLR) is considered in identifying research questions. The lack of prior knowledge about the research fields and hardly identify related papers and works for scientific dissemination is a major issue faced when conducting research. The research can be more convincing if there are researchers from similar fields sharing the same goals and issues to be solved to be taken as references. Hence, SLR is essential as it acts as a foundation for newcomers in the scientific research field. It gives new researchers a direction in an early state where status for the problems changes according to their prior or individual knowledge about the problems. It changes from "MY" current state of the problem to "THE" current state of the problem at the end of the entire process [9]. Furthermore, it is rather complex to search for keywords efficiently [9]. In other words, SLR allows newcomer researchers to rely on their understanding of that particular context and moving on to a more in-depth state promptly. The elements proposed by Hernández (2014) are objectives, research questions, justification of the study, feasibility of study and deficiencies in the knowledge of the problem [10]. The drafted questions provide a big picture of the research. The research questions of this study are as follows:

Research Question 1: How does technology relate to the teaching and learning process?

Research Question 2: How does the introduction of teachers' efficiency, students' performance and students' engagement into TPACK framework elements affect teaching and learning experience?

4 Results

4.1 Modified TPACK Framework

TPACK is known as technological pedagogical and content knowledge, which is a framework created by Mishra and Koehler in 2006 defining a total number of seven components regarding teachers' knowledge where teachers are required to teach with the implementation and integration of technology in conjunction with their knowledge for a specific area. These elements include technological knowledge (TK), pedagogical knowledge (PK), content knowledge (CK), pedagogical content knowledge (PCK), technological content knowledge (TCK), technological pedagogical knowledge (TPK) and technological pedagogical content knowledge (TPACK). As for basic elements, TK, PK and CK fall under this category whereas secondary elements are TCK, PCK, TPK and TPACK. As for the proposed TPACK framework, it has additional elements introduced, which are teachers' efficiency, students' performance and students' engagement (Fig. 1).

According to Fig. 2, teachers' efficiency is dependent on their technological knowledge with the implementation of the mobile interactive system, which is capable of conducting on-screen writing. As for students' engagement, it is dependent on teachers' pedagogical knowledge whereas students' performance is depending on teachers' content knowledge. In terms of TK, it is related to the technological devices and tools, which are available to be used by teachers in the classroom. The familiarity of teachers toward the technological devices and tools is taken into consideration as it will affect the efficiency of teachers operating the technological devices and tools as well as choosing the appropriate ones to cope with the lessons. In terms of PK, it is related to the methods that teachers implemented into the lessons, which best fits students' learnability in the classroom. Teachers need to consider which kind of strategies are well tailored for students meeting their needs and the requirements for teaching plan as well. Teachers are required to have a thorough understanding of best practices for teaching and proposing teamwork in a small group. As for CK, it is about the knowledge that teachers wanted to deliver to students. What to be taught in classrooms by teachers conducting lessons and teachers' knowledge on that particular subject are taken into consideration. As a teacher, one must have a deep and thorough understanding of the subject to be taught ensuring the validity and relevancy of knowledge to be delivered is sufficient and appropriate.

Moving on to the secondary elements, PCK is the combination of PK and CK, which is focusing on the delivery method and content, respectively. PCK is where teachers are having an understanding of which practices are best suited for teaching a certain subject or content. These practices are specially tailored for students who need special care rather than addressing it generally. As for TCK, it is the combination of technological devices and tools with the content to be delivered. Teachers are required to be familiar with the technological devices and tools, which will be used

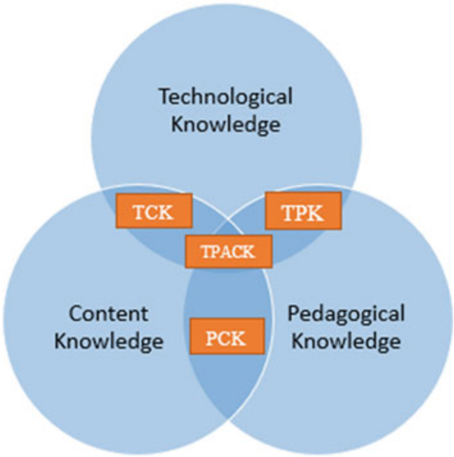


Fig. 1 TPACK framework

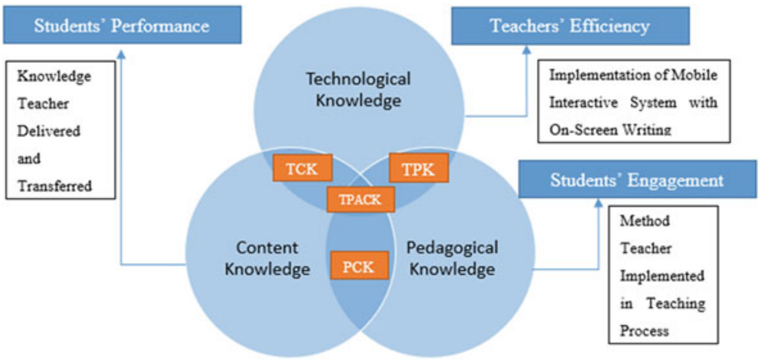


Fig. 2 Proposed TPACK framework

and integrated into a classroom activity. These technological devices and tools act as the medium for teachers driving students, meeting their expectations in terms of students' learning outcomes and the experience they wanted to achieve.

In terms of teachers' efficiency, it can be expressed as to whether the teachers are utilizing the technology that is integrated into the classroom based on their knowledge from the TPACK framework known elements and teaching plan as well as teaching materials prepared. In terms of students' engagement, it can be expressed as the interaction between students and teachers in the classroom, an alternation of classroom environment due to the presence of technology. In terms of students' performance, by the changes made from the aspect of the classroom environment, student eagerness to learn will be affected leading to an impact on their learning outcomes.

5 Discussion

In this study, the TPACK framework is being adopted due to the involvement of technology in the education sector nowadays. To address the three elements introduced earlier in this study, the TPACK framework introduced by Mishra and Koehler [7] is being further modified to fit this study. Hence, to relate technology with the teaching environment, an additional element, teachers' efficiency is being added into the existing TPACK framework. From the perspective of teachers, their level of understanding in terms of technology integration in the teaching environment is reflecting their efficiency of teaching with the use of technology. Furthermore, pedagogical knowledge is also vital as the way teachers teach their subjects is closely related to students' interactivity in the classroom. Therefore, students' engagement is added into the framework reflecting the engagement level of students in the classroom in conjunction with the pedagogical knowledge of their teacher for that particular subject. Students' knowledge gained throughout classes is depending on the relevancy of their teachers' teaching materials toward that particular subject. Hence the modified TPACK framework is formed addressing teachers' efficiency, students' performance and engagement in the teaching and learning environment and how these additional elements in the modified TPACK framework relate to existing TPACK framework elements. Teachers will be assessed to determine their efficiency in teaching by using the system. Furthermore, the integration of a mobile interactive system can boost up students' curiosity and add a fun factor to classes. As such, students' performance for that particular subject will be affected according to how teachers deliver content and knowledge of the subject using the mobile interactive system.

6 Conclusion

The study described the implementation of a modified TPACK framework with the introduction of new elements, which are teacher's efficiency, student's performance and students' engagement. The introduction of the mobile interactive system has the capability of enhancing the teaching and learning environment with better interactivity and visual impact. Expected outcomes of the proposed modified TPACK framework are using the elements of the TPACK framework to indicate the level of teachers' efficiency, students' performance and students' engagement. The nature of TPACK framework is specialized in accessing teachers' effectiveness in utilising the technologies in teaching. Teachers' efficiency is measured from the components of the TPACK framework that involves their technological knowledge, which are TK, TPK and TCK. Their technological knowledge serves as the foundation in conducting classes with the integration of the mobile interactive system. Asides from teachers' efficiency, the components PK, PCK and TPK involve the knowledge of pedagogy of the teachers, which is reflecting the students' engagement level during lessons. Furthermore, the content knowledge-related components from the TPACK framework are assessing the teachers' proficiency in the subject and reflecting the academic performance of the students. All in all, TPACK is the combination of TK, PK and CK as a whole in assessing the teachers whether they can conduct classes with the integration of technology and being interactive with the students. With the help of the modified TPACK framework having proposed mobile interactive system mapped, it is aimed to create a better teaching and learning environment with more interactivity between teachers and students. Moreover, teachers also able to help out students as they may be facing difficulties to catch up with the progression of the teaching plan. The proposed new framework allows teachers to get an overall insight into their students' performance with the adaption of appropriate technological tools and strategies.

Acknowledgements This research was funded and supported by Multimedia University and Fundamental Research Grant Scheme (FRGS), Malaysia.

References

1. Loo S, Chiam C, Aziz RC (2016) A case study: performance between open and distance learners and traditional classroom-based learners. In: Pan-Commonwealth Forum 8 (PCF8). Commonwealth of Learning (COL) and Open University Malaysia (OUM). http://hdl.handle.net/11599/2623
2. Sein-Echaluce ML, Fidalgo-Blanco Á, Alves G (2017) Technology behaviors in education innovation. Comput Human Behav. Elsevier Ltd. https://doi.org/10.1016/j.chb.2016.11.049
3. Gengle H, Abel M, Mohammed B (2017) Effective teaching and learning strategies in science and mathematics to improve students' academic performance in Nigeria. Br J Educ Soc Behav Sci 19(1):1–7. https://doi.org/10.9734/bjesbs/2017/26509

4. Harrell S, Bynum Y (2018) Factors affecting technology integration the classroom. Alabama J Educ Leadersh
5. Naziri F, Rasul MS, Affandi HM (2019) Importance of technological pedagogical and content knowledge (TPACK) in design and technology subject. Int J Acad Res Bus Soc Sci 9(1). https://doi.org/10.6007/ijarbss/v9-i1/5366
6. Vasodavan V, Dewitt D, Alias N (2019) TPACK in higher education: analysis of the collaborative tools used by lecturers. Asia Pac J Curric Teach
7. Koehler M, Mishra P (2009) What is technological pedagogical content knowledge (TPACK)? Contemp Issues Technol Teach Educ 9(1):60–70. Society for Information Technology & Teacher Education, Waynesville, NC USA. https://www.learntechlib.org/primary/p/29544/. Accessed 27 April 2020
8. Rahman SMM, Krishnan VJ, Kapila V (2017) Exploring the dynamic nature of TPACK framework in teaching STEM using robotics in middle school classrooms. In: ASEE annual conference and exposition, conference proceedings. https://doi.org/10.18260/1-2-28336
9. Torres-Carrion PV, Gonzalez-Gonzalez CS, Aciar S, Rodriguez-Morales G (2018) Methodology for systematic literature review applied to engineering and education. In: IEEE global engineering education conference, EDUCON, vol 2018-April. IEEE Computer Society, pp 1364–1373. https://doi.org/10.1109/EDUCON.2018.8363388
10. Hernández R, Fernández F, Baptista P (2014) Metodología de la investigación, 6th ditio. MacGraw-Hill/Interamericana, Mexico

Proposed Integration of Mobile Interactive System in the Classroom

Daniel Lai, Sook Ling Lew, and Shih Yin Ooi

Abstract In the twenty-first century, there are many technologies available to assist us in our daily lives. One of the most commonly owned technology is mobile technology. Due to its nature of high mobility and easier to be owned, it has become a preferred education technology. Furthermore, the conventional way of teaching is still the mainstream among other approaches, which is ineffective. Hence, the introduction of the proposed mobile interactive system is aimed to make a change in the classroom environment for altering both the teaching and learning process. Teachers' capabilities and abilities in facilitating classes using the mobile interactive system are somehow related to their prior technological knowledge, which is also one of the elements in the TPACK framework. Moreover, the TPACK framework is meant to assess a teacher's level of understanding about technology and how the knowledge is being utilized and applied in their teaching plan and teaching materials. As a solution, this study discovered the mapping of teachers' efficiency, students' engagement and students' performance into the TPACK framework can address the efficiency of the proposed mobile interactive system as an alternative mean for a teacher to deliver and transform knowledge to students. This study aims to integrate the proposed mobile interactive system in the TPACK framework and map with teachers' efficiency, students' engagement and students' performance.

Keywords Technology integration · Teaching and learning process · Interactivity and engagement · Classroom environment

D. Lai · S. L. Lew (✉) · S. Y. Ooi
Multimedia University, Melaka, Malaysia
e-mail: sllew@mmu.edu.my

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_77

1 Introduction

1.1 Research Background

In the digital era, technology is an element that is closely related to our daily lives. The presence and rapid growth of technology have led the education sector to transformed into a whole new phase. With the integration of technology into the education sector, a new term called 'educational technology' is introduced. It is absurd to not utilize the known technology and fused it into the classroom environment. In the traditional classroom environment, there is minimal to no involvement of technology where the teaching and learning process is leaning toward the teacher-centered approach [1]. Additionally, it has been proven that the relationship between students and their behavior toward learning via mobile technology is showing a positive attitude [2]. Therefore, the TPACK framework is adopted and modified in this study to assess the practicality of the proposed mobile interactive system being integrated into the classroom. Subsequent paragraphs, however, are indented.

1.2 Problem Statement

In today's classroom, the teacher-centered classroom is still existing in most schools regardless of the level of education. Through teacher-centered teaching and learning strategy, teachers are dominating in both teaching and learning activities [3]. As teachers are exposed to new types of knowledge due to the involvement of technology in education, their considerations on the use of technologies need their in-depth understanding of three aspects, which are technologies, traditional tools and resources [4]. On the whole, two major problems are focused in this study:

1. The readiness of teachers regarding the implementation and integration of technology in the modern classroom.
2. Classes that unable to gain students' attention and interest.

1.3 Objectives

In this study, a proposed mobile interactive system is integrated and introduced as an alternative teaching method. Due to the involvement of educational technology in the classroom, the TPACK framework is being further modified and the objectives of this study are as below:

1. Integrate the proposed mobile interactive system in the TPACK framework.
2. Mapping the research elements, which are teachers' efficiency, students' engagement and students' performance into the TPACK framework with the proposed mobile interactive system as the primary delivery method.

1.4 Manuscript Organization

This study comprises a total of six parts, which are the introduction, literature review, methods, results, discussion and conclusion.

2 Literature Review

2.1 Teacher-Centered Classroom

The teaching and learning process is an interaction between teachers and students, which are two different things but form a unity [5]. The use of traditional methods has received criticism for not creating an environment in the classroom to develop critical thinking and problem-solving skills [6]. Students are less engaged with teachers in the teacher-centered classroom because teachers have full control of the classroom. In a teacher-centered classroom, teachers and students are having minimum interaction where advantages of collaborative learning such as enhancing confidence, better academic achievement, enjoyment, proposing motivation and most importantly communication skills cannot be achieved [7].

As students are self-learning, their communication skills and teamwork are being overlooked. Therefore, the lack of students' engagement in the classroom is noticeable due to the stated nature and characteristics of a teacher-centered classroom. From the learning ability perspective, it is depending on students' intelligence, interest and motivation level where their learning rate will be varying from one another [8]. Likewise, one kind of teaching method cannot fit all students as it might not be entirely suitable for all students. Most importantly, classes conducted using a teacher-centered approach are usually boring. In such an environment, students tend to feel more pressure and learning can be dull resulting in students losing their interest [9].

2.2 Mobile Interactive System with On-Screen Writing Feature

In the modern era, mobile technology becomes very common and accessible for everyone. Hence, the implementation of mobile technology in the education sector has slowly become the norm. The adoption of mobile technology can be known as an innovative approach in education [10]. Teachers can integrate mobile technology into the teaching and learning process giving them a huge for potential boosting up student achievement [10]. Teachers are encouraged to use mobile technology in their teaching as it is a widely accepted technology especially for generation Z students [11]. Therefore, the increment of interactivity in the classroom can be done via mobile applications as well. Additionally, interactivity in the classroom encourages students to be participating actively in classroom activities [12]. The mobile application provides an alternative method for teachers and students to communicate with each other [13]. Moreover, a mobile interactive system can help teachers can facilitate classes with fun factors keeping students with each other as well as the learning materials [14]. Furthermore, an interactive whiteboard (IWB) allows teachers to have a better visualization in representing their teaching contents. In the meantime, IWB also enables students to have constructive discussions promoting collaborative learning [15]. However, the purchase and maintenance cost for the IWB is relatively higher compared with mobile interactive. The mobile interactive system provides teachers with greater mobility while teaching and less hardware dependent.

3 Method

In this study, a systematic literature review (SLR) is conducted as it is crucial in identifying research questions and justifying future research. SLR is important especially for neophytes in the scientific research field as the process conducted for the research can be complex to them. As neophytes, it is difficult to establish the keywords of search from a thesaurus, for filtering the results [16]. Generally, in SLR, the researcher is expected to begin with individual knowledge of the problem in the early phase and moving towards a comprehensive state. Research questions will be written and mentefacto conceptual will be drawn when the literature review has been conducted [17]. Speaking of research questions, the problem statement consists of research questions as one of the elements [18]. These questions give an overview of the research and serve as the protocol of the entire process to be followed. The research questions of this study are as follows:

Research Question 1: How does the proposed mobile interactive system relate to the teaching and learning process?

Research Question 2: How does the integration of the mobile interactive system with on-screen writing mapped into TPACK framework elements with research elements?

4 Results

4.1 Modified TPACK Framework

On-screen writing via the proposed mobile interactive system will be conducted in the classroom for additional interactivity and visual impact while carrying out classroom activities. As for teachers' efficiency, it can be referring to the effectiveness of teachers' integrating the technological approach such as the proposed mobile interactive system in the classroom [19]. Teachers are known to be the leaders of the classroom managing classroom atmosphere as it can affect students' learning efficiency [18]. In this context, it is referring to TK, TCK, TPK and TPACK. With the help of IoT, gender disparity can be overcome and enhancing teachers' efficiency and students' engagement [19]. Students' engagement can be improved with the assistance of technological tools creating effective teaching and learning environment resulting in getting feedbacks from students [20]. From the perspective of students' performance, it is related to CK, PCK, TCK and TPACK. With the presence of technology, the student's attitude is being affected positively influencing their academic performance [21]. In short, the relationship between TPACK framework elements with additional elements that are teachers' efficiency, students' engagement and students' performance can be illustrated in Fig. 1 and Table 1 as reference:

- PK: Interactive-based teaching.
- CK: Subject Proficiency.
- TK: Use of the proposed mobile interactive system.
- PCK: Teaching a subject through interactive-based teaching.

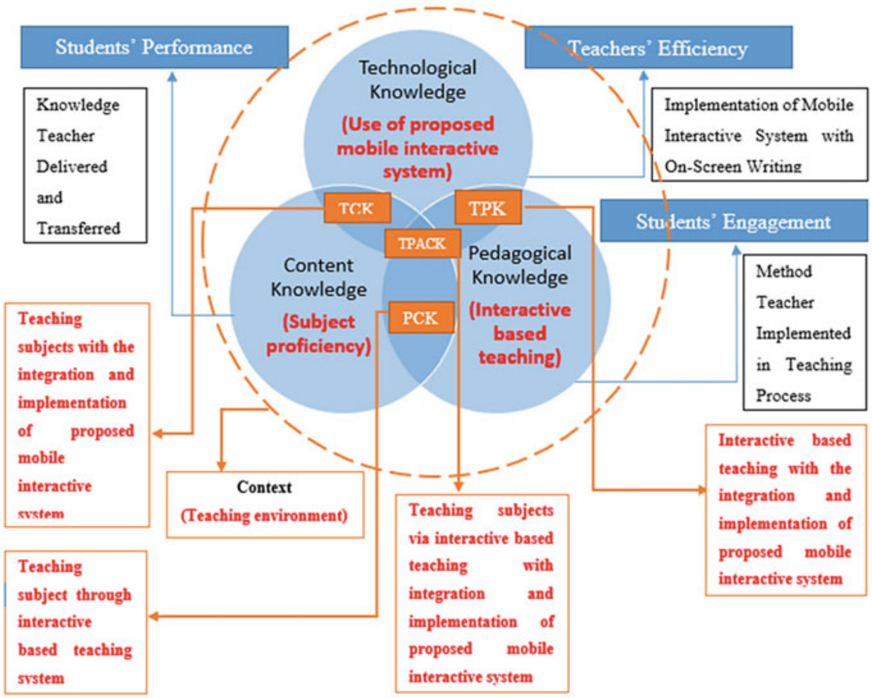


Fig. 1 Mapping of Proposed TPACK Framework with Teachers' Efficiency, Students' Engagement and Students' Performance

- TPK: Interactive-based teaching with the integration of the proposed mobile interactive system.
- TCK: Teaching subjects with the integration proposed mobile interactive system.
- TPACK: Teaching subjects via interactive-based teaching with the integration of the proposed mobile interactive system.

From Table 1, the list of hypotheses is formed based on the findings from Fig. 1, which is consisting of the mapping of the proposed mobile interactive system into the TPACK framework. In this context, the teacher's choice of pedagogy in conducting a class is an interactive-based teaching approach. Thence, the thoroughness of the teacher's understanding of pedagogical knowledge can affect the engagement between the teachers and the students in the classroom. For the lessons' content-wise, the teacher's proficiency level of the subject taught is critical in determining student's performance academically. As for technological knowledge, the use of the proposed mobile interactive system is involved. Teacher's understanding of the known technology is essential as it determines the efficiency of teaches in teaching. Since PK is focusing on pedagogy and CK is focusing on the teacher's proficiency

Table 1 List of hypotheses

Hypotheses (H)	Knowledge		Affected elements	
H1	TK	Significantly affect(s)	Efficiency of Teachers	In teaching and learning
H2	PK		Students' Engagement	
H3	CK		Students' Performance	
H4a	PCK		Students' Engagement	
H4b	PCK		Students' Performance	
H5a	TCK		Efficiency of Teachers	
H5b	TCK		Students' Performance	
H6a	TPK		Efficiency of Teachers	
H6b	TPK		Students' Engagement	
H7a	TPACK		Efficiency of Teachers	
H7b	TPACK		Students' Engagement	
H7c	TPACK		Students' Performance	

in that particular subject taught, PCK is hence, derived and referred to the implementation of an interactive-based teaching approach in teaching a subject. When it comes to the combination of TK and PK, TPK is formed where the integration of the proposed mobile interactive system is magnified in conjunction with the interactive-based teaching. As for TCK, it is slightly different from TPK in which it is focusing on the teacher's proficiency in the subject itself rather than the pedagogy in teaching the subject. Other than that, TCK has also involved the integration of the proposed mobile interactive system. All in all, TPACK is a combination of TK, PK and CK in which assessing teachers in teaching subjects via interactive based teaching with the integration of proposed mobile interactive system. Therefore, we can notice that the research elements that are represented as affected elements in Table 1 will be affected by the elements emphasized in the TPACK framework.

5 Discussion

In this study, integration of the mobile interactive system is proposed and the TPACK framework is adopted as well. The TPACK framework is introduced with a total of seven elements that are technological knowledge (TK), pedagogical knowledge (PK), content knowledge (CK), technological pedagogical knowledge (TPK), technological content knowledge (TCK), pedagogical content knowledge and technological pedagogical content knowledge (TPACK). In conjunction with research elements that are the efficiency of teachers, students' engagement and students' performance, a list of hypotheses have been made to find out how the known knowledge from the TPACK framework relates with the research elements as affected elements. From the teachers' perspective, their efficiency of teaching can be affected by their technological knowledge, technological content knowledge, technological pedagogical knowledge as well as technological pedagogical content knowledge. This knowledge is affecting the efficiency of teachers' teaching under the proposed mobile interactive system integrated and implemented in the teaching process. As a result of changing the teaching environment, the way students learn changes as well. With the use of the proposed mobile interactive system in facilitating classes, the engagement between teachers and students is affected. The integration of the proposed mobile interactive system opened up a new mean of communication offers greater flexibility and convenience. Moreover, since the classroom environment has been changed from both teaching and learning perspective, students' performance will be affected as well. Thence, these research elements are closely related to each other and are significantly affected by TPACK framework elements.

6 Conclusion

The study outlined the integration of a proposed mobile interactive system in the classroom with a mapping of a modified TPACK framework with the introduction of the teachers' efficiency, students' engagement and students' performance into the framework. A mobile interactive system will be used in assisting teachers in conducting classroom activities. The proposed mobile interactive system has the capability of enhancing the teaching and learning environment. Moreover, this study is aimed to create a better teaching and learning environment with more interactivity between teachers and students. Furthermore, the proposed mobile interactive system offers teachers an alternative way of delivering their knowledge more interactively. The system also grants teachers great mobility and flexibility when facilitating classes. Besides, their efficiency in conducting classes can vary due to their prior knowledge in the technological aspect. Like a chain reaction, teachers' efficiency in conducting classes using the proposed mobile interactive system can have an impact on students' engagement and students' performance. The proposed mobile interactive system creates a new way of communication for the students to give feedback and voice out their thoughts to their teachers. Hence, their interactivity and engagement with teachers in the classroom deter their level of interest in the subject taught, which affects students' performance academically.

Acknowledgements This research was funded and supported by Multimedia University and Fundamental Research Grant Scheme (FRGS), Malaysia with project code: FRGS/1/2018/SSI09/MMU/02/3.

References

1. Loo S, Chiam C, Aziz RC (2016) A case study: performance between open and distance learners and traditional classroom-based learners. In: Pan-commonwealth forum 8 (PCF8). Commonwealth of learning (COL) and Open University Malaysia (OUM). http://hdl.handle.net/11599/2623
2. Sein-Echaluce ML, Fidalgo-Blanco Á, Alves G (2017) Technology behaviors in education innovation. Computers in Human Behavior. Elsevier Ltd. https://doi.org/10.1016/j.chb.2016.11.049
3. Gengle H, Abel M, Mohammed B (2017) Effective teaching and learning strategies in science and mathematics to improve students' academic performance in Nigeria. Br J Educ Soc Behav Sci 19(1):1–7. https://doi.org/10.9734/bjesbs/2017/26509
4. Trgalová J, Clark-Wilson A, Weigand H-G (2018) Technology and resources in mathematics education. In: Developing research in mathematics education. Routledge, pp 142–161. https://doi.org/10.4324/9781315113562-12LNCS. Homepage. http://www.springer.com/lncs. Last accessed 21 November 2016
5. Putra CA (2018) Enhanced learning outcomes using interactive edutainment learning method. In: Proceedings international conference BKSPTIS
6. Serin H (2018) A comparison of teacher-centered and student-centered approaches in educational settings. Int J Soc Sci Educ Stud. https://doi.org/10.23918/ijsses.v5i1p164

7. Chawla S (2019) Collaborative learning strategies in software engineering course. In: CEUR workshop proceedings, vol 2506. CEUR-WS, pp 20–28
8. Kaviyarasi R, Balasubramanian T (2018) Exploring the high potential factors that affects students' academic performance. Int J Educ Manag Eng 8:15–23. https://doi.org/10.5815/ijeme.2018.06.02
9. Li MC, Mustakim SS, Kamarudin N (2019) The status and strategies of game teaching on Chinese course in primary schools. In: Graduate research in education seminar (GREduc 2019), 13 Dec. 2019, Faculty of Educational Studies, Universiti Putra Malaysia, pp 55–62
10. Cristol D, Al-Sabbagh S, Abdulbaki A, Majareh M, Tuhin SU, Gimbert B (2019) Innovative delivery of education in Bangladesh using mobile technology. In: Proceedings of world conference on mobile and contextual learning 2019, pp 37–45. Accessed 4 July 2020. https://www.learntechlib.org/p/210599/
11. Poláková P, Klímová B (2019) Mobile technology and generation Z in the English language classroom—a preliminary study. Educ Sci 9(3). https://doi.org/10.3390/educsci9030203
12. Siti AH, Fariza K (2019) Assessing interactivity and engagement in learning economics subject using a mobile application among pre-university students [conference session]. In: International conference of future education and advances, Miri, Sarawak
13. Sizova DA, Sizova TV, Adulova ES (2020) M-learning as a new interactive technology in education. Atlantis Press. https://doi.org/10.2991/assehr.k.200509.060
14. Kung-Teck W, Binti MM, Binti AN, Hamdan A (2020) Mobile-heutagogical practices among student teachers: its pedagogical affordances and challenges. Int J Interact Mobile Technol 14(2):130–143. https://doi.org/10.3991/ijim.v14i02.11819
15. Maryam B, Sören H, Gunilla L (2020) Putting scaffolding into action: preschool teachers' actions using interactive whiteboard. Early Childhood Educ J 48(1):79–92. https://doi.org/10.1007/s10643-019-00971-3
16. Torres-Carrion PV, Gonzalez-Gonzalez CS, Aciar S, Rodriguez-Morales G (2018) Methodology for systematic literature review applied to engineering and education. In: IEEE global engineering education conference, EDUCON. https://doi.org/10.1109/EDUCON.2018.8363388
17. Torres-Carrion PV, Gonzalez-Gonzalez CS, Aciar S, Rodriguez-Morales G (2018) Methodology for systematic literature review applied to engineering and education. In: IEEE global engineering education conference, EDUCON, vol 2018. IEEE Computer Society, pp 1364–1373. https://doi.org/10.1109/EDUCON.2018.8363388
18. Yu L (2018) How to manage classroom in college teaching [conference session]. In: Proceedings of the 2018 5th international conference on education, management, arts, economics and social science (ICEMAESS 2018). https://doi.org/10.2991/icemaess-18.2018.104
19. Muhammad S, Gulfam AS, Amtul H, Muhammad AM (2019) Promises and challenges of IoT in education [Paper presentation]. http://hdl.handle.net/11599/3425
20. Elmahdi I, Al-Hattami A, Fawzi H (2018) Using technology for formative assessment to improve students' learning. Turkish Online J Educ Technol TOJET 17(2):182–188
21. Sabado I (2018) Technology-assisted teaching and students performance in literature technology-assisted teaching and students' performance in literature. Int J Soc Sci Humanit 3(2, November):43–60. https://doi.org/10.5281/zenodo.1467798

WS-PDC: Persistent Distributed Channel-Based Web Services Applied on IFRS Data Processing and Loading

Noussair Fikri, Mohamed Rida, Noreddine Abghour, Khalid Moussaid, and Amina Elomri

Abstract In this paper, we are proposing a reporting approach based on dynamic data delivery using PDC (persistent distributed channels) and web services concept. This architecture has the same properties of an extract–transform–load-based system, except one different property, our architecture is an extract–transform–deliver-based system and has a distributed core for intensive data processing and also micro-batching. It uses a novel approach of PDC, which is based on the concept of share memory by communicating, and inspired from resilient distributed datasets of Apache Spark. Big Data approach helps us to build a fast and real-time processing system in order to produce up-to-date and on-demand reporting web service, where report data are delivered as HTTP responses for analysis. This approach is served as a solution for the lack of performance and memory consumption problems in ETL-based systems.

Keywords ETL · RDD · Real-time · Discretized · Big data · Spark · PDC · HTTP · Analysis · Reporting · Web · Services · Resilient · Distributed · Channels

1 Introduction

The extract–transform–load concept is used for ready-to-use data delivery [1]. Production data represent a structured mass of datasets, which drive us to project the 3'v challenges on our problem. In the volume challenge, ETL [2] must have the ability to deal with massive data processing. The variability challenge, each information system has its proper data schema, the reason why ETL must deal with business meta-models dynamically [3] to fit with foreign databases, and finally the velocity challenge [4], for fast data processing and ready-to-use data delivery. The best approach is Big Data processing models, following enhanced data processing interfaces. Among proposed Big Data processing platforms, Apache Hadoop and the next optimized generation, Apache Spark [5] based on resilient distributed datasets

N. Fikri (✉) · M. Rida · N. Abghour · K. Moussaid · A. Elomri
Faculty of Science, University Hassan II, Casablanca, Morocco

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_78

concept. The proposed architecture of ETL is based on clustered data processing, which delivers data in the form of data provider as a web service. Companies handle their production data using information system based on several modules depending on subject orientation, by using one compact system including main modules (products and services management, billing, receivables), in most cases, it can be SAGE, Microsoft ERP/CRM (Dynamics AX and Dynamics CRM) or SAP, as extensible systems. The proposed architecture delivers data as key/pair responses by distributed web services as cluster following subject orientation and handling nodes depending on data processing complexity. The proposed system is running on multiple nodes, with a single master node and dynamic number of worker nodes where each node has processing unit and web service unit. Data are extracted from different data sources (databases, flat files, etc) then delivered to Spark Workers as RDDs, which are represented in form of processing units. Data pairs are exposed to pre-set transformations through Spark Workers produce a ready to load RDDs for data warehouse. A parallel process is executed to provide in-memory data as a web service response. Web services are created following subject orientation, by using technical terms, for each DataMart, the worker node produces a web service that bases its request/response on dynamic RDD-based meta-models. On-demand data are available after data processing, started by user request. The system gets the available data from the request meta-model and builds the response from processed RDD. User's requests represent a group of parameters, which are in charge of transformation of data, and storage to specific in-memory DataMart. The proposed architecture covers all the facets of a classic ETL and resolves the following problematics:

- Processing time and response latency
- Request meta-model flexibility
- Reporting data availability
- Transformation pipeline simplicity
- Extensible programming interface.

In this article, we are going to discover all previous and related works of ETL concept and data warehouse basics, from extracting process to data loading, in order to explain the use of our architecture.

2 Background

2.1 Business Intelligence

The business intelligence concept is a necessity for companies, in the case of production of reports [1, 3], which describe a well-defined state that is the subject of a study of an eventual optimization. After extraction, it is necessary [1] to transform such data in an appropriate format by cleaning corrupted data and deduplication, verifying the integrity of the data and formatting, correcting and cleaning it up, and finally by applying joining, specific calculations and a definition of keys. The available known basic transformations are key restructuring, deduplication, format revision and cleaning. There is also an advanced transformation like aggregations, derivation, joining, summarization, validation, integration, filtering and splitting. An ETL usually passes through three phases, [1–3]: staging layer (for data extraction), integration layer of data (for data during processing) and access layer (for final data ready for exploitation). After execution of the three phases of an ETL, an ETL testing process is applied [6, 7] to verify the integrity of data loaded in data warehouse. The previous process produces a centralized database, ready for use by reporting and analysis tools. As data container, it has some specific features, it is subject-oriented [6–8] (it follows a functional schema that describes a specific subject on one or more operations of the company such as purchases, sales and stock movements, etc.), integrated (the data warehouse is a homogenized data centralization point resulting from a transformation of heterogeneous data from relational databases, or flat files, etc.), time-variant (all flows in the data warehouse are distinguished by time period) and represents a collection of non-volatile data (the basic rule of a data warehouse is [2, 4] non-deletion of data, new data acts as a new record or updating an existing item). A data warehouse is exploited by exploring or analyzing data with operations specific to OLAP concepts [7]. An OLTP database contains data at the gross and very detailed state, they are not yet ready to be exploited by the business entity: this is the work of ETL processes. An OLAP data warehouse is a multidimensional and summarized structure of data that simplify data retrieving, and accurate well-structured multi-model that can be schematized by three types of schemas [1]: Star schema, Snowflake schema and galaxy schema. Data repartition is the strength of a data warehouse, the best approach is to have a cubic view for data mining, its structure simplifies reading and gives more flexibility to access to data, and a multidimensional representation [8] is based on presenting each entity in the form of dimension. Each department within a company has a dedicated part in the data warehouse called DataMart [1, 8].

2.2 Big Data and Apache Spark

Big data is considered as the most coveted area of research in the last decade concerning the processing of important amount of data to have more meanings resulting from a thorough analysis of these data. Several technologies in the form of advanced tools and frameworks have been designed to meet the technical requirements of Big Data processing. These technologies are provided in the form of algorithms, optimized pipelines or programming interfaces for rapid data processing. After [5, 9, 10] Hadoop data processing approach for an emerging solution called Apache Spark solved the problem of data processing latency, based on a distributed and in memory approach. Apache Spark is a very powerful framework in terms of processing speed and has a fairly rich programming interface for data analysis in different domains. Spark rest on several libraries [9], it is a composite framework having Spark Core as the main component, the pillar of the entire Spark panoply, it provides several useful features, such as transformations and shared variables that help data propagation between its levels, memory optimization (LRU algorithm), scheduling, shuffling and interpretation, to perform the communication to sub-components such as the cluster handler for distributed computing and storage manager for optimal I/O movements [5]. Libraries on the most noteworthy layer of Spark have upheld the basics of Spark to deal with various workloads. We're taking a look at the Spark programming model, the RDD concept, a key/value group with fault tolerance capacity, which are paralyzed with a driver program's collection or an external data source like HDFS. It uses the pair's approach of key values to manage transformations. Each dataset state through the applied operations is stored in memory for quick access and processing through chosen algorithms. It uses two backup types, PERSIST and CACHE, they are defined by the following storage levels: MEMORY_ONLY, MEMORY_AND_DISK, MEMORY_ONLY_SER, MEMORY_AND_DISK_SER, DISK_ONLY, MEMORY_ONLY_2 and MEMORY_AND_DISK_2.

2.3 Persistent Distributed Channels

The architecture of PDC-driven platform is based on the concept of near-metal programming [11], given instructions are compiled and executed, without an interpreter like python or a virtual machine like (JVM) Java or layer that can slow down data processing. This concept is inspired from the basic layer of Apache Spark core, which is resilient distributed datasets of Spark and based also on Last-Recent Used algorithm. We used this architecture because of its efficiency in processing a large amount of data. It has three main levels:

- Partitioning
- Synchronization
- Channeling and processing.

The persistent distributed datasets particularity is the use of share memory by communicating concepts [11, 12]. For each worker, the given partition is delivered as channels of data for processing by using concurrency and also parallelism is the case of multi-core. For each routine, there is a set or array of data that is processed using a specific operation (e.g., transformations), as shown in Fig. 1, channels are shared, but not data, if data are used by a go routine, any other go routine can touch it.

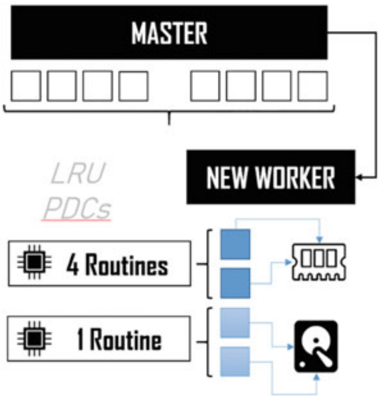


Fig. 1 Channeling and processing of PDCs

2.4 Reporting IFRS Driven Data

The international financial reporting standard is an update of IAS and a cover for straightforwardness of financial yearly activities in Europe and several countries but not in the USA, which use (GAAS) general accepted accounting standards [13–15]. IFRS is a set of mandatory statements as strategic policies for business description, as follows [16]:

- Statement of financial status
- Statement of comprehensive income
- Statement of changes equity
- Statement of cash flow
- Statement of profit and loss.

In each company, there is an information system called a general ledger system that summarizes the financial activities in form of general ledger from which its name. These activities are described as IFRS or GAAP statements. All financial details are structured following COA (chart of account) [17] in order to emphasize segments by BU (business units), CC (cost center) and others levels of classification. The general ledger is composed of sub-ledgers [14] that wrap sales, debts, billing, purchases and other activities. We call posting the aggregation of all records in sub-ledgers by financial segments defined in COA [14, 17]. We deal with two kinds of bookkeeping frameworks or systems, the ones that are ERP integrated and the others that are used as auxiliary systems.

2.5 Web Services

The main purpose of web service is to publish data in a simplified format. There are two types [18, 19] of web services, the REST and SOAP. The most common type of web services is the REST approach. The REST (Representational state transfer) [18] is consumed through the HTTP protocol, using the four methods POST, GET, PUT and DELETE, which constitute the essential manipulations of the services. SOAP (Simple Object Access Protocol) [19] is a protocol that gives a decentralized and distributed messaging using XML between peers. The input, output and service parameters are embedded in the web service configuration. A SOAP service can be consumed using WSDL (Web Service Description Language), and a REST service can be consumed using WADL (Web Application Description Language).

3 Method

Our approach is to build an architecture for data manipulation and delivery as PDSs (persistent distributed datasets) based REST request/response. By using extracting process, the proposed system retrieves data from single or multiple data sources. It can be a database, a flat file, a Microsoft Excel file or another data flow, coming from information systems or web API. After data retrieval, the system submits data as persistent distributed datasets to the transformation pipeline. This processing canal is in form of one or several workers, which are in charge of applying a basic or advanced transformation operation to set these data from production status to ready-to-use status. These workers can be deployed on virtual machines or containers depending on chosen architecture. This transformation pipeline uses a Golang programming interface. The transformed PDCs are submitted to RESTful web services as ready-to-process data with transformations to apply on it. After execution of transformation pipeline, users can retrieve data from these web services by using an HTTP request, based on JSON. These web services required the following parameters:

- The server on which is deployed.
- The requested DataMart for subject orientation.
- The requested dimensions.
- For each dimension, the requested attributes.

The following Fig. 2 shows the main architecture of our approach, from data sources to web services' delivery. The main component in this system is the master node, which acts as head of all operation running in sub-layers, and has a main web service that is in charge of acquiring data or a targeted data source depending on data injection strategy then transformations to apply on these data in the form of ETL instructions (calculations–aggregations). Input data are processed as persistent distributed datasets following instructions given to main web service. The distributed pipeline can have one or more workers, which process then expose its data on dedicated web service following a distribution rules for functional purpose or a random distribution for basic transformations. Then submitted to worker's web services, which are in charge of data publishing as RESTful response.

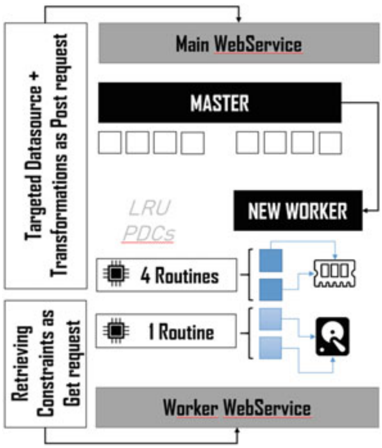


Fig. 2 Our approach: the PDC-based web services

The DATA POST and INSTRUCTIONS POST (for data and instructions injection) requests routed to these web services can be expressed by the following header model.

```
{
  "input":
  [{
    "worker": "localhost/accountPayable",
    "datastore": "temp",
    "datasources":
    {
      {
        "name": "supplier0",
        "type": "database",
        "url": "localhost",
        "dbtype": "postgresql",
        "username": "test",
        "password": "testpassword",
        "request": "SELECT * FROM SUPPLIER",
      }
    }
  }]
}
```

```
{
  "input":
  [{
    "worker": "localhost/accountPayable",
    "instructionsStore": "temp",
    "instructions":
    {
      {
        "name": "instruction0",
        "type": "joining",
        "constraints": "NO_REPLICAT",
        "datastores": "temp,_",
        "data":"supplier0,supplier1,supplier2",
        "output": " pdcws0,_",
      }
    }
  }]
}
```

The GET (for data retrieving) requests routed to these web services can be expressed by the following header model.

```
{
  "output":
  [{
    "webservice": "localhost/pdcws0",
    "request": "instruction0"
    }
  }]
}
```


References

1. Chaudhuri S, Dayal U, Narasayya V (2011) An overview of business intelligence technology. Commun ACM 54(8):88–98. https://doi.org/10.1145/1978542.1978562
2. Chaudhuri S, Dayal U (1997) An overview of data warehousing and OLAP technology. SIGMOD Rec 26(1):65–74. https://doi.org/10.1145/248603.248616
3. Su J, Tang Y (2017) Business intelligence revisited. In: 2017 fifth international conference on advanced cloud and big data (CBD), pp 1–6. https://doi.org/10.1109/CBD.2017.9
4. Obalı M, Dursun B, Erdem Z, Görür AK (2013) A real time data warehouse approach for data processing. In: 2013 21st signal processing and communications applications conference (SIU), pp 1–4. https://doi.org/10.1109/SIU.2013.6531245
5. Luu H (2018) Beginning Apache Spark 2: with resilient distributed datasets, Spark SQL, structured streaming and spark machine learning library. Apress
6. Berson A, Smith SJ (1997) Data warehousing, data mining, and Olap. McGraw-Hill, Inc.
7. Reddy GS, Srinivasu R, Rao MPC, Rikkula SR (2010) Data warehousing, data mining, OLAP and OLTP technologies are essential elements to support decision-making process in industries 02(09):9
8. OLAP Solutions (2019) Building multidimensional information systems, 2nd edn. Database and data warehousing technologies. General and introductory computer science. Subjects. Wiley. https://www.wiley.com/en-us/OLAP+Solutions%3A+Building+Multidimensional+Information+Systems%2C+2nd+Edition-p-9780471400301
9. Salloum S, Dautov R, Chen X, Peng PX, Huang JZ (2016) Big data analytics on Apache Spark. Int J Data Sci Anal 1(3):145–164. https://doi.org/10.1007/s41060-016-0027-9
10. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman S, Franklin MJ, Ghodsi A, Gonzalez J, Shenker S, Stoica I (2019) Apache Spark: a unified engine for big data processing. https://cacm.acm.org/magazines/2016/11/209116-apache-spark/abstract
11. Fikri N, Rida M, Abghour N, Moussaid K, El Omri A (2020) A near metal platform for intensive big data processing using a novel approach: persistent distributed channels. In: Proceedings of the 2020 5th international conference on big data and computing, New York, NY, USA, pp 1–5. https://doi.org/10.1145/3404687.3404692
12. Tang P (2010) Multi-core parallel programming in Go
13. Lessambo FI (2018) Consolidated financial statements. In: Lessambo FI (ed) Financial statements: analysis and reporting. Springer International Publishing, Cham, pp 269–276
14. Page M, Spira LF (2016) Special issue on "business models, financial reporting and corporate governance. J Manag Gov 20(2):209–211. https://doi.org/10.1007/s10997-016-9344-6
15. Cherny J (2014) Financial statements are messages that need a context to be better understood. Int J Discl Gov 11(2):161–176. https://doi.org/10.1057/jdg.2012.20
16. Lessambo FI (2018) IFRS and GAAP. In: Lessambo FI (ed) Financial statements: analysis and reporting. Springer International Publishing, Cham, pp 299–324
17. Adhitama MA, Sarno R, Sarwosri (2016) Account charting and financial reporting at accounting module on enterprise resource planning using tree traversal algorithm. In: 2016 international

conference on information communication technology and systems (ICTS), pp 20–25. https://doi.org/10.1109/ICTS.2016.7910266
18. Malik S, Kim D (2017) A comparison of RESTful vs. SOAP web services in actuator networks. In: 2017 ninth international conference on ubiquitous and future networks (ICUFN), pp 753–755. https://doi.org/10.1109/ICUFN.2017.7993893
19. A comparison of RESTful vs. SOAP web services in actuator networks—IEEE Conference Publication (2019). https://ieeexplore.ieee.org/document/7993893

Higher-Order Mode Tri-Band Stacked Patch Antenna for Mobile, Wi-Fi and Bluetooth Jamming Applications

Rayan Mina, Georges Zakka El Nashef, and Hassan Chreim

Abstract A triple-band stacked patch antenna working on high-order modes is presented in this paper. The proposed design consists of three layers, each having a different slot shape to cover three close range frequency bands (925–960 MHz), (1.805–1.88 GHz) and (2.4–2.483 MHz) intended for mobile, Wi-Fi and Bluetooth Jamming applications. The antenna is optimized such that each slotted patch layer contributes to the impedance matching of one of the bands. Excitation of the high order modes helps achieve an acceptable gain and a wider bandwidth for each resonance frequency. U-slot, S-slot and J-slots were implemented at the first, second and third layers, respectively, giving this structure a newly defined combination of slot shapes (U–S–J). The simulated results of the proposed antenna show that it has a peak gain of 4.7, 4.3 and 5.1 dBi and achieves a good 10 dB return loss bandwidth of 41, 119 and 92 MHz at 940 MHz 1.84 and 2.44 GHz, respectively. The antenna radiation characteristics satisfy modern generation wireless terminal systems.

Keywords Multi-band antenna · Mobile communications · Stacked patch · Wi-Fi · Bluetooth

1 Introduction

The radio frequency (RF) spectrum usage has been growing continuously during the past decades with new communications standards emerging rapidly. In modern wireless technologies, one standard could use multiple operating bands. For example,

R. Mina (✉) · G. Z. El Nashef
Saint-Joseph University of Beirut, B.P. 1514, Mar Roukoz, Lebanon
e-mail: rayan.mina@usj.edu.lb

G. Z. El Nashef
e-mail: georges.zakkanashefel@usj.edu.lb

H. Chreim
Dassault Systèmes, 10 Rue Marcel Dassault, 78140 Vélizy-Villacoublay, France
e-mail: hassan.chreim@3ds.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_79

fifth-generation mobile communication (5G) is planned to operate in ~50 different frequency bands ranging from ~400 MHz to ~7 GHz in range 1 and from ~25 to ~52 GHz in range 2 [1]. In practice, user equipment is never designed to be compatible with all these ranges simultaneously because they are not all implemented in the same geographical region. Usually, a wireless device hosts few antennas and is specified to work with a dozen of bands only resulting in 2–4 ranges being covered by each antenna. Therefore, designing multi-band antennas is crucial nowadays to cope with the recent technological trends, as well as to reduce the size, the cost and the coupling constraints of the wireless device.

It is quite challenging to achieve wideband and multiband performance, without increasing the size or using multi-layered structures [2]. An extensive amount of research work has been done in the past 10 years to design multiband antennas. Dual and triple-band capabilities using different slot shapes, layers and substrates have been demonstrated in [3–7]. However, all these antennas target sufficiently distant frequency ranges (offsets of 1.5–2 GHz in [4], and up to 3 GHz in [3, 5, 7]). The combination of low frequency separation between the target bands and a wide matching bandwidth (in each range) represents a challenge in patch antenna design.

This work is an integral part of a project that aims at developing a low-cost and low-power jamming unit to disable wireless communications within university campuses. The target frequency bands are 3GPP bands III (downlink 1805–1880 MHz) and VIII (downlink 925–960 MHz) as well as the ISM band (2.4–2.483 GHz), which can be easily used by students in classrooms to communicate via mobile, Wi-Fi and Bluetooth standards. Table 1 summarizes the fractional bandwidth (FBW) and frequency separations (offsets) requirements for this work.

The frequency separation between the second and third target bands is only 600 MHz while a little less than 100 MHz matching bandwidth is required for those frequency ranges. This aspect represents a design challenge.

Several antenna design techniques were used in the literature to widen the bandwidth of patch antennas. Defected ground plane is used in [8–10], while different feeding structures (e.g. aperture coupled slot, coplanar waveguide) are used in [11, 12]. Although they achieve very good results, such solutions require a significant effort during the design phase. Recently, several papers [13–15] have shown the benefits of using antennas working on high order modes to achieve wider bandwidth

Table 1 Frequency band requirements for this work

	Bandwidth (MHz)	FBW (%)	Separation to band VIII	Separation to band III	Separation to ISM band
3GPP band VIII	35	3.7	N/A	900 MHz	1.5 GHz
3GPP band III	75	4.1	900 MHz	N/A	600 MHz
ISM band	83	3.5	1.5 GHz	600 MHz	N/A

and acceptable gains at the cost of modified radiation patterns [15]. This work investigates an antenna design that revolves around the use of two techniques in order to target the triple-band requirements: (1) a multi-layer stacked patch structure with different slot shapes at each layer and (2) high order mode operation.

The paper is organized as follows: Section 2 describes the antenna structure with all the manufacturing details, the design procedure and the obtained S_{11} results. Section 3 shows the achieved radiation patterns and peak gain values for the developed antenna at all three bands. Section 4 concludes this work as well as describes future perspectives.

2 Antenna Design

2.1 Idea and Design Procedure

In the past 10 years, Microstrip patch antennas were designed with a U-slot to create single and multiband antennas in addition to achieving wideband performance [2]. The triple-layer stacked antenna proposed in this paper is designed using a U-slot in the first layer while two different slot shapes are cut in the second and third layers. The final design is a combination of U–S–J slots with each patch creating a different resonance.

The antenna is fed by a coaxial SMA probe connected to the first layer's radiating patch (U-slot). Excitation of the upper patches is realized by parasitic coupling through the dielectric substrates of the second and third layers (see Fig. 1). For this reason, the design process must be accurate and cannot be performed for each layer alone regardless of the others, since mutual coupling among all layers affects the matching performance significantly. At the same time, it is impossible to optimize

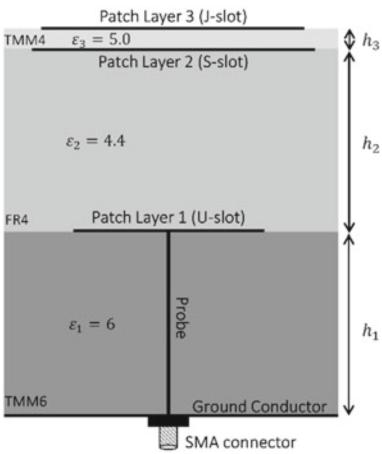


Fig. 1 Stacked patch antenna detailed structure

the whole structure that contains three different slot shapes at once, due to the high number of parameters in the design. Therefore, a balanced approach was adopted to obtain the final structure: the U-slot antenna is first designed to target the ISM band having the highest BW as per Table 1 requirements. Next, the S-slot patch is specifically dimensioned (i.e. rectangular dimensions, as well as length and width of all segments forming the slot) to create a resonance close enough to the next target frequency range (band III). The same procedure is applied to the J-slot to resonate at the last target frequency range (band VIII). Once the physical dimensions are fixed for all slots and layers, only the slots' positions within the second and third patches (S and J slots) are optimized together to achieve the optimal coupling among the antennas. This final step allows the structure to resonate at all bands with the required bandwidths to meet the specifications in this work for mobile, Wi-Fi and Bluetooth applications.

The first layer's thick substrate is made of thermoset microwave materials laminate TMM6 that has a dielectric constant of 6.0 and a low loss tangent of 0.0023. The second layer has also a thick substrate made of low-cost FR-4 material with a dielectric constant of 4.4 and a loss tangent of 0.017. The top layer's thin substrate is made of thermoset microwave materials' laminate TMM4 that has a dielectric constant of 5.0 and a low loss tangent of 0.0020. In the next paragraphs, details on each layer are given.

2.2 U-Slot Layer

A U-shaped slot is designed on the first rectangular patch to create a high order mode resonance around the ISM band (i.e. center frequency ~2.44 GHz). The detailed structure is shown in Fig. 2, and exact physical dimensions are given in Table 2.

The U-slot helps achieving a wider matching bandwidth by creating an additional resonance to the one created by the patch itself. The antenna was designed to resonate at a higher-order mode at the ISM band's center frequency (the lowest resonance mode is at ~460 MHz). Thanks to the thick substrate used, the U-slot has an additional benefit of creating a capacitance effect that compensates for the excessive inductive reactance of the feed pin's long probe.

The S_{11} performance over the studied frequency range is given in Fig. 3. The 10 dB return loss bandwidth at 2.4 GHz is 124 MHz and a narrow resonance can be seen for the lowest resonance mode at ~460 MHz and the next mode at ~960 MHz. The combination of the thick substrate, high-order mode excitation and the U-slot geometry has created a wide bandwidth resonance at the ISM band. Worth noting that, we do not need, at this stage, to target exactly the center frequency of 2.44 GHz, since top layers will also affect the matching as will be explained in the coming paragraphs.

The position of the feeding pin helps exciting high order modes on this patch. Narrow bandwidth resonances of other high order modes at 980 MHz and 2 GHz can also be seen in Fig. 3.

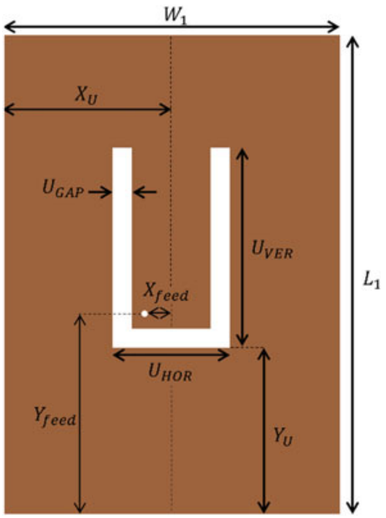


Fig. 2 First layer details of the proposed antenna

Table 2 Physical dimensions of the U-slot patch layer in mm

W_1	L_1	X_U	Y_U	X_{feed}	Y_{feed}	U_{HOR}	U_{VER}	U_{GAP}	h_1
84.22	120	42.11	41.8	8.0	49.38	29.2	50	4.72	12

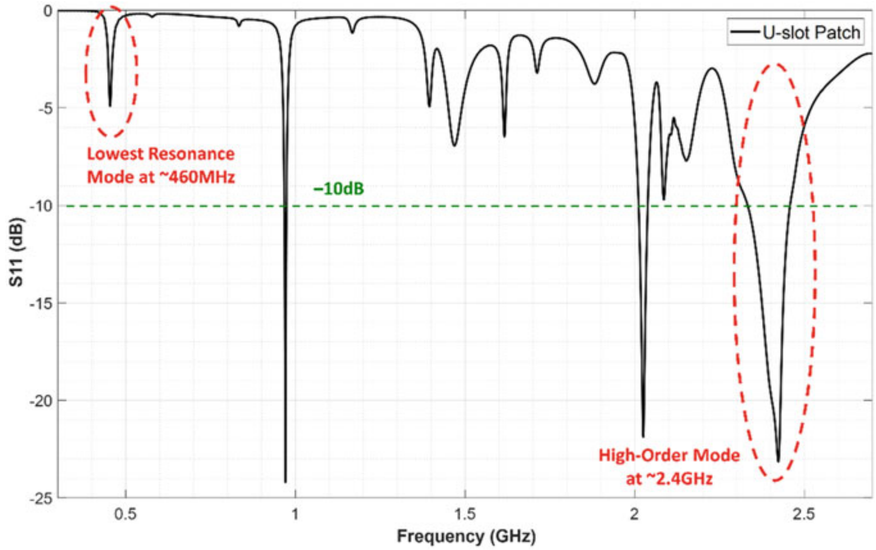


Fig. 3 S_{11} results of the first layer alone

2.3 S-Slot Layer

An S-shaped slot is dimensioned on the second patch layer to create a resonance at ~1.84 GHz, as the first layer alone cannot create a matching performance at this frequency. The slot's structure contains five segments that were designed independently as shown in Fig. 4. The physical dimensions are given in Table 3.

The S-slot patch layer is responsible for making multiple closely separated resonances at 1.84 GHz (see Fig. 5), which are not yet sufficient to achieve a good matching performance (below −10 dB) over the entire bandwidth. In addition, it creates some offset effects, due to parasitic coupling, which can be clearly seen in Fig. 5 (the ISM band resonance is shifted and narrowed compared to Fig. 3). The matching will be enhanced and the offsets will be corrected by adding the third patch layer that has the J-slot.

Several resonance modes exist at closely separated frequencies within the Band III frequency range, thus making a wide 10 dB return loss bandwidth possible. These are created by the second patch, with its thick substrate (h2) and the etched S-shaped slot. At this stage of the design process, two target bands are covered and a third layer will be added in the next paragraph for two purposes: (1) achieve a resonance at band VIII and (2) enhance the results obtained in band III to meet the specifications.

Fig. 4 Second layer details of the proposed antenna

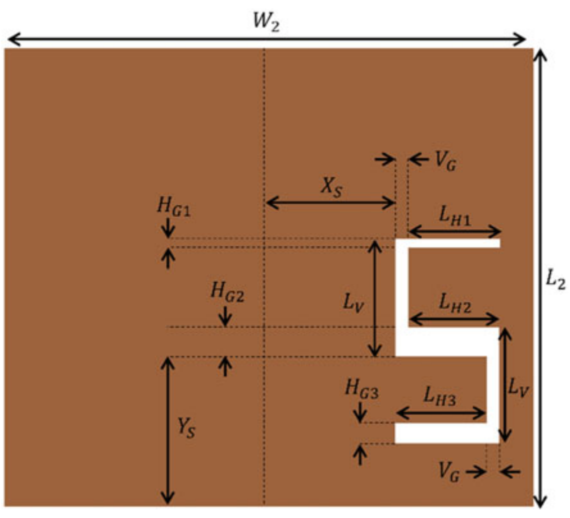


Table 3 Physical dimensions of the S-slot patch layer in mm

W_2	L_2	X_S	Y_S	L_V	V_G	H_{G1}	H_{G2}	H_{G3}	L_{H1}	L_{H2}	L_{H3}	h_2
125	110	30	36	28	3	2	7	5	22	19	22	12

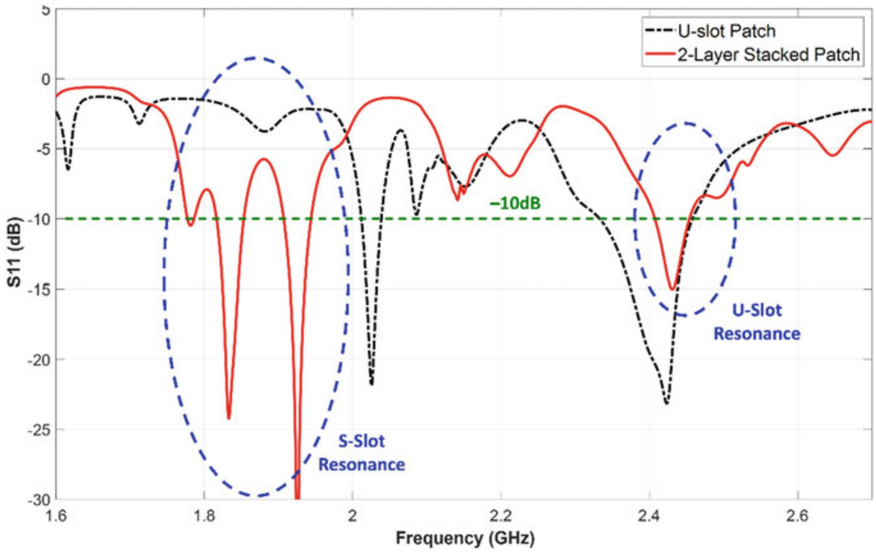


Fig. 5 S_{11} results of the first two layers

2.4 J-Slot Layer

A J-shaped slot is designed on the third patch layer to create a resonance at the lowest frequency range ~940 MHz. The slot's structure contains three segments that were designed as shown in Fig. 6. The physical dimensions are given in Table 4.

The dimensions of the J-slot were fixed first as described in Sect. 2 in order to create a close to 940 MHz frequency resonance. Then, the slot's position in the third

Fig. 6 Third layer details of the proposed antenna

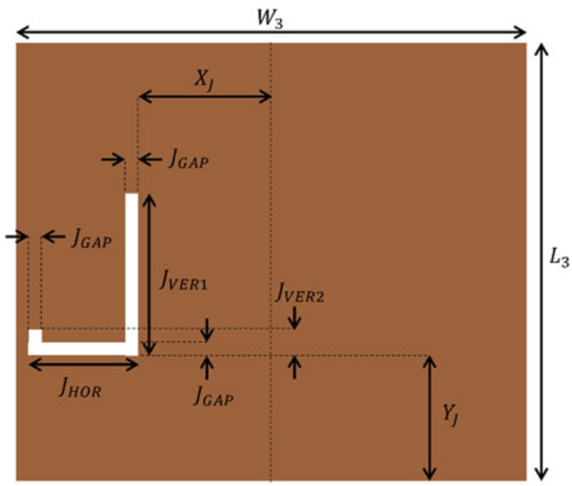


Table 4 Physical dimensions of the J-slot patch layer in mm

W_3	L_3	X_J	Y_J	J_{GAP}	J_{VER1}	J_{VER2}	J_{HOR}	h_3
117	100	30	28.5	3.0	37	3.0	25	0.3

layer (XJ, YJ) and the position of the S-slot in the second layer (XS, YS) along with the top substrate's height (h3) and dielectric constant (εr3) were tuned together using an optimization algorithm to achieve the three goals defined in Table 1, in terms of center frequencies as well as bandwidths.

The choice of these specific design parameters is motivated by several aspects: First, the degree of coupling between the lower and upper patches is mainly controlled by the amount of the overlap area, as well as by the spacing between them (i.e. substrate height) and to some extent the relative permittivity of the materials used. Second, the S and J slots' positions have a big impact in balancing the coupling between the first and the third layers, which is crucial for the excitation of the resonance modes. During optimization, several bounds were carefully imposed on the design variables being tuned to help the algorithm achieve a faster convergence. For example, the top patch (J-slot) dielectric constant has an upper bound to the first patch (U-slot) dielectric constant (εr3 < εr1) to ensure those two layers remain lightly coupled and hence avoid narrowing the impedance matching bandwidths.

The S_{11} matching performance of the stacked patch's complete structure over the entire studied frequency range is given in Fig. 7. The design was simulated using high-performance 3D EM analysis software (Simulia CST Studio Suite) and high-accuracy time-domain solver (transmission line matrix method, TLM) that utilizes

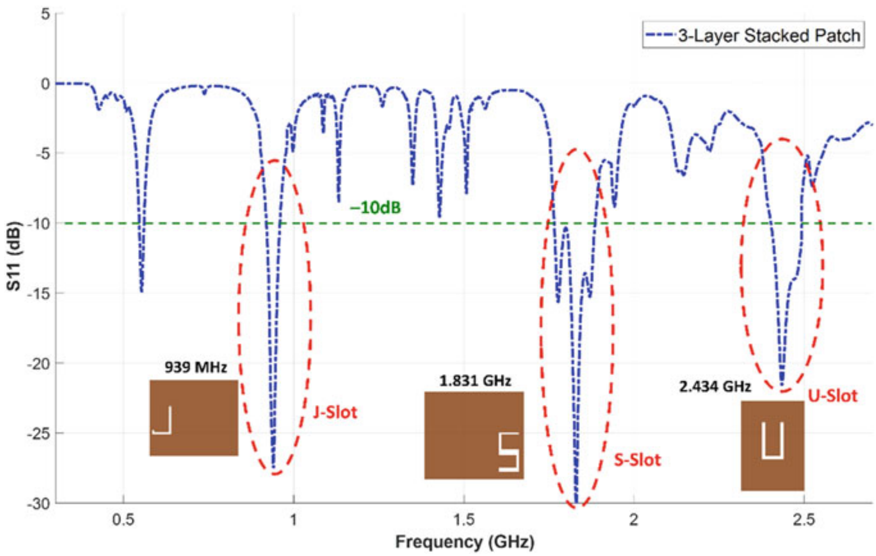


Fig. 7 S_{11} results of the proposed stacked patch antenna

multi-gridding technology for automatic mesh generation. Then, the results were verified for an additional consistency with a frequency-domain EM solver, the finite element method (FEM) using Keysight ADS. The obtained performance was sufficiently similar giving this structure a stable EM and radiating behavior. The SMA connector's exact model (detailed dimensions and dielectric material) was included in the simulation.

In the next section, the overall performance results related to the resonance frequencies and matching bandwidth (S_{11}), the radiation patterns and peak gain values of the proposed antenna are given and discussed.

3 Performance Results

The proposed stacked patch antenna has a 10 dB return loss bandwidth of 41, 119 and 92 MHz (Table 5) in the targeted bands at 940 MHz, 1.84 GHz and 2.44 GHz, respectively. Both frequency range and bandwidth requirements of Table 1 are fulfilled and the proposed tri-band antenna is fully compliant with the specifications.

The simulated far-field radiation pattern and peak gain of the proposed stacked patch at 0.94, 1.84 and 2.44 GHz in E-plane and H-plane are shown in Fig. 8.

The direction of maximum radiation in the E-plane occurs at $\theta = -63°$, $\theta = -55°$ and $\theta = 60°$ for bands VIII, III and ISM, respectively, with a corresponding gain of

Table 5 S_{11} results of the proposed antenna compared to specifications

	BW Spec (MHz)	FBW Spec (%)	BW result (MHz)	FBW result (%)	Compliance
Band VIII	35	3.7	41	4.4	Compliant
Band III	75	4.1	119	6.5	Compliant
ISM band	83	3.5	92	3.8	Compliant

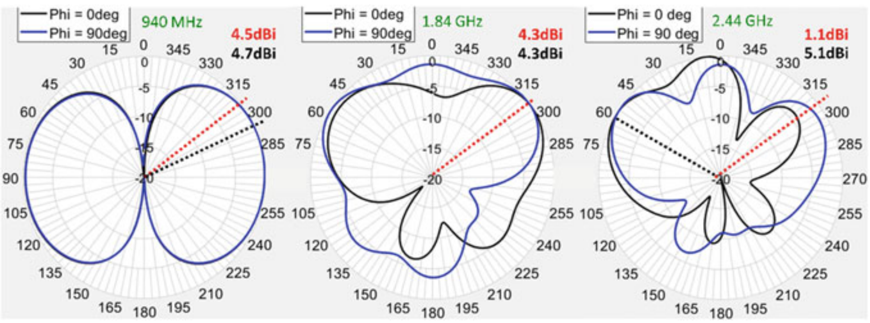


Fig. 8 Radiation patterns in E and H planes of the proposed antenna at all bands

4.7, 4.3 and 5.1 dBi. The gain values in the optimal direction $\theta = -55°$ are 4.5, 4.3 and 1.1 dBi at 940 MHz, 1.84 GHz and 2.44 GHz, respectively (red lines in Fig. 8).

4 Conclusion

In this paper, a triple-layer triple-band stacked patch antenna working at high-order modes is proposed. Optimization of some design parameters (slots positions) under strict bound rules was necessary to obtain an optimal coupling between all layers of the structure and achieve good results. The antenna achieved a 10 dB return loss bandwidth of 41, 119 and 92 MHz, and a peak gain of 4.7, 4.3 and 5.1 dBi at 940 MHz, 1.84 GHz and 2.44 GHz, respectively. It can be used in the front end of a low-power indoor jamming device to cover mobile, Wi-Fi and Bluetooth communications.

References

1. 3GPP TS 36.101 V13.19.1 (2020) Evolved universal terrestrial radio access (E-UTRA); user equipment (UE) radio transmission and reception (Release 13)
2. Darimireddy NK, Ramana Reddy R, Mallikarjuna Prasad A (2015) Design of triple-layer double U-slot patch antenna for wireless applications. J Appl Res Technol 13(5):526–534
3. Lin D, Zhen YL, Yong JX, Gao LN, Jun F (2010) A compact microstrip slot triple-band antenna for WLAN/WiMAX applications. IEEE Antennas Wirel Propag Lett 13:1178–1181
4. Wing CM, Sai HW, Kwai ML, Kai FL (2013) Single-layer single-patch dual-band and triple-band patch antennas. IEEE Trans Antennas Propag 61(8):4341–4344
5. Mathew S, Anitha R, Deepak U, Aanandan CK, Mohanan P, Vasudevan K (2015) A compact tri-band dual polarized corner truncated sectoral patch antenna. IEEE Trans Antennas Propag 63(12):5842–5845
6. Boukarkar A, Xian QL, Jia WY, Peng M, Yuan J, Yi QY (2017) Highly integrated independently tunable triple-band patch antenna. IEEE Antennas Wirel Propag Lett 16:2216–2219
7. Tian L, Zhenghui X, Weiming L, Wu R (2017) Stacked tri-band microstrip patch antenna. In: Proceedings of IEEE 5th international symposium on electromagnetic compatibility (EMC-Beijing)
8. Kaur A, Khanna RL (2015) A stacked sierpinski gasket fractal microstrip antenna for IEEE802.11a/b/UNII-1 and C band applications. In: Proceedings of IEEE 4th Asia-Pacific conference on antennas and propagation
9. Goyal A, Kumar R, Gupta K, Singh DK (2015) A triple band stacked patch antenna with slotted ground structure. In: Proceedings of 1st international conference on futuristic trend in computational analysis and knowledge management
10. Manouare AZ, El Idrissi A, Ghammaz A, Ibnyaich S (2015) Broadband triple-band CPW-Fed patch antenna for WLAN/WiMAX operations. In: Proceedings of international conference on wireless networks and mobile communications
11. Ghassemi N, Rashed-Mohassel J, Neshati MH, Ghassemi M (2008) Slot coupled microstrip antenna for ultra wideband applications in C and X bands. Prog Electromagn Res 13:15–25
12. Zhu Y, Min QL, Hong QH (2016) A compact dual-band monopole antenna for 4G LTE and WIFI utilizations. In: Proceedings of IEEE MTT-S international microwave workshop series on advanced materials and processes for RF and THz applications

13. Guha D, Banerjee A, Kumar C, Antar YM (2012) Higher order mode excitation for high-gain broadside radiation from cylindrical dielectric resonator antennas. IEEE Trans Antennas Propag 60(1):71–77
14. Wang D, Bo Ng K, Wong H, Chan CH (2014) Differentially-fed higher-order mode patch antenna at 60 GHz band. In: Proceedings of IEEE international workshop on electromagnetics
15. Khan QU, Ihsan MB, Fazal D, Malik FM, Sheikh SA, Salman M (2017) Higher order modes: a solution for high gain, wide band patch antennas for different vehicular applications. IEEE Trans Veh Technol 66(5):3548–3554

Traffic Accident Analyzer: A Visual Analytics Tool for Traffic Accidents Dataset

Aljawharah Almajyul and Nadia Al-Ghreimil

Abstract The number of traffic accidents in Saudi Arabia is huge, despite all the measures put in place. The current problem in traffic accidents can be best solved by visualization. It will enable users to see the analytics presented clearly and enable decision-makers to gain new insights about the road or identify new patterns faster. This research enables people to capture perceptions of traffic from the large volume of traffic dataset. Traffic Accident Analyzer is proposed as a tool that provides visual analytics for authorities, in order to understand traffic better and limit accidents. Interactive map, word-cloud, histogram, donut chart, and calendar are five visualization techniques that are integrated together into the tool. The approach is illustrated as a case study for traffic accident visualization systems, using a traffic accident dataset from the city of Riyadh for a period of 3 years. The preliminary evaluation of the Traffic Accident Analyzer Tool referred to that the tool is appropriate for the intended purpose, also, the tool is acceptable to users based on usability testing.

Keywords Information visualization · Traffic accidents · Data analysis

1 Introduction

The rapid increase in economic growth leads to urban expansion and the use of motorization in a lot of countries. Approximately 40% of the population spends a daily equivalent of 1 hour on the road [1]. As a result, there is a significant increase in the number of vehicles which leads to congestion in traffic and accidents [1, 2]. According to the World Health Organization 2010, 1.24 million people were killed on the road, and up to 50 million people were injured worldwide. In Saudi Arabia,

A. Almajyul (✉)
College of Computer Science and Information Technology, Shaqra University, Shaqra, KSA, Saudi Arabia
e-mail: aalmjawel@su.edu.sa

N. Al-Ghreimil
College of Computer and Information Sciences, King Saud University, Riyadh, KSA, Saudi Arabia

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_80

there are 19 deaths per day and 4 injuries per hour [3]. Authorities claim that at least one traffic accident occurs in the Kingdom every minute. Al-Zahrani, the secretary-general of the Traffic Safety Committee, says, "Saudi Arabia witnesses up to 7,000 deaths annually, and over 39,000 injuries per year". According to Al-Zahrani, the Kingdom has one of the highest death records in the world; for 100,000 people, it has about 21 death instances [4]. In 2015, the cost of damages was estimated at an amount of SAR 54.9 billion [5]. According to statistics, the number of registered and used vehicles in KSA increased from 8.4 million vehicles in 1990 to 14.9 million vehicles in 2011 [5], approximately 100% in 21 years. The number of cars has increased even more with the recent decision to allow women to drive. We believe there is a real problem in traffic in Saudi Arabia that needs examination. Authorities need a system that enables them to get insights into traffic accidents [6]. According to Thomas and Cook, "Visual Analytics is the science of analytical reasoning facilitated by interactive visual interfaces" [7], meaning it combines automated analysis techniques with interactive visualizations for effective understanding, reasoning, and decision-making based on large and complex spatio-temporal data [8, 9]. Riyadh, being the Capital of Saudi Arabia, has the largest number of traffic accidents compared to other cities in Saudi Arabia [5]. We used Riyadh as a sample to test the traffic accident analyzer and the tool can be used with any dataset with the same attributes, and we seek to answer the questions: Could visualizing the locations and timing of traffic accidents, using customizable views, help decision-makers to carry out the appropriate procedures? Would visualizing time and date of accidents lead to creating a better understanding of traffic accidents and the reason behind them? These questions aim to achieve the research goal: to gain overall insights into traffic accidents and their causes by visually exploring accident data combining different visualization techniques, providing a tool that helps decision-makers visualize accidents in two aspects, place and time, and making views customizable to help answer specific queries, for example, number of accidents that occurred on weekends between 2:00 and 3:00 p.m.

2 Literature Review

Traffic data has never been produced in such large quantities as today. Researchers believe that understanding the dynamics of traffic data can help to handle most of the traffic-related problems and determine the black spots: the locations that have more frequent traffic accidents [10, 11]. Globally, countries suffer from Traffic accidents. Many studies try to address this situation in Western Asia, countries like Iran, Saudi Arabia, and Turkey as well as Southeast Asian countries like Singapore besides a few European countries. Most of them use one or a few visualization techniques like a map and charts, and a few focus on studying data mining algorithms rather than visualization. The diversity includes visualization techniques, datasets, tools, and programming languages that have been used. To our knowledge, not many studies employ combined visualization techniques like combining heatmaps and calendars with the map while applying multifiltering at the same time.

2.1 Depending on Maps

One of several studies that have been conducted in Iran was in Mashhad by Shafabakhsh et al. The aim was to show hotspots using information modeling. Spatial temporary analysis and GIS were used to identify the location of accidents. They found that the high-crash-prone zones were located within the vicinity of the city [12]. Another study aimed to improve road safety was conducted in Turkey by Yalcin and Duzgun. They identify the vehicle types mostly involved in road accidents. The importance of the study was to demonstrate that road safety can be achieved by exploring and analyzing traffic data [13]. The two studies above used maps with colored circles to mark hotspots and severity degree. Both similarly used 3D for better visualization of the results. While these methods generated good results, it is important to point out that the techniques employed were by all accounts complicated [12, 13].

2.2 Combining Maps with Other Techniques

Igual and Segui [14] presented a visualized map of accidents in Barcelona. The map depicted accidents by the hour of the day and day of the week. They showed the efficiency of the heatmap to understand the underlying issues of road accidents. Lokal et al. [15] provided a method for avoiding potential accidents in rural and urban areas. Histograms and data charts were used to study the observed patterns. A simple method for visualizing the impact of road accidents was proposed by Anwar et al. Using Singapore as a case study, they developed a relatively simpler way of depicting road accidents. Color was used to indicate the traffic speed on the map. Also, the bars and timeline charts were used to predict the nature of accidents. The road accidents were represented in a novel manner depending on color [11]. To better understand road accidents, Pack et al. proposed a visual analytics tool. They combined histograms, two-dimensional plots and parallel coordinates plots, and interactive maps. The tool comprised filter, ranking panels, and tabular view. The results demonstrated that it is possible to improve road safety by the utilization of traffic data [16]. Ramos et al. [8] presented a new approach to identifying black spots of road accidents in Portugal using visual analytics techniques. The approach focuses on the link between road elements and black spots by analyzing the relationship between these two aspects and defining the major contributory causes of accidents. Picozzi et al. [10] presented a web tool that could be used to have a quick overview of traffic data in Oulu. The tool integrated Chart, Map, and Calendar, making it reliable. To reduce road accidents, they recommended the use of space and time visualization which will enable decision-makers to precisely analyze events. Al Omar et al. [17] used a web platform and different visualization to analyze road accidents. They mainly used dot-based heatmap, and region-based visualizations. Other techniques that were used include linear-time and periodic-time visualizations. The map techniques used were powerful tools to represent accidents. Wibisono et al. [18] visualized traffic conditions in the road using a heatmap. They were able to generate a virtual location on the map. Colors were used to depict changes in traffic flow. In summary, some of the research used a tool to manage and visualize data which help to analyze different types of traffic data using various visualization techniques. Other studies relied on GIS technologies to analyze spatial data. While GIS is a good tool for analyzing complex spatial data, it is not supported to visualize non-spatial data. Several researchers used different programming languages to provide a custom-made program for analyzing traffic data. [8, 10, 18, 11, 16, 17] employed different APIs. Map, line chart, and bar chart were the techniques used extensively since they provide a good representation of traffic. Therefore, we employ map, line chart, and bar chart alongside other visualization techniques like word cloud, calendar, and pie chart since they provide effective results that achieve our insights and meet the needs of the stakeholders.

3 System Framework

Two phases cover the three levels of system architecture: data preparation, data processing, and data visualization. Figure 1 shows all phases and transitions of the system. The first process is to understand the dataset and prepare it for visualization thereafter, applying aggregation technique to perform a calculation on a set of values and return a single value. The final part is presenting various visualization techniques to show the accident data efficiently to meet the user insights.

3.1 Traffic Accident Dataset Understanding

We obtained the dataset from the General Department of Traffic in Riyadh. It contained around 242,000 records of traffic accidents in 3 years. There were three main tables: accident, vehicle, and parties and the attributes were approximately 55. In the dataset, there are inaccurate data where most values in Age take zero. Also, there are many missing values, such as accident report time attribute and incorrect data in Age, e.g. 1435, −3, −75, and −1066. Some data was hidden for privacy such as ID number, name, and address. It was necessary to preprocess, clean, and rearrange the dataset according to the needed visualization.

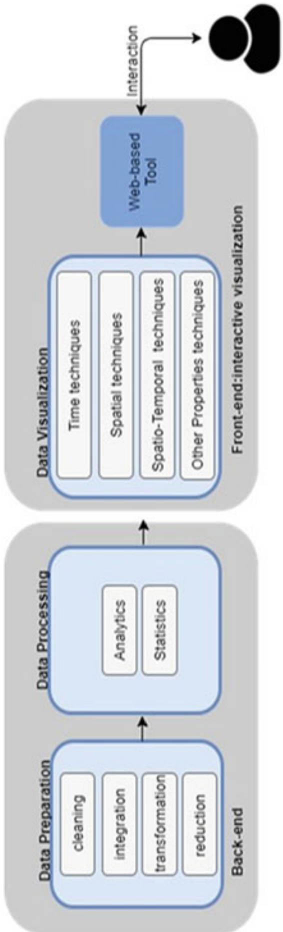


Fig. 1 System architecture

3.2 Data Preparation

As mentioned, the dataset had missing or duplicate values. The weather, for example, shows almost a fixed value (around 97% contained the same value). We removed special characters such as *! %?. And we separate the column that contained time and date into two columns, to handle each attribute separately. And we generated a new column for the Accidents' Days that has the days' names. The days' names are necessary to allow the user to select a specific day or to compare the weekend and other days. The initial format of accident coordinates in the dataset did not match the type of map we used. Therefore, we converted the coordinates from UTM to Decimal Degrees. Ultimately, we gained 246,814 records of accidents that have 19 attributes and various values; see Table 1.

Table 1 Attributes of the dataset

Features	Description
Type of accident	The kind of accident that occurred, e.g. bumped a moving vehicle, bumped a traffic signal, bumped a vehicle parked, and bumped animal
Accident zone	The accident zone (North–East–West–Middle)
Accident reasons	The reason behind the accident, e.g. cross the red signal, do not leave enough distance, lack of lights, preoccupation with driving, sleeping, and speeding
Street	The street numbers
Severity	The degree of accident severity, e.g. minor severity and death
X Coordinate	The Y coordinate in the DD system defining the accident location
Y Coordinate	The Y coordinate in the DD system defining the accident location
Neighborhoods	Neighborhoods of Riyadh such as Yarmook, Remal, and Aqeeq
Time	The time of accident in 24-h format
Dates	DD/MM/YYYY accident dates from 2013 to 2015
Age	Age of people involved in accidents
Nationality	The nationality of the driver of the vehicle
Weather	The weather condition when the accident occurred
Accident Angle	For example, face to face, side, and rear
Car Manufacturer	Car manufacturer of the vehicle involved in the accident
Vehicle Model	Year in which the vehicle was produced
Accident Day	The accident day, e.g. Saturday
Driver Gender	Male or female

3.3 Visualization Techniques

Data visualization is leveraging the human perceptual and offering interaction techniques to assist to answer queries [19]. We built the tool according to Shneiderman's principle (Overview, Zoom and Filter, Details on Demand) [10]. Traffic data visualization depends on four types of visualization techniques: Time, Spatial, Spatio-temporal, and other Properties [1]. The three types of techniques are combined into an interface, to allow users to make decisions and visually interpret the results. Map, Line chart, Word-cloud, Histogram, Pie chart, and Calendar are applied to explore the data and drill down charts.

4 Visual Analysis Views

4.1 Mapping

Through our questionnaire with the General Department of Traffic, we concluded that there is a need to know the black spots of accidents. A map is the main method to browse geographic data and an incredibly intuitive way to visualize data with longitude and latitude coordinates. First, we tried to apply a heatmap but it was not the best solution where we could not have the vision that we intended, and it was not compatible with the platform. After that, we applied a dot-based map and we faced an obstacle with the dimensions and spaces while using the zoom. Finally, we employed a Hexagonal binning map to show traffic accidents and make a density map. Using Hexbins, we can create a great mapping visualization and we must take into consideration the use of Hexbins sparingly and locally. The reason is that Hexbins allow data to be shown in areas of known size, not in arbitrary polygons, e.g. County and State. Also, we create a parameter to scale functions for mapping as follows:

```
[hexbinLong] = HexbixX(X*[Zoom]*10,Y*[Zoom]*10)/([Zoom]*10)
[hexbinLet] = HexbixY(X*[Zoom]*10,Y*[Zoom]*10)/([Zoom]*10)
```

We used a Zoom parameter to control the density of the hexagon grids to allow the user to control the binning size. The higher the zoom, the more binning in the view (see Fig. 2). The zoom parameter is called Scale in the user interface. Also, we multiply the parameter by 10 to give a pleasant grid of hexagons.

4.2 Histogram

Histograms are utilized to visualize many attributes such as age, time, gender, nationality, weather, and weekdays, and the user can exclude a certain item. For example, to visualize the age group and clarify which age group has the largest number of

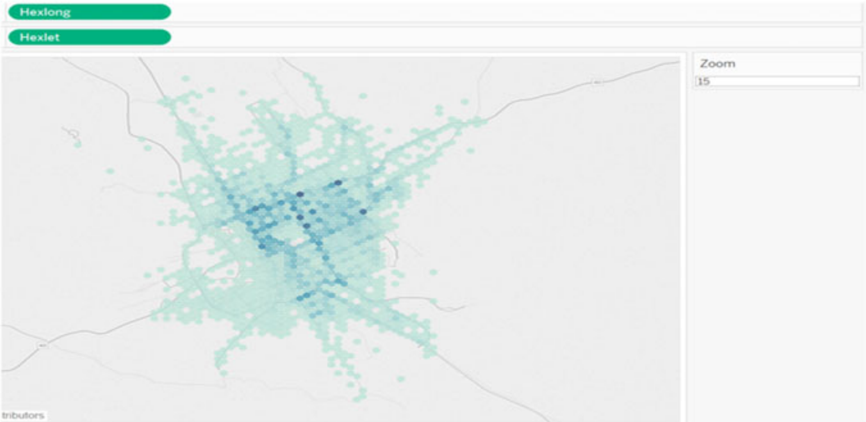


Fig. 2 Density map

accidents (see Fig. 7), we found the 25–34 age group is the most involved in traffic accidents when excluding 0. We used the countries' flags to refer to nationalities and we used X for the unknown; when excluding Saudi (see Fig. 7) we figure out easily Pakistanis were the most involved in the accidents. This may be because Pakistanis come at the top of nationalities that do driving and we lack the statistics that declare the number of licenses of each nationality.

4.3 Word Cloud

To have a big picture about the car manufacturers, we are conducting the word-cloud function, to know which cars are mostly involved in accidents. For instance, we can clearly see that the car Hyundai Elantra has the most accidents, therefore, it appears in a big font (see Fig. 7). We did not compare with statistics about the number of cars by make/model in Saudi Arabia.

4.4 Pie Chart (Donut)

We used a Pie Chart to visualizes the severity types (deaths, damage, minor/serious injuries) of accidents and show the bigger picture of how each severity type relates to the other (see Fig. 4).

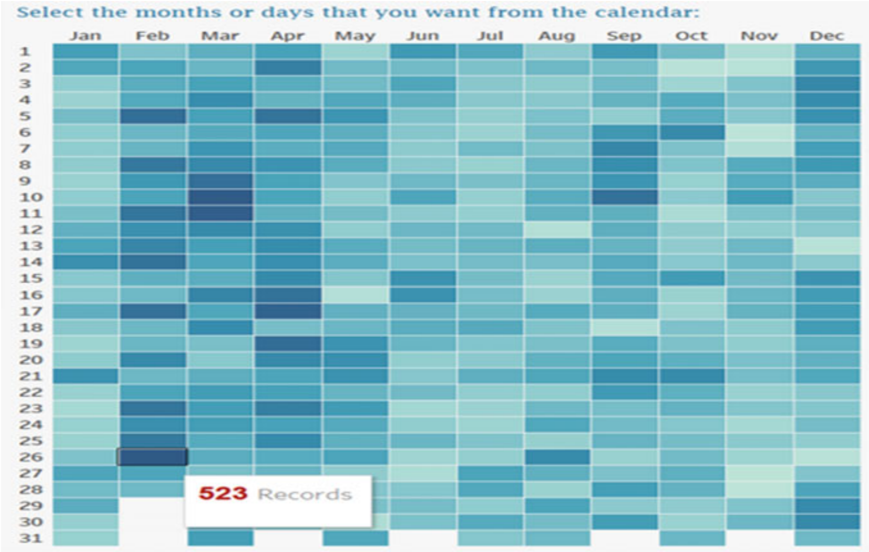


Fig. 3 Calendar

4.5 Calendar

Heatmap (calendar) was used to represent time series data of accidents and visualize tracked time over the 3 years. Calendar contributes to providing an overview of multiple years and provides the selection of a month and day for details-on-demand. Using color gradients, the user's eyes can easily pick out the darkest and lightest cells. The darkest cell refers to days with the most traffic accidents and vice versa (see Fig. 3). In the questionnaire, General Department of Traffic mentioned it is difficult to determine which time periods have the greatest number of traffic accidents. Thus, we think this type of visualization technique will help to address this difficulty.

5 Case Study

To demonstrate the effectiveness of Traffic Accident Analyzer, a usability test was conducted with 10 participants on 4 scenarios. The first scenario is a user wants to know locations of traffic accidents that occurred in April 2014, and the driver of the vehicle was a man. With a few settings, a result like in Fig. 4 can be achieved. On the map the locations of accidents can be seen and those with a higher number of accidents are clearly distinguished with darker color.

The second scenario is a user wants to know which nationality drivers caused most accidents during 2014–2015 and how the weather was. Pakistani came second

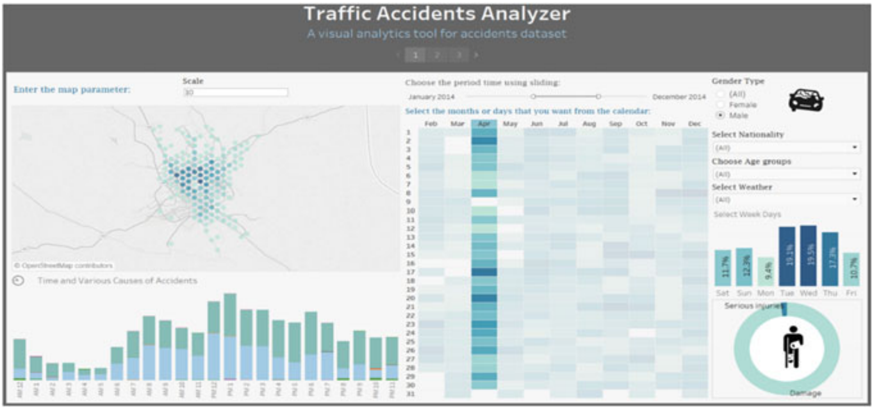


Fig. 4 Case1_Four steps to get the result: (1) Choose the page. (2) Adjust 2014 from the sliding bar. (3) Select April from the calendar. (4) From gender type filter, check on man option

after the Saudis as we mentioned before. Indians, Egyptians, and Yemeni came after that (see Fig. 5). The result is the weather was clear most of the times. Probably, the data entered was not accurate, or because the dataset represents the weather in Riyadh and absolutely the result will be different in other cities.

The third scenario is a user wants to select the location that has the greatest number of accidents during the 3 years and what the biggest causes of these accidents are. Figure 6 refers to speeding as the biggest cause of accidents during that time. The fourth scenario is a user desires to know which vehicle manufacturer has a high

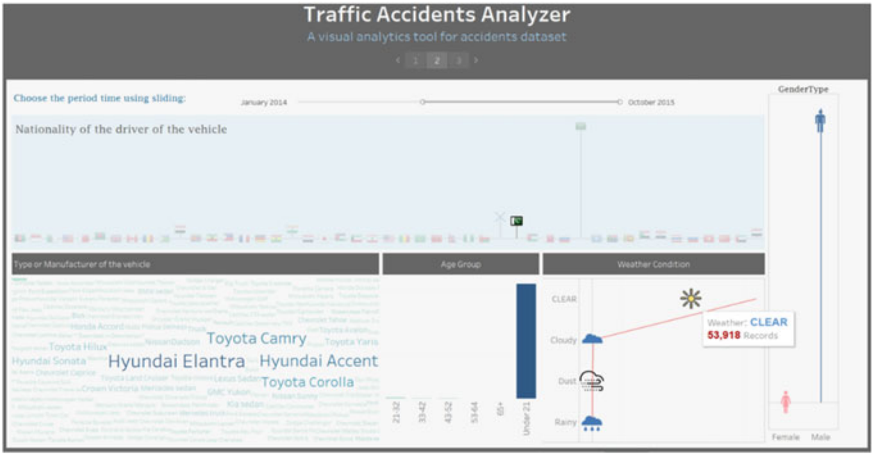


Fig. 5 Case2_Four steps to get the result: (1) Choose the page. (2) Drag the sliding bar to be between 2014 and 2015. (3) Pass the mouse on the country flag to know the nationality and the number of accidents. (4) Pass or click on the weather icons for more information

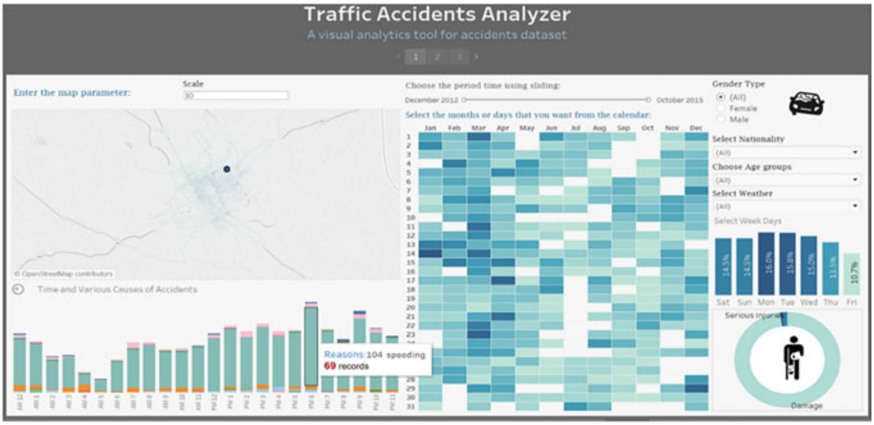


Fig. 6 Case3_ Four steps to get the result: (1) Choose the page. (2) Ensure the sliding bar extended along the length of years. (3) Select the darkest point on the map. (4) Pass the mouse on the stacked bar to know the accident's reason

number of accidents in 2013. Hyundai Elantra was the vehicle most recorded in accidents in that year (see Fig. 7).

Usability test was performed to measure efficiency, effectiveness, and user satisfaction. Most participants were workers in the traffic sector. The test has two groups where the first group used the system directly while the other online. Each participant had to fill pre/post-questionnaires. The pre-questionnaire is designed to collect background data about the participant experience. The post-questionnaire was to measure the user satisfaction with the system. Also, we tried to get the user opinion about the experience and if they had encountered any issues during the test.

Table 2 presents the result of eight scenarios performed by the users. We recorded

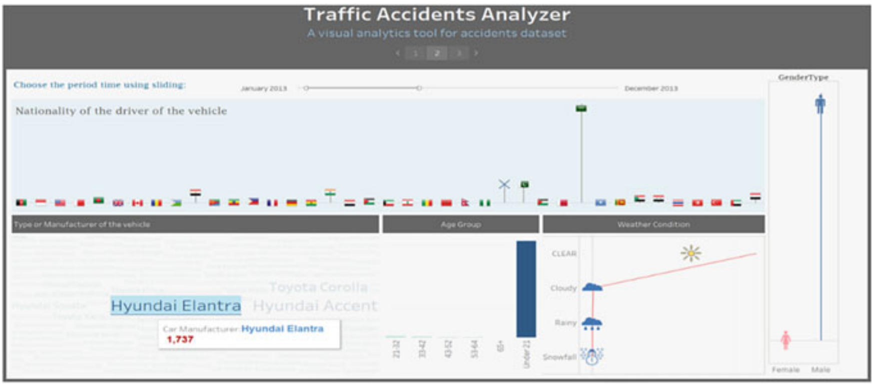


Fig. 7 Case4_ Three steps to get the result: (1) Choose the page. (2) Drag the sliding bar to set 2013. (3) Pass the mouse on the car manufacturer names to know the number of accidents

Table 2 The effectiveness results of using the system

Tasks	Number of errors of users										Avg.
	1	2	3	4	5	6	7	8	9	10	
Task 1	0	0	0	0	0	0	0	0	0	0	0
Task 2	0	0	0	0	0	0	1	0	0	0	0.2
Task 3	0	0	0	0	0	0	0	0	0	0	0
Task 4	0	0	0	0	1	0	0	1	0	0	0.4
Task 5	0	0	0	0	0	0	0	0	0	0	0
Task 6	0	0	0	0	0	0	0	0	0	0	0
Task 7	0	0	0	1	0	0	0	0	0	0	0.2
Task 8	0	0	0	0	0	0	0	0	0	0	0

errors that occurred either from the user or those caused by the system. As shown in the testing results, the number of errors when the users test the system tasks were very low and most tasks were completed within time. The errors that users encountered in the fourth task were about selecting more than one item at the same time. Table 3 demonstrates the efficiency result which shows the time consumed by the participant to perform each task.

We noticed that there was a significant difference between the two groups in the time consumed, where the time taken by the online group was significantly larger than the other group. Mostly, the reason is the speed of the Internet connection and the speed of the user's device. In general, the results indicate the time consumed is acceptable. To evaluate user satisfaction, we interviewed the users and asked them to fill the post-questionnaire. The post-questionnaire is based on the SUS questionnaire (System Usability Scale). Fortunately, Fig. 8 shows that all participants have a score of 70 and above and the average is 83, which means that they like the tool. Some of the participants pointed to the map as wonderful which shows the black spots of the accidents easily, it is also interesting.

Table 3 Efficiency results of using the system

Tasks	Time taken (in seconds)										Avg.
	1	2	3	4	5	6	7	8	9	10	
Task 1	6	5	5	4	5	9	10	10	9	11	7.4
Task 2	25	37	30	21	28	34	64	50	63	60	41.2
Task 3	85	60	35	29	11	37	45	43	50	62	45.7
Task 4	70	49	34	60	28	57	66	57	65	72	55.8
Task 5	37	15	21	16	11	24	35	30	32	38	25.9
Task 6	30	15	15	21	19	27	32	34	32	29	25.4
Task 7	25	20	25	19	21	26	34	28	36	30	26.4
Task 8	19	22	17	10	18	33	29	33	34	36	25.1

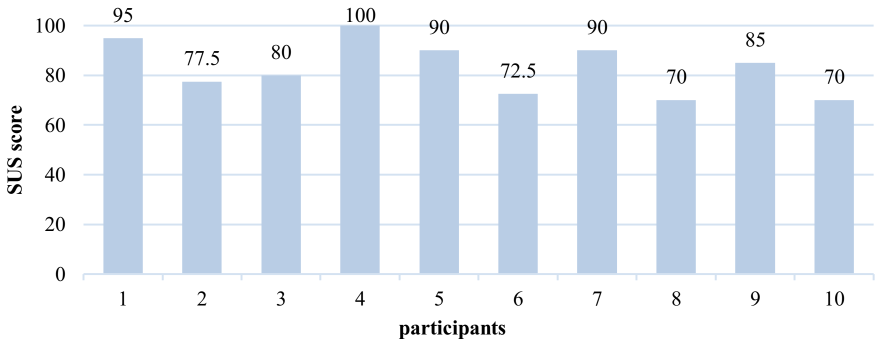


Fig. 8 SUS score for each participant

6 Conclusion

There are famous sayings: "The data is the new gold" and "A picture is worth a thousand words". Traffic Accident Analyzer derived its concept from these two meanings. In this paper, we built a system that enables users to analyze traffic data and present it clearly to allow decision-makers faster gain new insights into traffic accidents and identify new patterns using various visualization techniques. We faced some challenges in creating a map that shows black points effectively. The existence of missing values and features in the dataset of traffic accidents led to the reduction of system features. For example, we wanted to distinguish among all types of transport, heavy trucks buses, motorcycles… etc., this feature is not included in the dataset. Finally, the tool showed good results using the usability test. The research questions that we mentioned earlier were answered and it was as much as we intended. We could increase the effectiveness of the system by adding the forecasting into the system, including other datasets from other sectors such as hospital records and car insurance data for more accurate results. In the end, we would like to point out that we have learned the importance of having clean, correct, and reliable source data.

References

1. Chen W, Guo F, Wang FY (2015) A Survey of Traffic Data Visualization. IEEE Trans Intell Transp Syst 16(6):2970–2984
2. Guo H, Wang Z (2010) TripVista : triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. In: Vis. Symp. PacificVis.2011 IEEE Pacific, pp 163–170
3. Mansuri FA, Al-zalabani AH, Zalat MM, Qabshawi RI (2015) Road safety and road traffic accidents in Saudi Arabia. Saudi Med J 36(4):418–424
4. Traffic accidents occur every minute in KSA, http://www.arabnews.com/saudi-arabia/news/717596. Last accessed: 2017/9/26

5. Mohamed HA (2015) Estimation of socio-economic cost of road accidents in Saudi Arabia: willingness-to-pay approach (WTP). Adv Manag Appl Econ 5(3):19
6. R. Spence (2001) Information visualization, 3rd edn. Springer, Switzerland
7. Thomas JJ, Cook KA (2005) Illuminating the path: the research and development agenda for visual analytics. IEEE Comput Soc, 184
8. Ramos L, Silva L, Santos MY, Pires JM (2015) Detection of road accident accumulation zones with a visual analytics approach. Procedia Comput Sci 64:969–976
9. Viégas FB, Wattenberg M (2007) Artistic data visualization: beyond visual analytics. Online Commun Soc Comput 4564:182–191
10. Picozzi M, Verdezoto N, Pouke M, Vatjus-Anttila J, Quigley A (2013) Traffic Visualization—applying information visualization techniques to enhance traffic planning. In: Int. Conf. Comput. Graph. Theory Appl. Int. Conf. Inf. Vis. Theory Appl., 554–557
11. Anwar A, Nagel T, Ratti C (2014) Traffic origins: a simple visualization technique to support traffic incident analysis. IEEE Pacific Vis Symp, 316–319
12. Shafabakhsh GA, Famili A, Bahadori MS (2017) GIS-based spatial analysis of urban traffic accidents: case study in Mashhad Iran. J Traffic Transp Eng English Ed 4(3):290–299
13. Yalcin G, Duzgun HS (2015) Spatial analysis of two-wheeled vehicles traffic crashes: Osmaniye in Turkey. KSCE J Civ Eng 19(7):2225–2232
14. Igual L, Seguí S (2017) Barcelona accidents 2017, https://dieguico.carto.com/viz/50b06d8c-13ab-11e5-8619-0e4fddd5de28/embed_map, last accessed 2017/11/12
15. Lokala U, Nowduri S, Sharma PK (2017) Road accidents Bigdata mining and visualization using Support Vector Machines (SVM). In: 19th international conference on machine learning and applications
16. Pack M, Wongsuphasawat K, VanDaniker M, Filippova D (2009) ICE—visual analytics for transportation incident datasets. Inf. Reuse, 200–205
17. Alomar A, Alrashed N, Alturaiki I, Altwaijry H (2017) How visual analytics unlock insights into traffic incidents in Urban Areas. Soc Comput Soc Media Appl Anal 10283:378–393
18. Wibisono A, Jatmiko W, Wisesa HA, Hardjono B, Mursanto P (2016) Traffic big data prediction and visualization using Fast Incremental Model Trees-Drift Detection (FIMT-DD). Knowledge-Based Syst 93:33–46
19. What is data visualization? http://searchbusinessanalytics.techtarget.com/definition/data-visualization, last accessed 2017/12/02

Identifying the Effects of COVID-19 on Psychological Well-Being Through Unsupervised Clustering for Mixed Data

Katharina Lingelbach, Sabrina Gado, Doris Janssen, Daniela Piechnik, Martin Eichler, Dennis Knopf, Leopold Hentschel, Markus Schuler, Daniel Sernatinger, and Matthias Peissner

Abstract The COVID-19 pandemic has a strong worldwide impact on not only the health and economic sectors but also the (socio-)psychological functioning of individuals. Since psychological health is an important protective factor to prevent diseases, it is crucial to identify individuals with increased vulnerability during the crisis. 275 adults participated in a German online survey from April until August 2020 which investigated health-related, social, behavioral, and psychological effects

K. Lingelbach (✉)
Institute of Human Factors and Technology Management (IAT), University of Stuttgart, Stuttgart, Germany
e-mail: katharina.lingelbach@iat.uni-stuttgart.de

S. Gado · D. Janssen · D. Piechnik · M. Peissner
Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany
e-mail: sabrina.gado@iao.fraunhofer.de

D. Janssen
e-mail: doris.janssen@iao.fraunhofer.de

D. Piechnik
e-mail: daniela.piechnik@iao.fraunhofer.de

M. Peissner
e-mail: matthias.peissner@iao.fraunhofer.de

M. Eichler · L. Hentschel · M. Schuler
National Center for Tumor Diseases, University Hospital Dresden, Dresden, Germany
e-mail: martin.eichler@uniklinikum-dresden.de

L. Hentschel
e-mail: leopold.hentschel@uniklinikum-dresden.de

M. Schuler
e-mail: markus.schuler@uniklinikum-dresden.de

D. Knopf · D. Sernatinger
Seracom GmbH, Stuttgart, Germany
e-mail: dennis.knopf@seracom.de

D. Sernatinger
e-mail: daniel.sernatinger@seracom.de

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_81

of the COVID-19 pandemic. We here introduce an unsupervised clustering approach suitable for mixed data types combining the Gower distance with the Partitioning Around Medoids (PAM) algorithm k-Medoids. We were able to identify three clusters differing significantly in subjects' well-being, psychological distress, and current financial and occupational concerns. The clusters also differed in age with younger persons reporting greater financial and occupational concerns, increased anxiety, higher psychological distress, and reduced subjective well-being. Features with the strongest impact on the clustering were examined using a wrapping method and the feature importance implemented in the random forest. Particularly, answers regarding financial and occupational concern, psychological distress, and current well-being were decisive for the assignment to a cluster. In summation, the clustering approach can identify persons with weakened psychological protective factors allowing them to provide tailored recommendations for preventive actions based on the cluster affiliation, e.g., via a web application.

Keywords Psychological well-being · COVID-19 pandemic · Mental and public health · Machine learning · Mixed data clustering · k-Medoids · Gower distance · Partitioning around Medoids (PAM) algorithm · Data mining · Random forests

1 Introduction

The COVID-19 pandemic not only has a major worldwide impact on health and economic sectors but also on psychological well-being and quality of life. The study of these effects is particularly relevant, since social and psychological resources are known to be important protective factors for diseases [1]. The possibility to identify especially vulnerable subgroups in the population especially vulnerable during or due to the consequences of the COVID-19 pandemic would allow for fast implementation of targeted preventive interventions. Consequently, almost immediately after the outbreak several (inter-)national projects started investigating not only health-related and epidemiological effects but also the impact on the economic, social, and psychological status within longitudinal surveys [2–5]. Li et al. [6] examined the psychological effects of the COVID-19 pandemic by analyzing posts of 17,865 active members in the Chinese social network Weibo using machine learning (ML) methods. The authors observed that posts associated with negative emotions (e.g., anxiety or depression) and sensitivity to social risks increased, while emotionally positive posts and life satisfaction decreased after the onset of the pandemic. Jha et al. [7] used Bayesian networks and classical ML methods on data of a representative sample of 17,764 adults in the US to identify key factors for mental health during the COVID-19 pandemic. They were able to estimate the level of mental health with an accuracy of about 80%. Based on these findings, there is a great need for further research to confirm preliminary findings regarding the psychological, emotional, and social effects of the COVID-19 pandemic and to identify individuals vulnerable in their psychological well-being and with a special disposition for mental disorders during and due to the current living conditions under COVID-19.

ML approaches in particular have the ability to learn rules which identify complex relationships between variables, further called features, in large datasets. Cluster analyses are data-driven statistical classification techniques to reveal which elements in a dataset are similar with respect to previously selected features [8, 9]. Most clustering procedures assign elements to clusters based on their distance to the cluster centers called centroids. Many well-established algorithms are suited to cluster numerical and independent data [10]. However, real-world datasets often comprise mixed data types also including ordinal (categorical but in an ordered structure) and nominal (categorical without an ordered structure) features [11]. Major challenges for clustering techniques of large mixed datasets are to a) define an adequate measure of distance in a multi-dimensional space, b) choose an appropriate cluster center initialization method, and c) identify the optimal number of clusters [12].

The Gower distance is suggested to be a suitable metric to estimate the similarity between elements in mixed datasets [13]. It comprises several distance matrices, which are calculated depending on the data type ranging from 0 to 1 between each element. For numerical interval and ordinal data (ranked and adjusted for ties), the range-normalized Manhattan distance is typically used [13, 14]. For nominal data, the variables are dummy-coded first by assigning either a 0 or 1 for each category depending on the absence or presence of the qualitative attribute. In the second step, the Dice coefficient representing the distance between categories (0 = equal) is calculated [13, 15, 16].

The individual distances between the k variables of two elements calculated above are combined to the Gower distance. It is defined as the average similarity S_{ij} between the two elements i and j for each variable k [13]

$$S_{ij} = \frac{\sum_{k}^{n} \delta_{ijk} s_{ijk}}{\sum_{k}^{n} \delta_{ijk}} \tag{1}$$

where δ_{ijk} represents a quantity, which is equal to 1 if the kth variable can be compared for the elements i and j; otherwise, it is set to 0. The matrix of the Gower distance contains the scores of all pair-wise element comparisons. Combining the Gower distance with a Partitioning Around Medoids (PAM) algorithm such as the k-Medoids [17] is suggested to be well suited to identify evenly distributed clusters in mixed datasets [11, 18–20]. The medoid of a cluster in PAM algorithms represents the most centrally located object, which is the median, in a cluster. Consequently, the objects with the smallest sum of relative distances to all other objects are selected as initial medoids and the initial clusters are obtained by assigning each object to the nearest medoid [17].

In this study, an explorative data analysis is used to cluster subjects differing in their psychological well-being during COVID-19. For this purpose, we applied an unsupervised clustering approach combining the Gower distance with the PAM algorithm k-Medoids to a real-world mixed dataset.

2 Methods

2.1 Description of the Survey and Data Collection

The data was extracted from a still-running German online survey called *WIBCE* (German: *Was ich bei Corona erlebe*; English: *What I experience during Corona*) which started on April 1, 2020. Here, responses until August 24 were analyzed. Participants could voluntarily fill in the online questionnaire as many times as they liked even on a daily basis. However, some modules (e.g., items examining depression) of the questionnaire were provided at a defined time interval (e.g., with at least 7 days in between). Since answers to the questions were not mandatory, the number of provided data differs for each feature. The items of the online questionnaire comprise epidemiological and health-related information (general as well as COVID-19 specific, e.g., whether a person is at risk for severe illness), the behavior in everyday life during the pandemic, social factors (e.g., the number and form of contacts with others) as well as psychological factors. To access psychological well-being and related concepts, we used established scales and questionnaires, which are the Patient Health Questionnaire-4 (PHQ-4, [21]) for psychological distress including an anxiety [22] and depression [23] dimension as well as the EQ-5D-5L examination of health-related quality of life [24–26]. In addition, we asked participants to rate their current well-being, concerns regarding infection with SARS-CoV-2, and their future financial and occupational situation. To quantify participants' contact behavior, we computed a score based on contact-related items (e.g., numbers of contacts to other persons in the last 24 h and contacts at work). An overview of the features, their ranges, and their meaning is provided in Table 1. For the cluster analysis, only the first entry of a person and only participants providing answers to all variables related to psychological well-being were included.

2.2 Description of the Sample

A total of 173 participants (39.9% male) were included in the cluster analysis with an average age of $M = 44.81$ years ($SD = 13.31$; range: 18 to 74). The average household size was 2.9 persons and 37.0% reporting to live with under-age children. 5.8% of the participants reported symptoms associated with COVID-19 and 40.5% reported criteria assigning them to a group at higher risk of severe COVID-19 based on the definition of the Robert Koch Institute. The educational level was rather high compared to the average German population with 79.8% graduated at least high

Table 1 Overview of the features

Feature name	Interpretation
Features used in the cluster analyses	
Current well-being	Range: 0 (bad) to 4 (excellent)[a]
Occupational concern	Range: 1 to 4, where 1 indicates no concerns and 4 represents major concerns regarding the future occupational situation[a]
Financial concern	Range: 1 to 4, where 1 indicates no concerns and 4 major concerns[a]
Infection concern	Range: 0 (not concerned at all) to 4 (very concerned)[a]
Quality of life (EQ-5D-5L)	Range: 0 to 1, with 0 indicating low perceived quality of life[b]
Depression (PHQ-2)	Range: 0 to 6, higher values indicating greater perceived depression[a]
Anxiety (GAD-2)	Range: 0 to 6, higher values indicating greater perceived anxiety[a]
Additional features used in the prediction of cluster affiliation	
Age	Range: 18 to 74[a]
Sex	Dummy variable (0 and 1) with a value 0 is assigned to a female
Psychological distress (PHQ-4)	It is a sum score of the PHQ-2 and GAD-2 with a range 0 to 12. Higher values indicate greater psychological distress[a]
Education	Values range from 0 representing no school leaving certificate to 4 indicating a qualification for university entrance[a]
Income	Range: 0 to 8, where 0 indicates a net household income below 500€ and 8 a net household income above 7000€[a]
COVID-19 risk factor	Dummy variable (0 and 1). A value of 0 indicates that the person is not at a greater risk for a severe infection
Sum of risk factors	Range: 0 to 4. It is the sum of reported risk factors for one person[a]
Contact with others	Range: 2 to 25, higher scores indicate more active and frequent contact with other people[a]
Multiple participation in WIBCE	Range: 1 to 85 representing the number of participations of a person in the survey since April 2020[a]

[a]step size = 1
[b]step size = 0.01

school (here the German Abitur). 58.4% reported a net household income of more than 3,000 € per month.

2.3 Cluster Analysis

Cluster identification. To identify subsets of people with similar characteristics, an unsupervised cluster analysis was used. Since here the investigated dataset comprises not only numerical variables (e.g., age) but also ordinal variables (e.g., scores with a short range (0–5)) and nominal variables (e.g., gender), we used a clustering approach suitable for mixed datasets. Therefore, the Gower distance as a similarity measure was combined with the PAM algorithm k-Medoids. Numerical data was scaled using the MinMaxScaler from sklearn (0.23.2) in Python (3.8.6). The features were further z-standardized within the calculation of the Gower distance using the Gower package (0.0.5). Clusters were identified using the k-Medoids algorithm implemented in sklearn_extra (0.1.0b2).

Determination of number of clusters. To determine the number of clusters, several methods have been introduced [27]. Here, we compared the Elbow method and Silhouette coefficient implemented in sklearn (0.23.2). In the Elbow method, we calculated the sum of squared errors of the distance between the sample points in each cluster and the centroid of the cluster iteratively for the different number of clusters k. Smaller values indicate more homogeneous data within the clusters. The choice of k is typically done by inspecting the Elbow plot showing the overall error as a function of the cluster number. Typically, the error drops sharply for small cluster numbers and then suddenly only a little error reduction can be gained by further increasing the cluster number. This elbow point is used as the recommended number of clusters. As an alternative, the Silhouette coefficient is a measure of how similar an element (e.g., a person) is to the associated cluster compared to another adjacent one. The Silhouette coefficient ranges from -1 to $+1$, with a high value indicating that the element has a high similarity with its own cluster and a low similarity with the adjacent. In case of deviating results, we would have explored which number of clusters is more suitable for the dataset by analyzing the clusters' features statistically.

Cluster description. For the description of the resulting clusters and their properties, explorative statistical comparisons for the demographical and psychological features were performed. Differences between the clusters were examined using 5,000-fold bootstrapped 95% confidence intervals (CI) of the mean value with no overlap between the CI indicating a statistical significance of $p < 0.01$ and a partial overlap without including the means a statistical significance of $p < 0.05$ [28].

Prediction of the cluster affiliation. We were further interested in which features were particularly informative for the clustering and whether we can predict to which cluster an individual belongs based on the given responses in the questionnaire. Therefore, supervised ML algorithms were applied using the cluster affiliation as labels and the data of 156 persons (participants with missing values in any of the features were excluded). Four different ML algorithms were compared: (1) a k-nearest neighbor classifier, (2) a logistic regression, (3) a random forest, and (4) a support vector machine. The models were implemented in Python (3.8.6) using sklearn (0.23.2) [15]. The dataset was divided into a training and a test set with a stratified 80:20 split. Hyperparameters of the algorithms were optimized using the training data in a fivefold cross-validated grid search and evaluated with the performance measure accuracy balanced for the number of samples per class. The feature wrapping method sequential feature selection using the library mlxtend [29] with accuracy as an evaluation metric was used to reduce redundant features. The relevance of each feature in the prediction was calculated from an embedded method of the random forest accessible via the attribute *feature importance*. A fivefold cross-validation with balanced accuracy as a performance metric was chosen to evaluate the prediction. For the comparison of the model performance, we used the random level based on the number of classes ($100/n\_classes$) and a random level estimated by a dummy classifier [15] with the method *stratified* as a reference.

3 Results

3.1 Determination of Number of Clusters

We observed a consistent suggestion of $k = 3$ clusters for the adequate number of clusters with both, the Elbow method and Silhouette coefficient (see Fig. 1 and Fig. 2). Table 2 provides an overview of the descriptive analysis of the clusters regarding their demographic information. The clusters are comparable with respect to their size as

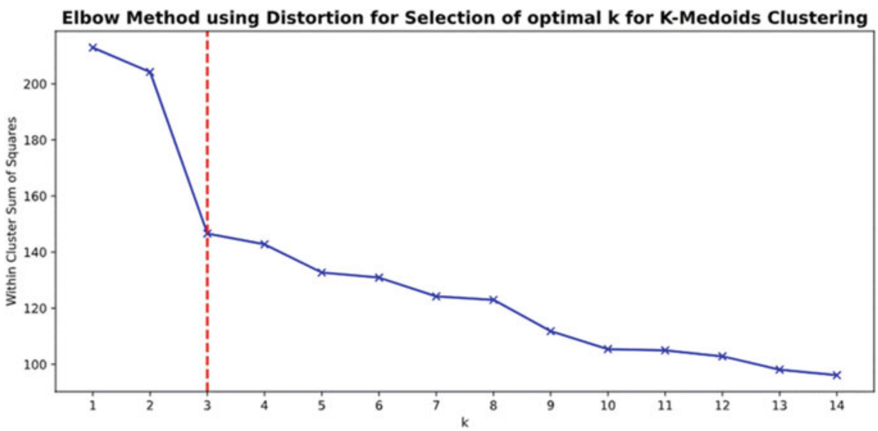


Fig. 1 Elbow plot shows the overall error using the within-cluster sum of squares (y-axis) as a function of the cluster number k (x-axis). A bend comparable to an elbow is visible at k of three indicating a suitable number of clusters

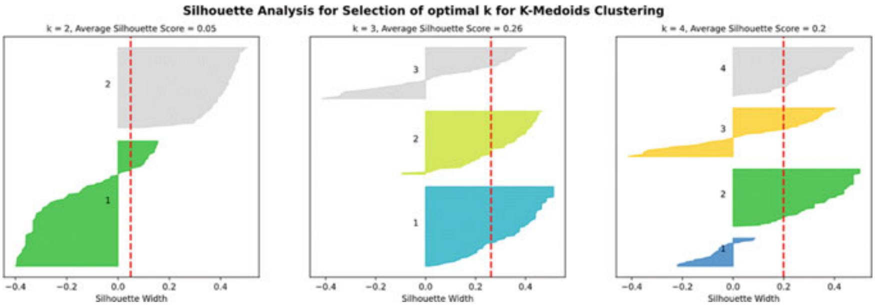


Fig. 2 The Silhouette analysis returns a score of 0.26 at $k = 3$ indicating that three clusters are a good choice. Here the Euclidean distance was used as metric

Table 2 Overview of the descriptive statistic for the clusters

Cluster	C1	C2	C3
N	71	56	46
Sex (male in %)	33.8	46.4	41.3
Mean age in years (SD)	47.85 (13.09)	44.05 (11.81)	41.04 (14.10)
Mean number of persons in household (under-age)	2.83 (0.72)	3.02 (0.79)	3 (0.5)
Belonging to a COVID-19 risk group in %	40.8	37.5	43.5

well as the number of persons in a household, gender, and belonging to a COVID-19 risk group.

3.2 Statistical Analysis of the Clusters

To explore the differences between the three clusters (C1, C2, and C3), cluster means and their 95% confidence intervals (CI) were calculated for each feature via bootstrapping with 5,000 repetitions. Results are provided in Fig. 3 and Table 3. A significant difference for age with $p < 0.05$ between C1 (older) and C2, and C3 (younger) C1 was found. Significantly, more female participants were assigned to C1 compared to C2 ($p < 0.05$) but not C3. There is no difference in the distribution of female and male participants between C2 and C3. The clusters differ noticeably in their ratings regarding the current well-being with the highest values in C1, decreased well-being in C2 (C1 vs. C2, $p < 0.05$), and lowest well-being in C3 (C3 vs. C1 and C2, $p < 0.01$). C1 reported the least concern regarding the occupational and financial situation followed by C2 and C3 (all with $p < 0.01$). C3 showed the significantly highest concern. Interestingly, C3 on average reported a significantly lower income than C1 and C2 (with $p < 0.05$). There was a difference between C3 and C1, C2 indicating higher concerns regarding an infection with the SARS-CoV-2 in C3 (with $p < 0.05$).

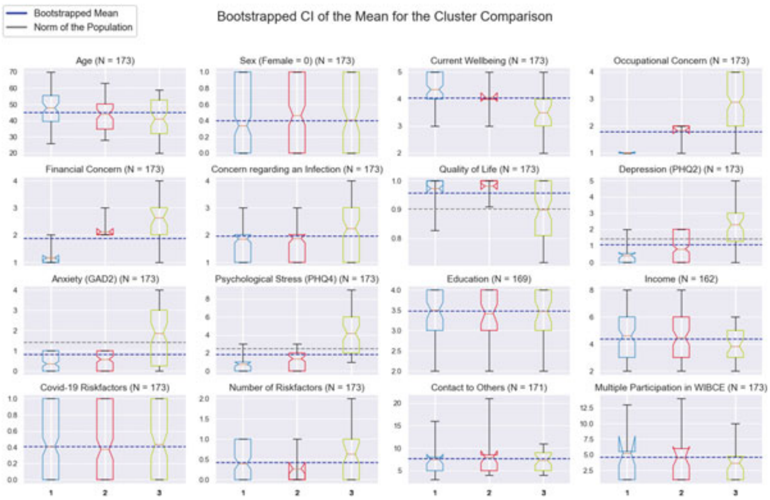


Fig. 3 Bootstrapped means and 95% confidence intervals (CI) of the clusters (blue, red, and green) for each feature. Notches in the boxes of the plot visualize the upper and lower boundaries of the CI with the orange line representing the mean of each cluster and the blue line representing the overall mean of the clusters. The gray line represents a population norm as reference with $M = 0.902$ for the EQ-5D-5L quality of life index [30], $M = 1.4$ for the PHQ-2 [23], $M = 1.4$ for the GAD-2 [22], and $M = 2.5$ for the PHQ-4 [21]. The box comprises 50% of the distribution from the 25th to the 75th quartile. The ends of the whiskers represent the 5th and 95th quartile of the distribution

Regarding the quality of life, there was no difference in the ratings between C1 and C2 but a significant difference between both and C3 (with $p < 0.01$) with the least quality of life in the latter. Analyzing the two dimensions (1) depression with the PHQ-2 and (2) anxiety with the GAD-2, we observed a significant difference between C1 and C2 ($p < 0.05$) as well as between C1 and C2 to C3 ($p < 0.01$). C1 reported the lowest values for depression and anxiety and C3 reported to be significantly more depressed and anxious. For the sum score PHQ-4 of the two dimensions indicating psychological distress, there was a significant difference among all clusters at $p < 0.01$ with the lowest values in C1 and highest psychological distress in C3. There are no differences between the clusters for education, belonging to a COVID-19 risk group, and contact with others. However, in case the participants reported to belong to a COVID-19 risk group, participants of C3 reported on average more factors indicating their belonging to the risk group compared to C2 ($p < 0.05$) but not C1. Finally, participants of the C1 filled in the survey more often compared to C3 ($p < 0.05$). To conclude, C3 revealed the least well-being and quality of life, higher depression, anxiety, and psychological distress. Furthermore, it reported higher concerns regarding (a) the occupational and (b) financial situation as well as (c) a possible infection with the SARS-CoV-2.

Table 3 Bootstrapped means and 95% CI with the lower and upper bounds for the features and clusters

Feature	C1	C2	C3
Age	47.84 [44.73, 50.92]	44.00 [40.98, 47.14]	41.03 [36.98, 45.09]
Sex	0.34 [0.23, 0.45]	0.46 [0.34, 0.59]	0.41 [0.26, 0.57]
Current well-being	4.35 [4.18, 4.51]	4.07 [3.91, 4.23]	3.50 [3.22, 3.78]
Occupational concern	1.01 [1, 1.04]	1.84 [1.71, 1.95]	2.89 [2.61, 3.15]
Financial concern	1.16 [1.07, 1.24]	2.11 [2, 2.23]	2.63 [2.35, 2.91]
Concern regarding an infection	1.84 [1.70, 1.99]	1.86 [1.70, 2.02]	2.24 [1.96, 2.54]
Quality of life (EQ-5D-5L Index)	0.97 [0.96,0.98]	0.98 [0.97, 0.99]	0.90 [0.86, 0.93]
Depression (PHQ-2)	0.40 [0.23, 0.58]	0.78 [0.57, 1]	2.30 [1.89, 2.72]
Anxiety (GAD-2)	0.35 [0.23, 0.48]	0.57 [0.43, 0.73]	1.87 [1.39, 2.37]
Psychological distress (PHQ-4)	0.75 [0.52,1]	1.36 [1.05, 1.70]	3.41[4.18, 5.02]
Education	3.50 [3.29, 3.70]	3.43 [3.17, 3.65]	3.48 [3.20, 3.72]
Income	4.60 [4.13, 5.07]	4.47 [3.92, 5.02]	3.84 [3.30, 4.40]
COVID-19 risk factor	0.41 [0.30, 0.52]	0,38 [0.25, 0.5]	0.43 [0.28, 0.59]
Number of risk factors	0.39 [0.25, 0.55]	0.27 [0.14, 0.43]	0.63 [0.39, 0.91]
Contact with others	7.53 [6.52, 8.68]	7.97 [6.73, 9.40]	7.36 [6.27, 8,6]
Multiple participation in WIBCE	5.25 [3.30, 8.00]	4.54[3.16, 6.32]	

3.3 Prediction of the Cluster Affiliation

For an explainable machine learning approach, it is of particular value to know the features' relevance in the clustering. Therefore, we applied supervised machine learning to examine whether we can predict the cluster affiliation with a high accuracy and use a feature wrapping method and the feature importance implemented in the random forest. Table 4 provides the performances of the explored algorithms and their hyperparameters chosen by the grid search. The chance level estimated by the dummy classifier was set to 0.313.

The feature wrapping method selected (1) *current well-being*, *concern* regarding the (2) *occupational* and (3) *financial situation,* and (4) *psychological distress* (PHQ-4) as informative features for almost all algorithms. Only the logistic regression algorithm chose an additional feature (5) *age.* For the prediction of cluster label affiliation to the three clusters, the random forest performed best with a high accuracy score of 94.4% (SD = 0.07). The estimated feature importance ranked *concerns regarding the occupational situation* as a feature with the highest impact on the prediction (score of 0.392), followed by *concerns regarding the financial situation*

Table 4 Overview of the average model performance estimated from the cross-validation with the balanced accuracy metric for the training and test set for each algorithm for the prediction of cluster affiliation. KNN = k-Nearest Neighbor, LR = Logistic Regression, RF = Random Forest, and SVM = Support Vector Machine. Change level of the dummy classifier: 0.313

Algorithm	*M* (SD) training	*M* (SD) test	Hyperparameter
KNN	0.845 (0.052)	0.878 (0.194)	n_neighbors = 4, leaf_size = 8, algorithm = ball_tree
LR	0.867 (0.081)	0.644 (0.103)	C = 0.234, penalty = l1, solver = saga
RFC	0.851 (0.057)	0.944 (0.070)	max_depth = 7, max_features = auto, n_estimators = 200, criterion = gini
SVM	0.879 (0.032)	0.889 (0.141)	C = 1, gamma = 0.001, kernel = linear

(score of 0.254), *psychological distress* (PHQ-4; score of 0.226), and *current well-being* (score of 0.128).

4 Discussion

By combining the Gower distance with the PAM algorithm k-Medoids, we could identify three clusters differing significantly in subjects' well-being, psychological distress, and current financial and occupational concerns during the COVID-19 pandemic. Further, our cluster comparison revealed that younger persons were assigned to the cluster reporting greater concerns, increased anxiety, higher psychological distress, reduced well-being, and quality of life. Especially the financial and occupational concerns, psychological distress, and current well-being were decisive variables for the cluster affiliation. Since classifiers' accuracies were high, we can assume that the features used to predict the cluster labels were also most informative for the clustering. While the sequential feature selection extracts the relevant features independently from each other, the random forest's feature importance includes interdependencies in its selected subset [31]. The clustering approach proposed here performed well on the mixed dataset. However, it might be less suitable for larger datasets due to a high time complexity of $O(K(N-K)^2 I)$ and related scaling issues [32]. Further research could investigate other clustering algorithms (e.g., CLARA, CLARANS, and fuzzy k-modes) and non-random initialization methods. Previous studies reported increased mental disorders, distress, and decreased mental health [4, 7], especially in individuals between 18 and 34 years [4, 7], women [4, 7], persons already diagnosed with mental disorders [7], and people living with young children [4]. We observed comparable findings in our C3. In comparison to reference values of a norm population [21, 23], the scores for depression (PHQ-2) and psychological distress (PHQ-4) were significantly higher in C3 but not in C1 and C2. The increased concerns regarding the financial and occupational situation might be explained by the prevailing poor hiring opportunities and the economic recession in Germany. Interestingly, the clusters did not differ in their number of social contacts with other persons. Hence, either the reduced contact with other persons due to social distancing actions did not have a strong influence on psychological health and could not explain differences in psychological health between the clusters, or the C3 suffered particularly strongly due to the social distancing. However, whether the low psychological well-being and increased perceived distress in C3 is a consequence of the COVID-19 pandemic or a phenomenon observed in the young population in general is not clearly distinguishable.

5 Conclusion

Our unsupervised clustering approach could identify individuals with reduced psychological protective factors and increased psychological distress. This allows providing cluster-tailored recommendations, e.g., via a web application, for preventive actions potentially helping to alleviate concerns, decrease perceived distress, and improve psychological well-being as well as the quality of life [33].

References

1. Valtorta NK, Kanaan M, Gilbody S, Ronzi S, Hanratty B (2016) Loneliness and social isolation as risk factors for coronary heart disease and stroke: systematic review and meta-analysis of longitudinal observational studies. Heart 102:1009–1016
2. Kühne S, Kroh M, Liebig S, Zinn S (2020) The need for household panel surveys in times of crisis: the case of SOEP-CoV. Surv Res Methods 14:195–203
3. COSMO Germany: German COVID-19 Snapshot MOnitoring (COSMO Germany). PsychArchives Preprints. https://www.psycharchives.org/handle/20.500.12034/2398 (2020)
4. Pierce M, Hope H, Ford T, Hatch S, Hotopf M, John A, Kontopantelis E, Webb R, Wessely S, McManus S, Abel KM (2020) Mental health before and during the covid-19 pandemic: a longitudinal probability sample survey of the UK population. Lancet Psychiatry 7:883–892
5. Panchal N, Kamal R, Orgera K, Cox C, Garfield R, Hamel L, Munana C, Chidambaram P (2020) The implications of COVID-19 for mental health and substance use. Kaiser Familiy Foundation
6. Li S, Wang Y, Xue J, Zhao N, Zhu T (2020) The impact of COVID-19 epidemic declaration on psychological consequences: a study on active weibo users. Int J Environ Res Public Health 17
7. Jha IP, Awasthi R, Kumar A, Kumar V, Sethi T (2020) Learning the mental health impact of COVID-19 in the United States with explainable artificial intelligence. medRxiv
8. Clatworthy J, Buick D, Hankins M, Weinman J, Horne R (2005) The use and reporting of cluster analysis in health psychology: a review. Br J Health Psychol 10:329–358
9. Romesburg HC (2004) Cluster analysis for researchers. Lulu Press, S.l.
10. Nagpal A, Jatain A, Gaur D (2013) Review based on data clustering algorithms. In: 2013 IEEE conference on information & communication technologies (ICT 2013). Thuckalay, Tamil Nadu, India, 11–13 April 2013. IEEE, Piscataway, NJ, pp 298–303
11. van de Velden M, Iodice D'Enza A, Markos A (2019) Distance-based clustering of mixed data. WIREs Comput Stat 11, e1456
12. Ahmad A, Khan SS (2019) Survey of state-of-the-art mixed data clustering algorithms. IEEE Access 7:31883–31902

13. Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871
14. Grané A, Salini S, Verdolini E (2020) Robust multivariate analysis for mixed-type data: novel algorithm and its practical application in socio-economic research. Socio-Econ Plan Sci
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
16. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302
17. Park H-S, Jun C-H (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36:3336–3341
18. Madhulatha TS (2011) Comparison between K-means and K-medoids clustering algorithms. In: Wyld DC, Wozniak M (eds) Advances in computing and information technology. First international conference; proceedings, vol 198. Springer, Berlin, pp 472–481
19. Budiaji W, Leisch F (2019) Simple K-medoids partitioning algorithm for mixed variable data. Algorithms 12:177
20. Foss AH, Markatou M, Ray B (2019) Distance metrics and clustering methods for mixed-type data. Int Stat Rev 87:80–109
21. Kroenke K, Spitzer RL, Williams JBW, Löwe B (2009) An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 50:613–621
22. Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Löwe B (2007) Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 146:317–325
23. Löwe B, Kroenke K, Gräfe K (2005) Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res 58:163–171
24. Ludwig K, Graf von der Schulenburg J-M, Greiner, W (2018) German value set for the EQ-5D-5L. Pharm Econ 36, 663–674
25. Janssen MF, Szende A, Cabases J, Ramos-Goñi JM, Vilagut G, König HH (2019) Population norms for the EQ-5D-3L: a cross-country analysis of population surveys for 20 countries. Eur J Health Econ: HEPAC 20:205–216
26. Huber MB, Reitmeir P, Vogelmann M, Leidl R (2016) EQ-5D-5L in the general german population: comparison and evaluation of three yearly cross-section surveys. Int J Environ Res Public Health 13:343
27. Yuan C, Yang H (2019) Research on K-value selection method of k-means clustering algorithm. J-Multidiscip Sci J 2:226–235
28. Cumming G, Finch S (2005) Inference by eye: confidence intervals and how to read pictures of data. Am Psychol 60:170–180
29. Raschka S (2018) MLxtend: providing machine learning and data science utilities and extensions to python's scientific computing stack. JOSS 3:638
30. Cabases J, Szende A, Janssen B (2014) Self-reported population health: an international perspective based on EQ-5D. Springer
31. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res, 1157–1182
32. Mannor S, Jin X, Han J, Zhang X (2010) K-Medoids Clustering. In: Sammut C, Webb GI (eds) Encyclopedia of machine learning. Springer, US, Boston, MA, pp 564–565
33. Janssen D, Lingelbach K, Piechnik D, Gado S, Maurer P, Eichler M, Knopf D, Hentschel L, Schuler M, Sernatinger D, Peissner M (2021) WIBCE—a web application helping people to reflect their infection risk and psychological well-being and act accordingly during the COVID-19 pandemic. In: Proceedings of the 12th international conference on applied human factors and ergonomics and the affiliated conferences, New York, USA

BERT-Based Tagging Method for Social Issues in Web Articles

Tokutaka Hasegawa and Shun Shiramatsu

Abstract In recent years, public collaboration to address social challenges has become increasingly important for the sustainable development of local and global communities. When discussing how to tackle social challenges, it is essential to investigate relevant existing activities. This study aims to collect web articles on various social issues to support such discussions and research. For this purpose, we have developed an automatic tagging system for web articles on social issues and evaluated it using the Bidirectional Encoder Representations from Transformers model (BERT), Wikidata, and Wikipedia. Specifically, our method classifies each sentence in a web article into social issue tags using BERT, which was trained beforehand using the Japanese Wikipedia. We constructed a training corpus using Wikidata and Wikipedia and extracted candidate social issue tags from the "Social issues" class of Wikidata. Hierarchical tagging is now possible, thanks to the structure of Wikidata. The results of our evaluation showed that the proposed method obtained F1 scores of 0.94 for zeroth-level classifications, 0.85 for first-level classifications, and 0.85 or better for all second-level classifications.

Keywords Public Collaboration · BERT · Automatic Tagging

1 Introduction

In recent years, approaches based on civic tech, open data, and open source have been actively pursued globally to resolve social issues through cross-regional citizen collaboration. However, at present there is no mechanism for enabling people to share examples of potential solutions. Only exceptional cases have been shared, such as winning an award in a contest related to the use of open data or sharing content

T. Hasegawa (✉) · S. Shiramatsu
Nagoya Institute of Technology, Nagoya, Japan
e-mail: t.hasegawa@srmtlab.org

S. Shiramatsu
e-mail: siramatu@nitech.ac.jp

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_82

on social networking sites by influential people. In this work, we have developed a method for automatically tagging web articles related to social issues that enables various cases of addressing social issues to be shared and used for events in which people can participate. In a previous work, we developed a web application named MissionForest that can structure public activities for addressing social issues and publish them as linked open data (LOD). However, we have not been able to collect case studies from people other than users of MissionForest. To enable the use of case studies from other regions in public events, it is necessary to develop a mechanism that can automatically collect information on the approaches people are using for various social issues. Our proposed method automatically tags social issues using the Bidirectional Encoder Representations from Transformers model (BERT) in web articles and structures cases with hierarchical social issue tags by utilizing Wikidata entities as tags. In this paper, we discuss our development of the method and report the results of evaluations to determine its accuracy.

In Sect. 2, we give a brief overview of related research. Section 3 describes our development of the social issue corpus. In Sect. 4, we propose our automatic tagging method for social issues and report the evaluation results. The results are discussed in more detail in Sect. 5. We conclude in Sect. 6 with a brief summary and mention of future work.

2 Related Research

2.1 Language Representation Models BERT and ALBERT

BERT [1] and A Lite BERT (ALBERT) [2] are general-purpose natural language processing models based on the Transformer deep learning model. They are typically used to transfer learning for various tasks such as comprehension. In this study, we use BERT and ALBERT for tagging.

2.2 Fine-Grained Propaganda Detection

The NLP4IF workshop held at EMNLP-IJCNLP in 2019 included a contest to tackle the extraction of propaganda from news articles [3]. This contest featured two tasks: one for letter-based extraction (FLC) and the other for sentence-based extraction (SLC). As most of the teams that obtained high F-values used the BERT method, we believe that pre-trained general-purpose language models such as BERT and ALBERT are effective in extracting social issues from resources such as news articles on the Web.

2.3 Web Annotation Data Model

The Web Annotation Data Model [4] is a specification recommended by the W3C for sharing web annotations. We aim to annotate web articles on social issues, both manually and by using a machine, and publish the data in this format.

2.4 LOD for Resolving Social Issues

Shiramatsu et al. [5] proposed a data model to convert social issues into linked open data (LOD). Our method to promote civic technology by matching problem solvers also uses this LOD. In the future, we plan to design a schema of cases related to social issues using this data model as a reference.

2.5 MissionForest

MissionForest [6, 7] is a web application that enables the sharing of collaborative activities. Multiple people can edit tasks in the form of a work breakdown structure (WBS) to accomplish a task and solve a common problem.

2.6 Automatic Tagging

Several studies have examined ways of automatically tagging labels to unlabeled text data. Saito et al. [8] proposed an automatic labeling method that uses a distributed representation to extract words with similar meanings from each category of sentence vector and assign them as labels. The distributed representation consists of a two-layer neural network that includes the linguistic nature of the words and phrases, and similar words can be trained to have similar vectors. While Saito et al. used automatic tagging for just five news articles of different types, in our study, we target news articles about social issues, which include about 30 candidate tags. This makes the classification difficult and complex.

3 Development of Social Issues Corpus

Each sentence in a web article is automatically tagged by determining whether it refers to a social issue, and if so, its type is automatically determined. In this section, we create a training corpus for this purpose. Specifically, we use the hierarchical

```
SPARQL query
select distinct ?label ?sitelink where {
 ?s wdt:P279 wd:Q1920219;
   rdfs:label ?label.
  filter(lang(?label) = "ja")
  BIND(?s AS ?wikipedia).
  ?sitelink schema:about ?wikipedia .
  FILTER
REGEX(STR(?sitelink), "ja.wikipedia.org/wiki/") .
 }
```

Fig. 1 Example of SPARQL query

structure of social issues in Wikidata and the Japanese Wikipedia articles associated with the structure. Watanabe et al. [15] developed a social issue ontology using Wikipedia's category hierarchy, but articles other than those about social issues were often mixed in as noise. If we use Wikidata's "subclass of" property (wdt:P279), there is a much smaller risk of mixing such noise. Therefore, we use the entities obtained by following the subclasses of Wikidata's "Social issues" entities (wd:Q1920219) as social issue tags. An example of the SPARQL query we created is shown in Fig. 1 However, since there were only two subclasses of social issues in 2018, we developed a hierarchical structure of social issues in 2019 by manually adding a subclass of properties to the entities (e.g., "Crime" and "Disaster"). We believe such hierarchical tagging will enable users to browse web articles describing social issues in an exploratory manner. The entities in Wikidata that are subclasses of social issues are hereinafter called "Social issue tags". To create a training corpus using the Japanese Wikipedia articles, we used links to articles on social issues. First, since each Wikipedia article is associated with one Wikidata entity, we can extract a list of Wikipedia articles corresponding to social issue tags by tracing the tags in Wikidata to the associated articles. Next, the text of a Wikipedia article that has no correspondence with a social issue tag is extracted as a sentence with a link to the Wikipedia article corresponding to that tag. Such sentences can be regarded as positive cases that should be given the social issue tag. We use the SPARQL inquiry service for extracting the hierarchical structure of Wikidata and use the MediaWiki API for extracting main text articles from Wikipedia. An example of the social problem corpus we created is shown in Fig. 2. We chose sentences that were at least 30 words long and contained words that matched the class. The source page is a page with links to Wikipedia articles for the disaster class itself and its subclasses (e.g., tsunami and earthquake). Therefore, even if the word "disaster" is not directly included in the text, it can be labeled as "Disaster". We created a corpus in the same way for the other classes as well. To classify social issues with high accuracy, it is necessary to include negative cases that do not describe social issues. For this purpose, we randomly selected articles from the Japanese Wikipedia and defined negative cases as articles with tags other than "Social issues" that did not have social

Tsunami Earthquake

Such earthquakes do not produce strong motions, but the source area of the tsunami spreads faster than the tsunami itself, within a few minutes, and thus the tsunami is larger. If the rupture propagation speed is much slower than this, the tsunami will spread before the source area spreads and no large tsunami will be generated.

Tsunami earthquakes generate large tsunamis even though the shaking itself is small, making immediate evacuation difficult and potentially devastating. The 1896 Meiji Sanriku earthquake, a notable example of a tsunami earthquake, killed more than 20,000 people.

In the 2011 Tohoku-Pacific Ocean Earthquake, strong short-period seismic motions were generated by slippage in the deep landward side of the plate boundary, and long periods of seismic motions and strong tsunamis were generated by slippage in the shallow oceanic trench side offshore.

Fig. 2 Example of social issues corpus

problems tags such as “Crime”, “Disaster”, and “Terrorism” and that did not have morphological elements commonly used at the end of social problem tags such as “Life”, “Crime”, and “Medical care”.

4 Automatic Tagging Method for Social Issues

4.1 Task Settings

When constructing the corpus, our aim is to determine which kind of social issue a statement is referring to. In this study, we set up the three tasks listed below. We compared several multi-label classification models and used the generic language model BERT or ALBERT for the tasks. Both the pre-training model and the Sentencepiece model [9], which is used to tokenize sentences, were trained using data from the Japanese Wikipedia. These models were then used to fine-tune BERT and ALBERT for classification.

(1) Zeroth-level classifications. We set up a task to determine if a given sentence was about a social issue. This task is necessary to extract social issue articles that are not included in the first level or second level. The role of the zeroth level is similar to that of the first-level and second-level “Others” tags to account for parsing errors in the actual text of social issues.

(2) First-level classifications. We selected a number of specific first-level classes of the “Social issues” class (“Crime”, “Disaster”, “Abuse”, “Discrimination”, “War”, "Pseudoscience", "Harassment", "Environment issues", and "Persecution") and classified articles of the Japanese Wikipedia related to these subclasses.

(3) Second-level classifications. We selected second-level classes directly linked to the first-level classes, and classified articles of the Japanese Wikipedia related to these subclasses. For example, the "Crime" subclass was divided into "Terrorism", "Sexual abuse", "War crimes", "Torture", "Scams", and "Murder".

4.2 Model Selection

Percentage of negative cases

To classify social issues with high accuracy, it is necessary to include negative cases that do not describe social issues. For this purpose, we randomly selected articles from the Japanese Wikipedia and created negative cases by labeling articles with tags that do not contain social issues as "Others". Figure 3 shows the prediction accuracy of the BERT automatic tagging model for the second-level subclass "Crime" when the percentage of negative cases in the training data was varied. As we can see, there was a trade-off between precision and recall in the predictive model. When the proportion of "Others" was low, recall was high, and conversely, when the proportion of "Others" was high, precision was high. In this context, we found that the optimal case was when the percentage of "Others" was between 25 and 30%, which is the intersection point of the precision and recall curves. Therefore, we trained the negative cases of "Others" at about 25% in this automatic tagging model for all subclasses in the first and second levels. However, due to the dichotomous classification of the zeroth level, the percentage of training data was 50% for the cases in the zeroth level.

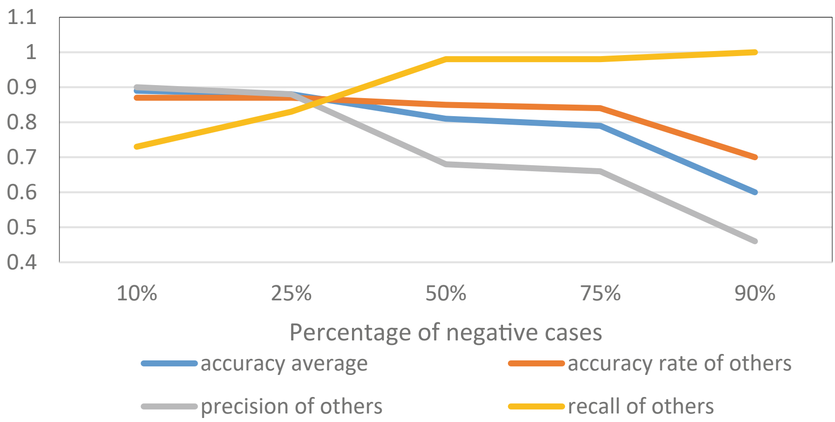


Fig. 3 Prediction accuracy in negative case percentage change

Comparison of multi-label classification models

We compared four ML algorithms for multi-label classification (BERT, Random Forest [10], Naive Bayes [11], and K-Nearest Neighbor [12]) to automatically tag and classify web news articles. The second-level subclass “Crime” was used for the corpus to compare the accuracy of the models. We divided the data into 90% for training data and 10% for test data to ensure a fair comparison. We compared the average percentages of ten of the correct responses of the test data [13]. Figure 4 shows the predictive accuracy of each classification model. BERT performed best, exceeding 0.8 in F1 score, recall, and precision. We also examined whether there was any statistically significant difference between the BERT model and the other models using McNemar’s test [14]. We assumed a null hypothesis that no difference between the two systems is correct and checked if there was a significant difference. If the null hypothesis holds, there is a 50% chance of being correct in either of the two models. Therefore, the probability distribution of the number of successes in one system follows a binomial distribution with probability 0.5. Equation (1) shows how this metric is calculated, where n is defined as the number of correct answers for either one of the two models.

$$P = \sum_{k}^{n} \binom{n}{k} 0.5^{k} (1 - 0.5)^{n-k} \tag{1}$$

Results showed that $p = 1.48\text{e}{-06}$ when comparing BERT and Random Forest, $p = 1.50\text{e}{-20}$ when comparing BERT and K-Nearest Neighbor, and $p = 6.13\text{e}{-44}$ when comparing BERT and Naive Bayes. In none of these cases did this value exceed the significance level.

Comparison of ALBERT and BERT

ALBERT is a general-purpose language model that reduces the weight of BERT and speeds up the learning process. We tested it by changing the batch size and number of training steps to see if it would work in our automatic tagging task. Table 1 shows

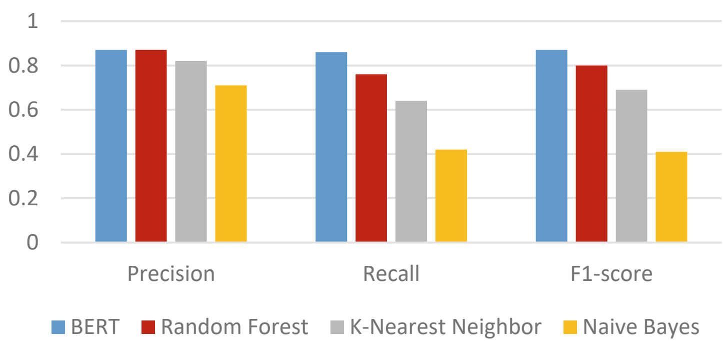


Fig. 4 Predictive accuracy of each classification model

Table 1 Prediction accuracy of BERT and ALBERT in each condition

	Batch size	Steps	Accuracy
BERT	4	7800	0.87
ALBERT 1	4	2200	0.40
ALBERT 2	4	11,000	0.61
ALBERT 3	8	11,000	0.64
ALBERT 4	8	38,000	0.76
ALBERT 5	8	110,000	0.80

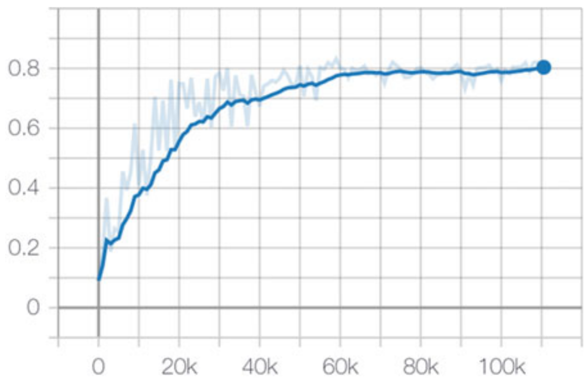


Fig. 5 Accuracy of ALBERT model in condition 5

a comparison of the prediction accuracy between BERT and ALBERT. While the original BERT model had a batch size of 4 and 7800 training steps, the ALBERT model had a batch size of either 4 or 8, and the training steps were verified in five levels from 2200 to 110,000. The training task was a second-level "Crime" classification task. The results showed that the ALBERT model improved accuracy up to 0.80 as the batch size and training steps were increased. Figure 5 shows the accuracy of the ALBERT model in condition 5, where the accuracy converged to around 0.8 when the training steps were around 60,000. These results mean it is necessary to examine other factors to improve the accuracy of the ALBERT model. Therefore, we used the BERT model in this study for the first-level classification task and all other hierarchies.

4.3 Accuracy Verification

The task discussed in Sect. 4.1 was trained using the BERT model. Tables 2 and 3 show the prediction accuracy of the zeroth-level classification problem in 4.1(1), Tables 4 and 5 show the prediction accuracy of the first-level classification problem

Table 2 Results of prediction accuracy in news article (1)

News articles				
	Precision	Recall	F1 score	Support
Others	0.95	0.92	0.93	241
Social issues	0.92	0.95	0.94	241
Accuracy			0.94	482

Table 3 Results of prediction accuracy in news article (2)

News articles				
	Precision	Recall	F1 score	Support
Others	0.99	0.91	0.94	1000
Social issues	0.71	0.95	0.81	241
Accuracy			0.91	1241

Table 4 Results of prediction accuracy in "Social issues" (1)

Social issues				
	Precision	Recall	F1 score	Support
Crime	0.93	0.76	0.84	2571
Disaster	0.97	0.82	0.89	1773
Abuse	0.83	0.73	0.77	248
Discrimination	0.89	0.75	0.81	255
War	0.91	0.82	0.86	667
Pseudoscience	0.94	0.87	0.91	264
Harassment	0.87	0.75	0.80	166
Environment issues	0.92	0.82	0.86	283
Persecution	0.79	0.45	0.58	42
Others	0.74	0.97	0.84	3090
Accuracy			0.85	9359

in 4.1(2), and Table 6 shows the prediction accuracy of the second-level prediction in 4.1(3).

Zeroth-level classification

First, the zeroth-level classifications were trained with the social issues corpus described in Sect. 3, but most of them were classified as "Others". Therefore, we used our manually created corpus to train the classification model and validate the classification accuracy. Tables 2 and 3 show the difference in the proportion of the "Others" class used for validation. The results of the evaluation showed an F1 score of 0.94 for the zeroth-level classifications. In reality, the percentage of sentences that

Table 5 Results of prediction accuracy in "Social issues" (2)

Social issues				
	Precision	Recall	F1 score	Support
Crime	0.35	0.49	0.41	51
Disaster	0.74	0.89	0.81	35
Abuse	1.00	0.17	0.29	30
Discrimination	0.75	0.39	0.51	31
War	0.67	0.06	0.12	31
Pseudoscience	1.00	0.54	0.70	28
Harassment	0.79	0.45	0.58	33
Environment issues	1.00	0.19	0.32	32
Persecution	0.90	0.29	0.44	31
Others	0.57	0.96	0.71	163
Accuracy			0.60	465

Table 6 Results of prediction accuracy in "Crime"

Crime				
	Precision	Recall	F1 score	Support
Others	0.88	0.83	0.86	256
Terrorism	0.90	0.86	0.88	42
Sexual abuse	0.92	0.89	0.91	269
War crimes	0.79	0.83	0.81	64
Torture	0.83	0.83	0.83	60
Scam	0.98	0.89	0.93	161
Murder	0.76	0.91	0.83	189
Accuracy			0.87	1041

should be tagged as "Social issues" in news articles is not considered to be 50%. When the percentage of "Social issue" tags in the validation data is set to 20%, the percentage of false judgments is high. Therefore, we need to be careful when we actually operate this system.

First-level classification

Table 4 shows the validation of classification accuracy with data not used for the training of the classification model among the social problems explained in Sect. 3. Table 5 shows the validation of classification accuracy using news articles written about social issues that we manually collected (Sect. 4.3.1). The results for the social issues corpus showed that the accuracy was relatively high except for the "Persecution" class, which had less support. However, validation with actual news articles showed that the accuracy was reduced in most classes. Figure 6 shows the

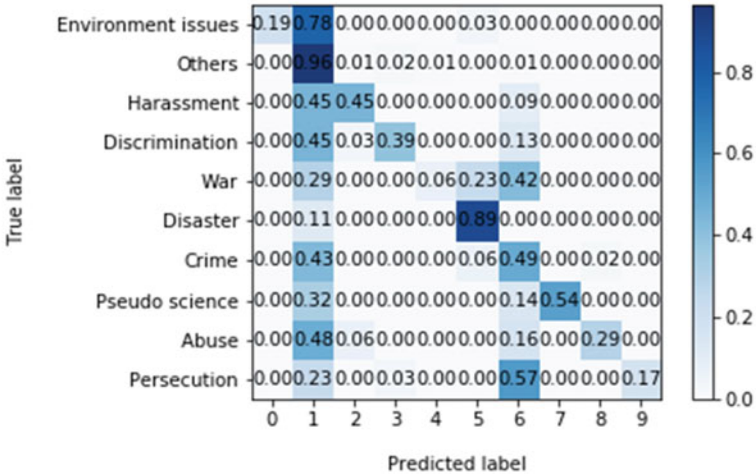


Fig. 6 Confusion matrix of first-level classification

confusion matrix of the verification with the actual news articles. We can see here that the classification model misclassified many classes as “Others”. Also, many of the “War” and “Persecution” classes were classified as “Crime” classes, which was not entirely wrong. These results indicate that it is sometimes necessary to add multiple tags to a single sentence.

Second-level classification

The average accuracy of each tagging model was generally high in the system, exceeding 0.87 for all models. However, as discussed in Sects. 4.3.1 and 4.3.2, if the accuracy verification is performed using the same proportions and sentences for each class as in the actual news articles, the accuracy might be less than what is described here. To improve the accuracy of these predictions, we need to re-learn the data by under-sampling or increasing the number of samples. As an example, Table 6 shows the results for “Crime” among the models of second-level classification.

5 Discussion

Regarding the automatic tagging of social questions with our method, the percentage of correct answers for all classes was above 0.85, which suggests it could be used effectively if we manually correct the wrong answers. On the other hand, the result for the “Persecution” class was 0.58, which is lower than the other classes. After checking the confusion matrix, we found that the “Persecution” class tended to be combined with the “Crime” class and was sometimes misclassified as “Others”. The corpus of the “Others” class is likely to contain information of other classes, which may cause problems with the quality of the corpus. In fact, the percentage of sentences that should be tagged as "Social issues" in a news article is unlikely to be 50%; it is likely to vary depending on the website where the news article was published. Therefore, we plan to change the proportion of "Others" in the future when verifying the accuracy of the survey. Also, in Wikidata, "Persecution" is not a subclass of "Crime", but there is a possibility that it is embedded or included in the actual web article. To resolve this issue, we feel it is necessary to add multiple tags to a single sentence. It may be possible to assign multiple labels to the probability output from the classification model by assigning weights or setting threshold values. In the future, we plan to manually evaluate the quality of the automatically constructed corpus. In addition, although both the training and validation data were extracted mechanically from the Wikipedia page, we believe it is necessary to deal with news articles that actually exist on the Web.

6 Conclusion

In this paper, we proposed an automatic tagging method using the Bidirectional Encoder Representations from Transformers model (BERT), Wikidata, and Wikipedia. To automatically tag social issues, we also constructed a training corpus using Wikidata. We found that this automatic construction method can greatly reduce the amount of manual labor. Our model uses the raw text of Wikipedia for pre-training to perform automatic tagging, and as a result, it achieved F1 scores of 0.94 for zeroth-level classifications, 0.85 for first-level classifications, and 0.85 or better for all second-level classifications. Overall, we achieved an accuracy of at least 0.85, which indicates that we can use the model effectively if we manually correct the errors. However, there is still room for improvement regarding the labeling of negative cases. In the future, we plan to improve the accuracy of the automatic tagging by manually evaluating the quality of the constructed corpus. We will also evaluate our system by using real-world examples such as co-creative projects conducted in the Special Interest Group on Crowd Co-creative Intelligence (SIG-CCI) of the Japanese Society for Artificial Intelligence and civic tech activities conducted by Code for Nagoya.

Acknowledgements This work was supported by NEDO (JPNP20006), JST CREST (JPMJCR15E1, JPMJCR20D1), and JSPS KAKENHI (17K00461).

References

1. Devlin J, et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805

2. Lan Z, et al (2019) Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942
3. Da San Martino G, Barron-Cedeno A, Nakov P(2019) Findings of the nlp4if-2019 shared task on fine-grained propaganda detection.In: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda
4. Sanderson R, Ciccarese P, Young B (2017) Web Annotation Data Model (2017)
5. Shun S, Teemu T, Tadachika O, Toramatsu S (2016) Developing a goal matching service using linked open data of social issues and their solutions as public goals. Trans Jpn Soc Artif Intell 31:29–39, 01
6. Masaru W, Shun S, Akihisa S (2017) A study of the Co-creation Support System "Mirailab" and the Task Structure System "MissionForest" to support the continuity of co-creation. In: Proceedings of the 3rd Annual Conference of the Japanese Society for Artificial Intelligence
7. Yasuaki G, Masaru W, Shun S (2017) MissionForest: Prototype of a task-structured system for supporting collaboration inside and outside of an organization In: Proceedings of the 79th national convention of the information processing society of Japan, vol 6ZA-02, no 1, pp 355–356
8. Saito T, Osamu U (2017) Automatic labeling for news article classification based on paragraph vector. In: 2017 9th international conference on information technology and electrical engineering (ICITEE). IEEE
9. Yohei K (2019) BERT pretrained model trained on Japanese Wikipedia articles: GitHub, GitHub repository (2019). https://github.com/yoheikikuta/bert-japanese
10. Vladimir S, Andy L, Christopher T, Chiristopher Culberson J, Sheridam RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR Modeling: J Chem Inf Comput Sci 43(6):1947–1958
11. McCallum A, Kamal N (1998) A comparison of event models for naive bayes text classification.In: AAAI-98 workshop on learning for text categorization, vol 752, No 1
12. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244
13. Ron K (1995) A study of crossvalidation and bootstrap for accuracy estimation and model selection. In: International joint conference on artificial intelligence IJCAI
14. McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2):153–157
15. Watanabe S, Shiramatsu T, Sengoku A (2017) Co-creation continuity support by mutual cooperation between co-creation support system "Mirilab" and task structuring system "MissionForest". In: Proc. of the 3rd annual conference of Japan society for artificial intelligence

Multidimensional Blockchain Security Analysis

Ilya Shilov and Danil Zakoldaev

Abstract Scaling and securing data exchange in distributed systems are important directions for scientific research. Known solutions provide certain level of security but lack some qualities required for wide usage. Among the obstacles are scaling distributed solutions and constructing exchange between separate robust distributed ledgers. The paper presents multidimensional blockchain— an approach to building robust distributed ledgers based on the notion of blockchain. The solution can be implemented in two approaches: block mode and state machine mode. A general security analysis, mathematical security analysis, and GUC-framework analysis are performed. Mathematical analysis is based on existing works on security of one-dimensional blockchain and consensus algorithms. A GUC-model is created to prove security of multidimensional blockchain. It is proven that multidimensional blockchain securely implements robust distributed ledger and provides secure inter-system exchange given that search and verification protocol is secure. The work is a basis for further research towards implementation of scalable robust distributed ledger.

Keywords Multidimensional Blockchain · Persistence · Liveness · UC framework · Transaction · Security proof

1 Introduction

Cryptographic currencies and related technologies have appeared recently but have already made a great impact on the development of information technologies and implementing distributed systems. An important notion of modern distributed systems is ledger. Distributed ledger is a ledger maintained by two and more machines

I. Shilov (✉) · D. Zakoldaev
ITMO University, Saint Petersburg, Russia
e-mail: ilia.shilov@itmo.ru

D. Zakoldaev
e-mail: d.zakoldaev@itmo.ru

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_83

[13, 14]. For ledger controlled by one node, persistence and liveness are fulfilled by default. It is way more difficult to achieve these qualities for distributed ledgers functioning in an unreliable environment in presence of adversaries working against the protocol [5, 12]. Blockchain solves this exact problem.

It is worth mentioning that blockchain is not deprived of disadvantages. While some have been overcome by recent research [4, 19], several are still critical and slow down the development of this technology. Significant disadvantage of blockchain is the increase of data storage size. Blockchain implies replication of complete block sequence on every node of the network. Although attempts to solve this problem were undertaken, in practice, a working solution was not presented. Among the most complicated operations is the transmission of funds between fiduciary currency systems and cryptocurrencies or between cryptocurrencies. Nowadays, such operations require creation of so-called sidechains or attraction of third parties, which also solve the problem only partially.

Multidimensional blockchain is a proposed approach to building robust distributed ledgers. Its concept is based on idea of sidechains used when transferring funds between separate non-compliant solutions in cryptocurrency world. This technology is meant to solve such problems as data transition between separate systems based on blockchain, uniting blockchains with different operating rules into unified system and decreasing the volume of data storage on nodes, thanks to scaling. The aim of this work is multidimensional blockchain security proof.

2 Multidimensional Blockchain

Multidimensional blockchain is a system which consists of a set of blockchains each of which except for the first one follows registration procedure in one of existing blockchains. Registration means storing information on genesis block and, possibly, some features of blockchain in different blockchain. The concept of genesis block requires special attention. This is the first block in blockchain. The concept is same for blockchains in multidimensional blockchain. Genesis blocks are explicitly registered in existing blockchains.

Figure 1 shows general view of multidimensional blockchain which unites several blockchains into one system. It is supposed that every blockchain implements distributed ledger—this assumption allows to disengage from concrete operation mode. Therefore, it is not explicitly mentioned how exactly blockchain registration is performed: in block of special type or in internal data structure. Both modes are considered later.

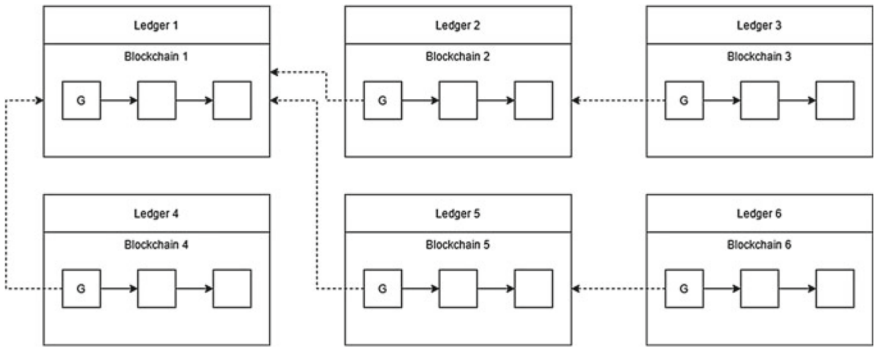


Fig. 1 Multidimensional blockchain

3 Mathematical Evaluation of Multidimensional Blockchain Security

There are two main directions for security analysis of multidimensional blockchain. On one hand, scaling causes change of settings for separate robust distributed ledgers. This causes change of successful attack probability for underlying attacker, who controls constant number of nodes. On the other hand, when constructing exchange via registration, the probability of attack for each blockchain depends not only on attacker's actions, but also on probability of breaking security of all ledgers on the way up to the root of blockchain tree.

During scaling the number of nodes in each ledger changes. Thus, Eq. (1) holds (N—number of nodes, N_A—number of attackers, k—current number of blocks on the blockchain, and L—the number of blockchains).

$$p_a = F(N, N_A, k) \Rightarrow p_a^{'} = F\left(\frac{N - N_A}{L} + N_A,\ N_A, k\right) \tag{1}$$

Following formula can be applied for registration process:

$$p_a = \left(1 - \prod_{i=0}^{L} q_i\right) = 1 - \prod_{i=0}^{L}(1 - p_i) \tag{2}$$

where p is probability of successful attack and q is probability of secure functioning. Consequently, the probability depends on the probability that none of L blockchains in the hierarchy shall be compromised. In practice the target is equivalent to calculation of probabilities in a tree where each node is associated with probability and complete probability for each node depends on all parent nodes.

3.1 Proof of Work Analysis. Nakamoto Approach

The first evaluation of blockchain security (in fact—proof of work security) was given in the first paper on Bitcoin [1]. An equation was presented, which allowed evaluating secure depth of block in the blockchain that allows taking the probability of successful attack as negligible. For Bitcoin, secure depth is about six blocks. The probability of successful attack depends on overall power of the adversary (in fact on the relation between adversaries' power and the power of honest nodes).

If the number of miners in each blockchain is same, multidimensional blockchain does not influence the security of separate blockchains: each blockchain operates on its own in sense of consensus mechanism. If the number of miners decreases (while building child blockchains) the attack simplifies, provided the adversary concentrates his resources on attack of just one specific blockchain in multidimensional blockchain. Figure 2 demonstrates the increasing of the probability of successful attack on last six blocks depending on the adversaries' power. Equation (3) denotes the dependence if N is the number of blockchains (all variables are same as in original paper).

Thus, when creating new blockchains for scaling, it is necessary to consider that the number of miners in the system cannot increase significantly in a short time. Such evaluation is to be taken every time when application of multidimensional blockchain is considered.

$$P_a = 1 - \sum_{k=0}^{z} \frac{\lambda^k e^{-\lambda}}{k!} \left(1 - \left(\frac{q}{p/N}\right)^{(z-k)}\right) \tag{3}$$

Evaluation the probability of attack on multidimensional blockchain during registration is also based on the equations given for scaling. However, it does not use the number of nodes in each blockchain as a parameter. General formula of attack probability is given in Eq. (4).

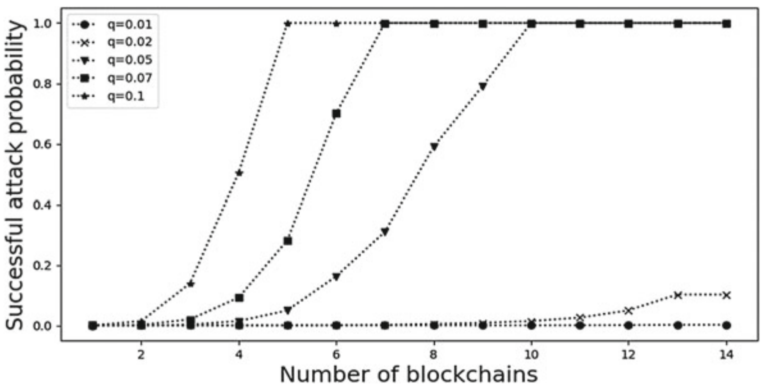


Fig. 2 Probability of successful attack with decreasing number of nodes

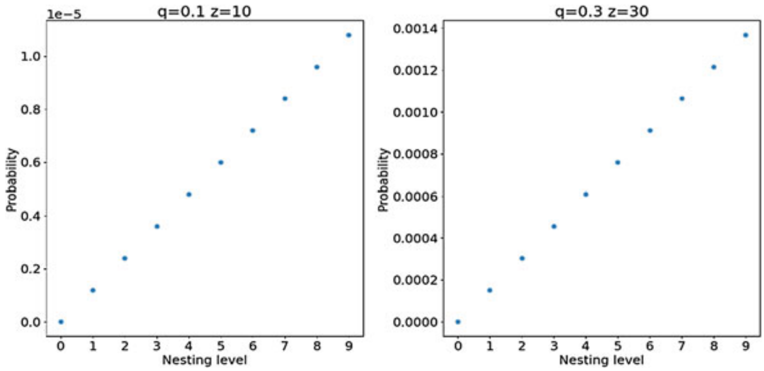


Fig. 3 The probability of successful attack on blockchain like Bitcoin depending on the nesting level in multidimensional blockchain

$$P_a = \left(1 + \prod_{i=0}^{L}\left(\sum_{k=0}^{z} \frac{\lambda^k e^{-\lambda}}{k!}\left(1 - \left(\frac{q_i}{p_i}\right)^{(z-k)}\right)\right)\right) \quad (4)$$

For simplified representations, consider Bitcoin and the results for ledger in a specific unit of time. The calculation is performed with Eq. (2). Figure 3 shows the result of calculations.

3.2 *Proof of Work Analysis. GHOST Approach*

An important shortcoming of Nakamoto security analysis is the absence of transition delay among the model parameters. Paper [18] contains evaluation of successful attack probability with this value accounted. Besides, the same work contains a theorem which proves that networks with delays are not less (essentially even more) secure than networks without delays. For networks with delays, the evaluation is hardened due to the absence of information on network structure. Using TPS instead of "blocks per second" is a specific feature of the paper. A general security requirement is

$$\beta\left(\lambda_{rep}\right) \geq \frac{q}{1-q}\lambda_{rep} = \frac{q}{p}\lambda_{rep} \quad (5)$$

Here β is block acceptance rate, λ_{rep} is observed block creation rate, q is the percentage of blocks created by attacker (probability that next block is created by attacker), p is the percentage of blocks created by honest nodes (probability that next block is created by an honest node).

Splitting miners into groups when creating multidimensional blockchain (or creating blockchains inside multidimensional blockchain) causes change of relation between p and q, i.e., probabilities that the next block is created by an attacker or honest node. Thus, the evaluations also change. As with the longest chain rule, it is necessary to evaluate the probability of successful attack to find out whether it is secure to create a new blockchain inside multidimensional blockchain.

It is possible to examine the change of relation between parameters of a given equation. It is reasonable to evaluate only scaling as connecting blockchains in order to organize secure information exchange does not cause change of parameters used in GHOST approach. The change of relation between p and q is given below

$$\frac{q'}{p'} = \frac{N_A}{\left(\frac{N-N_A}{L} + N_A\right) - N_A} = \frac{N_A}{N - N_A} * L = \frac{q}{p} * L \tag{6}$$

Block inclusion rate does not change over time, thanks to target recalculation mechanisms. However, block creation rate generally changes if attacker and remaining honest nodes do not change their behavior. Generally, block generation depends on the number of nodes. The dependence is proportional

$$\left(\frac{q}{p} * \lambda_{\text{rep}}\right)' = \left(\frac{q}{p} * \lambda\right) * (1 + q(L - 1)) \tag{7}$$

Thus, increasing the number of ledgers causes increase of the lower bound for Eq. (5) which might break security of some protocols and consensus mechanisms.

3.3 Proof of Work Analysis. IOHK Approach

Scaling blockchain by transformation into multidimensional analog causes change of some parameters defined in papers on Bitcoin security. Security proofs and analyses are based on the notions of persistence and liveness [13] and the features of blockchain which allow to achieve the requirements of these notions: common prefix property (CPP), chain quality property (CQP), and chain growth property (CGP). Scaling security analysis is equivalent to analysis whether these requirements are still fulfilled when nodes are split into groups. Security analysis for registration is more complicated and is not presented in this paper as it significantly depends on formal search and verification protocol for blocks and transactions which is not observed in this paper.

In papers [13, 14], security proofs are given as theorems which use a set of parameters. Change of parameters might lead to security breach. Thus, scaling requires evaluating that persistence and liveness are fulfilled in all inner ledgers. The theorems do not contain probabilities which prevents evaluation of registration security with Eq. (2). This, security of parent ledger must be taken as a prerequisite: a ledger is insecure if its parent is not persistent. Liveness is less important as registration is complete only when the record of registration goes deep enough into the chain of blocks.

In paper [13], the main requirement for the model is the fulfillment of following requirement (honest majority assumption):

$$t \leq (1-\delta)(n-t) \text{ and } \delta \geq 2f + 2\varepsilon \tag{8}$$

Lemma 21 has a parameter k—depth required to take the block for irreversible. It depends on cryptographic hash-sum length and the round to block translation value η. Value f changes over time—it is the probability that at least one honest node creates a block in a round. Its parameters: n—number of nodes, t—number of attacking nodes, p—probability of successful request to oracle, and q—number of requests per round for node.

$$\frac{f'}{f} = \frac{pq\left(\frac{N}{L}-t\right)}{1+pq\left(\frac{n}{L}-t\right)} * \frac{1+pq(n-t)}{pq(n-t)} = \frac{\frac{1}{n-t}+pq}{\frac{1}{\frac{n}{L}-t}+pq} < 1 \tag{9}$$

In other words, increasing number of blockchains causes decreasing the probability that at least one honest node creates a block during round. Thus, lower bound of parameter δ decreases which strengthens requirement for honest majority. Consequently, scaling requires fulfillment of required condition.

3.4 Proof of Stake Analysis

As for the end of 2020, the only proof of stake consensus mechanism with proven security is Ouroboros [2, 11, 15]. This work inherits the principles defined in papers on Bitcoin security. However, an important proof of stake feature is independence of physical system maintenance. According to the chain selection rule in Ouroboros Genesis, recreation of correct chain adopted by honest majority is possible from the first block, i.e., Genesis block (bootstrapping from Genesis). Then security of the system does not depend on the real number of maintaining nodes as it is defined by number of accounts in the system and relation between stakes of honest accounts and accounts controlled by the adversary. Thus, theorems defined in papers on Ouroboros security remain correct for multidimensional blockchain, provided that honest majority requirement is fulfilled in Genesis block of a new blockchain.

3.5 Byzantium Security Analysis

Finally, it is necessary to define Byzantium protocols security inside multidimensional blockchain. Byzantium protocols are consensus protocols in distributed systems. All these protocols (if no authentication used) have common constraint: correct operations are possible only if the number of honest nodes is greater than 2/3 of cumulative number of nodes [17]. Including such system into multidimensional blockchain and performing further scaling requires following this constraint. Then similarly to proof of work all previous results and security proofs remain correct provided that the constraint is followed.

4 GUC-Security of Multidimensional Blockchain

Universal composability framework is an approach frequently used in protocol security proofs [3, 6, 7]. This framework is a bases for security proofs for some cryptographic primitives (digital signatures, key exchange, etc.) and for complete systems – consensus mechanisms [9, 10].

In the framework, the system is defined as a set of interconnected interactive Turing machines (ITM) which in addition to ordinary Turing machines functionality have tapes for interactions with other ITM. There are several ways to construct security proofs with UC framework: simulation, numerical approach, and hybrid models. Any of these approaches restricts the model with several constraints which define the borders for security proofs.

Multidimensional blockchain security proof is based on multidimensional blockchain UC model. Construction of the model requires considering some features of the technology. In multidimensional blockchain, there is a set of blockchains which possibly use common settings—it means that, it is necessary to use version of UC framework with global setup (GUC-framework) [8]. Several assumptions for security proofs are used instead of admissible environment notion [16]:

1. Distributed ledgers can use common settings, i.e., to call common ideal functionalities.
2. Specifics of separate consensus mechanisms are not examined.
3. Robust ledgers used in multidimensional blockchain are identical to ideal functionalities implementing robust distributed ledger.
4. All assumptions on functioning of one-dimensional blockchains are followed.
5. An attacker is dummy and only transfers environment requests to system nodes.

During security proofs, modified models of one-dimensional blockchains are used. They differ from original with existence of external transactions. Search and verification protocol is out of scope which means that similarly to [13] V(.) predicate is used to present external transactions.

It is worth mentioning that different models are applied for security analysis of scaling and registration. Figure 4 presents a GUC-model of multidimensional

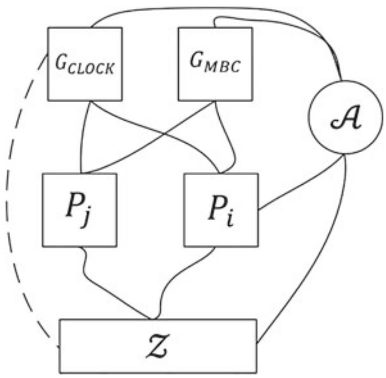


Fig. 4 GUC-model of multidimensional blockchain for scaling

blockchain. Multidimensional blockchain here works as a wrapper for ordinary one-dimensional blockchains. Verification of external transactions is performed using special ideal functionality.

Statement 1. Substituting one-dimensional blockchain with multidimensional blockchain persistence and liveness remain provided search and verification protocol is secure.

To prove this statement, it is sufficient to show that execution with multidimensional blockchain is equivalent to execution with one-dimensional blockchain from the environmental point of view. The case with one-dimensional blockchain inside multidimensional blockchain is trivial: multidimensional blockchain operates as a mediator and does not influence any protocols or security mechanisms. When several ledgers exist, multidimensional blockchain retranslates requests to the ledgers manually. Transactions are created by environment. Difference between external and ordinary transactions was presented earlier.

Hence, for environment systems with one-dimensional and multidimensional blockchains do not differ as both systems implement same set of messages and security of nested blockchains and ideal functionality G_{VERIFY} is given as an assumption.

Consider protocol implementing multidimensional blockchain. A model of node maintaining this protocol is required for it. Model is based on code of one-dimensional robust distributed ledger, given, for instance, in [13], with some extensions needed to work with external transactions. Among extensions there is G_{VERIFY} mentioned earlier. Transactions are also modified with fields required for external addressing. All the other interfaces must be equal to corresponding interfaces of G_{LEDGER} to construct security proof.

For model simplification, an approach with wrapper is used. Ledger-Protocol$_{q,D,T}$(p) from [13, 14] is extended with calls to G_{VERIFY} when using ValidTX. Besides, a handler for obtained transactions is defined:

1. If transaction is external, it is added to buffer and sent to G_{VERIFY} for check. After response, transaction is either added to a set of ready for inclusion to a block or is rejected.
2. If transaction is internal, it is added to set of transactions ready for inclusion to a block immediately.

Consider behavior of node maintaining multidimensional blockchain protocol MBC-Protocol. During initialization, the node registers itself into necessary ideal functionalities. Besides, the node obtains information on maintained ledgers from environment—it is emulating environment for internal Ledger-Protocol, i.e., it is executing code for several other processes inside its own process. Also, it performs registration in ideal functionalities for them. Finish of code execution in each round happens after finish of code execution in each internal protocol.

Requests CLOCK-READ and CLOCK-UPDATE are handled by controller and notifications are delivered to internal nodes. On SUBMIT the request is delivered to internal ledger. The node independently makes decisions on participation in ledger maintenance—by sending MAINTAIN-LEDGER to internal node. With each internal activation a request to G_{VERIFY} is made to obtain information about all external transactions not yet included to blocks.

Procedures ExtendState and FetchInformation are also important. When new state is created information about all new irreversible external transactions in new state is delivered to G_{VERIFY}. Irreversible state is a state created by a block which got deep enough to become irreversible by adversary. Additionally, G_{VERIFY} gets updates on state depth and accepts only transactions which went deep enough into the ledger.

Statement 2. MBC-Protocol UC-implements multidimensional blockchain provided secure search and verification protocol exists.

Proof of this statement is given with hybrid models. For this, components of protocol model are sequentially substituted by components of model based on ideal functionalities. A security proof if constructed with this gradual transition from one equivalent model to another, i.e., via emulation. Original and target models are given at Fig. 5.

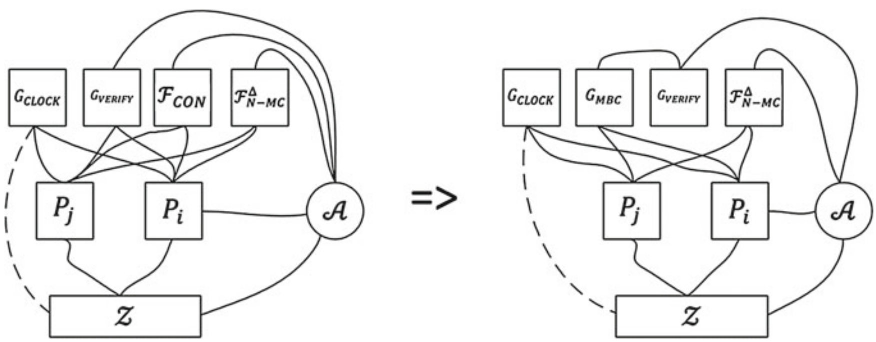


Fig. 5 Hybrid models for security proof

Next steps demonstrate intermediate models and their equivalence.

HYB0 is original model. All nodes run MBC-Protocol given previously. External transactions are verified with requests to search and verification protocol for blocks and transactions. Consensus mechanism is embedded into F_{CON}. HYB1 is a model where nodes communicate with ideal functionalities maintaining one-dimensional ledgers. Equivalence, if achieved, thanks to applying universally composability theorem [8]. Hence, each node independently redirects requests to ideal functionalities. HYB2 is a model where each node updated information about external transactions independently. Each G_{LEDGER} interacts with G_{VERIFY} independently. The models are equivalent as with protocol functioning notification is sent when transaction becomes irreversible for all honest nodes. Analogically, ledger notifies G_{VERIFY} when transaction is accepted as irreversible. HYB3 is a model with a wrapper which retranslates requests. In this case, all requests on transactions from nodes are retranslated to one node which forwards them to ideal functionalities working inside it. The approach differs from HYB2 only with its structure.

Last model is equivalent to a wrapper described previously. Hence, protocol MBC-Protocol UC-implements multidimensional blockchain. It is worth mentioning that implementations remain correct only provided search and verification protocol is secure.

Statement 3. MBC-Protocol UC-implements robust distributed ledger.

Proof for this statement follows from universally composability theorem. External interface of multidimensional blockchain is equivalent to external interface of one-dimensional blockchain. MBC-Protocol UC-implements multidimensional blockchain. Hence, it also UC-implements robust distributed ledger which from the environment point of view works equivalently to ideal functionality which implements multidimensional blockchain.

The approach with multidimensional blockchain for interactions between ordinary one-dimensional blockchains is implemented with a model presented on Fig. 6.

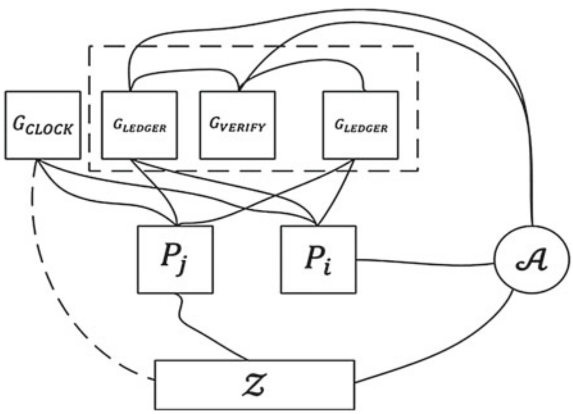


Fig. 6 A model for registration in multidimensional blockchain

Statement 4. Persistence and liveness of one-dimensional blockchain are not broken in multidimensional blockchain given UC-secure search and verification protocol exists.

Persistence and liveness are required and sufficient condition for distributed ledger to be robust. Persistence means that when transaction dives deep enough into the ledger, it becomes part of state for all nodes. Liveness means that honest node transaction is included into ledger in a finite time. In multidimensional blockchain, work of internal and outgoing external transactions does not differ from work of ordinary transactions in ordinary distributed ledger. This implies that persistence and liveness remain if equivalent ledger without external transactions is robust.

External transactions can break persistence if transaction gets inverted in initiating ledger. To prevent this, G_{VERIFY} functionality is used. If search and verification protocol works correctly, such situation does not occur. Liveness can be broken only if search and verification of external transaction is performed too long or is broken with probability greater than negligible. Analogously, with correct search and verification protocol, transactions are always verified in a finite time.

Statement 5. Multidimensional blockchain allows to implement secure information exchange between robust distributed ledgers provided search and verification protocol is UC-secure.

This statement is a corollary of Statement 4. With correct search and verification protocol, each ledger with external transactions remains robust. Thanks to receipt of information about external transactions, it is possible to implement secure exchange.

5 Conclusion

The paper presents analysis of multidimensional blockchain security. This is solution to scaling problem for robust distributed ledgers, which simultaneously allows to perform secure information exchange between separate robust distributed ledgers managed by different persons. Multidimensional blockchain is based on extension of one-dimensional blockchain notion. Main results of work which have scientific and practical value:

1. Formal description of multidimensional blockchain in block and state model.
2. General security evaluation of multidimensional blockchain security.
3. Mathematical evaluation of multidimensional blockchain security.
4. Statements proving:
    a. The possibility of secure information exchange between robust distributed ledgers built on basis of multidimensional blockchain.
    b. The possibility of applying multidimensional blockchain as a basis for building robust distributed ledger (or scaling existing ledgers).

The results can be used as basis for further research in sphere of robust distributed ledgers. An actual problem is designing of secure protocol for information exchange about external transactions—search and verification protocol for blocks and transactions. Currently, proof of security is based on suboptimal approach when all nodes relate to each other. A perspective research is experimental verification of multidimensional blockchain functioning and comparison to basic characteristics of other systems' functioning. An important scientific objective is implementation of consensus mechanisms which allow to reject mining or using system tokens for consensus. Advanced cryptography research is devoted to zero-knowledge algorithms and protocols. A non-solved problem is embedding these protocols into multidimensional blockchain.

References

1. Antonopoulos A (2014) Mastering bitcoin: unlocking digital crypto-currencies. O'Reilly Media Inc., Sebastopol
2. Badertscher C, Gaži P, Kiayias A, Russell A, Zikas V (2018) Ouroboros genesis: composable proof-of-stake blockchains with dynamic availability. In: ACM conference on computer and communications security—ACM CCS 2018, pp. 913–930
3. Badertscher C, Maurer U, Tschudi D, Zikas V (2017) Bitcoin as a transaction ledger: a composable treatment. In: Katz J, Shacham H (eds), Advances in cryptology—CRYPTO 2017. CRYPTO 2017. LNCS, vol 10401. Springer, Cham, pp 324–356
4. Bentov I, Gabizon A, Mizrahi A (2016) Cryptocurrencies without proof of work. In: Clark J, Meiklejohn S, Ryan P, Wallach D, Brenner M, Rohloff K (eds), Financial cryptography and data security. FC 2016. LNCS, vol 9604. Springer, Berlin, Heidelberg, , pp 142–157
5. Cachin C, Guerraoui R, Rodrigues L (2011) Introduction to reliable and secure distributed programming. Springer, Berlin, Heidelberg
6. Canetti R (2001) Universally composable security: a new paradigm for cryptographic protocols. In: Proceedings 42nd IEEE symposium on foundations of computer science. Newport Beach, CA, USA, pp 136–145
7. Canetti R (2014) Universally composable signatures, certification, and authenticated communication. In: Proceedings of 17th computer security foundations workshop (CSFW). IEEE Computer Society, Washington, DC
8. Canetti R, Dodis Y, Pass R, Walfish S (2007) Universally composable security with global setup. In: Vadhan SP (eds), Theory of cryptography. TCC 2007. LNCS, vol 4392. Springer, Berlin, Heidelberg, pp 61–85
9. Canetti R, Shahaf D, Vald M (2016) universally composable authentication and key-exchange with global PKI. In: Cheng CM, Chung KM, Persiano G, Yang BY (eds), Public-key cryptography—PKC 2016. PKC 2016. LNCS, vol 9615. Springer, Berlin, Heidelberg, pp. 265–296
10. David B, Dowsley R, Larangeira M (2019) ROYALE: a framework for universally composable card games with financial rewards and penalties enforcement. In: Goldberg I, Moore T (eds), Financial cryptography and data security. FC 2019. LNCS, vol 11598. Springer, Cham, pp 282–300
11. David B, Gaži P, Kiayias A, Russell A (2018) Ouroboros Praos: an adaptively-secure, semi-synchronous proof-of-stake blockchain. In: Nielsen J, Rijmen V (eds), Advances in cryptology—EUROCRYPT 2018. EUROCRYPT 2018. LNCS, vol 10821. Springer, Cham, pp 66–98

12. Duan S, Meling H, Peisert S, Zhang H (2014) BChain: byzantine replication with hight throughput and embedded reconfiguration. In: principles of distributed systems—OPODIS 2014, LNCS, vol 8878. Springer, Cham, pp 91–106
13. Garay J, Kiayias A, Leonardos N (2015) The Bitcoin backbone protocol: analysis and applications. In: Oswald E, Fischlin M (eds), Advances in cryptology—EUROCRYPT 2015. EUROCRYPT 2015. LNCS, vol 9057. Springer, Berlin, Heidelberg, pp 281–310
14. Garay J, Kiayias A, Leonardos N (2017) The Bitcoin backbone protocol with chains of variable difficulty. In: Katz J, Shacham H (eds), Advances in cryptology—CRYPTO 2017. CRYPTO 2017. LNCS, vol 10401. Springer, Cham, pp 291–323
15. Kiayias A, Russell A, David B, Oliynykov R (2017) Ouroboros: a provably secure proof-of-stake blockchain protocol. In: Katz J, Shacham H (eds), Advances in cryptology—CRYPTO 2017. CRYPTO 2017. LNCS, vol 10401. Springer, Cham, pp 357–388
16. Pass R, Seeman L, Shelat A (2017) Analysis of the blockchain protocol in asynchronous networks. In: Coron JS, Nielsen J (eds), Advances in cryptology—EUROCRYPT 2017. EUROCRYPT 2017. LNCS, vol 10211. Springer, Cham, pp 643–673
17. Pease M, Shostak R, Lamport L (1980) Reaching agreement in the presence of faults. J ACM (JACM) 27:228–234
18. Sompolinsky Y, Zohar A (2013) Accelerating Bitcoin's transaction processing fast money grows on trees, not chains. IACR Cryptology ePrint Archive
19. Vukilic M (2017) Rethinking permissioned blockchains. In: Proceedings of the ACM workshop on blockchain, cryptocurrencies and contracts (BCC '17), pp 3–7

Varidation of Indoor Localization Method by CNN Using RSSI

Makoto Uehara, Masashi Kimura, and Hideyuki Kobayashi

Abstract A Global Positioning System (GPS) is widely used as a method of outdoor localization. However, indoor localization system is required because the GPS signals are disturbed indoors. Although the fingerprinting method is a simple and inexpensive method of indoor localization, it is affected by environmental factors such as multipath effects. This paper proposes a Convolutional Neural Networks (CNN)-based localization model that consists of pooling layers for Wi-Fi fingerprinting indoor localization. These pooling layers of CNN can extract features of RSSI that contribute to the localization. Experimental results show that the proposed model reduces MAE about 16%. Furthermore, the coefficient of determination improves by 0.5 compared to the fingerprinting method.

Keywords Received signal strength indicator · Convolutional neural network · Fingerprint-based localization · Indoor localization.

1 Introduction

Localizations for people or objects are necessary for navigation systems. GPS is a main outdoor localization method. It is also used in wireless devices. However, the GPS cannot be used indoors because of the inability to receive satellites signals. On the other hand, indoor localization methods using Wi-Fi access points are proposed, which because plenty Wi-Fi access points exist in indoor environments. Two major

M. Uehara (✉) · H. Kobayashi
National Institute of Technology, Sendai College, 4-16-1 Ayashichuo Aoba-ku, Sendai, Miyagi 989-3128, Japan
e-mail: a1911501@sendai-nct.jp

H. Kobayashi
e-mail: kobayashi@sendai-nct.ac.jp

M. Kimura
Convergence Lab. Inc., 1-9-5 #806 Kanda-awajicho, Chiyoda-ku, Tokyo 101-0063, Japan
e-mail: kimura@convergence-lab.com

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_84

methods of indoor localization using Wi-Fi signals have been proposed. One of them is based on a Triangulation. The other is based on a Scene Analysis. Triangulation includes Time Of Arrival (TOA), Time Difference Of Arrival (TDOA), Angle Of Arrival (AOA) [1–4], and so on. However, these technologies require the location of the access points in advance or additional hardware. Thus, if these method uses for indoor localization, it leads to high equipment costs. The Scene Analysis, represented by the fingerprinting method, is a cheaper and simpler method of indoor localization. Fingerprinting method uses the Received Signal Strength Indicator (RSSI) which is the received signal strength from a Wi-Fi access point [5]. The fingerprint consists of a set of RSSIs and access points' location. Localization using this method has two phases. The first phase is the offline acquisition phase which creates a fingerprint database. The second phase is online phase called a matching phase. In this phase, the user measures the real-time RSSI for their location and find mostly matches fingerprint from the database. The fingerprinting method finds the location using the two phases. However, RSSI is affected by environmental factors such as multipath effects. Especially, in indoors, environmental factors are significant because many objects exist, such as fingerprinting method degrades when estimating using RSSI affected by multipath.

This paper aims to reduce the impact of environmental factors such as multipath effects during the fingerprint acquisition process. Accuracy of fingerprint methods depends on the RSSI. Therefore, the fingerprint methods need to avoid using RSSIs which are affected by environmental factors such as multipath effects or shadowing.

Machine learning algorithms such as RandomForest have the ability of estimate the location without using RSSI affected by multipath. However, machine learning algorithms cannot select features of RSSI for localization automatically.

Thus, we propose a Convolutional Neural Networks (CNN)-based localization model that consists of pooling layers in this paper. These pooling layers can extract features of RSSI. Specifically, each RSSI is weighted in the neural network to determine importance in localization. Experimental results show that our model outperform traditional fingerprinting method and machine learning methods in indoor localization accuracy. Moreover, the proposed methods outperform in fewer access point environment.

2 Related Works

2.1 Localization Algorithm

Two major methods of indoor localization using Wi-Fi signals have been proposed. One of them is based on a Triangulation. The other is based on a Scene Analysis.

Triangulation. Triangulation estimates people or objects by using several reference points with known coordinates. This scheme can be further divided into the lateration or the angulation.

Lateration. The lateration estimates the position by measuring distance from several reference points to device. The lateration needs to measure the distance from at least three reference points. The distance is generally inferred from time of arrival (TOA), time difference of arrival (TDOA). These approaches can be founded in literature. Schauer et al. [6] concludes TOA-based Wi-Fi has been not yet feasible on off-the-shelf devices because of the need for a more accurate Wi-Fi clock. Li et al. [7] investigates TDOA in wireless LANs. However, all transmitters and receivers must be precisely synchronized. In addition, a timestamp needs to be labeled in the transmission signal so that the receiving device can compute the travel distance of the signal.

Angulation. The angulation calculates the angle of arrival (AOA) relative to multiple reference points. Angulation does not require synchronization between transmitter and receiver devices. How et al. [8] proposes AoA-based wireless indoor localization method using Wi-Fi access points with antenna arrays. However, this approach needs antenna arrays or directional antennas, which makes the receiver complex. Moreover, the estimation accuracy decreases as the target device moves away from the reference points. A small angular error leads to a large error in the estimation [4].

Scene Analysis. Fingerprinting method classified as Scene Analysis collects location-specific fingerprints in offline phase, and compares the fingerprints obtained instantly in the offline phase with those obtained in the offline phase. The fingerprint is defined by the RSSI from each access point at a specific location. The RSSI indicates the strength of the signal received by the wireless devices from signal sending devices such as Wi-Fi access points. RSSI is inversely proportional to the square of the distance. Therefore, each location has a unique fingerprint.

2.2 Deep Learning

Machine learning algorithms such as RandomForest and LightGBM are able to make predictions excluding outliers. Specifically, data including outliers are grouped together to reduce the impact on estimation. However, these algorithms cannot extract features automatically from the input.

Outliers are values that are significantly out of line in a dataset. Deep learning is a method for optimizing a network of layered functional modules represented by parameters using the gradient descent method [9, 10]. The result of affine transformation of the input of each layer is further transformed nonlinearly by the activation function to the output of the value. These are stacked to form a complex nonlinear transformation. Deep learning can reduce the impact of outliers on estimation by not advancing the learning when an outlier is received as network input.

An important aspect of deep learning is that the features of the input are learned from the data using a generic learning procedure. Traditionally, constructing a pattern recognition or machine learning system has required engineering and expertise to design a feature extractor. Deep learning allowed the system to automatically detect and classify the input patterns.

CNN [11] utilizes a convolutional and pooling layer in the middle layer. The pooling layer can feature extraction [12]. Therefore, CNN is used as a standard model for solving image recognition and various other tasks.

Activation Functions. In the neural networks used in deep learning, input is fed to the neurons in the input layer. Each neuron has a weight. Multiplying the input by the weight gives the output of the neuron. The activation function is used to scale the output in determining the output of each layer. The transformation of the activation function needs to be a nonlinear transformation to use gradient descent which known as backpropagation. Nonlinear transformations allow the network to provide a more complex representation and accurate predictions. Rectified Linear Units (ReLU) [13] is commonly used in CNN, and GeLU is an improved version of it.

ReLU has replaced classical activation functions such as Sigmoid and Tanh [14, 15]. ReLU is given as follows:

$$ReLU(x) = \begin{cases} 0 & (x \leq 0) \\ x & (x > 1) \end{cases} . \tag{1}$$

From the equation (1), if the input x is less than 0, the output becomes 0. Everything greater than zero become its own value. If classical activation functions are used in deep networks, vanishing gradient problems occurs. ReLU solves the problems. In addition, ReLU is fast to learn because the output increases as the value of the input increases.

Gaussian Error Linear Unit (GeLU) [16] is an activation function used in OpenAI GPT [17] and BERT [18]. This function is defined by the following equation:

$$GeLU(x) = x\Phi(x) . \tag{2}$$

where $\Phi(x) = P(X \leq x), X \sim N(0, 1)$ is the cumulative distribution function of the standard normal distribution. It can be approximated by

$$GeLU(x) \approx 0.5x \left(1 + tanh\left(\sqrt{2/\pi}(x + 0.044715x^3)\right)\right) . \tag{3}$$

In Eq. (2), GeLU multiplies the input by 0 with a probability of $\Phi(x)$ and multiplies the input by 1 with a probability of $(1 - \Phi(x))$. GeLU takes into account a probabilistic element in ReLU. ReLU deterministically multiplying the input by zero. GeLU is determined probabilistically by multiplying the input by zero or one.

2.2.1 Objective Functions.

The goal of deep learning is to minimize the error between the target value t and the prediction y. A neural network computes this error and adjusts the parameters according to the task. In this section, Mean Square Error (MSE) loss and Smoot L1 norm loss are discussed.

MSE loss calculates the sum of the errors for target and prediction value and divides it by the number of data. MSE loss is given as follows:

$$MSEloss = \frac{1}{N}\sum_{n=1}^{N}(t_n - y_n)^2 \ . \tag{4}$$

where N is the number of samples, y_n is the predicted value for the n-th data, and t_n is the target value for the n-th data.

Girshick [19] proposed Smooth L1 norm loss. The equation is given as follows:

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & (|x| \leq 1) \\ |x| - 0.5 & otherwise \end{cases} \ . \tag{5}$$

If the x is less than 1, Smooth L1 norm loss uses the squared term. Otherwise, the L1 term is used. Since the squared term squares the value, the cost of the outlier increases exponen-tially. L1 loss is highly robust because it only takes absolute values. Thus, smooth L1 norm loss is less sensitive to outliers than MSE loss.

2.3 *Dataset for Indoor Localization*

Miskolc IIS Hybrid Indoor Positioning System Dataset [20–22] contains more than 1500 measurements in a three-story building and covers about 50% of it. The measurements were recorded with the same type of Android device in order to reduce the influence of the variety of hardware. In addition to RSSI from 32 access points, it also includes Bluetooth and magnetic sensor values. The collection location of each measurement is given in coordinates (x, y, z) from a specific location in the building. Each position has one data; therefore, the number of data means the number of measurement points. This dataset was available in the UCI Machine Learning Repository [22]. In this paper, we will refer to this dataset as the Miskolc Indoor Dataset.

Wi-Fi RSSI Indoor Localization [23, 24] was measured at third floor of University of Victoria. Autonomous robot was used to collect the Wi-Fi fingerprint. The robot has multiple sensors including wheel odometer, an inertial measurement unit (IMU), a LIDAR, sonar sensors, and a color and depth (RGB-D) camera. It can move to the target location and collect fingerprints automatically. The localization accuracy of this robot is 0.07mÂ±0.02m. 6 Wi-Fi access points and five of them provide also 5 GHz radio waves. Since the dataset contains a total of 11 access points, RSSIs from each access point was measured at 345 points. The biggest difference from Miskolc Indoor Dataset is that, about 50 fingerprints are collected for each location. Since this data is for one floor only, there are two coordinate axes: x and y. In this paper, we will refer to this dataset as the Victoria Indoor Dataset.

3 Proposed Method

In this paper, we propose a Convolutional Neural Networks (CNN)-based localization model that consists of a pooling layer for indoor localization using RSSIs. The CNN consists of an input/output layer, convolutional and pooling layers for feature extraction. Figure 1 shows the network structure. Two hidden layers, consisting of a convolutional layer and a pooling layer. AdamW is used for the gradient descent method. For the activation function, we use the Gaussian Error Linear Unit (GeLU). This function is an improved version of the Rectified Linear Unit (ReLU) commonly used in CNNs. In order to investigate the effect of outliers, we create a model using smoothed L1 norm loss and mean squared error (MSE) loss as the objective function. To avoid overfitting, early stopping method is applied which stop the training when the validation errors increases.

In experiments, we compare with fingerprinting methods and machine learning algorithms using accuracy metrics. The Mean Absolute Error (MAE) is used as an accuracy metric. In addition, the coefficient of determination (R^2) is calculated to determine the contribution of RSSI in localization. R^2 is a measure of the correlation between the predicted values and the actual values in the database.

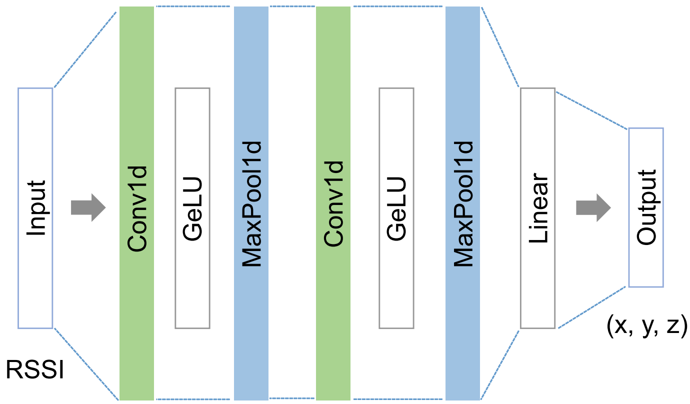


Fig. 1 Structure of Network

4 Experiments

4.1 Accuracy Evaluation Experiment

To investigate the effectiveness of the proposed model in localization, we evaluate MAE and R^2 of the fingerprinting method, two machine learning algorithms, and the proposed model.

An evaluation dataset is Mishkolz Indoor Dataset. We use 35 features, including the coordinates and RSSI from all access points. Location names and device IDs were removed because they do not affect the estimation.

In the fingerprinting method, 20% of the data use as real-time RSSI measured in the online phase. Rest of 80% use as a fingerprint database. Fingerprint method is based on the least-squares method. The coordinate (x, y, z) of the instance with the smallest error is the estimated position.

We compare the proposed model with two machine learning algorithms, RandomForest and LightGBM. 80% of the dataset use for training and rest 20% use for testing. The data used in early stopping are also used to adjust the hyper parameters. Thus, it affects the termination of learning. Therefore, 20% of the training data use as the validation data for early stopping.

Furthermore, we investigate the impact of outliers on the estimation. We created a model using two objective functions, Smoothed L1 norm and MSE loss. Smoothed L1 norm has an effective property for outliers on the dataset.

4.2 Investigation of the Number of Fingerprinting at Each Location

To investigate the variation in accuracy with the number of access point, we evaluate another dataset. In addition to Mishkolz Indoor Dataset used in 4.1, we use Victoria Indoor Dataset. This dataset has fewer number of access points than Mishkolz Indoor Dataset. However, this dataset has more data than Mishkolz Indoor Dataset.

In this experiment, each model is evaluated by calculating MAE and R^2 using two datasets. In addition to all the data from each location (n = 50, n = the number of data of each location), we take the average of all the data and set the number of data to 1 (n = 1) in Victoria Indoor Dataset.

5 Results and Discussions

5.1 Results of Accuracy Evaluation Experiment

Experimental results from Mishkolz Indoor Dataset are shown in Table 1. The proposed model with Smoothed L1 norm loss has the lowest MAE is 1.38 and the highest R^2 is 0.93. The high R^2 indicates that the proposed model can well capture the characteristics of the RSSI at each location. Furthermore, the MAE of Smoothed L1 norm loss is lower than that of MSE loss. This is attributed to the outlier-resistant property of the former. It shows that the effect of outliers can be minimized by changing the objective function in the proposed model. The fingerprinting method is second highest after the proposed model with MAE is 1.64. However, R^2 is 0.89. These results show the lowest among all models. The fingerprinting method refers to all the RSSIs in the database. Therefore, unreliable RSSIs may be used for estimation. Thus, the large variance of the RSSI makes it difficult to use the fingerprinting method for an estimator.

5.2 Results of Investigation of the Number of Fingerprinting at Each Location

The results of experiments comparing the two patterns are shown in Table 2 and 3. The accuracy of the proposed model was lower than other models when the number of data at each location was 50 (n = 50). This is attributed to the low representation of the model. The results for the number of data (n = 1) are shown in Table 3. The proposed model has lowest MAE; however, each model had the same level of accuracy.

Table 1 MAE and R^2 for Each Model in Miskolc Indoor Dataset

	Fingerprinting	Machine Learning Algorithm		Proposed Model	
	Method	RandomForest	LightGBM	MSELoss	SL1Loss
MAE	1.64	2.51	2.53	1.54	1.38
R^2	0.89	0.91	0.92	0.93	0.94

Table 2 MAE and R^2 for Each Model (5.2, n=50)

	Machine Learning Algorithm		Proposed Model	
	RandomForest	LightGBM	MSELoss	SL1Loss
MAE	0.10	0.23	0.26	0.23
R^2	0.99	0.99	0.99	0.99

Table 3 MAE and R^2 for Each Model (5.2, n=1) Take the average of all the data and combine them into one

	Fingerprinting Method	Machine Learning Algorithm		Proposed Model	
		RandomForest	LightGBM	MSELoss	SL1Loss
MAE	0.71	0.71	0.72	0.69	0.70
R^2	0.96	0.97	0.97	0.98	0.98

5.3 Overall Discussions

The proposed model achieved a higher accuracy than the other models in Mishkolz Indoor Dataset. Mishkolz Indoor Dataset has less building coverage and less information than Victoria Indoor Dataset. Experimental results show that the proposed model is highly accurate even if the information of RSSI in the building is missing.

Throughout the two experiments, the accuracy is not good when the measurement was only once at each location. However, as we can see from Table 2, the results get better as the number of data increases. In fact, the number of data in Victoria Indoor Dataset was more than ten times greater than in Mishkolz Indoor Dataset. As a result, the number of collections at each location is important in the localization. Increasing the number of measurements will improve the accuracy.

6 Conclusions

In this paper, we proposed a CNN-based localization model including a pooling layer that can extract features of RSSI. Through comparison with related methods, the proposed method outperforms fingerprinting method. Especially, the proposed model shows the highest accuracy in Mishkolz Indoor Dataset which covers only about 50% locations of the building. Increasing the number of measurements at each location is expected to improve an accuracy of the estimates. With the increase in the number of data, the model needs to add more layers and parameters. In future work, we will improve the model to allow more types of data. In addition, we will collect data in crowded and uncrowded environments. Through experiments using these data, we will investigate the effect on the estimation of changes in RSSI.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number 19K14045.

References

1. Ito S, Kawaguchi N (2006) Wireless lan based hybrid positioning system using bayesian inference and access point selection. IEEJ Trans Electron, Inf Syst 126(10):1212–1220
2. Liu H, Darabi H, Banerjee P, Liu J (2007) Survey of wireless indoor positioning techniques and systems. IEEE Trans Syst, Man, Cybern, Part C (Appl Rev) 37(6):1067–1080. https://doi.org/10.1109/TSMCC.2007.905750
3. Chen C, Chen Y, Han Y, Lai H, Liu KJR (2017) Achieving centimeter-accuracy indoor localization on wifi platforms: a frequency hopping approach. IEEE Internet Things J 4(1):111–121. https://doi.org/10.1109/JIOT.2016.2628701
4. Xia S, Liu Y, Yuan G, Zhu M, Wang Z (2017) Indoor fingerprint positioning based on wi-fi: an overviewISPRS Int J Geo-Inf 6(5):135
5. Takehiro Iwadate HK (2019) Investigation of gathering fingerprints method using multiple terminals considered irregulars rssi. IEICE Tech Rep 118(466):313–318
6. Schauer L, Dorfmeister F, Maier M (2013) Potentials and limitations of wifi-positioning using time-of-flight. In: International conference on indoor positioning and indoor navigation, pp 1–9. https://doi.org/10.1109/IPIN.2013.6817861
7. Xinrong L, Pahlavan K, Latva-aho M, Ylianttila M (2000) Comparison of indoor geolocation methods in dsss and ofdm wireless lan systems. In: Vehicular technology conference Fall 2000. IEEE VTS fall VTC2000. 52nd vehicular technology conference (Cat. No.00CH37152), vol 6, pp 3015–3020, vol 6. https://doi.org/10.1109/VETECF.2000.886867
8. Hou Y, Yang X, Abbasi Q (2018) Efficient aoa-based wireless indoor localization for hospital outpatients using mobile devices. Sensors 18(11):3698
9. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
10. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536
11. Lecun Y, Haffner P, Bengio Y (1999) Object recognition with gradient-based learning. In: Shape, contour and grouping in computer vision. Springer-Verlag, Berlin, Heidelberg, p 319
12. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems, vol 25. Curran Associates, Inc., pp 1097–1105. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
13. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on international conference on machine learning, ICML'10. Omnipress, Madison, WI, USA, pp 807–814
14. LeCun Y, Bottou L, Orr GB, Müller KR (1998) Efficient backprop. In: Neural networks: tricks of the trade, this book is an outgrowth of a 1996 NIPS Workshop. Springer-Verlag, Berlin, Heidelberg, pp 9–50
15. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Gordon G, Dunson D, Dudík M (eds) Proceedings of the fourteenth international conference on artificial intelligence and statistics, *Proceedings of Machine Learning Research*, vol 15, pp 315–323. JMLR Workshop and Conference Proceedings, Fort Lauderdale, FL, USA. http://proceedings.mlr.press/v15/glorot11a.html
16. Hendrycks D, Gimpel K (2016) Gaussian error linear units (gelus)
17. Radford A (2018) Improving language understanding by generative pre-training
18. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding
19. Girshick R (2015) Fast r-cnn
20. Tamas J, Toth Z (2018) Classification-based symbolic indoor positioning over the miskolc iis data-set. J Locat Based Serv12(1):2–18
21. Tóth Z, Németh R, Tamás J. Data model for hybrid indoor positioning systems
22. Hybrid indoor positioning dataset from wifi rssi, bluetooth and magnetometer data set. http://archive.ics.uci.edu/ml

23. Hoang MT, Yuen B, Dong X, Lu T, Westendorp R, Reddy K (2019) Recurrent neural networks for accurate rssi indoor localization. IEEE Internet Things J 6(6):10639–10651. https://doi.org/10.1109/JIOT.2019.2940368
24. Westendorp, M.T.H.X.D.T.L.B.Y.R.: Wifi rssi indoor localization (2019). https://doi.org/10.21227/1yd5-rn96. https://dx.doi.org/10.21227/1yd5-rn96

Applicability of AutoML to Modeling of Time-Series Data

Ajanta Kancharla and N. Raghu Kishore

Abstract More and more businesses have taken to adopting Machine Learning (ML) and Deep Learning methods for solving problems in the application domain. One such area is Anomaly detection. While research in recent years has yielded better accuracies, the main hurdle for businesses remains managing the high cost of resources required for their implementation. Traditional ML/Deep Learning techniques require expert knowledge, and thus presently rely on human efficiency. Automated machine learning (AutoML) has emerged as a solution to the above scenario. In this paper, we evaluate the performances of traditional anomaly detection techniques used in time-series forecasting, namely, statistical and deep learning models, and compare their results in terms of efficiency and performance with AutoML. We use H2O driverless AI (H2o.ai automl documentation, http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html, accessed: 2020-06-10) as our platform for experiments on AutoML. All the experiments are performed on the Numenta Anomaly Benchmark's streaming time-series dataset (Lavin A and Subutai A in 2015 IEEE 14th international conference on machine learning and applications (ICMLA), pp 38–44).

Keywords Anomaly detection · Time series · H2O AutoML · Convolutional LSTM · Recurrent neural network · LSTM autoencoders · ARIMA · Univariate · Real-time computing

1 Introduction

AutoML proposes a way of building efficient machine learning models with minimal effort and machine learning expertise [1]. It claims to reduce the time and effort spent on tedious parts of the AI life cycle. AutoML services are being offered by different

A. Kancharla (✉) · N. Raghu Kishore
Mahindra Ecole Centrale, Hyderabad, India
e-mail: ajanta160506@mechyd.ac.in

N. Raghu Kishore
e-mail: raghukishore.neelisetti@mechyd.ac.in

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_85

platforms and enterprises which can not only reduce dependency on data scientists and analysts but also significantly speed up the process of setting up machine learning pipelines for even seasoned data science practitioners. One such area where AutoML lends its services is in the forecasting of time series and in the detection of abnormalities in time-series data [2, 3].

Any system which detects anomalies must ensure that the anomalous behavior is detected to prevent disruptions in business processes and troubleshooting. The data used by enterprises and businesses is often real-time streaming data where it is necessary that anomalies are identified and relevant models are updated before the next time step [4]. With growing interest in automation, many businesses and organizations have started moving to AutoML to experience a significant rise in productivity. With the precarious nature of the anomaly detection problem itself and the buzz around AutoML, a natural question arises, how well do existing, traditional anomaly detection methods compare against AutoML?

In this paper, we evaluate the performances of different deep learning architectures and statistical algorithms known to perform well on anomalous data, namely LSTM AutoEncoders, RNN, Convolutional LSTM, and ARIMA, against AutoML for the detection of anomalies in Numenta Anomaly Benchmark's (NAB) [5] four datasets. For our experiments, we used H2O AutoML. H2O was selected for our experiments because its platform provides functionality for time-series data. In [6], it was found that H2O's AutoML performed slightly better than other AutoML platforms used for comparison, quickly converging to optimal results.

In this paper, the results achieved using the different methods are evaluated and compared on the basis of better performance and efficiency.

1.1 Data Preprocessing

H2O AutoML [2] offers the ability to detect basic data types, like numerical, categorical, and time-series data. It handles missing data present during training, through different tree-based algorithms like XGBoost, LightGBM, and RuleFit, and also accounts for gaps and forecast horizon in the dataset.

1.2 Modeling Approach

Feature Tuning: This stage combines random hyperparameter tuning with feature selection and generation. H2O uses gradient boosting algorithms, generalized linear models, genetic algorithms, and neural networks for time-series-specific feature engineering. In each iteration, features are updated using variable importance from the previous iteration as a probabilistic prior to deciding what new features to create [7]. The best performing model and features are then passed to the feature evolution stage. The selection of a model is mostly based on validation accuracy achieved on

Table 1 Part of data used in experiments

Ambient temperature system		
Index	Timestamp	Value
0	2013-07-04 00:00:00	69.880835
1	2013-07-04 01:00:00	71.220227
2	2013-07-04 02:00:00	70.877805

the window. **Feature evolution**: This stage uses a genetic algorithm to find the best set of model parameters and feature transformations to be used in the final model. The results of this phase is a set of features arrived after transformation on the dataset, which are to be used in the final model.

2 Dataset

The dataset used for all the experiments is NAB [8] corpus comprising real-world, time-series data files in which the anomalous periods of behavior in each time-series dataset are labeled. We used four of NAB's real-world datasets with known causes for our experiments, ambient temperature system failure, CPU utilization, ec2 request latency system failure, and NYC Taxi. According to Numenta instructions [8], the first 15% of the dataset was used as a probationary period, to learn normal patterns of behavior. The anomalous periods detected by the different anomaly detection methods are cross-checked against the labels provided (Table 1).

3 Real-Time Anomaly Detection Methods and Algorithms

In this section, we define the different anomaly detection methods which have been used for comparison against H2O AI's AutoML. A mixture of deep learning and statistical methods is experimented on and compared against H2O on the basis of better efficiency and performance.

Recurrent Neural Network (RNN): Recurrent neural networks have been successively implemented for detecting anomalies in time-series data [9]. A recurrent neural network (RNN) is a case where the objective is predicting a future unseen instance in the sequence of observations with respect to the window of observations seen earlier in the sequence. They are recurrent because they repeat the same task for every element of a sequence. They make use of historical data to learn from earlier stages where the hidden layers act as memory or storage for capturing information in the sequence [10]. RNN suffers from the famous vanishing gradient problem, where the information captured in the earlier steps is not retained [11]. Thus, RNNs are not suitable for remembering longer sequences of data. This problem is solved by the use of Long Short-Term Memory which handles it by the use of additional memory cells called "memory line" which store memory states. Hence, they are more powerful for time-series forecasting [10].

Convolutional Neural Network (CNN): Convolutional Long Short-Term Memory architecture has been shown to produce good results in time-series forecasting [12] and anomaly detection [13]. Convolutional neural network (CNN) takes a sequence of past observations as input and uses this to output a future observation. It does so by using convolutional filters, which extract informative features from the data series [12]. In a time series of length n, with a number of variables k, the kernel assumes the width of the time series with a variable length and moves from the beginning of the time series to the end performing convolutions.

LSTM AutoEncoders: Long Short-Term neural autoencoders have been shown to be effective for time-series learning, including for outlier detection [14]. An autoencoder consists of three parts, an encoding function, a decoding function, and a distance function; mathematically, they can be defined as transition ϕ and ψ such that

$$\begin{aligned}
&\phi : X \to F \\
&\psi : X \to X \\
&\arg\min_{\psi,\phi} = \|X - (\psi \times \phi)X\|^2
\end{aligned}$$

Autoencoders encode high-dimensional input data to the hidden layers using relevant non-linear activation functions and then try to reconstruct the non-linear combination of input features through the decoding layer [14]. Autoencoder ensembles rely on the availability of multiple recurrent neural network autoencoders with different network connection structures [14]. They are trained with a backpropagation algorithm which tries to minimize the reconstruction loss.

ARIMA: The steps followed for building an ARIMA model are (1) Model Identification, (2) Parameter Selection, (3) Modal Validation, and (4) Model use [15, 3]. In the model identification phase, the stationarity of a given data series is determined. A stationary time series is one in which statistical properties of a process do not change with time, or values of a variable vary around a constant mean and variance. Based on this property, variable "d" is chosen in order to remove the temporal dependence of data. As the acronym suggests, ARIMA (p, d, q) captures the three key elements of the model, **A**uto-**R**egression, a regression model that uses the dependencies between an observation and number of lagged observations (p). It can be represented mathematically by the equation:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} \ldots + \beta_p Y_{t-p} + \varepsilon_t$$

where $Y_{t-1}$is the lag1 of the series, β_1 is the coefficient of lag1 that the model estimates, and α is the intercept term. Similarly, **M**oving **A**verage [16] is an approach that takes into account the dependency between observations and their residual error terms, and thus only depends on lagged forecast error. It can be represented mathematically by the equation:

$$Y_t = \alpha + \varepsilon_1 + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} \ldots + \phi_q \varepsilon_{t-p}$$

where the error terms are the errors of the autoregressive models of the respective lags. The ARIMA model combines both AR and MA terms and can be shown as an equation:

$$\mathbf{Y_t} = \alpha + \beta_1 \mathbf{Y_{t-1}} \ldots + \beta_\mathbf{p} \mathbf{Y_{t-p}} + \varepsilon_\mathbf{t} + \phi_1 \varepsilon_{\mathbf{t}-1} .. + \phi_\mathbf{q} \varepsilon_{\mathbf{t-p}}.$$

4 Experimental Setup

All experiments were conducted on Python 3 Jupyter Notebook, running Ubuntu 16.04, TensorFlow was used as our backend, and Keras as our core model development library in the deployment of deep neural network models. For AutoML experiments, H2O Docker was used, with a system memory of 64 GB, 4 CPUs, and 1 GPU as provided in H2O's Driverless AI.

For our experiments, in each dataset useful features such as hour of the day, day/night, and weekend/weekday are extracted from the timestamp column and are stored as features on the data-frame. We follow the approach used in Numenta, where the first 15% of the dataset is used as a probationary period for learning normal or expected behavior of the time series. 10% split is used for the validation set, and the rest of the dataset is used for testing. The data is normalized to a range from 0 to 1 and is reshaped into a format suitable for input into an LSTM network. The data is then split into sequences of historical data; depending on the dataset, a suitable size is chosen.

For the implementation of **RNN**, three layers are used. The first layer is a Simple RNN layer with 64 units. The next layer is a dropout layer with rate 0.3, to avoid overfitting. And the final layer is a dense layer with 1 neuron and linear activation. For the implementation of **Convolutional LSTM**, the architecture remains similar to that of RNN, the first layer is changed to a Convolution layer with 64 filters having kernel size 1 and activation **"relu"** and the next layers are a maxpooling layer and dense layer with 30 and 1 neurons with tanh activation. For the implementation of **LSTM autoencoder** [15], 5 layers are used. The first layer is an LSTM layer with 64 units. The second layer is a dropout layer with rate 0.2. The next layer is a RepeatVector layer that simply repeats the input n times. We use the Adam optimizer with loss function as Mean Absolute Error. The number of epochs for each of the models is set to 40.

For **ARIMA**, the p, d, and q values are determined with the help of Pyramid's auto.arima. Pyramid determines the best set of parameters according to a given information criterion and the stepwise algorithm which is less likely to overfit compared to an extensive grid search.

H2O AutoML provides functionality for the selection of accuracy/time/interpretability settings before the training of the dataset. We chose the parameters based on the recommended settings provided by H2O. Accuracy is set in between 7 and 10, time, between 1 and 3, and interpretability is set to 5 on the H2O Driverless AI user interface.

All the methods are evaluated on the basis of anomalies detected in a small anomalous window (see Fig. 1) instead of discrete points. The size of the anomaly windows, centered around a ground truth anomaly, is taken as proposed by Numenta Benchmark [5], 10% length of data divided by the number of ground truth anomalies in the data file. All anomalies detected within an anomalous window are taken to be one; anomalies falling outside the window are considered to be false negatives. As has been proposed by Numenta [8], a threshold is chosen such that it outputs the best results on the dataset, true anomalies detected.

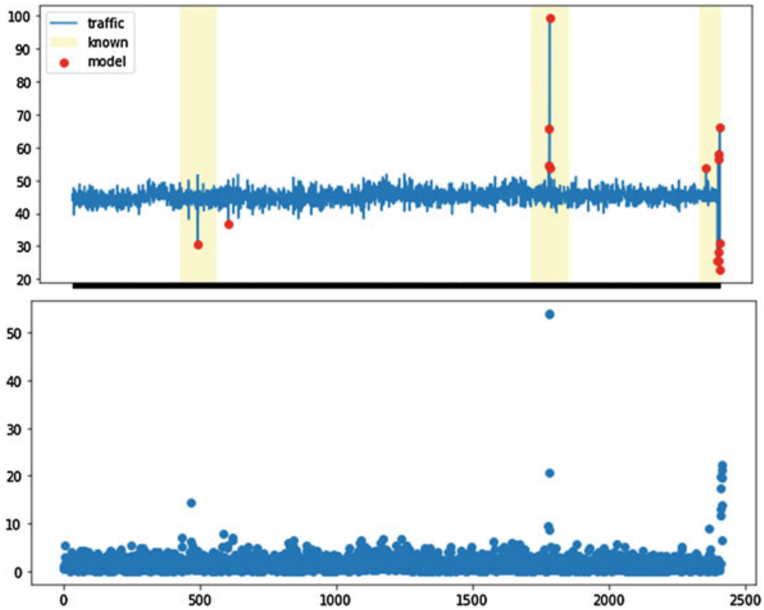


Fig. 1 Top plot: Anomalies visualized on Matplotlib using the predictions generated by H2O. The axis spanned in yellow depicts the anomaly window size around ground truth anomalies. The red points represent the anomalies detected by using a suitable threshold on the precision error, which gives the best results across all the windows on ec2 latency request. Bottom plot: prediction error for all the predicted values generated; higher prediction error denotes the presence of outliers

5 Results

We use standard metrics of Precision and Recall to compare the performances of different methods. Table 2 presents the execution time, and number of TP, FP, and FN observed by each model on a given dataset. It can be observed that all the methods perform reasonably well in detecting anomalies, with almost no false negatives, and a few false positives. ARIMA implemented using auto-arima was observed to be computationally bulky taking a longer time for training. H2O AutoML achieves good recall and precision (Table 3) across all datasets; its performance is on par, sometimes better than the results achieved on popular neural architecture frameworks and ARIMA for anomaly detection. In our experiments, the time taken by the last step of the AutoML pipeline, the final **pipeline training** (see Table 4) is taken as the execution time for comparison in Table 2. It was done because it wouldn't be fair to compare the total time taken by AutoML, including all the steps in the pipeline like data preprocessing, model selection, and feature evolution, against the training time taken by a single model/method on a dataset with defined parameters as set from human expertise.

In H2O AutoML, the time taken for feature evolution was observed to take the maximum time (approximately 75–85% of the total execution time of pipeline). Table 4 provides an example of the time breakdown for the training of different models at each stage on H2O in ec2 request latency. Driverless AI publishes an auto-report after each experiment which contains useful information obtained such as performance overview—MAE/RMSE on training and test dataset, experimental methodologies, algorithms used, feature transformations observed and their importance, assumptions, and limitations for selection of the final model which outputs optimal results. They can be useful in improving existing models; for instance, in the report published for latency requests, some of the feature transformations were mean, median, and sum of target lags, which achieved relative importance of greater than 0.5. Had they been engineered as features for training in the traditional methods, better performance could have been realized.

6 Conclusions

In this paper, we evaluated and compared the performances of traditional methods with AutoML for anomaly detection in time-series data. It was observed that H2O AutoML Driverless AI performed well across all datasets and achieved good precision and recall. Its performance was on par, sometimes better than the performance achieved on popular, traditionally used methods for a delicate problem like anomaly detection in time series. We believe, AutoML can indeed be used in tandem with human judgment to realize better methods in future.

Table 2 Results

Detector	Ambient temperature			NYC taxi demand			Ec2 request latency			CPU utilization		
	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN
AutoEncoder	2	4	0	2	3	3	3	0	0	2	4	0
Conv. LSTM	2	2	0	5	3	0	3	1	0	2	2	0
RNN	2	3	0	5	4	0	3	1	0	2	3	0
ARIMA	2	2	0	5	4	0	3	1	0	1	2	1
H2O	1	2	1	4	0	1	3	1	0	2	3	0

Table 3 Precision, Recall, and Execution time

Detector	Ambient temperature			NYC taxi demand			ec2 request latency			CPU utilization		
	P	R	Time	P	R	Time	P	R	Time	P	R	Time
AutoEncoder	0.33	1	78.8 s	0.4	0.4	149.08 s	1	1	37.98 s	0.33	1	1401.92 s
Conv. LSTM	0.5	1	36.82 s	0.62	1	120.86 s	0.75	1	25.30 s	0.5	1	1221.02 s
RNN	0.4	1	12.86 s	0.55	1	131.88 s	0.75	1	42.82 s	0.4	1	303.72 s
ARIMA	0.5	1	62.93 min	0.55	1	140.33 min	0.75	1	36.76 min	0.33	0.5	165 min
H2O	0.33	0.5	61.13 s	1	0.8	107.42 s	0.75	1	69.27 s	0.4	1	108.06 s

Table 4 Timing breakdown in H2O driverless AI

Stage	Time (s)	No. of models
Data preparation	13.12	0
Model and feature tuning	531.09	77
Feature evolution	**2,176.34**	373
Final pipeline training	69.27	1

References

1. Feurer M, et al (2015) Efficient and robust automated machine learning. In: NIPS'15 proceedings of the 28th international conference on neural information processing systems, vols 2 and 28, pp 2755–2763
2. Ahmad S, Scott P (2016) Real-time anomaly detection for streaming analytics. ArXiv Preprint ArXiv:1607.02480
3. Ranjan KG, et al (2021) Review of preprocessing methods for univariate volatile time-series in power system applications. Electr Power Syst Res 191:106885
4. Canizo M, et al (2019) Multi-head CNN–RNN for multi-time series anomaly detection: an industrial case study. Neurocomputing 363:246–260
5. Numenta. 2017. Anomaly labeling instructions. https://drive.google.com/file/d/0B1XUjaAXeV3YlgwRXdsb3Voa1k/view
6. Truong A, et al (2019) Towards automated machine learning: evaluation and comparison of AutoML approaches and tools. In: 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI), pp 1471–1479
7. H2o.ai automl documentation. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html. Accessed 10 June 2020
8. Lavin A, Subutai A (2015) Evaluating real-time anomaly detection algorithms—the numenta anomaly benchmark. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA), pp 38–44
9. Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. Internat J Uncertain Fuzziness Knowl-Based Syst 6(2):107–116
10. Greff K, et al (2017) LSTM: a search space odyssey. IEEE Trans Neural Netw 28(10):2222–2232
11. Medel JR, Savakis AE (2016) Anomaly detection in video using predictive convolutional long short-term memory networks. ArXiv Preprint ArXiv:1612.00390
12. Hoff JC (1983) A practical guide to box-Jenkins forecasting
13. Shi X, et al (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: NIPS'15 proceedings of the 28th international conference on neural information processing systems, vols 1 and 28, pp 802–810
14. Kieu T, et al (2019) Outlier detection for time series with recurrent autoencoder ensembles. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence, pp 2725–2732
15. David FN, Peter W (1951) Hypothesis testing in time series analysis. Biometrika 114(4):579–579
16. MovingAverage. Retrieved February 28, 2017https://cran.rproject.org/web/packages/smooth/vignettes/sma.html
17. H2o.ai automl github," https://github.com/h2oai/h2o-3. Accessed 10 June 2020

18. Essien A, Giannetti C (2019) A deep learning framework for univariate time series prediction using convolutional LSTM stacked autoencoders. IEEE Int Symp Innov Intell Syst Appl (INISTA) 2019:1–6

Obtaining a ROS-Based Face Recognition and Object Detection: Hardware and Software Issues

Petri Oksa, Tero Salminen, and Tarmo Lipping

Abstract This paper presents solutions for methodological issues that can occur when obtaining face recognition and object detection for a ROS-based (Robot Operating System) open-source platform. Ubuntu 18.04, ROS Melodic and Google TensorFlow 1.14 are used in programming the software environment. TurtleBot2 (Kobuki) mobile robot with additional onboard sensors are used to conduct the experiments. Entire system configurations and specific hardware modifications that were proved mandatory to make out the system functionality are also clarified. Coding (e.g., Python) and sensors installations are detailed both in onboard and remote laptop computers. In experiments, TensorFlow face recognition and object detection are examined by using the TurtleBot2 robot. Results show how objects and faces were detected when the robot is navigating in the previously 2D mapped indoor environment.

Keywords ROS · Ubuntu · Object detection · Face recognition · 3D sensor · LiDAR

1 Introduction

ROS is a versatile software framework that can also be utilized in many other applications than in conventional robotic solutions. As a noteworthy example, a mobile robot equipped with a machine learning software platform can bring added value and purpose of use when connected along with environment mapping and autonomous navigation. In recent novel software frameworks, object detection and face recognition are possible to build on an open-source platform.

Google TensorFlow offers free software libraries mainly supported for machine learning applications. Today, TensorFlow offers a new opportunity for researchers and developers to utilize it in many ROS-based solutions [1–5]. In [6] author presents an extensive ROS toolbox for object detection, tracking and face/action recognition

P. Oksa (✉) · T. Salminen · T. Lipping
Tampere University, Computing Sciences, Pohjoisranta 11, 28100 Pori, Finland
e-mail: petri.oksa@tuni.fi

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_86

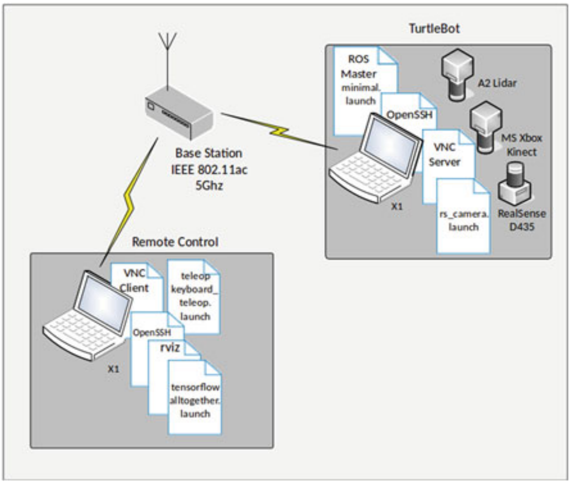


Fig. 1 AutoRobo system overview

with 2D and 3D support enabling the robot to understand the environment. At its simplest way to adapt this, only a web camera and a ROS computer are needed.

The AutoRobo project (Autonomous Robot Ecosystem) consists of an open-source cloud-computing platform, software frameworks and a physical multi-robot environment for the automation or assisting of preprogrammed work processes. Among others, healthcare and hospice operations retain processes which can be assisted by mobile robots. The service architecture model developed in Tampere University, Pori, consists of a novel mobile cloud robotic platform in the support of patient work. The case study presented in this article is implemented in the system environment illustrated in Fig. 1.

Compatibility of software repository packages to a certain system architecture, ROS topic subscribing, Ubuntu-TensorFlow packages and ROS distribution may cause issues. Guidelines and instructions are available in several Q&A forums and public source repositories, for instance, in GitHub/developer pages. Nevertheless, they can propose to be incompatible for the users' ROS distribution and versions of package dependencies [7]. To overcome such difficulties that can emerge in a system environment represented in Fig. 1, guidelines for coding, required software package installations and robot hardware setup are given. The aim of this research was to find solutions for issues that can arise in open-source object-/face recognition and sensors software installation for the ROS robot platform and also find solutions to overcome them.

The rest of the paper is organized as follows. Section 2 provides TurtleBot2 Montado package installation and remarks. Section 3 describes Ubuntu sensors installation for Real Sense and A2 LiDAR. In Sect. 4, TensorFlow package installation guidelines are given. Section 5 presents experiments considering the whole system functionality. Finally, results and conclusions are discussed in Sect. 6.

2 TurtleBot2 Package Installation and Remarks

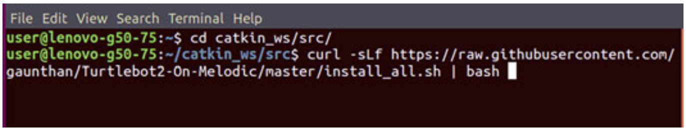
TurtleBot2 Debian installation option is not supported for ROS Melodic on Ubuntu 18.04 distribution. Therefore, the only way to install the TurtleBot2 package is to download a package and build it from the source code. For previous ROS distributions, like Kinetic, the Turtlebot2 package can be installed via Debian packages. In our system environment, ROS Melodic distribution is formerly installed, so a Turtlebot2 package installation follows the source code compiling. The complete ROS Melodic installation instructions can be found on the ROS.org web page at http://wiki.ros.org/melodic/Installation/Ubuntu.

The first thing is to build the TurtleBot2 workspace. In the terminal, the following command directs to a catkin workspace directory.

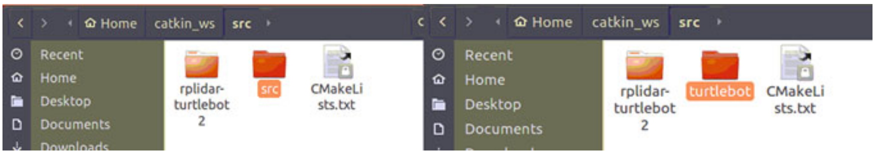
```
$ cd catkin_ws/src
```

After this, the software package should be downloaded from https://github.com/gaunthan/Turtlebot2-On-Melodic. The following command (inside the root of catkin workspace) builds up the running environment for Turtlebot2 [8].

```
$ curl -sLf https://raw.githubusercontent.com/gaunthan/Turtlebot2-On-Melodic/master/install_all.sh | bash
```



After downloading the package, the ‘src’ folder should be renamed to ‘turtlebot’ as depicted below.



The following command (inside the root of catkin workspace) builds up the running environment for Turtlebot2.

```
$ catkin_make
```

In case of the following error, the joystick package should therefore be installed.

```
CMake Error at /opt/ros/melodic/share/catkin/cmake/catkinConfig.cmake:83 (find_p
ackage):
  Could not find a package configuration file provided by "joy" with any of
  the following names:

    joyConfig.cmake
    joy-config.cmake

  Add the installation prefix of "joy" to CMAKE_PREFIX_PATH or set "joy_DIR"
  to a directory containing one of the above files.  If "joy" provides a
  separate development package or SDK, be sure it has been installed.
Call Stack (most recent call first):
  turtlebot/turtlebot/turtlebot_teleop/CMakeLists.txt:5 (find_package)
```

The next two command lines install the missing joystick package.

```
$ sudo apt install ros-melodic-joy
$ catkin_make
```

Connecting a Turtlebot2 to the onboard computer and running the launch file below brings up the robot.

```
$ roslaunch turtlebot_bringup minimal.launch
```

The following error might occur during the installation, but it does not cause any disadvantages and can be ignored in this phase. The robot should be now fully controllable and ready to go.

```
ERROR: cannot launch node of type [laptop_battery_monitor/laptop_battery.py]: la
ptop_battery_monitor
ROS path [0]=/opt/ros/melodic/share/ros
```

If TurtleBot2 is equipped with, e.g., Microsoft Xbox Kinect sensor,*.bashrc* must then configured. The following line should be added to it:

```
$ export TURTLEBOT_3D_SENSOR=Kinect
```

3 Sensors Installation and Issues

This section presents the robot onboard sensors installation used in the experiments. The sensor installation procedure is quite straightforward but can be complicated in the case of the older ROS-compatible robot types such as TurtleBot2 (Kobuki).

3.1 Real Sense D435

In our robotic platform, the Real Sense sensor is dedicated to object detection and face recognition and Kinect for indoor environment mapping and navigation. We start with ROS driver installation. Note that in addition to sensor driver installation, *librealsense* is also needed.

At first, the system must be updated by typing `sudo apt-get update` in a terminal. Then, the software package has to be downloaded by running the following command in terminal.

```
$ sudo apt-get install ros-melodic-realsense2-camera
```

When the download is completed, libraries and keys must be installed by running the next four command lines.

```
$ sudo apt-key adv --keyserver keys.gnupg.net --recv-key
F6E65AC044F831AC80A06380C8B3A55A6F3EFCDE || sudo apt-key
adv --keyserver hkp://keyserver.ubuntu.com:80 --recv-key
F6E65AC044F831AC80A06380C8B3A55A6F3EFCDE
$ sudo add-apt-repository "deb http://realsense-hw-public.s3.ama
zonaws.com/Debian/apt-repo bionic main" -u
$ sudo apt-get install librealsense2-dkms
$ sudo apt-get install librealsense2-utils
```

When the installation is completed and debugged without any errors, the computer should reboot by running sudo reboot in the terminal to make all changes effective. Now the Real Sense sensor should be ready. To test it, the following launch file starts the sensor.

```
$ roslaunch realsense2_camera rs_camera.launch
```

The camera stream coming from the sensor can be selected by opening it in its own *rqt* window. We use the next command line operation to open it.

```
$ rosrun rqt_image_view rqt_image_view
```

When choosing the *image_raw* from the drop-down menu, a camera image should come into view as shown in Fig. 2.

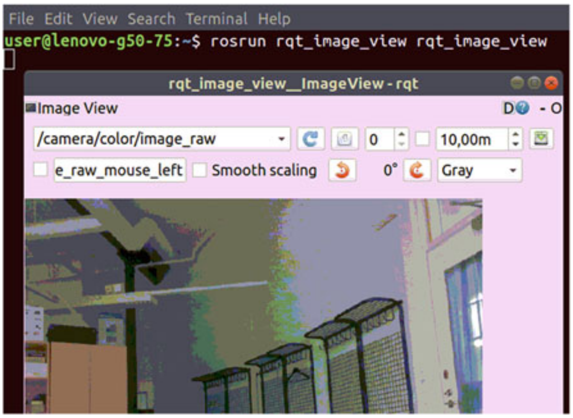


Fig. 2 Camera stream from real sense

3.2 A2 LiDAR

An additional sensor to accurate TurtleBot2 SLAM (Simultaneous Localization and Mapping) *gmapping* in the experiments is A2 LiDAR [9]. LiDAR improves the environment mapping by scanning the environment 360 degrees besides the other sensors. It is placed on top of a TurtleBot2; see Fig. 5 for sensor fittings.

Slamtec A2 drivers are available at http://wiki.ros.org/rplidar and https://github.com/roboticslab-fr/rplidar-turtlebot2. In the terminal, the command `cd ~/catkin_ws/src` opens the directory into the catkin workspace. The following command downloads the package.

```
$ git clone https://github.com/roboticslab-fr/rplidar-turtlebot2.git
```

After the download is completed, the following commands build the package:

```
$ cd ~/catkin_ws
$ catkin_make
```

For ensuring complete build, update is recommended.

```
$ cd src/rplidar-turtlebot2
$ git pull
$ catkin_make
```

Next, USB settings should be changed.

```
$ cd src/rplidar-turtlebot2/rplidar_ros/scripts
$ chmod +x create_udev_rules.sh
$ ./create_udev_rules.sh
```

After USB configurations, the *setup.bash* source is run by the command `source devel/setup.bash`. Now the sensor should be ready. To test it, the following launch file starts the sensor:

```
$ roslaunch rplidar_ros view_rplidar.launch
```

The resulting ROS visualization (*rviz*) view should look something as shown in Fig. 3. If the *roslaunch* does not work properly, USB unplug/plug-in helps to wake up the sensor.

4 TensorFlow Package Installation

We install TensorFlow version 1.14 as it is compatible with our laptop software/hardware architecture and Ubuntu distribution. Graphics Processing Unit (GPU) requirements must also be taken into account. TensorFlow uses Compute Unified Device Architecture (CUDA) which means that NVIDIA GPUs are

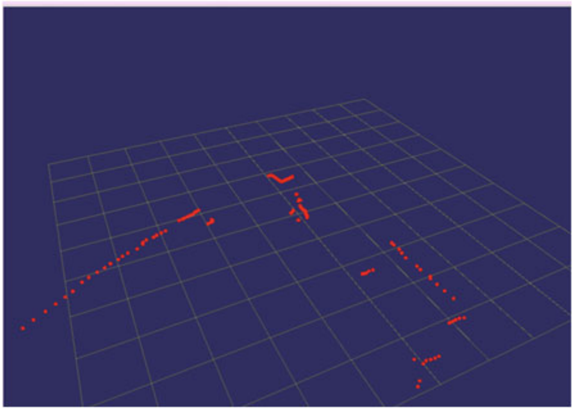


Fig. 3 A2 LiDAR view in rviz

supported. In case of GPU incompatibility, TensorFlow can be used with Central Processing Unit (CPU) support only as used in our installation method.

Preliminaries prior to TensorFlow package installation are that both ROS Melodic and camera (Real Sense) are successfully installed. Note that TensorFlow should be installed before the ROS/TensorFlow recognition package installation. ROS/TensorFlow package is downloadable at https://github.com/cagbal/ros_people_object_detection_tensorlow. Then, system requirements should be checked via TensorFlow (Ubuntu) webpage: https://www.tensorflow.org/install/pip.

At first, we start with pip (Python Package Installer) and python-dev installations.

```
$ sudo apt update
$ sudo apt install python-dev python-pip
```

then, TensorFlow and TensorFlow hub installation.

```
$ pip install tensorflow==1.14
$ pip install tensorflow-hub==0.7.0
```

After the installations above, it is recommended to debug the code by running it in a Python shell.

```
$ python
$ import tensorflow as tf
$ import tensorflow_hub as hub
```

If any error messages do not show up in shell, the installation had gone successfully. As we aim to use ROS and TurtleBot2 robot for object detection and face recognition, TensorFlow Application Programming Interface (API) for ROS is needed. To obtain that, we use the repository available at https://github.com/cagbal/ros_people_object_detection_tensorflow. This repository uses a number of open-source projects to work properly:

- [Tensorflow]
- [Tensorflow-Object Detection API]

- [Tensorflow Hub]
- [ROS]
- [Numpy]
- [face_recognition] https://github.com/ageitgey/face_recognition
- [dlib]
- [cob_perception_common] https://github.com/ipa-rmb/cob_perception_common.git
- [protobuf]

For Tracker part:

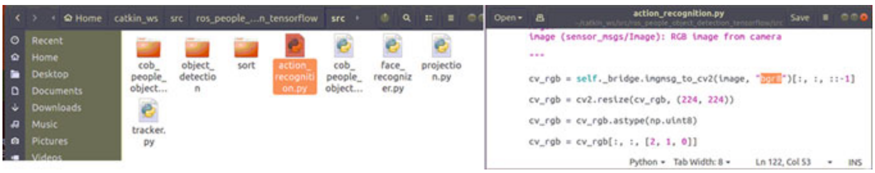
- scikit-learn
- scikit-image
- FilterPy

Firstly, it is important to notice that the repository [cob_perception_common] compatibility depends on ROS distribution. For ROS Melodic, it is *ipa320* as shown in the following repository cloning. Next, TensorFlow should be installed on the system by running the following commands one by one in the terminal.

```
$ cd catkin_ws/src
$ git clone --recursive https://github.com/cagbal/ros_people_obj
ect_detection_tensorflow.git
$ git clone https://github.com/ipa320/cob_perception_common.git
$ cd ros_people_object_detection_tensorflow/src
$ protoc object_detection/protos/*.proto --python_out=.
$ cd ~/catkin_ws
$ rosdep install --from-path src/ -y -i
$ catkin_make
$ pip install face_recognition
```

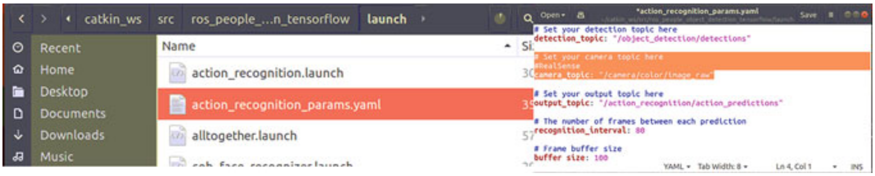
ROS TensorFlow API package is now successfully downloaded and installed. The next phase is to change directories so that they match the destination system directory structure. Python files (*.py*) in directory path `/catkin_ws/src/ros_people_object_detection_tensorflow/src` should be changed as follows:

`action_recognition.py` [replace"passthrough" to "bgr8″ and save].

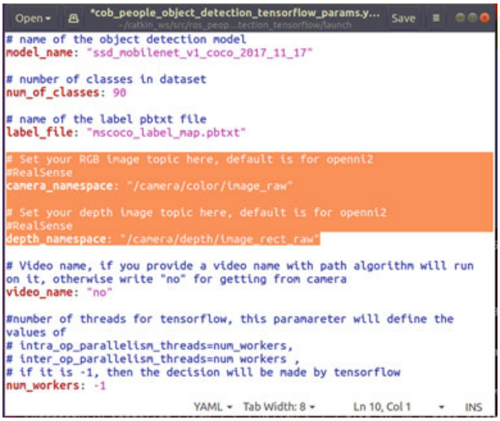


Same filename change should be done to `cob_people_object_detection_tensorflow.py`, `face_recognizer.py` and `projection.py`. Then *.yaml* files in directory path `/catkin_ws/src/ros_people_object_detection_tensorflow/launch` should be changed as follows:

`action_recognition_params.yaml` [change"camera_topic" to"Real Sense" and save].



Same camera_topic change should also be done to `cob_face_recognizer_params.yaml`, `projection_params.yaml` and `cob_people_object_detection_tensorflow_params.yaml`. In the latter.*yaml* file, *rgb* should be used and the *depth_image* set to topics as follows.



5 Experiments

In this section, the entire system is experimented and analyzed. All the experiments in this article were done on a Lenovo Thinkpad Carbon X1 laptop detailed in Table 1. Both onboard and remote computers are similar laptop computers.

Table 1 Onboard computer used in experiments

Memory	Processor	Graphics	OS architecture	Hard disk	Ubuntu distribution
15,2 GiB	Intel® CoreTM i5-835OU CPU @ 1.70 GHz × 8	Intel® UHD Graphics 620 (KBL GT2)	64-bit	503,0 GB	Ubuntu 18.04.5 LTS

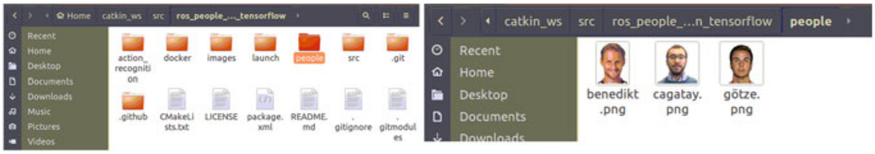


Fig. 4 Face recognition database

5.1 *Face Recognition*

When applying face recognition, only face images in "people" folder are recognized. Thus, this folder should include all the face images that are going to be recognized. All the images should be in.*png* format and file named according to the person as shown in Fig. 4.

Adding new images into this folder database is simple. To make recognition as accurate as possible, the white background of the image gives the best results for the recognition. In this way, framing outline between the faces and background are more accurate reducing the delays in recognition.

5.2 *Testing the System*

In order to test the system functionality, the following preliminaries should be done. In the first, we bring up the Real Sense sensor by running the following launch file:

```
$ roslaunch realsense2_camera rs_camera.launch
```

By opening the second terminal, we set up the recognition. There are options to start all at once or separately in their own terminals for object detection and face recognition. To launch everything, run the following launch file in the terminal.

```
$ roslaunch cob_people_object_detection_tensorflow allto-
gether.launch
```

Launching only object detection.

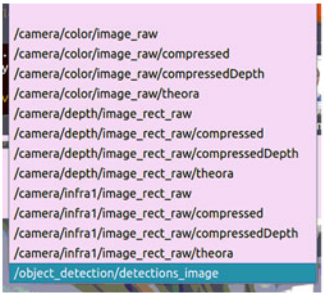
```
$ roslaunch cob_people_object_detection_tensorflow
cob_people_object_detection_tensorflow.launch
```

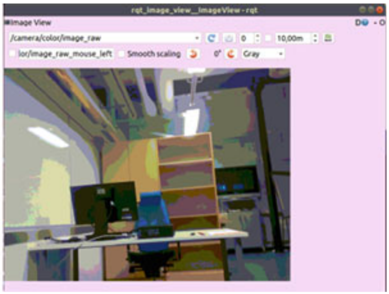
Launching only face recognizer.

By opening rqt_image_view, it is possible to visually monitor and follow all recognition operations.

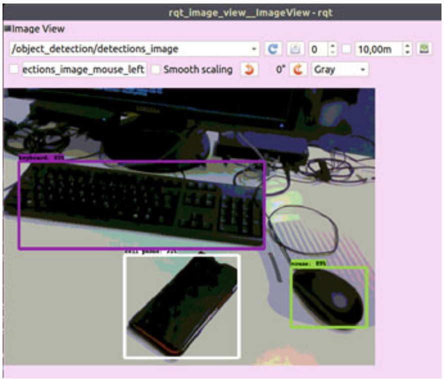
```
$ rosrun rqt_image_view rqt_image_view
```

After running the *rqt_image_view* command, a new window opens showing a video stream from a Real Sense camera.





When choosing "/object_detection/detections_image" instead of "/camera/color/image_raw" from the left upper corner menu bar, all recognitions should appear in the view as shown below.



6 Results and Discussion

The aim of this research was to find solutions for issues that can arise in TensorFlow open-source object-/face recognition and sensors software installation for the ROS robot platform. To prove these issues, experiments conducted with the TurtleBot2 robot were carried out. Practical and detailed solutions for each installation issue were given step by step. These instructions are kept hands-on type answering to the most crucial case-driven issues when setting up ROS and TensorFlow communication. Available source codes and software packages are utilized, such as several ROS packages from GitHub public repository, reducing the complexity of further programming.

On its whole, the entire system is proved functional even though getting all hardware working seamlessly with the ROS platform remains quite complex. Figure 5 shows the TurtleBot2 physical structure after sensors are fitted on the robot. Kinect sensor is located on the middle shelf; Real Sense and LiDAR are on the top shelf.

Fig. 5 Robot platform

In the following, Fig. 6 left-hand side upper corner is the window of object detection, right-hand side upper corner shows the face recognition window and on the lower right-hand side is the video stream from Kinect. The indoor environment shown in the figure is previously mapped by using TurtleBot2's *gmapping* algorithm.

Figure 7 shows the CPU load when `bringup minimal.launch`, `rviz_launchers view_navigation.launch`, `teleop keyboard_teleop.launch`, `rs_camera.launch` and `tensorflow alltogether.launch` are all launched together.

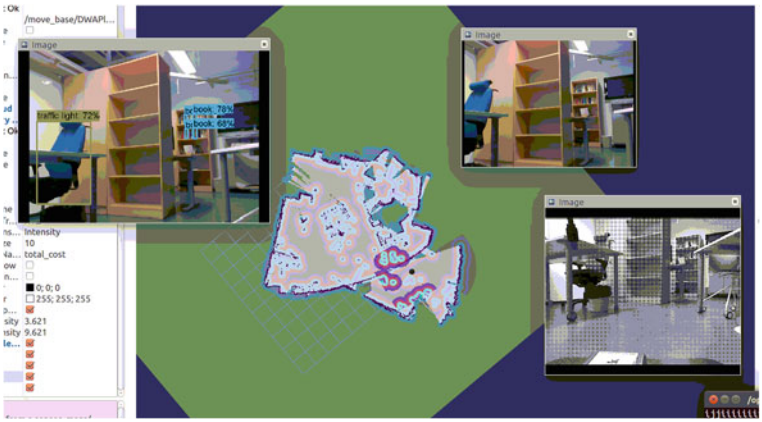


Fig. 6 Recognition

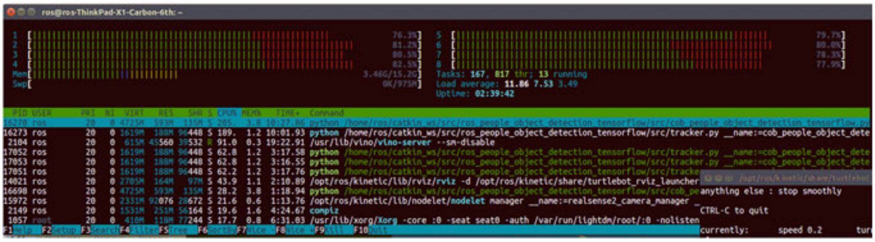


Fig. 7 CPU load

Percentages 205 and 189% in Fig. 7 are both TensorFlow programs causing high CPU load. To solve the aforementioned problem, the main solutions would be.

- To extend computation capacity of onboard computer (especially RAM disk space);
- To make sure that all images are in applicable image format;
- Same outlines and white background color in all images located in people folder.

The motivation for using TensorFlow with ROS is the need for object detections and face recognition implemented in ROS-based moving robots. The primary goal is an open-source robotic platform for all users that have the interest to develop the proposed system platform further. In the near future, we continue by combining both Kinect and A2 LiDAR topics into a one *rviz* window producing a better *gmapping* outcome. While still experimental, the entire system environment is fully functional, as demonstrated by the previous experiments.

References

1. Kouzehgar M, Tamilselvam YK, Heredia MV, Elara MR (2019) Self-reconfigurable façade-cleaning robot equipped with deep-learning-based crack detection based on convolutional neural networks. Autom Constr 108
2. Tongloy T, Chuwongin S, Jaksukam K, Chousangsuntorn C, Boonsang S (2017) Asynchro-nous deep reinforcement learning for the mobile robot navigation with supervised auxiliary tasks. In: IEEE 2017 2nd international conference on robotics and automation engineering (ICRAE), pp 68–72
3. Millan-Romera JA, Perez-Leon H, Castillejo-Calle A, Maza I, Ollero A (2019) ROS-MAGNA, a ROS-based framework for the definition and management of multi-UAS coop-erative missions. In: IEEE 2019 international conference on unmanned aircraft systems (ICUAS), pp 1477–1486
4. Reddy PP (2019) Driverless car: software modelling and design using Python and Tensorflow, No. 1446, EasyChair
5. Gaifullin R, Ivanou M, Gazizov R (2019) Natural Human-Robot Interaction Toolkit. In: International conference on human interaction and emerging technologies. Springer, Cham, pp 196–200
6. ROS People Object Detection & Action Recognition Tensorflow, https://github.com/cagbal/ros_people_object_detection_tensorflow, last accessed 2020/10/30

7. Cervera E, Del Pobil AP (2019) Roslab: sharing ros code interactively with docker and jupyterlab. IEEE Robot Autom Mag 26(3):64–69
8. Make Your TurtleBot2 run on ROS Melodic (Ubuntu 18.04), https://github.com/gaunthan/Turtlebot2-On-Melodic, last accessed 2020/10/30
9. Akhmetzyanov A, Yagfarov R, Gafurov S, Ostanin M, Klimchik A (2019) Exploration of Underinvestigated Indoor Environment Based on Mobile Robot and Mixed Reality. In In-ternational Conference on Human Interaction and Emerging Technologies. Springer, Cham, pp 317–322

Increase the Quality of Treatment with Medical Apps Through Remote Compliance Testing

Janina Sauer, Alexander Muenzberg, Franz Reisewitz, Andreas Hein, and Norbert Roesch

Abstract Medical Apps are increasingly gaining in importance. But comprehensive testing of the app is difficult and there is no single method. For example, compliance has been ignored so far in most cases. A general remote method has been developed to test the compliance behavior when using medical apps, as well as to take them into account in later treatment. With the help of timestamps at appropriate locations, frequency, times and periods of use can be measured and (automatically) compared with predefined values. Furthermore, this method provides information about the efficiency and the learning phase of the users, as well as the usability, which is of great importance for the evaluation of compliance. With the help of different smartphone sensors, it is possible to evaluate the data quality. By taking other factors into consideration, such as the quality of experience and external circumstances, as well as the determination of data quality, the medical app can be evaluated comprehensively. This allows user profiles to be created . This resulting information can be entered into the treatment and the quality of the treatment can be increased .

Keywords Medical App · Remote testing · Compliance · Adherence · Usability · Data quality · SUDS · Stress measurement

J. Sauer (✉) · A. Muenzberg · A. Hein
Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
e-mail: janina.sauer@uni-oldenburg.de; janina.sauer@hs-kl.de

A. Muenzberg
e-mail: alexander.muenzberg@uni-oldenburg.de

A. Hein
e-mail: andreas.hein@uni-oldenburg.de

J. Sauer · A. Muenzberg · F. Reisewitz · N. Roesch
University of Applied Sciences Kaiserslautern, Kaiserslautern, Germany
e-mail: franz.reisewitz@hs-kl.de

N. Roesch
e-mail: norbert.roesch@hs-kl.de

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_87

1 Introduction

1.1 Medical Apps and the Importance of Patient Compliance

Smartphone applications in the medical context (medical apps) are becoming more important. They enable health professionals and patients to stay in closer contact regardless of location. Furthermore, medical apps inform the patient and give him or her a comprehensive picture of the illness and/or treatment, thus increasing the patient's self-administration. This alone may already have an effect on compliance [1].

An accurate assessment of compliance behavior is necessary for effective and efficient treatment planning and to ensure that changes in health outcomes can be traced back to the recommended scheme, especially in the case of chronic diseases.

1.2 Measurement of Compliance

An ideal measurement approach would be unobtrusive, unnoticed observation, which is difficult to implement under real-life conditions [2].

One of the measurement approaches is to survey medical staff and patients about their subjective assessments of compliance behavior. This survey is prone to error in many respects. Objective strategies also lead to drawbacks in assessing compliance.

Each disease and treatment has its own compliance challenges and thus its own nature and complexity of measurability [3].

2 Methods

2.1 Test and Measure Compliance Remotely

For most apps, it is advisable to use individual events, timestamps and counters. These events are often buttons that are used to initiate or terminate functions in the app, for example, to save a data entry [4].

Timestamps and counters can be very easily integrated into the existing program code with little additional effort. By remote testing, the mobile application can be evaluated in everyday situations without the test subjects having to come to a usability laboratory. All functions are naturally tested for their usability in real-life situations [5].

This method can be adapted to measure compliance. Thus, it shows when the patient uses the app, when and which entries are made and how often. It can be compared whether these times, periods and numbers correspond to expectations.

If necessary, it is possible to influence the use and thus the treatment by directly contacting the user.

Furthermore, this also offers the possibility to compare the number of entries with the previously defined values. If these values are below the defined threshold value, a reminder can be set automatically.

The counter of the start event and the counter of the end event should eventually have the same value. Thus, every input that was started was completed successfully. If again the start counter is higher than the save counter, this is a clear sign that the input was started often but not successfully completed for unknown reasons.

In order to be able to track patient activity even more precisely, the time required to make an input can also be checked by timestamping each start and end points of an action. In this way, the duration of an input can be checked and compared with target values.

If the entry is permanently incomplete, a reminder can be set at the appropriate time. Permanent reminders often lead to the fact that they are ignored by the user and the use of the app is perceived as annoying.

The data already entered is subjected to a plausibility check, which in this case means that the values are within a usual range.

After a longer period of use in line with expectations, the inputs should show a success. This depends on the reason to use the app in the first place.

The remote compliance measurement method is useful for testing the app before it is released to the market. This way it can be tested and evaluated in everyday situations and if necessary, the app can be optimized.

On the other hand, the compliance can be checked permanently, even after the app has been launched. This way, the supervising health professional has an overview at all times whether the patient is using the app as expected and can intervene and adjust the treatment if necessary. From the point of view of long-term compliance, possible learning curves and habits can be identified, after which the treatment and the associated use can also be adjusted. However, the Quality of Experience must be considered in any case.

2.2 Data Quality of Remote Compliance Measurement

A disadvantage of remote testing is the local separation of the test leader and test subject. When using a smartphone, observation by the front camera to interpret facial expressions makes little sense and data protection is complex. A way must therefore be found to ensure that the data quality is adequate.

By measuring patient activity, statements can be made about data quality. If the entries are made faster than usual (compared to the previously defined standard value of the individual patient), this suggests a hectic, stressful situation, which leads to errors.

Biofactors are another point of measurement for data quality. Common biofeedback point for stress is the pulse [6]. The increase in pulse in stressful situations can be measured via the built-in smartphone camera. If the pulse is elevated from the beginning, the entry is made in stress.

The built-in accelerometer measures the smartphone's movement. Thus, it can be determined whether the input happened in rest or motion. This also indicates the stress level of the patient.

Whether a measurement of the pulse, acceleration or both is appropriate depends strongly on the app and what is considered good compliance behavior in the specific case. Since the measurement of movement is inconspicuous, this method is suitable in most cases.

2.3 Compliance in the Context of Usability

The usability or user experience has a great influence on the evaluation of compliance and vice versa. A good usability leads to a higher compliance [7]. Therefore, a comprehensive measurement of compliance also includes the measurement of usability. There are various standards for this, in the form of user surveys, e.g., the SUS [8], the usability lab or also in a remote variant using timestamps [5].

With the remote variant, efficiency (time-on-task, clicks, lostness) and learnability (happy path, learning curve, complexity of the action) can also be measured. The quantity and quality of the findings of the two methods Usability Laboratory (state of the art) and remote testing are fully comparable [5].

By means of different case studies, it was shown how user needs are revealed and thus a better usability for the user is created, which leads to an increase in compliance and finally to an increase in the quality of treatment [9].

2.4 Compliance Profiling

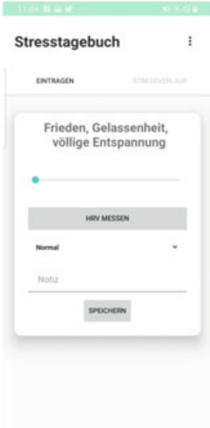
Users can be categorized according to various factors. The most common are positive and negative behavior and positive and negative attitude. Attitude describes the user's attitude toward the type of treatment, including the attitude of the user toward the technology used in general [10].

By collecting and evaluating the remote (test-) data, users can deduce profiles that confirm or contradict the impressions of the patient and the health professional. In any case, the compliance profile is extended in this way.

3 Results

The described remote test method for medical apps will be tested in a study. This test concept, which was developed specially for mobile, medical applications, has

Screenshot 1 Input entry



already been published in parts [4, 5]. The mobile app to be tested is a stress diary app developed especially for this purpose.

All students and employees of the University of Applied Sciences Kaiserslautern, Germany (UASK) are eligible as test persons. The study started in December 2020.

3.1 Measuring Stress

Subjective stress levels are entered via simple sliders of test persons (see Screenshot 1). Optional comments are possible for each entry.

To measure the stress level of the tester, the Subjective Units of Distress Scale (SUDS) is used. In this measurement, the subjects themselves indicate their perceived stress on a scale of 0–10 [11]. Entries can be made as often as desired and reminders as push notes are possible at an individual time.

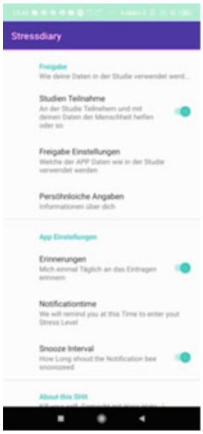
Stress curve and entries can be viewed, deleted and edited. It is also optionally possible to measure the pulse via the camera (see Screenshot 2).

3.2 Data Protection

The data protection officer checked the app and its privacy. This is followed by the creation of a data protection declaration, which the potential testers receive in the information mail for testing the app and within the app itself. The following data protection measures have already been taken (see Screenshot 3):

– Use of the app is possible without participation in the test phase.
– The app can be used without sharing sensor data.
– The app can be used without passing on the pulse measurement.

Screenshot 2 View of the entries



Screenshot 3 Enabling the sensor data

– Use of the app without passing on the notes possible.
– The UASK is responsible for the data acquisition.
– Anonymization of voluntary, personal information.

Each user is given a unique identification number (Unix Time Stamp) when they first open the app, so they can act completely anonymously. Optionally, additional personal information is possible, e.g., Age, Student/Employees, Course of studies, Semester and Department.

3.3 *Purpose of Data Collection*

"How stressful is my university?" Many students experience academic stress at universities, e.g., due to coursework and financial. This can lead to sleep disorders, eating disorders and burnout, among others [12–16]. The special situation of the COVID-19 pandemic has further unknown influences on the students and employees of the UASK. A special factor in this respect is stress. In order to get an overview of the members of UASK, it should be measured as simply as possible without causing additional stress through use. The simple structure of the app makes it easy to maintain and update. Thus there is the possibility to repeat the study in order to compare the results during the COVID-19 pandemic. However, the duration of this study phase is initially limited to one year. The study phase is initially limited to one year.

Thus, among others, the following periods are covered and would be available for comparison: Lecture period with digital courses, Possible lecture period with attendance events, Holidays, Lecture free time, Examination phase preparation, Testing phase and Semester break.

The knowledge gained can be incorporated into the further design of the studies and/or work at the university.

3.4 *Evaluation Concept of Usability and Compliance*

The remote evaluation concept for the stress diary is divided into the following parts:

3.4.1 Clickstream

Several buttons have been timestamped and these are sorted by the Happy Path. The Happy Path also contains timespans. Thus, actions of the test persons can be compared with the desired procedure of the developers. The entire behavior of the test person within the app can be displayed.

3.4.2 Smartphone Sensors

During use, the accelerometer of the smartphone records the motion data. The app can be used without sending the sensor data. Thus, acceptance can be assessed in this regard at the end of the evaluation.

3.4.3 Biofeedback

The determination of the pulse can also provide information about the data quality. However, an increased pulse can also be the normal condition of a test person. So this information is only useful and can only be evaluated if the pulse has been measured several times. This measurement is not mandatory.

3.4.4 Compliance Profiles

One entry should be made every day. It is expected that a test person will use the app for at least one week. If these two factors are met, the user is considered compliant.

Therefore, the compliance profiles were adapted to the specific requirements of this app (see Table 1).

This quantification of user compliance can be used in the later evaluation of the data and for comparability with the data obtained. It is possible that the daily number of entries will not remain the same. Therefore, it must be averaged over the duration of use. In addition to the duration of use and frequency, additional factors are recorded:

- Are the entries always made at the same time?
- Is the app opened via the push notification or are entries made with a reminder?
- Does the daily usage frequency of working days change compared to the weekend?
- Does the number of daily entries fluctuate strongly?

Table 1 Special compliance profiles

Duration of use		Frequency of use		Compliance profile	
Under 1 week	−1	Once daily	1	Sample user	0
Under 1 week	−1	2–5 times daily	2	Intensive sample user	1
1–2 weeks	1	Daily	1	Short-term user	2
Under 1 week	−1	Over 5 entries daily	3	Very intensive sample user	2
1–2 weeks	1	2–5 times daily	2	Sample intensive users	3
Over 2 weeks	2	Once daily	1	Long-term user	3
1–2 weeks	1	2–5 times daily	2	Short-term intensive users	3
Over 2 weeks	2	2–5 times daily	2	Long-term intensive users	4
1–2 weeks	1	Over 5 entries daily	3	Short-term high intensive user	4
Over 2 weeks	2	Over 5 entries daily	3	Long-term high intensive user	5

3.4.5 Problems

The following problems were encountered in the conceptual design of the study:

Selection Bias: The available testers are UASK members. The majority of the students are of typical age and have grown up with a smartphone.

Information Bias: There is no way to verify the correctness of the data entered. Not only deliberate wrong entries but also unconscious wrong entries are possible.

4 Conclusion and Outlook

It can be assumed that medical apps will gain in importance over the next few years; patient compliance will also become more important. So, compliance should be included in the testing.

The described remote testing method is easy to implement and individually adaptable to the app Therefore, the compliance profiles were adapted to the specific requirements of this app.

The usability and compliance are strongly correlated. Therefore, the usability should also be tested.

Furthermore, the described method can be used after the product has been placed on the market to evaluate the long-term compliance of the user. For this purpose, a point system is to be developed in further research in order to simplify the measurement of compliance and to be able to display it simply and directly to the practitioner so that no additional expense is incurred.

With the help of different measuring point statements about the data quality can be made. An increase in pulse, as well as hectic movement of the smartphone and the comparison of the input duration with the expected values, can determine a stress level and thus put the data input into a context.

Future research will evaluate whether the determination of user profiles can also be done remotely. It should also be investigated what consequences it has if the user is categorized into a certain user group.

The described remote test concept is to be comprehensively tested in a study at the University of Applied Sciences Kaiserslautern using a stress diary app. The results of this study will be incorporated into the concept and refined.

References

1. Albrecht U-V (2016) Chances and risks of mobile health apps (CHARISMHA). Medizinische Hochschule Hannover, Hannover. urn:nbn:de:gbv:084-16040811153
2. Vermeire E, Hearnshaw H, Van Royan P, Denekens J (2001) Patient adherence to treatment: three decades of research. A comprehensive. J Clin Pharm Ther 26:331–342. https://doi.org/10.1046/j.1365-2710.2001.00363.x
3. World Health Organization (2003) Adherence to long-term therapies—evidence for action

4. Sauer J, Muenzberg A, Hein, Roesch N (2019) Simplify testing of mobile medical applications by using timestamps for remote, automated evaluation. In: 2019 international conference on wireless and mobile computing, networking and communications (WiMob), pp 203–206. https://doi.org/10.1109/WiMOB.2019.8923241
5. Sauer J, Muenzberg A, Siewert L, Hein, Roesch N (2020) Remote testing of usability in medical apps. In: 8th EAI international conference, MobiHealth 2019, Dublin, Ireland, 14–15 Nov 2019, Proceedings, pp 3–17. ISSN 1867-8211
6. Tullis T, Albert B (2008) Measuring the user experience. Elsevier, Burlington, United States, pp 87–97/183–189
7. Patwardhan M (2016) Assessing the impact of usability design features of an mHealth App on clinical protocol compliance using a mixed methods approach. Arizona State University, Tempe
8. Brooke (1986) SUS—a quick and dirty usability scale, United Kingdom
9. McCurdie T, Taneva S, Casselman M, Yeung M, McDaniel C, Ho W, Cafazzo J (2012) mHealth consumer apps: the case for user-centered design. Biomed Instrum Technol 46(s2):49–56. https://doi.org/10.2345/0899-8205-46.s2.49
10. Hogan TP, Awad AG, Eastwood R (1983) A self-report scale predictive of drug compliance in schizophrenics: reliability and discriminative validity. Psychol Med. https://doi.org/10.1017/s0033291700050182
11. Wolpe J (1969) The practice of behavior therapy. Pergamon Press, New York. ISBN 0080065635
12. Stallman H (2010) Psychological distress in university students: a comparison with general population data. Aust Psychol 45(4):286–294
13. Debeuf T, Verbeken S, Van Beveren M-L, Michels N, Braet C (2018) Stress and eating behaviour: a daily diary study in youngsters. Front Psychol 9(2657)
14. Paula Amaral A, João Soares M, Margarida Pinto A, Telma Pereira A, Madeira N, Carvalho Bos S, Marques M, Roque C, Macedo A (2018) Sleep difficulties in college students: the role of stress, affect and cognitive processes. Psychiatry Res 260:331–337
15. Vaez M, Laflamme L (2008) Experienced stress, psychological symptoms, self-rated health and academic achievement: a longitudinal student of Swedish university students. Soc Behav Pers 36(2):183–196
16. Patel V, Flisher A, Hetrick S, McGorry P (2007) Mental health of young people: a global public-health challenge. Lancet 369(9569):1302–1313

Linking the Linguistic Resources Using Graph Structure for Multilingual Sentiment Analysis

Mohamed Raouf Kanfoud and Abdelkrim Bouramoul

Abstract Sentiment Analysis aims to analyse a large amount of data to infer and deduce the different opinions expressed therein. It would be interesting to link several linguistic resources and tools to combine their advantages and obtain a result as close as possible to reality. Moreover, creating a new lexicon either manually or automatically is time-consuming and needs considerable efforts. This paper proposes a new approach to build a polar lexicon by using the existing resources to generate a rich and time-saving one. Nevertheless, not all linguistic resources have the same structure. It is then challenging to use them together at once. We propose to map the resources into the graph structure, where the nodes represent the words, and the edges represent the relations between the nodes, thus the words. The obtained results are promising, F1-measure was 0.82 for the English language, and accuracy ranged from 0.76 to 0.98 for all other languages used in the experiment.

Keywords Sentiment analysis · Linguistic resources · Relationship · Multilingual · Lexicon-based approach · Graph-based

1 Introduction

Sentiment Analysis (SA) is a task that aims to determine the polarity of what people say about the products, persons, events, or anything else. The reviews are published on different websites and social media networks such as commercial websites, movie review websites, Facebook, and Twitter. One of the approaches used in SA is the lexicon-based approach. It aims to get the word's polarity either positive or negative. The words are attributed a subjectivity scores ranging from -5 to $+5$, for example. Positive words express a desirable state. Meanwhile, negative words express

M. R. Kanfoud (✉) · A. Bouramoul
MISC Laboratory, Constantine 2 University, Constantine, Algeria
e-mail: raouf.kanfoud@univ-constantine2.dz

A. Bouramoul
e-mail: abdelkrim.bouramoul@univ-constantine2.dz

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_88

an undesirable state. The creation of these lexica is not easy; it is a laborious task. Besides, there is not a lexicon for each language. In addition, lexicon-based methods depend on the availability of the lexicon for the language in question, and having a lexicon that covers all words in the language is almost impossible. Another faced problem is that the linguistic resources do not follow the same structure and then are mutually incompatible. The contribution of this paper consists of two parts: first, the creation of multilingual lexicon by linking different linguistic resources in several languages; second, the proposition of a new process to find the word in the lexicon, which is based on Dijkstra's algorithm [7].

The remainder of the paper is organized as follows: Sect. 2 presents related work. Section 3 describes the proposed approach along with foundations and the process. Then, the evaluation and the results are described in Sect. 4. Meanwhile, Sect. 5 sums up the remarks and observations over the obtained results. Finally, Sect. 6 provides some conclusions and future work.

2 Related Work

The creation of a lexicon automatically started earlier and has been done for multiple languages. One of the earliest works is [11]; the authors have used a seed set of 1336 adjectives hand-labelled (657 positive, 679 negative), and to enlarge the list they have used the intuition if two adjectives are conjoined by the word "and" they probably have the same polarity, and if they are conjoined by the word "but" it implies that they have different polarity. In [13], another process of creation has been presented. It is based on synonyms in WordNet to measure the semantic orientation of adjectives; using this relation between the synsets aids to propagate the list of polar words. Meanwhile, [8] describes creating the first version of SentiWordNet (SWN), also using WordNet as a base resource. They get the orientation (PosScore and NegScore) of words in the function of the gloss of synsets. Furthermore, the authors published the version 3.0 which is an enhanced version of SWN. The SWN3.0 was released with WordNet version 2.0 [1]. SentiWordNet is constructed to cover English words because it is based on WordNet. Other researchers prefer to create new lexicons adapted to their purposes or translate existing lexica to the languages used in their work. In [4], the authors presented an open multilingual WordNet, combining the WordNet with open licences, data from Wiktionary, and the Uni-code Common Locale Data Repository. The authors in [2] have based on linking the four resources: EWN, AWN, ESWN, and SAMA, to create a large-scale Arabic sentiment lexicon for Arabic opinion mining. Researchers in the field widely use the sentiment analysis approaches based on the lexicon [17]. They adopted a lexicon-based approach to get the word's subjectivity in the text, besides applying some adjustments such as the intensifiers and shifters to calculate the final score of word's subjectivity. The authors of [5] investigate the keyword extraction based on the graph structure. The graph is constructed by a textual representation where the nodes represent the words and the edges represent the co-occurrence of those words. After that, the authors applied the degree centrality measure to select the top five hundred words to treat them as the final keywords used to determine the polarity of the review pattern. Another work [18] proposed semi-supervised learning for text classification using a graph representation along with label propagation. The authors claim that using a graph representation can overcome the shortcomings of the bag-of-words representation, especially for short documents.

Accordingly, we propose to develop an approach to exploit all previously created resources (Lexica, Dictionaries, Thesauri) to avoid creating another new one for each research and bypass the translation pre-process for the word where no polar lexicon exists. Besides, our goal is to analyse texts in different languages without detecting the language of the word before getting its polarity from an adequate lexicon.

3 The Proposed Approach

The proposed approach, namely `GraLexi` (**GRA**ph **LEXI**con), is based on two processes. The first one is the generation of the multilingual lexicon by linking and mapping different existing linguistic resources. The second is the process of finding the word in the previously generated lexicon.

3.1 GraLexi Foundations

It is important to mention that we have adopted the structure of the graphs [9] to unify and represent the different linguistic resources on the proposed approach, besides being benefitted from the mathematical features on which the graphs are based. In the following, we define the different components and vocabulary used in `GraLexi`. Firstly, we have three types of nodes: (i) **Node** represents the words (words are extracted from the resources). (ii) **Polar node** comprises the polarity (e.g., positive/negative). (iii) **Link node** links the sub-graphs. To clarify, we call Labelled Node the node directly connected with a Polar Node, otherwise, we call it Unlabelled Node. The edges (arcs) represent the relation between the nodes. In the proposed approach, we distinguish two types of relations: (i) **Direct relation** keeps the same polarity between the connected nodes. (ii) **Indirect relation** inverses the polarity of the previous word. Note that we used "words" and "terms" interchangeably.

The new added component to the graph in our contribution is called the **Token**. The Token is a mobile component in the graph representing the input word (entered by the user) for which we want to find the polarity. It will get through the graph, node by node, to find a similar node. We distinguish two types of Tokens. The first one is called the *original Token* (the term at input, i.e., the word extracted from the text we search its polarity). The second one is called the *sibling Token*, generated by the Workers (c.f. Generation step in Sect. 3.2). The sibling Tokens are generated in order to increase the chance to find at least one of the Tokens (original or siblings).

Obviously, the probability of finding m elements with m $>$ 1 in a set of n objects is higher than finding one element in n objects $\frac{m}{n} > \frac{1}{n}$.

3.2 GraLexi Process

In this subsection, we develop the `GraLexi` process, from the generation of the graph (transform the linguistic resource to the graph structure) until obtaining the word's polarity. The proposed process illustrated in Fig. 1 is composed of three steps.

1. **Mapping** generates the graph from linguistic resources. With this intention, we extract the terms and relate them with a type of relationship defined in the linguistic resource in question such as the polarity for lexicon, translation found in dictionary, and synonym and antonym in thesaurus.
2. **Generation** generates the sibling Tokens from the original Token by means of the Workers; see Fig. 1 step 2. The Worker is the modifier of the original Token. Hence the Token generated by the Worker is called sibling Token. Also Fig. 1 step 2 shows the relations between the original Token, the Workers, and the sibling Tokens, in addition to Worker's level.
3. **Finding** finds the term's polarity. It is the essential step in the proposed approach, in which we apply the search algorithm, in purpose to find the similar nodes to the sibling Tokens in the graph. Therefore, it calculates the polarity of the original Token. The polarity calculation relies on the path (or sub-path) that connects the labelled node with the found node (node has the same value of the sibling Token); see Fig. 2.

The general formula to calculate the final polarity is introduced in (1).

$$Polarity = P_k \times \prod_{j=k}^{n-1} t_j \qquad (1)$$

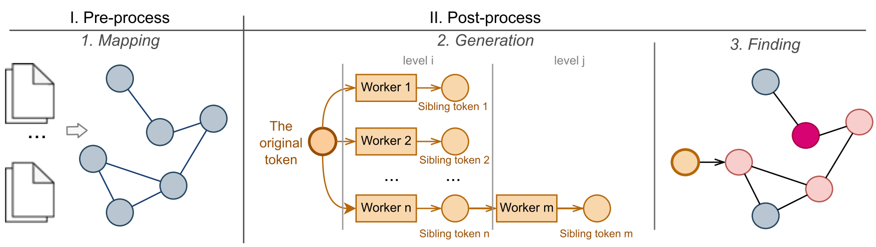


Fig. 1 Mapping the exiting linguistic resources, generating the sibling tokens using the Workers, and finding at least one of the tokens in the graph

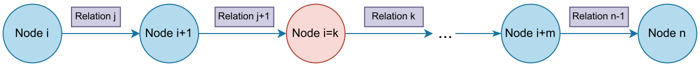


Fig. 2 A sample of the path from the labelled node (has the order k) to the found node (has the order n)

where k represents the order of the labelled node, P_k the polarity of the labelled node in the path which has the order k, n the order of the similar found node (similar to the sibling token), and t_j the type of relation j (whether direct or indirect) between two linked nodes.

We propose two search strategies: (i) **Bottom-Up**, starts from random node to the searched Token. For this there are two variants: the first one is *Forward sufficient*, if the taken path has at least one labelled node. Otherwise *Rollback needed* starts from the found node to the nearest labelled node. (ii) **Top-Down** it is the reverse one of Bottom-Up. It starts by a Polar Node until finding the similar node (similar to the sibling token).

As mentioned before, the generation of more than one sibling Tokens aims to improve the chance to find the original Token's polarity. One question should be asked: which one among the sibling Tokens generated for the same objective (finding the original Token's polarity) is to be taken? To answer this question, three strategies of choice are proposed and detailed below: (i) **By order**:According to the First Come First Serve (FCFS) strategy, it chooses the polarity of the first sibling token, returned by the search algorithm. (ii) **By priority**: It chooses the polarity of the sibling token generated by the Worker that has a high priority over other Workers (the priority of Worker is defined by the user beforehand). (iii) **By dominance**: It chooses the polarity of the sibling token belonging to the majority class.

3.3 Example

This example illuminates the `GraLexi` process. Take the Spanish word "Bueno", for instance. The goal is to infer its polarity (positive or negative) as illustrated in Fig. 3. We used three Workers with different configurations: Neutral, Translator (English), and Translator (French) to generate three sibling Tokens (i.e., Bueno, Good, and Bien). Next, in the finding step, Dijkstra's algorithm was applied to search them. In this example only "Good" and "Bien" were found. The found token generated by the Worker with the highest priority will be used to calculate the polarity of the original Token (i.e., "Good"). Then, based on the following: (i) The taken path from Root to node of the word "Good" (c.f. Fig. 3 The path). (ii) The Worker type (the type of Translator Worker is direct). (iii) Applying the proposed Eq. (1). We infer the polarity of the Spanish word "Bueno", in that case was positive.

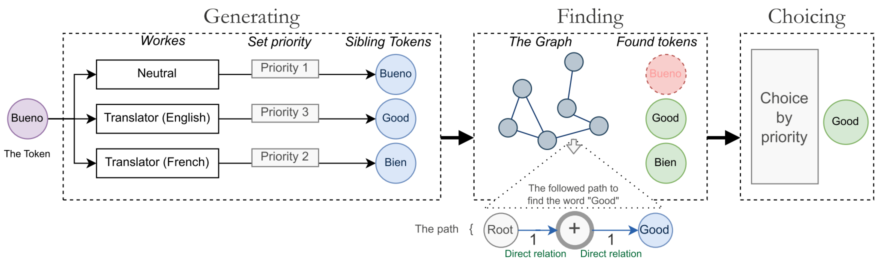


Fig. 3 The process of Finding and Choosing the token in the graph

4 Experimentation

In order to evaluate the proposed approach, we implemented the graph using Python programming language[1] along with a Python package NetworkX[2] [10], which provides a programming interface and graph implementation. For the resources, we chose the following: Arabic WordNet (AWN) [3] for Arabic words, French WordNet (WOLF) for French words [15], WordNet 3.0 (WN) for English words [14], and for polarity (words bearing sentiment) we chose the two resources: SentiWordNet3.0 (SWN3.0) [1] and MPQA Subjectivity Lexicon (MPQA) [19].

4.1 Experimental Setup

The constructed graph should encompass labelled words (the word's polarity is known either positive or negative) where the graph created by these words is called a polar graph (for polarity graph nature). To do that, two resources are used in parallel. The SWN3.0 then MPQA. In view of the SWN3.0 structure, we have selected words with the following conditions:

- Words that have PosScore ≥0.1 or NegScore ≥0.1;
- Words that have different PosScore and NegScore;
- If two similar words have different scores, we take the word that has the maximum difference between the two scores (i.e., PosScore and NegScore).

Secondly, we mapped WOLF and linked the Synsets with the corresponding one in WN3.0 to create the sub-graph G_WOLF. Thirdly we redo the same mechanism for AWN to get G_AWN. Finally, we combined the generated sub-graphs above to get one Graph used in the evaluation. Figure 4 illustrates an extracted part of the graph. Besides, we repeated the previous process after replacing SWN3.0 with MPQA. We linked the positive words with the Positive Polar Node and the negatives words with

[1] Python version 3.7.

[2] NetworkX version 2.4.

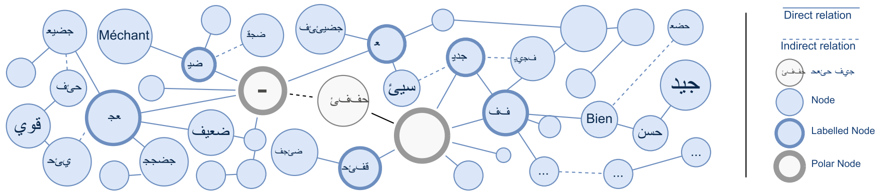


Fig. 4 A sample extracted from the graph used in the evaluation

the Negative Polar Node in the graph. At the end, we got two graphs: the first one based on SWN3.0 and the second is based on MPQA.

We utilized a **Top-Down** procedure (discussed in Sect. 3.2) and we adopted the Dijkstra's algorithm [7] to find the shortest path between the Polar Node and the Similar Node (the searched word). Also, we used two Workers: (i) **Neutral** (the neutral Worker keeps the same value of the Token in input). We used it along with English, French, and Arabic languages. (ii) **Translator** (translate the Token's value to a specified language). We used it along with Spanish and German languages.

4.2 Experimental Results

To evaluate and show the benefits of the proposed approach on multilingual and language-independent cases, a list of words from five languages (English [12], French [6], Arabic [16], Spanish [6], and German [6]) are used. As some list contains less than 2000 words, we selected 1000 words from each one to have portions equitably. That makes 2000 words (1000 positive words and 1000 negative words) and in total, 10000 words from the five languages.

We ran three times the Finding process. For each run 10000 words are randomly selected. The accuracy, recall, and F1-measure are calculated and outlined in Table 1. Then, we will discuss the results in the next section.

5 Discussion

The proposed approach gives accuracy (0.98) and recall (0.70). At the same time, the effectiveness depends on some specifications which we report in the following.

The effectiveness of the proposed approach depends on the choice of labelled words. We have noticed considerable enhancing on accuracy, by changing the polar graph from G_SWN to G_MPQA. The accuracy increased from 0.77 to 0.98 (+0.21) for the English language. Besides, we recorded an increasing of +0.12 for French, an increment of +0.16 for Spanish, and +0.15 for German.

Table 1 Experimental results

Language	Graph	#Nodes	#Edges	Accuracy	Recall	F1-score
English	G_SWN	056198	177911	0.77	0.72	0.74
	G_MPQA	013234	029997	0.98	0.70	**0.82**
French	G_SWN + G_WOLF	127646	304862	0.71	0.65	0.68
	G_MPQA + G_WOLF	095849	159632	0.83	0.64	**0.72**
Arabic	G_MPQA + Undia-critized	038548	198808	0.75	0.24	**0.37**
	G_MPQA + Diacritized	032695	216203	0.76	0.15	0.25
Spanish	G_MPQA + Fr	095849	159632	0.80	0.43	0.56
	G_MPQA + En	013234	029997	0.90	0.58	**0.70**
German	G_SWN	056198	177911	0.75	0.63	**0.68**
	G_MPQA	013234	029997	0.90	0.46	0.61

The choice of the Worker also could affect performance. For Spanish, we used two Workers of type Translator (English/French). We notice that the accuracy increased from 0.80 to 0.90 (+0.10) and the recall increased from 0.43 to 0.58 (+0.15) by using two different Workers.

Using a rich lexicon as SWN enhances the recall of the system. Meanwhile, it could decrease the accuracy due to the overlap of sentiment words between their relation (Synonyms, Antonyms). As shown in the German language, the recall increased from 0.46 to 0.63 (+0.17).

The characteristics of language influence the proposed approach's performance, as studied with the Arabic language. When we used only undiacritized words, we have increased the value of recall from 0.15 to 0.24 (+0.09).

Before concluding, it is worth noting that we got the words' polarity of French, Arabic, Spanish, and German languages without using any polar lexicon for these languages. We just based the investigation on one polar lexicon (i.e., SWN or MPQA) and related the different words respecting some relations (e.g., Synonyms, Antonyms, Equivalent in other languages).

The proposed approach `GraLexi` aims to find the words' polarity regardless of the language, which is very useful, especially in the multilingual sentiment analysis field.

6 Conclusion

This paper proposed a new approach to create a multilingual lexicon for sentiment analysis by linking and exploiting the different existing linguistic resources. In our experiment we mapped the following resources, WN, AWN, WOLF, SWN, and MPQA into one graph to unify their structures. Next, we applied the Dijkstra's algorithm to find the words from word lists of five languages (English, French, Arabic, Spanish, and German) to show the language independent of `GraLexi`. For the English language, we reported an F1-score attaining 0.82. Meanwhile, accuracy ranged from 0.76 to 0.98. Then we discussed the results obtained in the experimentation and how the choice of resources and the characteristics of the language can affect the final result. Given the encouraging results, we tend to continue developing the proposed approach by considering the word's context, Part of Speech tags (PoS), and enlarging the graph to be cross-domain. Some extensions can be done on the graph's structure, such as adding weights to the edges (relations between words), besides using the intensifiers (weaken, strengthen) to adjust the polarity. We also suggest applying a pattern matching between terms instead of using exact matching to increase the Recall.

References

1. Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC, vol 10, pp 2200–2204
2. Badaro G, Baly R, Hajj H, Habash N, El-Hajj W (2014) A large scale Arabic sentiment lexicon for Arabic opinion mining. In: Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP), pp 165–173
3. Black W, Elkateb S, Rodriguez H, Alkhalifa M, Vossen P, Pease A, Fellbaum C (2006) Introducing the Arabic wordnet project. In: Proceedings of the third international WordNet conference. Citeseer, pp 295–300
4. Bond F, Foster R (2013) Linking and extending an open multilingual wordnet. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1352–1362
5. Bordoloi M, Biswas SK (2017) E-commerce sentiment analysis using graph based approach. In: 2017 international conference on inventive computing and informatics (ICICI). IEEE, Nov 2017
6. Chen Y, Skiena S (2014) Building sentiment lexicons for all major languages. In: Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics
7. Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1(1):269–271
8. Esuli A, Sebastiani F (2006) Sentiwordnet: a publicly available lexical resource for opinion mining. In: LREC, vol 6. Citeseer, pp 417–422
9. Gross JL, Yellen J (2005) Graph theory and its applications. CRC Press
10. Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds) Proceedings of the 7th Python in science conference, Pasadena, CA, USA, pp 11–15

11. Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics
12. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining—KDD '04. ACM Press
13. Kamps J, Marx M, Mokken RJ, De Rijke M et al (2004) Using wordnet to measure semantic orientations of adjectives. In: LREC, vol 4. Citeseer, pp 1115–1118
14. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to WordNet: an on-line lexical database. Int J Lexicogr 3(4):235–244
15. Sagot B, Fišer D (2008) Building a free French wordnet from multilingual resources. In: Ontolex 2008, Marrakech, Maroc
16. Salameh M, Mohammad S, Kiritchenko S (2015) Sentiment after translation: a case-study on Arabic social media posts. In: Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, Denver, Colorado. Association for Computational Linguistics, pp 767–777, May–June 2015
17. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Comput Linguist 37(2):267–307
18. Widmann N, Verberne S (2017) Graph-based semi-supervised learning for text classification. In: Proceedings of the ACM SIGIR international conference on theory of information retrieval. ACM, Oct 2017
19. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing, pp 347–354

Chinese Social Media (Weibo) as a Tool to Advance Participatory Management During the Pandemic Period

Xiaoxu Liang, Naisi Hua, and Yu Zhang

Abstract Holistic cultural heritage management is one of the most important components of sustainable urban development strategy, especially in the rapid urbanization context. In line with it, social media have been carried out as an effective tool to enhance community engagement in the decision-making process of the urban heritage renewal project. However, researches on how social media could support people-centered heritage management and evaluate based on the collected data are still limited. Taking Lijiang, Pingyao, Suzhou, Macau, and Kulangsu as case studies, this article proposes an assessment framework to analyze the users' posts' information, such as geo-location, texts, and photos on Weibo. Besides, it offers an insight into the context of the COVID-19 pandemic period. This research further explored the form of the online community aiming to reveal the current state of online participatory practices in China. As shown in the result session, the participatory degree, to a great extent, remains in the informing and consulting levels. The article concludes that the public participation in the inclusive heritage management process is still at a nascent level. More digital tools should be encouraged to apply to the cultural heritage management field.

Keywords Cultural heritage management · Inclusive governance · Historic urban landscape approach · Social media · Weibo

X. Liang
Polytechnic University of Turin, Viale Mattioli, 39, 10125 Turin, Italy
e-mail: Xiaoxu.liang@polito.it

N. Hua · Y. Zhang (✉)
School of Architecture, Harbin Institute of Technology, Harbin, China
e-mail: Yu.zhang@edu.hit.cn

N. Hua
e-mail: 19B934018@stu.hit.edu.cn

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6_89

1 Introduction

The development of the Authorized Heritage Discourse in China is accompanied by new global aspirations [1]. Effective actions and holistic approaches are undertaking to contain the cultural identity, the sense of place, the value of the heritage and human environment, and the notion of authenticity [2]. The introduction of the HUL approach to China stands out as a "bottom-up" expression of social values and social choice. A series of relevant activities were organized by the World Heritage Institute for Training and Research in the Asia-Pacific Region (WHITRAP) on promoting HUL approach since 2014. During this COVID-19 epidemic period, China set a paradigm in the worldwide level to face this sudden global crisis. Some effective and efficient approaches even have been taken by the Chinese government to enhance public participation in the urban renewal process taking advantage of this quarantine period. However, researches on evaluating and illustrating the participatory governance are still limited, and effective online tools to involve residents in decision-making have yet to be explored adequately.

Following the introduction, the literature review states the contribution and development of HUL in China; the role of social media in the heritage protection process, and the possible assessment methods on community engagement. The article continues with the construction of the assessment framework from data acquisition, collection to the analysis process. The results detailed and further explained the statistics concerning general information of the Weibo posts and users, evaluation of the normalized texts, and mapping of the online-posted pictures. The conclusion part ends the article with several reasonable suggestions and thinking.

2 Background

2.1 The Historic Urban Landscape Approach

Historic Urban Landscape (HUL) is a sustainable and holistic approach to cultural heritage management, promoted by UNESCO since 2005 with the announcement of Vienna Memorandum [3]. Applying the HUL approach in the local context is explained along with highlighting the value of community in 2011 [4]. Since then, community Engagement, as one of the four tools of the HUL approach, gains increasing attention in the international heritage conservation field [5]. HUL can better recognize cultural diversity and the dynamic nature of urban heritage in the context of rapid globalization and urbanization [6]. It encourages to consider not only the attributes and values of various resources but also the heritage site itself [7].

The role and values of the community are reconsidering, accompanied by the growing awareness of the people-centered concept and collaborative planning [8, 9]. Various stakeholders are categorized into two main groups as core communities and broader communities [10]. Residents who are living inside or near the heritage sites are recognized as the core community. At the same time, the definition of the broader community is far more inclusive, such as professionals, scholars, governments, real estate companies, and people who care about this topic [11].

Chinese practices in the cultural heritage properties also attract the attention of worldwide scholars with heated debate. Bandarin and Van Oers have noticed that the Luoyang method in China set a good example for the following urban heritage protection projects by effectively avoiding the damage to underground construction [12]. On the contrary, the loosening of the regulatory system in the rural heritage projects due to the passive position of economic discourse could partially result in negative endings [13]. The importation of the HUL approach in China occurred a couple of years later than the Western countries. It is believed that improving the living conditions of residents and the authorization of sharing benefits from urban heritage conservation could benefit the sustainable urban development in China in a long-term view.

2.2 Social Media and Cultural Heritage

Social media offers a platform to involve more stakeholders, such as grassroots and experts with distance, to take part in the cultural heritage conservation process [2]. It plays a core role in the digital era to promote communication and collaboration among those profit mattered community. Online community as a new form of community is born in the digital landscape heritage context. Internet users gather for a specific common goal or topic to organize their practices, collect opinions, or exchange relevant information [14]. Online community differs from the so-called traditional community due to the latitude of the time, places, occupations, ages, and other possible barriers. It enables stakeholders from different groups to discuss and communicate far easier.

Majority of researches on the utilization of social media in cultural heritage conservation topics is Western based. The hue and brightness of the tweeted images (pictures posted on Twitter) are illustrated to map the smart city [15]. The role of Facebook played in the two-way communication between residents and governments is discussed in the urban planning project of Jordan [16]. Narrative studies are done with the aid of Facebook in collecting unofficial histories as a kind of digital archive [17]. Geotagged Instagram posts are visualized to reassembling the city of Amsterdam [18]. Pictures, tags, and real-time geo-location from Flickr are documented to reflect locals' and tourists' preferences and the political and religious issues in the historic city of Tripoli, Lebanon [19]. A new way of collaborating with the Chinese local government to stop removing the Dafo temple in Guangzhou is explored by tracing the information flow of a complaint post on Weibo [20]. There is an urgent need to develop social-media-involved cultural heritage studies in the Chinese rapid urbanization context.

2.3 Assessment for Public Participation

The judgements and comments on the community engagement of the urban heritage renewal projects in China are easy to be found. A community-government collaboration paradigm is the restoration project of Tianzifang, Shanghai; the living environment is significantly improved [21]. While in the Gulou renewal project (Beijing) in 2012, the action of forced relocating of more than 100 families is considered as a negative sample in the urban heritage conservation field [22]. The on-site workshops in the renewal process of Shuangwan village, Jiangsu, are seen as a media to make an effort on the right of discourses from a bottom-up way [13].

The value and transformation of communities should be evaluated and monitored in the urban resource management, where different resources have different levels of value [23]. The opinions on the structure of the assessment framework to measure and valorize the participatory level remain divergent in the worldwide level. A framework is built to evaluate the quality of life of citizens to indicate the level of smartness of citizens in a residential area of Turin, Italy [24]. The E-governance participatory defining system can be introduced to the cultural heritage management field. Similarly, IAP2 as a guide to describing stakeholders' engagement process can be shaped to evaluate the online public participation as well.

3 Methodology

The study attempts to gain a deeper understanding of the online community engagement degrees by analyzing continuous data from the beginning of 2020 when the COVID-19 went from outbreak to stabilization in China. Data is collected from Weibo, one of the most widely used social media in China by accessing the Application Programming Interface (API). The brief roadmap can be demonstrated from four main steps. First of all, targeted heritage objects are obtained from the World Heritage List (WHL). Following that, a series of data is extracted from Weibo by advanced searching with the city-level attribution of selected sites and cultural heritage synonymously linked with Proximity Operators in Chinese. Afterwards, a self-examination process is carried out to weed out the duplicated posts, and an active purification is supplemented to acquire a sufficient database. The final step is the analysis and visualization of the filtered posts in terms of its geo-location, text, and images.

3.1 Case Determination

The study is designed and further developed within the world cultural heritage sites in China. Up to 2020, China ranked the first place (paralleled with Italy) in the WHL by owning 55 inscribed world heritage properties, in which 37 are of cultural heritage, along with 15 natural heritage and 4 mixed properties. In terms of focusing more on urban heritage, we set up a filter with the location in the urban area with a certain number of residents to ensure the sufficiency of the data volume and the diversity of the samples. As a result, cases that meet the requirements are listed as the Old Town of Lijiang, Yunnan (1997), the Ancient City of Pingyao, Shanxi (1997), the Classical Gardens of Suzhou, Jiangsu (1997, 2000), the Historic Centre of Macau (2005), Macau, and Kulangsu, Fujian (2017).

3.2 Data Acquisition

The screening process is first defined by a series of keywords which should contain both the geographical information and the cultural heritage attribution. Proximity Operators are applied to show the relationships with all the individual words. The draft of searching string is translated from Chinese as follows: (Lijiang or Pingyao or Suzhou or Macau or Kulangsu) and (((Urban or Architectural) and Heritage) or (Historic and (City or District or Building)) or (Urban and (Renewal or Regeneration)) or (Traditional and (Architecture or District)) or (Historical Relic Protection) or (Heritage Sites Conservation) or (Old and (City or District or Building or Houses))). It seems a bit redundant in English due to the completely different language system between English and Chinese. Once the samples are extracted, it is first necessary to remove the duplications by adding a line of specific code into the programming.

Unlike most of the social media platforms, Weibo composes of three original websites: weibo.cn; weibo.com; m.weibo.com. Among those, weibo.cn is considered with a broader threshold of posts' documentation and users' statistics. The study tests and obtains the API interface and associated interface parameters from the Chrome browser, using Python as the programming language and using cookies to simulate browser access. Requests are sent via the Python re-quests library. The returned data is further parsed by the Python libraries LXML, Beautifulsoup, Regular Expression, etc. The time complexity for this operation is O(n).

3.3 Data Selection

Three steps are needed in the selection process on programming and by manual filtering to purify the acquired posts. Firstly, 180 duplicated posts with identical field content and the same poster are excluded automatically by programming. Then, according to the attribution of the posters, 336 posts from the local are picked out. In this step, the users who are registered in the same province with the targeted heritage sites are considered as local citizens. Following this step, the retained posts are normalized manually based on contents which are classified into 17 categories according to their purpose. Posts are only included in the analysis if they are sufficiently relative with heritage preservation or holistic cultural heritage management.

Table 1 The process of selection and purification of the relevant posts on Weibo

No.	Cultural heritage site location	Year listed	Step 1	Step 2	Step 3	Step 4
1	Old Town of Lijiang	1997	227	190	33	11
2	Ancient City of Pingyao	1997	182	175	46	23
3	Classical Gardens of Suzhou	1997; 2000	372	367	100	29
4	Historic Centre of Macau	2005	231	212	22	13
5	Kulangsu: a historic international settlement	2017	481	369	135	43
		Sum	1493	1313	336	119

Posts which are weighted as the record of daily activities, report on restoration project, collective memory, and advertisement are screened out. Both the texts and pictures of the rest 110 posts were considered to have a strong association with participatory management and could be used for further analysis. Table 1 gives a summary of the data selection process.

3.4 Data Analysis

The posts were analyzed in three aspects: general data, texts, and images. They are recounted and analyzed for posting time (monthly), genders of posters, posts from the local/tourists, posts with/without images, and posts by celebrities/common users.

The texts are manually divided according to the aim of the contents into 14 groups: heritage value sharing, lectures, exhibitions, official announcement, archive, complaint, feedback on questioning, suggestions, workshops, collaborative planning, collective memory, the record of daily activities, advertisement, and report on restoration projects. The posts that belonged to the last four groups are removed since they were considered irrelevant to the participatory activities. The remaining ten areas are adapted to a framework with five degrees of public participation: inform, consult, involve, collaborate, and empower [6, 25]. This grade was taken from IAP2 Spectrum of Public Participation to give a systematic description of how public participation happens on social media in the Chinese context.

The subjects sort images into two groups: Spaces and Taken Scenarios. The category of Space contains the following elements: City, Urban Landscape, Street, Courtyard, Architecture, Architectural Details, Interior Space, Interior Decoration, Natural Landscape, Marks, and Gardening Elements. At the same time, the category of Taken Scenarios is further separated as the conference, exhibition, tourism, food, and others. Duplicated images are also remarked. By mapping the posted pictures, the main focus of the local posters who are considered taking part in participatory management.

4 Result

The result mainly visualized the relevant number of posts of the intended heritage sites, examined the content and amount of selected posts which are categorized into five participatory levels, and sorted the main object of chosen pictures into two main areas.

4.1 The General State of Weibo Posts and Posters

1013 Weibo posts (Lijiang = 190, Pingyao = 175, Suzhou = 367, Macau = 212, and Kulangsu = 369) are obtained with their corresponding descriptive information. The garden of Suzhou and the historic city of Kulangsu attract the most attention of Weibo users and rank first places. Macau, Lijiang, and Pingyao appear far behind the first group with almost 50% fewer posts. The financial situation is likely to affect the result since Suzhou and Kulangsu are located in the coastal economic developed region. Although Macau is in the Pearl River Delta Economic Zone, which means a good financial condition, the difficulty of requesting a tourism visa may obstruct the way of visitors.

The posters who shared their opinions online concerning the selected urban heritage sites comprise a specific online community. The bar chart in Fig. 1 illustrates the number of men and women, the defined local citizens and tourists, and the celebrities and common users. Men make up 60% of this community with only a small difference from female posters. In contrast, it is clear that the majority of users are visitors from another region, and only around 25% of users are considered as local users. Meanwhile, nearly a third of participants own more than 20,000 followers, which is far more than the registration proportion (around 3‰).

Figure 2 shows a clear upward trend in the number of Weibo posts over the past 10 months. Due to the strict lockdown policy facing the pandemic of COVID-19, the data keeps at a low level in the first half of the year. Following a small peak in June, the total number of posts shrank noticeably from 82 to 68 in July. It is corresponding with the short relaxing period with the loosening restrictions of intra-urban traffic movement. The number increases dramatically in the following period along with the opening of regional boundaries. By the end of October, the number of posts reached

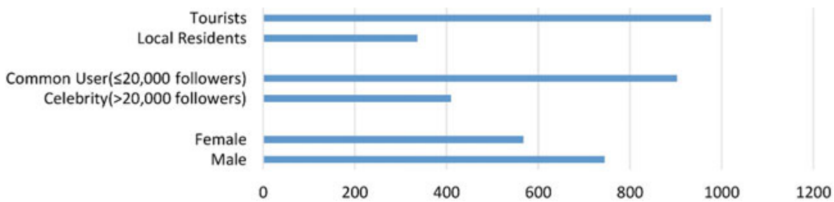


Fig. 1 Components of posters under the urban heritage conservation topic

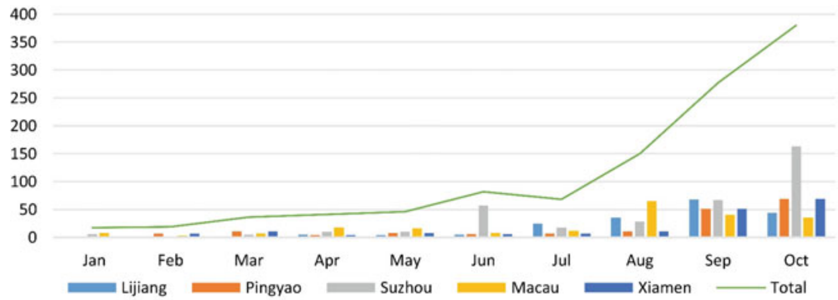


Fig. 2 Monthly Weibo posts during the pandemic period in China (01/01/2020–31/10/2020)

almost five times in July and 20 times in January. Up to now, people still are allowed to commute freely between different destinations with the personal health code.

4.2 *Evaluating Participatory Degrees According to the Texts*

The texts are examined manually by content analyzing and categorized into the five participatory levels. It is cheerful to have some findings in the involving and collaborating level which is the relative advanced phase of participation. Although the proportion of involving and collaborating takes only 6% of the total, it has a significant meaning as a double-way communication. Posting activities focus more on sharing heritage values to broader audiences and passively receiving the official announcement from the government-controlled ID and the education-oriented information from other celebrities. Nine out of ten users remain in the one-way communicative level, which is the most fundamental degrees of this assessment framework (Fig. 3).

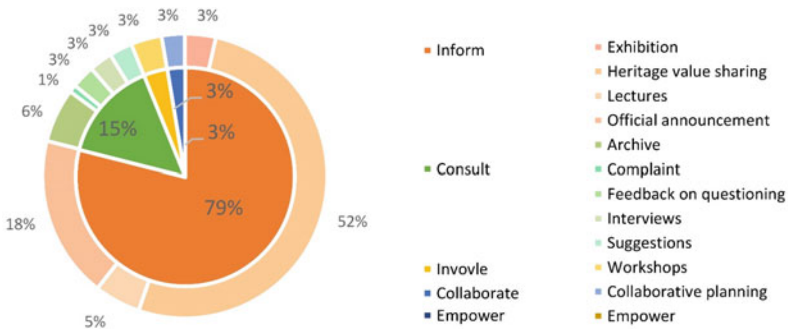


Fig. 3 Participatory degree assessment in the cultural heritage management process in China

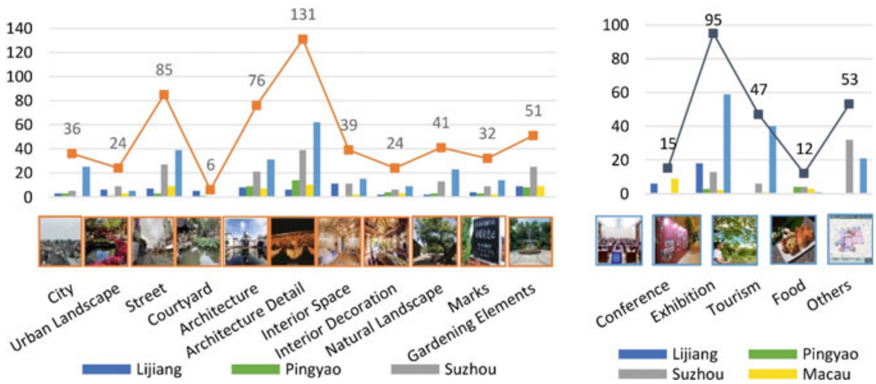


Fig. 4 Classification of Weibo posted images (01/01/2020–31/10/2020)

It is a pity that after such a long searching and checking procedure, no post has been found to represent the empowering level of participation. The reasons could be the remained reliance on paper documents, the exclusive operation environment to older people, and a still top-down process of engagement.

4.3 *Reading and Mapping of the Pictures*

The abstracted information from images reflects the public's focus and interests in the urban heritage set (Fig. 4). It is apparent from this table that local participants are most interested in architectural detail; the specific components of the building façade such as doors, windows, and bricks are most often photographed. The attention on the buildings and streets is preceded only by the architectural detail with a nearly halved number of pictures. The other spatial elements are almost in the same range of concern, while the courtyard received the least attention as the design element that appeared the least frequently among the five sites. It seems that physical spaces and elements close to the human body in scale are more popular.

The illustration of the picture-taken scenarios shows that exhibitions have considerable attractions to the local users. Tourism, Food, and Conference are likely not so important to their social life and online community. The result from the analysis of image categories shows that the poster's involvement in heritage preservation exists on multiple levels, even in unconscious situational interactions.

5 Conclusion

Although the public participation toward inclusive heritage management in China is still in the early stage, the status quo still reflects a bright future to improve due to the popularization of social media and smartphones. The article offers a new way to evaluate and present the public participation level through social media in China. The result conducted that the number of participants may largely associate with the financial situation of the city. Urban heritage sites which are located in the more developed area have attracted noticeably more attention. Besides, the general trend of posting activities is tightly linked with the regulations made against the COVID-19 pandemic.

Knowing from personal social media posts, the participatory level in the World Cultural Heritage site in China is near informing and consulting. The focuses of the citizens are mainly on the human-friendly elements, such as the details of the built environment. The exhibition could be an effective tool to promote heritage protection due to the great attention from the online community. The virtual exhibition could also gain a great number of audience and should be encouraged to be held. Digital tools should be encouraged to apply to offer frequent access to the establishment of a systematic heritage preservation mechanism. The outcome of this research leads a new perspective to heritage professionals and researchers to further understand the cultural diversity and public participation in the decision-making process of the heritage management process.

References

1. Svensson M (2018) Heritage 2.0: maintaining affective engagements with the local heritage in Taishun. In: Svensson M, Maags C (eds) Chinese heritage in the making. Amsterdam University Press, Amsterdam, pp 269–292
2. Liang X (2020) Participatory management for cultural heritage: social media and Chinese urban landscape. In: Stephanidis C, Antona M (eds) HCI international 2020—posters. HCII 2020. Communications in computer and information science, vol 1226. Springer, Cham
3. Vienna memorandum on world heritage and contemporary architecture—managing the historic urban landscape. UNESCO World Heritage Centre, Vienna, Austria
4. New life for historic cities: the historic urban landscape approach explained. UNESCO World Heritage Centre
5. Bandarin F, van Oers R (2014) Reconnecting the city: the historic urban landscape approach and the future of urban heritage. Wiley, Hoboken
6. Li J, Krishnamurthy S, Pereira Roders A, van Wesemael P (2020) Community participation in cultural heritage management: a systematic literature review comparing Chinese and international practices. Cities 96:102476
7. Veldpaus L (2015) Historic urban landscapes: framing the integration of urban and heritage planning in multilevel governance. Bouwstenen
8. Borona G, Ndiema E (2014) Merging research, conservation and community engagement. J Cult Herit Manag Sustain Dev
9. Spacca S, Dellapiana E, Sanna A (2018) Promoting industrial cultural heritage by augmented reality: application and assessment. Open Cybern Syst J 12

10. Poulios I (2014) Discussing strategy in heritage conservation: living heritage approach as an example of strategic innovation. J Cult Herit Manag Sustain Dev 4:16–34
11. Li J, Krishnamurthy S, Pereira Roders A, van Wesemael P (2020) Informing or consulting? Exploring community participation within urban heritage management in China. Habitat Int 105:102268
12. Bandarin F, van Oers R (2012) The historic urban landscape: managing heritage in an urban century. Wiley, Hoboken
13. Verdini G, Frassoldati F, Nolf C (2017) Reframing China's heritage conservation discourse. Learning by testing civic engagement tools in a historic rural village. Int J Herit Stud 23:317–334
14. Giaccardi E (2012) Heritage and social media: understanding heritage in a participatory culture. Routledge, London
15. Rose G, Willis A (2018) Seeing the smart city on Twitter: colour and the affective territories of becoming smart. Environ Plan D. https://doi.org/10.1177/0263775818771080
16. Rania Q (2017) Using social hub media to expand public participation in municipal urban plans. Procedia Eng 198:34–42
17. Purkis H (2017) Making digital heritage about people's life stories. Int J Herit Stud 23:434–444
18. Boy JD, Uitermark J (2017) Reassembling the city through Instagram. Trans Inst Br Geogr 42:612–624
19. Ginzarly M, Pereira Roders A, Teller J (2019) Mapping historic urban landscape values through social media. J Cult Herit 36:1–11
20. Deng Z, Lin Y, Zhao M, Wang S (2015) Collaborative planning in the new media age: the Dafo Temple controversy, China. Cities 45:41–50
21. Yung EHK, Chan EHW, Xu Y (2014) Sustainable development and the rehabilitation of a historic urban district—social sustainability in the case of Tianzifang in Shanghai. Sustain Dev 22:95–112
22. Bideau FG, Yan H (2018) Historic urban landscape in Beijing: the Gulou project and its contested memories. In: Maags C, Svensson M (eds) Chinese heritage in the making. Amsterdam University Press, Amsterdam, pp 93–118
23. Roders AP, Bandarin F (2019) Reshaping urban conservation: the historic urban landscape approach in action. Springer, Berlin
24. De Filippi FD, Coscia C, Guido R (2019) From smart-cities to smart-communities: how can we evaluate the impacts of innovation and inclusive processes in urban context? IJEPR 8:24–44
25. Stakeholder participation: IAP2 public participation spectrum. https://i2s.anu.edu.au/resources/stakeholder-participation-iap2-public-participation-spectrum/. Accessed 30 Nov 2020

Author Index

X.-S. Yang et al. (eds.), *Proceedings of Sixth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 235,
https://doi.org/10.1007/978-981-16-2377-6