

Shahram Latifi *Editor*

ITNG 2022 19th International Conference on Information Technology-New Generations

Advances in Intelligent Systems and Computing

Volume 1421

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Indexed by DBLP, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST).

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

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Advisory Board

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing, Universidad Central de Las Villas, Santa Clara, Cuba

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

Hani Hagrass, School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University, Győr, Hungary

Vladik Kreinovich, Department of Computer Science, University of Texas at El Paso, El Paso, TX, USA

Chin-Teng Lin, Department of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan

Jie Lu, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW, Australia

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute of Technology, Tijuana, Mexico

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro, Rio de Janeiro, Brazil

Ngoc Thanh Nguyen, Faculty of Computer Science and Management, Wrocław University of Technology, Wrocław, Poland

Jun Wang, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

More information about this series at <https://link.springer.com/bookseries/11156>

Shahram Latifi
Editor

ITNG 2022 19th International Conference on Information Technology-New Generations

 Springer

Editor
Shahram Latifi
Department of Electrical and Computer
Engineering
University of Nevada
Las Vegas, NV, USA

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-030-97651-4 ISBN 978-3-030-97652-1 (eBook)
<https://doi.org/10.1007/978-3-030-97652-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chair’s Message	ix
Part I Software Engineering	
1 Cross-Platform Blended Modelling with JetBrains MPS and Eclipse Modeling Framework	3
Malvina Latifaj, Hilal Taha, Federico Ciccozzi, and Antonio Cicchetti	
2 A Conceptual Framework for Software Modeling of Automation Systems ...	11
Mohammad Ashjaei, Alessio Bucaioni, and Saad Mubeen	
3 Virtual Reality Multiplayer Interaction and Medical Patient Handoff Training and Assessment	17
Christopher Lewis, Daniel Enriquez, Lucas Calabrese, Yifan Zhang, Steven J. Ambro, Ramona A. Houmanfar, Laura H. Crosswell, Michelle J. Rebaleati, Luka A. Starmer, and Frederick C. Harris, Jr.	
4 A Tool for Syntactic Dependency Analysis on the Web Stack	25
Manit Singh Kalsi, Kevin A. Gary, Vasu Gupta, and Suddhasvatta Das	
5 Using Software for Computational Fluid Dynamics and Molecular Dynamics	35
Jeena Shetti, Stefan Pickl, Doina Bein, and Marian Sorin Nistor	
6 Blended Modeling Applied to the Portable Test and Stimulus Standard	39
Muhammad Waseem Anwar, Malvina Latifaj, and Federico Ciccozzi	
7 An Evaluation Framework for Modeling Languages Supporting Predictable Vehicular Software Systems	47
Enxhi Ferko, Igli Jasharllari, Alessio Bucaioni, Mohammad Ashjaei, and Saad Mubeen	
8 A Model-Based Approach for Quality Assessment of Insulin Infusion Pump Systems	57
Tássio Fernandes Costa, Álvaro Sobrinho, Lenardo Chaves e Silva, Leandro Dias da Silva, and Angelo Perkusich	
9 Narrowing the Gap Between Software Engineering Teaching and Corporate Environment	65
Marcelo A. M. da Conceicao, Oswaldo S. C. Neto, Andre B. Baccarin, Luan H. S. Dantas, Joao P. S. Mendes, Vinicius P. Lippi, Gildarcio S. Gonçalves, Adilson M. Da Cunha, Luiz A. Vieira Dias, Johnny C. Marques, and Paulo M. Tasinaffo	
10 API-First Design: A Survey of the State of Academia and Industry	73
Nicole Beaulieu, Sergiu M. Dascalu, and Emily Hand	

Part II Data Science & Engineering

- 11 A Quality Dimension Analysis of Open Government Data in Undergraduate Public Funding in Brazil** 83
 Marcelo Moreira West and Glauco de Figueiredo Carneiro
- 12 A Survey of Real-Time ETL Process Applied to Data Warehousing Environments** 91
 Flávio de Assis Vilela and Ricardo Rodrigues Ciferri
- 13 Participatory Modeling: A New Approach to Model Graph-Oriented Databases** 97
 Luis A. Neumann, Enzo Seraphim, Otávio A. O. Carpinteiro,
 and Edmilson M. Moreira
- 14 Graph-Based Hierarchical Record Clustering for Unsupervised Entity Resolution** 107
 Islam Akef Ebeid, John R. Talburt, and Md Abdus Salam Siddique
- 15 Semantic-MDBScan: An Approach to Assign a Semantic Interpretation to Behavior Changes Detected in Data Stream Scenarios** 119
 Eldane Vieira Júnior, Rita Maria Silva Julia, and Elaine Ribeiro Faria
- 16 A Study on Usability Assessment of Educational Systems** 129
 Heber Miranda Floriano, Mario Jino, and Ferruccio de Franco Rosa

Part III Cybersecurity

- 17 Gesturing with Smart Wearables: An Alternate Way to User Authentication** 137
 Khandaker Abir Rahman, Avishek Mukherjee, and Kristina Mullen
- 18 Software Optimization of Rijndael256 for Modern x86-64 Platforms** 147
 Nir Drucker and Shay Gueron
- 19 Cybersecurity Ethics Education: A Curriculum Proposal** 155
 Ping Wang
- 20 Performance Evaluation of Online Website Safeguarding Tools Against Phishing Attacks; a Comparative Assessment** 161
 Rama Al-Share, Fatima Abu-Akleek, Ahmed S. Shatnawi, and Eyad Taqieddin

Part IV Blockchain Technology

- 21 Blockchain Based Trust for the Internet of Things: A Review** 171
 Dina Shehada, Maryam Amour, Suadad Muammar, and Amjad Gawanmeh
- 22 The Use of Blockchain Technology in Electronic Health Record Management: An Analysis of State of the Art and Practice** 179
 Henrique Couto, André Araújo, Rendrikson Soares, and Gabriel Rodrigues
- 23 Blockchain for Security and Privacy of Healthcare Systems: A Protocol for Systematic Literature Review** 187
 Saadia Azemour, Meryeme Ayache, Hanane El Bakkali, and Amjad Gawanmeh
- 24 Single Sign-On (SSO) Fingerprint Authentication Using Blockchain** 195
 Abhijeet Thakurdesai, Marian Sorin Nistor, Doina Bein, Stefan Pickl,
 and Wolfgang Bein

Part V Health Informatics

- 25 A Detection Method for Early-Stage Colorectal Cancer Using Dual-Tree Complex Wavelet Packet Transform** 205
Daigo Takano and Teruya Minamoto
- 26 Visualizing 3D Human Organs for Medical Training** 211
Joshua Chen, Paul J. Cuaresma, Jennifer S. Chen, Fangyang Shen, and Yun Tian
- 27 An Information Management System for the COVID-19 Pandemic Using Blockchain** 219
Marcelo Alexandre M. da Conceicao, Oswaldo S. C. Neto, Andre B. Baccarin, Luan H. S. Dantas, Joao P. S. Mendes, Vinicius P. Lippi, Gildarcio S. Gonçalves, Adilson M. Da Cunha, Luiz A. Vieira Dias, Johnny C. Marques, and Paulo M. Tasinaffo
- 28 Machine Learning for Classification of Cancer Dataset for Gene Mutation Based Treatment** 229
Jai Santosh Mandava, Abhishek Verma, Fulya Kocaman, Marian Sorin Nistor, Doina Bein, and Stefan Pickl

Part VI Machine Learning

- 29 Performance Comparison Between Deep Learning and Machine Learning Models for Gene Mutation-Based Text Classification of Cancer** 237
Fulya Kocaman, Stefan Pickl, Doina Bein, and Marian Sorin Nistor
- 30 Stock Backtesting Engine Using Pairs Trading** 245
Rahul Chauhan, Marian Sorin Nistor, Doina Bein, Stefan Pickl, and Wolfgang Bein
- 31 Classifying Sincerity Using Machine Learning** 255
Rachana Chittari, Marian Sorin Nistor, Doina Bein, Stefan Pickl, and Abhishek Verma
- 32 Recommendation System Using MixPMF** 263
Rohit Gund, James Andro-Vasko, Doina Bein, and Wolfgang Bein
- 33 Abstractive Text Summarization Using Machine Learning** 269
Aditya Dingare, Doina Bein, Wolfgang Bein, and Abhishek Verma
- 34 Intelligent System for Detection and Identification of Ground Anomalies for Rescue** 277
Antonio Dantas, Leandro Diniz, Maurício Almeida, Ella Olsson, Peter Funk, Rickard Sohlberg, and Alexandre Ramos

Part VII Human-Computer Interaction

- 35 An Application for Interaction Comparison Between Virtual Hands and Virtual Reality Controllers** 285
Daniel Enriquez, Christopher Lewis, Sergiu M. Dascalu, and Frederick C. Harris, Jr.
- 36 LDAT: A LIDAR Data Analysis and Visualization Tool** 293
Andrew Muñoz, Chase Carthen, Vinh Le, Scotty D. Strachan, Sergiu M. Dascalu, and Frederick C. Harris, Jr.
- 37 Social Media User Study** 303
Autumn Cuellar, Yifan Zhang, Sergiu M. Dascalu, and Frederick C. Harris, Jr.

38	Software Interfaces for New Vehicle Operating Cost Models Used in Economic Analysis of Transportation Investments: A User Study	311
	Arjun V. Gopinath, Hudson Lynam, Rami Chkaiban, Elie Hajj, and Sergiu M. Dascalu	
39	Microservice-Based System for Environmental Science Software Applications	321
	Vinh Le, Connor Scully Allison, Mitchell Martinez, Sergiu M. Dascalu, Frederick C. Harris, Jr., Scotty D. Strachan, and Eric Fritzinger	
Part VIII Networks		
40	Semantic Interoperability in the Internet of Things: A Systematic Literature Review	333
	Pedro Lopes de Souza, Wanderley Lopes de Souza, and Ricardo Rodrigues Ciferri	
41	IoT Machine Learning Based Parking Management System with Anticipated Prediction of Available Parking Spots	341
	Grzegorz Chmaj and Michael Lazeroff	
42	Channel State Information Spectrum Gap Filling Using Shallow Neural Networks	351
	Avishek Mukherjee, Beata Hejno, and Manish Osti	
Part IX Potpourri		
43	Unveiling a Novel Corporate Structure in World-Class Business, Merging Digital-Physical Environment in Hyper Famili Incorporation	363
	Mohammad Khakzadeh, Fatemeh Saghafi, Seyed Milad Seyed Javadein, Mohammad Hossein Asmaie, and Masoud Matbou Saleh	
44	Developing an Affective Audio Toolbox for Audio Post-production	371
	Harrison Ridley, Stuart Cunningham, and Richard Picking	
45	Boundary Approximation and External Visibility	379
	Laxmi Gewali and Samridhi Jha	
46	Detection of Strictly L3-Live Structures by Structural Analysis of General Petri Net Using SAT-Solver	387
	Yuta Yoshizawa and Katsumi Wasaki	
47	Space Abstraction and of PetriNets Using the Submarking Method Quasi-home States	393
	Tomoki Miura and Katsumi Wasaki	
	Index	399

Chair's Message



Welcome to the 19th International Conference on Information Technology – New Generations – ITNG 2022. Due to the continued global pandemic and the problems associated with traveling and in-person meeting, we are running the conference virtually. We hope and trust that in the following year, we will be able to meet in person when the audience feel safer and more comfortable to travel and participate in this event. The ITNG 2022 attracted quality submissions globally. The papers were reviewed for their technical soundness, originality, clarity, and relevance to the conference. The conference enjoyed expert opinion of over 40 author and non-author scientists who participated in the review process. Each paper was reviewed by at least two independent reviewers. At the end, 48 papers were accepted for presentation and publication in the ITNG 2022 program.

The chapters in this book address the most recent advances in such areas as machine learning, big data analytics, cybersecurity, blockchain technology, data mining, e-health, IoT & CPS, software engineering and social computing. In addition to technical presentations by the authors, the conference features two keynote speakers on Monday and Tuesday.

Several people contributed to the success of this year's conference by organizing technical tracks. Dr. Doina Bein served in the capacity of conference vice chair. We benefited from the professional and timely services of major track organizers and associate editors, namely Drs. Azita Bahrami, Christian Barria, Doina Bein, Allesio Bucaioni, Glauco Carneiro, Sergiu Dascalu, Luiz Alberto Vieira Dias, Fred Harris, Ray Hashemi, Kashif Saleem, Fangyan Shen, David Cordero Vidal, and Hossein Zare.

Others who were responsible for solicitation, review, and handling the papers submitted to their respective tracks/sessions include Drs. Wolfgang Bein, Poonam Dharam, and Mei Yang.

The help and support of Springer in preparing the ITNG proceedings is specially appreciated. Many thanks are due to Michael Luby, Senior Editor and Supervisor of Publications, and Brian Halm, Production Editor at Springer, for the timely handling of our publication order. We also appreciate the efforts made by the Springer Project Coordinator, Olivia Ramya Chitranja. Olivia spent much time looking very closely at revised chapters to make sure they are formatted correctly according to the publisher's guidelines.

We also thank the technical assistance of Prabhas Kumra in setting up the conference program and Zoom communication for us. Finally, the great efforts of the conference secretary, Ms. Mary Roberts, who dealt with the day-to-day conference affairs, including timely handling of emails, are acknowledged.

I hope that you all enjoy the ITNG 2022 program and find it technically and socially fulfilling.

The ITNG General Chair

Shahram Latifi

ITNG 2022 Reviewers

Andro-Vasko, James	Latifi, Shahram
Abbas, Haider	Le, Vinh
Alahmad, Yanal	Liu, Siming
Alasady, Majida	Machado, Ivan
Alwadi, Ali	Marques, Johnny
Bahrami, Azita	Mascarenhas, Ana
Barría, Cristian	Mialaret, Lineu Fernando Stegr
Bein, Doina	Mohammadi-Koushki, Negar
Bein, Wolfgang	Mohd-Ali, Nursabillilah
Cagnin, Maria	Monteiro, Miguel
Carneiro, Glauco	Muhanna, Muhanna
Cirstea, Silvia	Nguyen, Tin
Collazos, Cesar	Quiroz, Juan
Cordero, David	Redei, Alex
Cunha, Adilson Marques da	Resende, Antonio
Cunningham, Stuart	Saleem, Kashif
Dascalu, Sergiu	Scully-Allison, Connor
David, José	Sharma, Sharad
Dharam, Poonam	Shen, Fangyang
Dias, Luis Alberta Vieira	Silva, Paulo Caetano
Garcia, Vinicius	Soares, Michel
Gawanmeh, Amjad	Suzana, Rita
Gawanmeh, Amjad	Trausan-Matu, Stefan
Grout, Vic	Vieira-Dias, Luiz-Alberto
Harris, Frederick	Yang, Mei
Hashemi, Ray	Zare, Hossein
Hazzazi, Noha	

Part I

Software Engineering

Cross-Platform Blended Modelling with JetBrains MPS and Eclipse Modeling Framework

Malvina Latifaj, Hilal Taha, Federico Ciccozzi, and Antonio Cicchetti

Abstract

Modelling tools traditionally focus on one specific editing notation (such as text, diagrams, tables or forms), providing additional visualisation-only notations. For software-intensive systems with heterogeneous components and entailing different domain-specific aspects and different stakeholders, one editing notation is too little and voids many modelling benefits. Seamless blended modelling, which allows stakeholders to freely choose and switch between graphical and textual notations, can greatly contribute to increase productivity as well as decrease costs and time to market. In this paper we describe our work in bridging two powerful (meta) modelling platforms: Eclipse Modeling Framework, for the definition of tree-based and graphical DSMLs and models conforming to them, and JetBrains MPS, for the description of textual DSMLs and the projectional manipulation of textual models via multiple views. The possibility to visualise and edit the same information in these two platforms, otherwise disjoint, can greatly boost communication between stakeholders, who can freely select their preferred notation or switch from one to the other at any time.

Keywords

Blended modelling · Graphical modelling · Textual modelling · Model transformations · Language engineering · Model-driven engineering · Language workbenches · MPS · EMF · Ecore

M. Latifaj (✉) · H. Taha · F. Ciccozzi · A. Cicchetti
School of Innovation, Design and Engineering, Malardalen University,
Vasteras, Sweden
e-mail: malvina.latifaj@mdu.se; hilal.taha@mdu.se;
federico.ciccozzi@mdu.se; antonio.cicchetti@mdu.se

1.1 Introduction

The complexity of modern software systems is continuously growing. It is enough to mention the increasing trend of *smartifying* applications by augmenting them with self-adaptation and customisation features. With this trend, software tends to cross-cut and integrate multidisciplinary domains [4], posing the problem of keeping consistent the properties of the system from multiple viewpoints [8]. In any non-trivial development scenario, consistency management is impracticable if performed at the implementation code level of abstraction. Or more precisely, it could be possible, but the consequences of detected malfunctions are typically expensive analyses to understand where a problem originates and how it could be fixed [13].

Model-Driven Engineering (MDE) methodology proposes to use models as the main development artefacts [12]. In this way, it is possible to reduce the complexity of systems development, and notably validate the integration earlier in the process thanks to a higher-level of abstraction view of the result; in turn, these analyses potentially save the costs of modifications because they can significantly anticipate the implementation phases. Moreover, when dealing with multidisciplinary domains, MDE state-of-the-art recommends to use Domain-Specific Modelling Languages (DSMLs): as the name suggests, DSMLs are languages tailored for a specific domain, hence providing experts with modelling concepts closer to their knowledge areas [13]. The set of modelling concepts included in a DSML is defined by means of a metamodel, that can be considered as a grammar defining the set of legal productions for a language. In this respect, language engineers usually leverage workbenches to create DSMLs [10]; in fact, based on metamodel specifications language workbenches

provide effective features for a DSML provision, notably the automatic derivation of model management processes, conformance checks, the support for the definition of custom editors, and so forth. Depending on the technical space [3], the preferred underlying implementation technologies, etc., there exists a multitude of alternative language workbenches. Even if from a conceptual perspective they are all based on similar metamodel specifications, only in rare cases it is possible to exchange metamodels, and corresponding models, between workbenches.

This paper discusses our efforts in bridging two language workbenches, JetBrains MPS (later also referred to as simply MPS) [10] and the Eclipse Modeling Framework (EMF) [7]; more specifically, the work is devoted to the mapping of metamodels defined in MPS towards metamodels conforming to the EMF specification language, namely Ecore. Technically, we firstly contribute with an Ecore language specification for MPS; based on it, users can create metamodels by using the MPS language workbench features and possibly create models conforming to such metamodels. Alternatively, we contribute with a transformation for mapping the metamodels defined in MPS as Ecore metamodels usable in EMF. In this latter scenario, users can leverage EMF plug-ins e.g. to generate default tree editors, create a custom concrete syntax, use the metamodel as part of a model transformation chain, and so forth.

The paper is organised as follows: Sect. 1.2 describes the motivation and intended contribution, while Sect. 1.3 compares related research in the area with respect to this contribution. The details of the developed solution and its evaluation are provided in Sects. 1.4 and 1.5 respectively. Eventually, the paper discusses concluding remarks and draws future investigation directions in Sect. 1.6.

1.2 Motivation and Contribution

Modelling tools traditionally focus on one specific editing notation (such as text, diagrams, tables or forms). This limits human communication, especially across stakeholders with varying roles and expertise. Moreover, engineers may have different notation preferences; not supporting multiple notations negatively affects their throughput. Besides the limits to communication, choosing one particular kind of notation has the drawback of limiting the pool of available tools to develop and manipulate models that may be needed. For example, choosing a graphical representation limits the usability of text manipulation tools such as text-based diff/merge, which is essential for team collaboration. When tools provide support for both graphical and textual modelling, it is mostly done in a mutual exclusive manner. For systems with heterogeneous components and entailing different domain-specific

aspects and different types of stakeholders, mutual exclusion is too restrictive and voids many of the MDE benefits [5]. Therefore, modelling tools need to enable different stakeholders to work on overlapping parts of the models using different modelling notations (e.g., graphical and textual). We have previously defined the notion of *blended modelling* [6] as:

the activity of interacting seamlessly with a single model (i.e., abstract syntax) through multiple notations (i.e., concrete syntaxes), allowing a certain degree of temporary inconsistencies.

Seamless blended modelling, which allows stakeholders to freely choose and switch between graphical and textual notations, can greatly contribute to increase productivity as well as decrease costs and time to market, as we have preliminary shown in [2].

In this paper we describe our work in bridging two powerful (meta) modelling platforms: EMF, for the definition of tree-based and graphical DSMLs and models conforming to them, and MPS, for the description of textual DSMLs and the projectional manipulation of textual models via multiple views. The possibility to visualise and edit the same information in these two platforms, otherwise disjoint, can greatly boost communication between stakeholders, who can freely select their preferred notation or switch from one to the other at any time. The research contribution is twofold: (i) definition and implementation of a textual language in MPS for representing Ecore's meta-metamodel; this enables the creation of Ecore-based meta-models directly in MPS, and (ii) definition and implementation of a mapping mechanism for Ecore-based meta-models created in MPS to be imported in EMF, enabling the automated exchange of these meta-models across the two platforms.

1.3 Related Work

In MPS it exists a mechanism to import and export Ecore-based languages (metamodels). Metamodels are imported as language structures in MPS [1], allowing the user to create models conforming to it and add structural features to imported metamodels. However, this does not allow to create Ecore-based metamodels in MPS, which is instead our main goal rather than simply importing them. The main difference is that we reproduce Ecore entirely in MPS, instead of transforming Ecore-based metamodels into MPS languages, which would lead to loss of conformance to the Ecore metamodel itself in case of modifications.

EMFText is an approach that allows the definition of textual syntaxes for languages described in terms of Ecore. More specifically, it allows the developers to define DSMLs

using a native textual syntax in EMF [9]. Being part of EMF, this approach does not provide any of the projectional benefits offered by MPS, the reason why we chose MPS as textual modelling target platform. Scheidgen [11] provides techniques to combine textual and graphical modelling by embedding textual editors into graphical ones. Users open an embedded text editor by clicking on an element within the graphical host editor. This work can enhance the effectiveness of graphical modelling with languages that partially rely on textual representations. The author focuses on adding functionalities to graphical editors by embedding textual models. Nevertheless, most benefits of the projectional approach provided by MPS would not be leverageable.

Related to the switching between graphical and textual syntaxes, Wimmer et al. [14] propose to bridge Grammarware and Modelware to ease transformations of models containing both graphical and textual elements. In Grammarware, a mixed model is exported as text. In Modelware [14], a model containing graphical and textual content is transformed into a fully graphical model. Although related to our work in terms of blending graphical and textual notations, our aim is to bridge two well-established frameworks providing

rather different approaches to modelling in order to exploit the benefits of both and we do it by a common language, Ecore.

1.4 Bridging MPS and EMF

To connect the textual-based world of MPS and the diagram-based world of EMF, we propose a solution that enables the textual definition of metamodels conforming to the Ecore meta-metamodel in MPS, and the exchange of these meta-models from MPS to EMF. Ecore is the language used in EMF for describing models, which are serialized and persisted as XMI files [7]. Throughout the rest of the text, we refer to metamodels created in EMF as *EcoreEMF metamodels* (.ecore extension), and the ones created in MPS as *EcoreMPS metamodels* (.sandbox.mps extension). Implementation and evaluation of this solution are carried out by following the process illustrated in Fig. 1.1. The process consists of two main steps.

Step 1. Recreate the Ecore meta-metamodel, as defined in EMF, as a language (EcoreLanguage) in MPS and evaluate whether the EcoreMPS metamodel conforming to

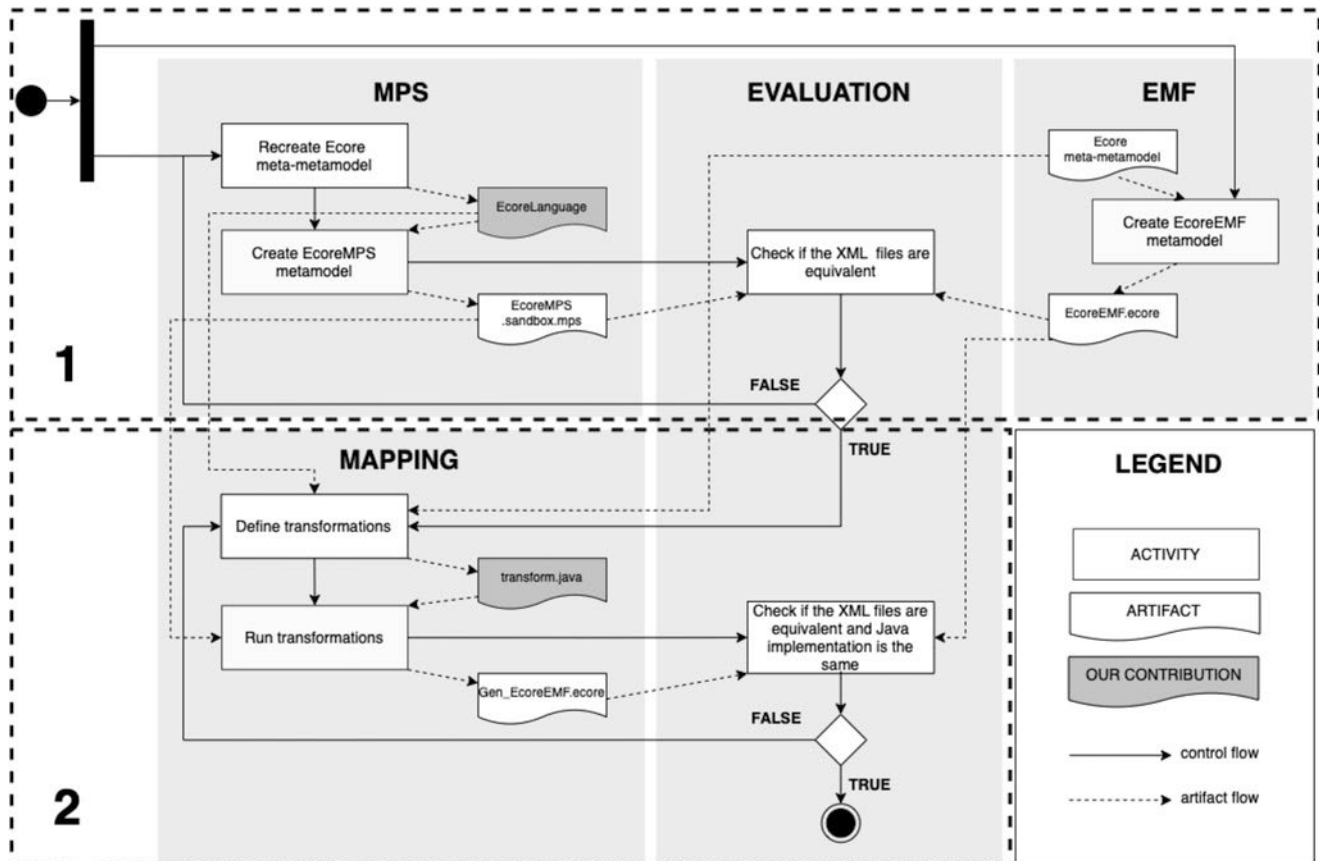


Fig. 1.1 Process of bridging MPS and EMF, and evaluation

EcoreLanguage is equivalent to the EcoreEMF metamodel conforming to the Ecore meta-metamodel in EMF. By equivalent, we mean that the metamodels contain the same concepts and hierarchical structure, while the order by which metaconcepts persist in the metamodel file might differ.

Step 2. Define automated mechanisms (i.e., model transformations) to transform EcoreMPS metamodels to Gen EcoreEMF metamodels and evaluate whether the latter is equivalent to the corresponding EcoreEMF metamodel and whether it can be correctly loaded and used in EMF. In addition, the generated Java classes from both metamodels (i.e., Gen EcoreEMF metamodel and EcoreEMF metamodel), should be the same. The implementation of EcoreLanguage in MPS and the transformations from EcoreMPS metamodels to EcoreEMF metamodels are described in Sects. 1.4.1 and 1.4.2, respectively, while the evaluation is described in Sect. 1.5.

1.4.1 EcoreLanguage: Implementing Ecore in MPS

The first step towards providing a bridge between EMF and MPS is recreating the Ecore meta-metamodel as an EcoreLanguage in MPS. The structure of the EcoreLanguage consists of *concepts*, *concept interfaces*, and their corresponding *children*, *properties*, and *references*, as found in the Ecore meta-metamodel. It is noteworthy to mention that concept interfaces, similarly to interfaces in Java, cannot be instantiated in the EcoreMPS metamodel. Thus, the concept interfaces that need to be instantiated, are implemented as concepts. For each concept of the language, there is an editor that facilitates the manipulation of the abstract syntax tree (AST) and provides intuitive interaction. The constraints aspect is used to express advanced constraints that cannot be covered by the language structure. An example of that is the name of the metamodel. In Ecore, the name of a metamodel cannot contain characters such as underscores and white spaces. If they were used in the name of an EcoreMPS metamodel, the transformation would output an invalid Gen EcoreEMF metamodel, that would not correctly load in EMF. To mitigate this risk, we add constraints that limit the choice of valid characters in MPS. Moreover, to allow the initialisation of some properties/references/children to default values when a concept instance is created, we use concept constructors and rely on the behaviour language aspect in MPS. Upon complete implementation, the language is packaged as a plugin, that can be distributed to users. Fig. 1.2 illustrates how the implemented EcoreLanguage is included in the list of available languages to choose from when adding a new solution to an MPS project. By importing this language the user can start defining EcoreMPS metamodels.

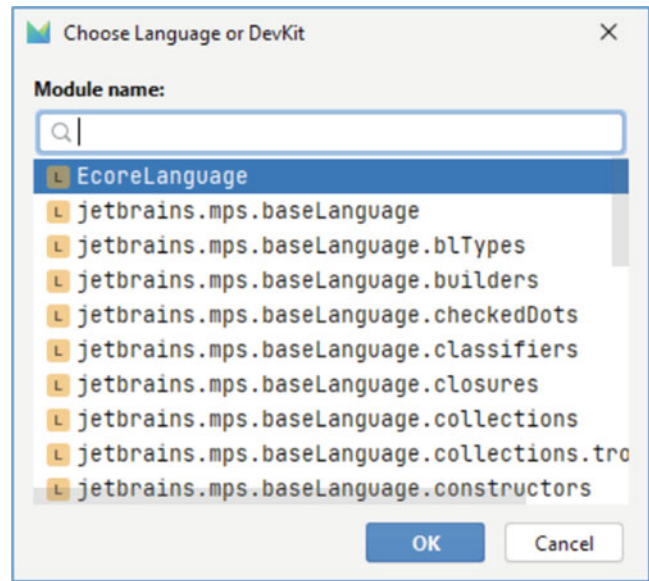


Fig. 1.2 Ecore language plugin in MPS

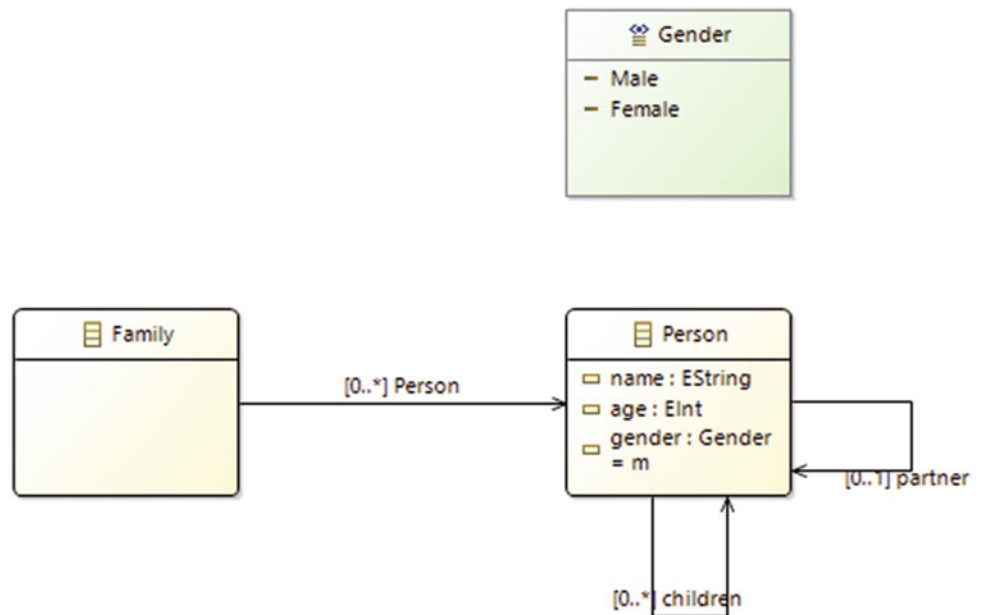
1.4.2 Automatic Export from MPS to EMF

The implementation of the Ecore meta-meta model in MPS as EcoreLanguage is pivotal for cross-platform modelling across EMF and MPS. However, even though it allows the textual definition of EcoreMPS metamodels, the latter still remain “isolated”, as they cannot be used outside the MPS environment. To build a bridge and enable the exchange of metamodels between MPS and EMF, EcoreMPS metamodels are transformed to Gen EcoreEMF metamodels that can be correctly loaded and used in EMF. The transformations are defined in Java and driven by an implicit mapping that is used to define correspondences between elements of the source (i.e., EcoreLanguage) and target (i.e., Ecore meta-metamodel) languages. While defining these correspondences, it is important to fully understand the structure of both languages, thus, in the following we provide code excerpts from the definition of a metamodel, both in EMF and MPS. To simplify the reading, we describe the procedure by its instantiation on a specific example, the Family metamodel (depicted in Fig. 1.3 in terms of EcoreEMF).

Listing 1.1 details the eClassifier Gender of type EEnum, whose values are restricted to eLiterals Male and Female from the family.ecore file. If we make a reference to Fig. 1.1, family.ecore represents the EcoreEMF.ecore artefact.

Listing 1.2 details the Gender concept, from the familymodel.sandbox.mps file that represents the EcoreMPS.sandbox.mps artifact in Fig. 1.1, and it consists of two parts.

1. **Language definition:** Defines the languages used for the definition of EcoreMPS metamodels. For this specific

Fig. 1.3 Family EcoreEMF metamodel

case, we have used a built-in language from MPS (line 3–5) that includes the BaseConcept and INamedConcept, and EcoreLanguage (line 6–11). EcoreLanguage reflects the Ecore meta-metamodel, where each concept can contain references, properties, and children, and they are all assigned randomly generated index values. Line 7 details the EEnum concept, while line 10 the EEnumLiteral concept.

- Metamodel description:** Defines the instances of concepts (i.e., nodes) that are described in the language. Each node stores a reference to its declaration, its concept. The node in line 15 stores a reference to EEnum, while the node in line 17, stores a reference to EEnumLiteral.

To connect the language definition and metamodel definition, we use hash-tables as data structures that can map keys to values.

Listing 1.1 XML Version of family.ecore

```

1 <eClassifiers name="Gender" xsi:type="ecore:
  EEnum">
2   <eLiterals literal="m" name="Male"
     value="0"/>
3   <eLiterals name="Female" value="1"/>
4 </eClassifiers>
  
```

Listing 1.2 XML Version of familymodel.sandbox.mps

```

1 <!-- language definition -->
2 <registry>
3   <language id="ceab5195-25ea-4f22-9b92
     -103b95ca8c0c" name="
     jetbrains.mps.lang.core">
4   ...
5   </language>
  
```

```

6   <language id="45e9c502-be8d-4b95-92c9
     -8ad2f7c494aa" name="
     EcoreLanguage">
7     <concept id="5921274573544802721"
         name="EcoreLanguage.
         structure.EEnum" flags="ng"
         index="1BB5TV">
8       <child id="5921274573544802722"
         name="eLiterals"
         index="1BB5TS" />
9     </concept>
10    <concept id="5921274573544831328"
        name="EcoreLanguage.
        structure.EEnumLiteral" flags="ng"
        index="1BBqUU" />
11  </language>
12 </registry>
13
14 <!-- metamodel definition -->
15 <node concept="1BB5TV" id="5qTU7U3AdSP"
     role="3Lc430">
16   <property role="TrG5h" value="Gender" />
17   <node concept="1BBqUU" id="5qTU7U3AdT0"
     role="1BB5TS">
18     <property role="TrG5h"
         value="Male" />
19     <property role="1BBqUN" value="0"
     />
20     <property role="1BBqUK"
         value="male" />
21   </node>
22   <node concept="1BBqUU" id="5qTU7U3AdT4"
     role="1BB5TS">
23     <property role="TrG5h" value="Female"
     />
24     <property role="1BBqUN" value="1" />
25     <property role="1BBqUK"
         value="female" />
26   </node>
  
```

Starting from the language definition, we iterate through all the elements of the language, and store the element's *index* as key, and the element's *name* as value. However, if we compare *node* and *property/reference* in the metamodel definition, we notice that they store references to their declarations differently. Nodes store references in the attribute *concept*, while properties and references store references in the attribute *role*. If we were to search for a node's reference to its declaration in the role attribute instead of concept, that would raise an exception, because the role attribute does not exist. Thus, to mitigate this risk, we define two hash-tables; one for conceptElements and one for propertyElements (used both for properties and references). Even in the event of changes to the EcoreLanguage, the implementation would detect the changes, and automatically update the hash-tables.

The next step consists in defining a recursive function that leverages the tree-like structure of XMI files and traverses the nodes, starting from the root node in the XMI file and branching to the leaves. First, we access the *get(Object key)* method of the conceptElements hash-table, which returns the value to which the key is mapped in this hash-table. Recall that the key in the conceptElements hash-table is the *index* of the concept, while the value, is the *name* of the concept. Depending on the meta-object that equals the returned value, the implementation outputs an XML file that follows the same template as an EcoreEMF metamodel.

Listing 1.3 details a code excerpt of the analyzeNode function. In lines 2 and 3 we have conditional statements that perform different computations, depending on the meta object that equals the returned value of the *get()* method. In the listing we only illustrate the "EPackage" and "EEnum", but the complete implementation also includes "EEnumLiteral", "EClass", "EReference", "EAttribute", "EAnnotation" and "EOperation" meta objects. Lines 4-15 provide more in-depth details regarding the computations that take place when the conditional statement that checks whether the returned value of the *get()* method equals "EEnum", evaluates to true. The first step in this computation consists in creating an eClassifier and adding the attribute that identifies an EEnum in XMI to it. Next, we iterate through the childNodes of the node we are currently analyzing. If the childNode's name is equal to "node", the childNode is passed as a parameter to the analyzeNode function, and the *appendChild()* method is used to append this childNode to the list of children of the node under analysis. Else, we access the *get()* method of the propertyElements hashtable, where we pass as a parameter the value of the childNode's role attribute. If the returned value equals to "name", then we set the value of the childNode's value attribute on the name attribute of the element.

Listing 1.3 analyzeNode Function

```

1 public static Node analyzeNode(Node concept) {
2     if (conceptElements.get(concept.
      getAttributes().getNamedItem(
      "concept").getNodeValue()).equals
      ("EPackage")){...}
3     else if (conceptElements.get(concept.
      getAttributes().
      getNamedItem("concept").getNodeValue()).
      equals("EEnum")) {
4         element = eclipseEcoreXML.
          createElement("eClassifiers");
5         element.setAttribute("xsi:type",
          "ecore:EEnum");
6         for( int i=1;i
          <concept.getChildNodes().
          getLength();i=i+2) {
7             if(concept.getChildNodes().
              item(i).getNodeName().
              equals("node")){
8                 element.appendChild
                  (analyzeNode(concept.
                    getChildNodes().item(i)));
9             }
10            else if (propertyElements.
              get(concept.getChildNodes().
              item(i).
              getAttributes().
              getNamedItem("role").
              getNodeValue()).equals("name"))
11                {
                  element.setAttribute("name",
                  concept.getChildNodes()
                  .item(i)
                  .getAttributes().
                  getNamedItem("value").
                  getNodeValue());
12            }
13        }
14        return element;
15    } ...
16 }

```

After all the nodes of the XMI file are visited, we run the transformations, which use the family.sandbox.mps file (corresponds to EcoreMPS.sandbox.mps) as input and generate the family.ecore file (corresponds to Gen EcoreEMF.ecore) described in Listing 1.4 as output.

Listing 1.4 XML Version of the Generated family.ecore

```

1 <eClassifiers xsi:type="ecore:EEnum"
      name="Gender">
2     <eLiterals name="Male"/>
3     <eLiterals name="Female" value="1"/>
4 </eClassifiers>

```

1.5 Evaluation

Discrepancies between the `EcoreLanguage` defined in MPS and its corresponding `Ecore` meta-metamodel in EMF, as well as faulty transformations from MPS to EMF, would hamper the exchange of `Ecore` metamodels from MPS to EMF. Thus, the definition of the `EcoreLanguage` and the implementation of the exchange transformations are determinant factors for enabling cross-platform modelling across EMF and MPS and, as such, it is of crucial importance to ensure their correctness. In an attempt to achieve the latter and separate eventual issues related to the implementation of the `EcoreLanguage` and the transformations, we carried out an evaluation process conceiving two major steps, as illustrated in Fig. 1.1.

Step 1 concerns the correctness of `EcoreLanguage`, which is validated via conceptual and structural comparison of `EcoreMPS` and `EcoreEMF` versions of a same metamodel. In case we identify inconsistencies between the two metamodels, `EcoreLanguage` is adjusted accordingly. The advantage of this evaluation step is two-fold. First, it assures that the `EcoreLanguage` is well-defined and the artefacts that are used as input to the transformations are correct. Second, validating the implementation in an iterative manner reduces time and effort in case of errors, as it facilitates the identification of the erroneous artefact. If the evaluation were only performed at the end of Step 2, it would be extremely challenging to identify the erroneous artefact (i.e., `EcoreLanguage` or transformations). Related to the example used in Sect. 1.4 (the `Family` metamodel), in this case we needed to compare the metamodel EMF definition in Listing 1.1 to the metamodel MPS definition in Listing 1.2. Both metamodels, include the `EEnum Gender` that contains two `ELiterals` (i.e., `Male` and `Female`) as children, thus we consider them equivalent, as they contain the same concepts and hierarchical structure.

Step 2 focuses on the correctness of the transformation implementation. For the transformations to be considered correct, the following conditions need to be fulfilled: (i) the `Gen EcoreEMF` metamodel should be correctly loaded in EMF, (ii) the XMI of `Gen EcoreEMF` metamodel and the XMI of `EcoreEMF` metamodel need to be equivalent, and (iii) the generated Java classes from the `genmodel` of each metamodel need to be equivalent. If any of these conditions is not fulfilled, the transformations need refinement. Considering the `Family` metamodel, we needed to compare Listing 1.1 to Listing 1.4. As it can be seen, both Listings contain the same `eClassifiers` and `eLiteral`, as well as the same structural hierarchy. The order of elements

and attributes might differ, but that does not affect the output, since the generated Java classes (implementing the metamodel in the modelling ecosystem as editors and resources) are the same for both metamodels. It is important to stress the fact that, although in the paper we leverage the sole `Family` metamodel for exemplification purposes, in the actual evaluation process it was only the simplest metamodel that we accounted for. Several metamodels (e.g., `SmartHome` and `Airport`), with varying complexity in terms of number of metaelements, were used for evaluation purposes too. The interested reader can download the implementation at <https://github.com/hilalsoft/exchanging> `ecore` model MPS as well as watch a demo of the solution at work at <https://play.mdh.se/media/t/04qpus1y0>.

1.6 Conclusions and Future Work

Blended modelling is expected to enhance the experience of diverse stakeholders interacting with modelling tools and among themselves, and to maximise their development's throughput. In this respect, this paper presented an approach to bridge two powerful (meta) modelling language workbenches, namely `JetBrains MPS` and `Eclipse Modelling Framework`. The aim was to provide a solution to support modelling activities by blending notations as provided by MPS and EMF. In fact, a metamodel conforming to `Ecore` could be (partially) defined in MPS and possibly used to derive languages and models in MPS, notably by leveraging its projectional editor features. At the same time, this work enables the automatic export of `Ecore`-based MPS metamodels into EMF; in this way other stakeholders could complete the metamodel in EMF as well as use modelling editors provided by Eclipse such as form-based, tree-based, or graphical editors. This work puts the basis for seamless bridging of MPS and EMF (modelling) language workbenches. However, there exists still work to be done that is planned as future research.

In particular, at now the bridging is partial: the backward export, from EMF to MPS, is not supported yet. This limits the collaboration between stakeholders, since metamodel edits made in EMF cannot be automatically mapped back to MPS. Another interesting aspect to investigate is the bridging of models. In fact, this feature would enable a smoother collaboration between stakeholders working with the two different modelling platforms taken into account.

Acknowledgment This work was supported by Vinnova through the ITEA3 BUMBLE project (nr. 18006) and the Knowledge Foundation through the SacSYS and HERO projects.

References

1. MPS Main Page., <https://www.jetbrains.com/mps/>. Accessed 28 Jan 2021
2. L. Addazi, F. Ciccozzi, Blended graphical and textual modelling for UML profiles: A proof-of-concept implementation and experiment. *J. Syst. Softw.* **175**, 110912 (2021)
3. J. B'ezivin, On the unification power of models. *Softw. Syst. Model.* **4**(2), 171–188 (2005)
4. F. Ciccozzi, D. Di Ruscio, I. Malavolta, P. Pelliccione, J. Tumova, Engineering the software of robotic systems, in *Proceedings of ICSE-C* (2017), pp. 507–508
5. F. Ciccozzi, R. Spalazzese, Mde4iot: Supporting the internet of things with model-driven engineering, in *Intelligent Distributed Computing* (2017), pp. 67–76
6. F. Ciccozzi, M. Tichy, H. Vangheluwe, D. Weyns, Blended modelling-what, why and how, in *Proceedings of MODELS-C. IEEE* (2019), pp. 425–430
7. E.M. Dave Steinberg, F. Budinsky, *Eclipse Modeling Framework* (Addison-Wesley Professional, Upper Saddle River, 2008), p. 12
8. J. El-khoury, O. Redell, M. Torngren, A tool integration platform for multi-disciplinary development, in *31st EUROMICRO Conference on Software Engineering and Advanced Applications* (2005), pp. 442–449
9. F. Heidenreich, J. Johannes, S. Karol, M. Seifert, C. Wende, Derivation and refinement of textual syntax for models, in *Model Driven Architecture – Foundations and Applications*, ed. by R. F. Paige, A. Hartman, A. Rensink, (Springer, Berlin/Heidelberg, 2009), pp. 114–129
10. V. Pech, *JetBrains MPS: Why Modern Language Workbenches Matter* (Springer, Cham, 2021), pp. 1–22
11. M. Scheidgen, Textual modelling embedded into graphical modelling, in *Model Driven Architecture – Foundations and Applications*, ed. by I. Schieferdecker, A. Hartman, (Springer, Berlin/Heidelberg, 2008), pp. 153–168
12. D.C. Schmidt, Guest editor's introduction: Model-driven engineering. *Computer* **39**(2), 25–31 (2006)
13. B. Selic, Model-driven development: Its essence and opportunities, in *Ninth IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC'06)* (2006), p. 7.
14. M. Wimmer, G. Kramler, Bridging grammarware and modelware, in *Proceedings of MoDELS* (Springer, 2005), pp. 159–168

A Conceptual Framework for Software Modeling of Automation Systems

2

Mohammad Ashjaei, Alessio Bucaioni, and Saad Mubeen

Abstract

In this paper, we propose a conceptual framework to facilitate the design and development of an automation system in which time-sensitive networking (TSN) is utilized for the backbone network and OPC UA is used for modeling of data exchange over TSN. As the configuration of OPC UA over TSN in a large automation setup can be a challenging task and requires specific expertise, we propose to add an abstract modeling layer that adopts the concepts of model-based development and component-based software engineering to facilitate the development of these systems. The proposed conceptual model can be automatically translated to the OPC UA modeling format. Such a modeling view will significantly reduce the complexity of OPC UA configurations, specially in large automation systems. Another benefit of the proposed framework is that the engineers, who do not have high levels of expertise in OPC UA, will be able to easily configure the OPC UA nodes in the automation system that utilize TSN for backbone communication.

Keywords

TSN · OPC-UA · Software modeling · Automation system · Industrial IoT

2.1 Introduction

The automation systems traditionally follow a hierarchical pyramid that was devised as a reference model [1] to structure

and develop these systems [2]. This model contains four layers, including the field layer or shop floor, control layer, supervisory layer, and management layer. The lowest layer in this model, the field layer, contains sensors and actuators. The next layer is the control layer that incorporates control units, e.g., programmable logic controllers (PLCs). The layer above the control layer provides supervisory control and data acquisition (SCADA) systems. Finally, the top-most layer supports management of these systems and consists of typical supervisory systems such as the manufacturing execution systems (MES). Figure 2.1 illustrates such a hierarchical design for an automation system.

As it can be observed in the automation pyramid, data exchange between the layers are an essential element to construct an automation system, where the medium to exchange the data is a network. There is a plethora of industrial network protocols that are customized based on the needs of each layer in the pyramid. For instance, the communication in the field layer is supported by industrial communication systems such as PROFINET,¹ EtherCAT,² and POWERLINK.³ On the other hand, the communication in the higher layers is supported by the standard switched Ethernet over TCP/IP protocol as there are no strict timing requirements specified in those layers.

The drastic acceptance of the Industrial Internet of Things (IIoT) paradigm in the past few years has also inspired how the automation systems are modeled. This paradigm proposes a flat model in contrast to the automation pyramid model. The IIoT model brings several benefits, including limiting the number of gateways that are used between different network protocols as well as unifying the communication among all devices and systems. This, in turn, reduces the complexity of the network configuration and management.

M. Ashjaei (✉) · A. Bucaioni · S. Mubeen
Mälardalen University, Västerås, Sweden
e-mail: mohammad.ashjaei@mdu.se; alessio.bucaioni@mdu.se;
saad.mubeen@mdu.se

¹<https://www.profinet.com/technology/profinet>.

²<https://www.ethercat.org/>.

³<https://www.ethernet-powerlink.org/>.

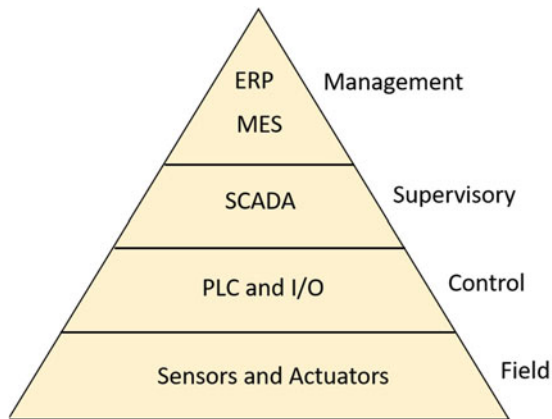


Fig. 2.1 The automation pyramid

The IEEE Time-Sensitive Networking (TSN) standards [3] have emerged as an attractive backbone network solution to establish the IIoT paradigm. Besides the TSN-based networks, Open Platform Communication-Unified Architecture (OPC UA)⁴ has evolved over the years as a data modeling approach to facilitate the data exchange in factory communication. Therefore, several works have addressed the utilization of OPC UA over TSN to mitigate the network management in the context of IIoT [4]. Nevertheless, (re-)configuration of OPC UA models over a large TSN network requires a special expertise as the OPC UA specifications are massive.

The main contribution of this paper is to propose a conceptual framework for designing network communication for an automation system. The communication is based on the TSN technology as the backbone network together with the OPC UA data modeling approach to enhance the data exchange over the network. First, we propose the conceptual network infrastructure. Then we present a conceptual modeling approach to facilitate the utilization of OPC UA and mitigate the design complexity of automation systems.

2.2 Technologies Comprising the Framework

This section presents various technologies that are used for the information exchange in automation systems. Furthermore, the section discusses the languages and models that can be used for model- and component-based software development of automation systems.

2.2.1 Time-Sensitive Networking

The focus of this paper is on a set of standards that provide different features to exchange data-intensive information with low latency over the network. These standards are currently considered for the backbone communication in many domains [5]. Among the features of TSN, we present different scheduling policies, including credit-based shaper, time-aware shaper and preemption, which have been included in the IEEE 802.1Q-2018 standard [3]. A *credit-based shaper* (CBS) algorithm is defined to prevent the transmission of traffic bursts. This algorithm applies to a set of traffic classes, known as stream reservation (SR) classes, by allocating a credit to each class. There can be up to 7 traffic classes in a network. An SR traffic class transmission should be done only if there is zero or positive credit is available, otherwise the traffic should wait in a queue while the credit increases. During the transmission the credit decreases with a constant rate, known as *sendSlope*. The rate of credit increase is constant, known as the *idleSlope* that can be configured by the network administrator.

The TSN standards also propose enhancements to provide temporal isolation for *scheduled traffic*. The scheduled traffic are planned offline, i.e., their transmission time slots are configured during the design time (before running the network). In particular, *transmission gates* are associated with each queue of a switch port, and transmission from a queue is only allowed if the relevant gate is open. Through the gate mechanism we can prevent the transmission of a traffic class in favor of an urgent traffic class. The standards also support frame preemption [6]. According to this mechanism each traffic class can be set as *express* or *preemptable* traffic classes. Express traffic can preempt the preemptable traffic, but it cannot be preempted itself. The preemptable traffic cannot preempt other classes. Preemption support can be combined with the CBS and the gate mechanisms.

2.2.2 OPC UA

OPC UA is a higher-layer communication protocol to securely exchange information between two devices. The lower-layer communication protocols can be Ethernet-based networks or a wireless communication channel. The communication within the OPC UA is based on the client-server method in which the server contains a set of information and the client sends requests to retrieve the

⁴<https://opcfoundation.org/>.

information from the server. The information is structured as an object model, known as the *node* that can be defined as three types, being a variable, a method or an event. A variable type can store a value with different attributes, while the event type is used to transmit an event that occurs in the system. The method represents a function call in the OPC UA node that returns after its execution is completed. It can be used to perform a specific task on a data before its transmission. A set of nodes that a server contains is referred to as an address space [7]. A set of services are defined in order to get access to the information residing in a node. For instance, in order to access the attributes of a node, the *read/write* services are used.

Recently, a new update of OPC UA is proposed in which the data exchange model is based on the publisher-subscriber method, which is complimentary to the client-server method. Within the OPC UA specification, the new version is known as the *PubSub* model [8]. In this model, a publisher is responsible to provide a data set that contains the information and the subscriber will receive the data that is subscribed for.

2.2.3 Languages and Models for Software Development

The paradigms of Component-Based Software Engineering (CBSE) [9] and Model-Based Engineering (MBE) [10] have been effectively employed for the development of software systems in various domains such as automotive [11, 12]. For instance, in the automotive domain, there are several modeling languages [5] that support modeling various types of networks ranging from low-bandwidth networks [13, 14], e.g., Controller Area Network to high-bandwidth networks [15–17], e.g., Switched Ethernet. However, such a level of support for model- and component-based software development, especially at various abstraction levels during the development, is missing from the automation domain.

The IEC 61131-3 standard [18] is proposed by the International Electrotechnical Commission (IEC) for the development of software on Programmable logic controllers (PLC) in the automation domain. This standard defines a set of languages for modeling of automation systems. These languages are available both in textual and graphical format. The standard consists of two parts: (i) common elements, and (ii) programming languages. The first part provides definitions for data types, variables, resources, run-time entities and internal organization of the software. The second part provides four languages for the software development on PLCs: (i) Instruction List (IL), (ii) Structured Text (ST), (iii) Ladder Diagram (LD), and (iv) Function Block Diagram (FBD). In addition to these languages, a state machine definition language, called the Sequential Function Chart (SFC) is supported by the standard. Among these languages, FBD is a

graphical language that is an attractive candidate that can be incorporated into the model- and component-based software development paradigm for automation systems.

2.3 Proposed Framework

In this section, we present the proposed conceptual framework for model- and component-based software development of automation systems that utilize TSN as the backbone network and OPC UA as the application layer communication over TSN.

2.3.1 The Conceptual Communication Infrastructure

An automation system consists of two domains with different requirements. The field layer that contains the field devices, such as sensors, actuators, controllers, robots and even mobile robots are part of a domain commonly known as the Operation Technology (OT) domain. In the OT domain, there are hard real-time and safety-critical requirements that should be taken into account during the design phase. For instance, the communication between the sensors and controllers should be done within the milliseconds range. Moreover, the information should be delivered within a specified deadline to ensure that the timing requirements specified on the system are satisfied. Therefore, industrial networks are usually used in the OT domain to fulfill such strict requirements, such as PROFINET and EtherCAT.

On the other hand, the higher layers of automation systems consist of several management systems, data collection and acquisition systems as well as statistical analysis systems. These systems, including MES and SCADA systems, help to assure the continuous quality of the production in manufacturing. At this layer, which is commonly known as the Information Technology (IT) domain, the timing and safety requirements specified on the system are less stringent. For example, the response times of the monitoring systems and management systems are often in the range from few seconds to few minutes. In the case of using cloud computing for manufacturing, due to the nature of the cloud connections via the Internet, the requirements on the timing are more relaxed. The networks, thus, do not need to be industrially-approved networks and commonly TCP/IP over standard switched Ethernet is used for the communication.

As the timing and safety requirements in different domains of an automation system are different, the seamless utilization of various network technologies, including industrial Ethernet, wireless, and TCP/IP Ethernet, and the resource management of the entire network becomes a daunting task. Any changes in the system require a full configuration of

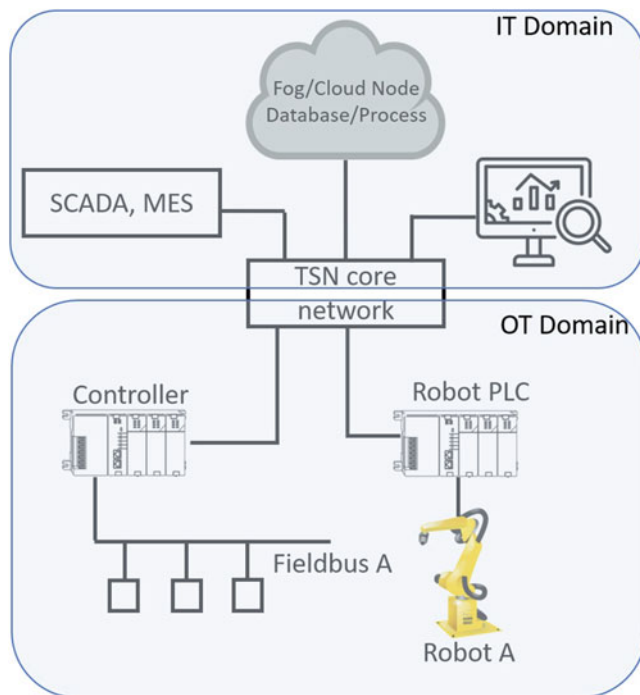


Fig. 2.2 The TSN-based network concept for automation systems

the network, which can affect the other parts of the network. Therefore, a desired solution would be to consolidate the network protocols and unify them with a suitable network technology. Recently, due to the advancement in the TSN standardization, it is considered as a promising solution that can support both time-critical and non-critical traffic transmissions in the same network while preserving the timing requirements of both domains. Figure 2.2 illustrates such a design in which the TSN switches are used as a backbone network where the devices from the IT and OT domains can be connected to each other. The TSN switches can support time-critical transmission using time-aware shaper and frame preemption, while credit-based shaper and best-effort transmission can be used for transmission of low-critical traffic in the IT domain.

Given the features of TSN and support for both domains, TSN technology can be a promising solution to unify the network infrastructure of automation and in turn mitigate the complexity of the network management. However, designing such a network in particular considering the legacy networks and legacy computers/devices in the automation systems is a challenging task. There are several solutions to overcome this complexity by providing TSN configuration tools [19], predictability analysis tools [20], and legacy support tools [21], yet the optimum configuration utilizing various TSN features is a non-trivial task. One of the recent solutions to reduce the configuration complexity is to use OPC UA as the application layer of the network for management of structured data

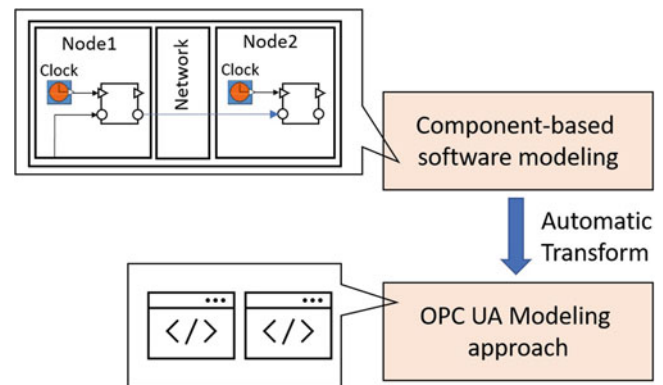


Fig. 2.3 The overall approach

exchange. Most of the proposed works address utilization of OPC UA over TSN for the OT domain, e.g., [22], while OPC UA can be used in a holistic view for both domains, including communication to the fog/cloud and the management systems. Based on the proposed conceptual communication infrastructure, several control functions in PLCs or field-level controllers can be moved to the cloud as the network can support low-latency transmission, which in turn reduces the utilization of PLCs and controllers. In this regard, some existing works provide promising results [23]. However, these works do not utilize TSN and OPC UA.

2.3.2 Holistic Modeling Approach

As the configuration of OPC UA over TSN can be extremely complex and requires special expertise, we propose a component-based software modeling approach that can be translated automatically to the OPC UA modeling format. Figure 2.3 illustrates the overall approach in which the OPC UA nodes and the communication are modeled at a high level of abstraction without knowing the low-level information of the system. After verifying the correctness of the model at this abstraction level, an automatic translation will be performed to obtain the OPC UA models.

We propose a hierarchical modeling approach for the component-based software modeling with essential components to model automation systems that utilize TSN network for backbone communication while the computing nodes are supported by OPC UA client and servers. Figure 2.4 depicts such a hierarchical modeling approach. The top level element is the automation system model that encapsulates the components of both IT and OT domains. For example, the system can be a control system that resides in the cloud and receives the status data from the robot's PLC via the TSN network. The system contains one or more models of applications and networks. The application represents the OPC UA client or

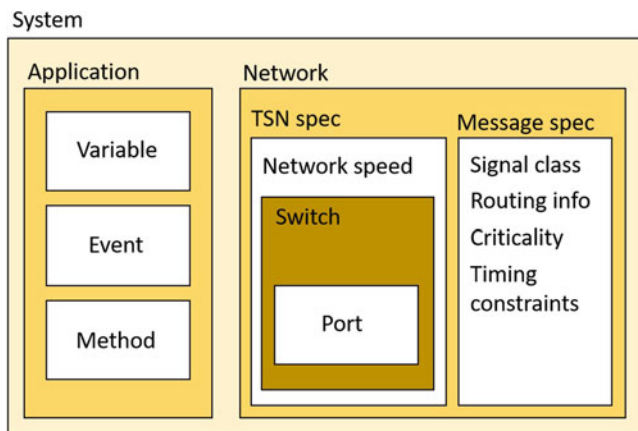


Fig. 2.4 Proposed hierarchical model of automation systems that utilize OPC UA over TSN

server application that is installed on a computing unit, e.g., a server installed on the robot's PLC. As it is discussed before, each application may have OPC UA nodes with different types, such as variable, event or method, each of which can have several attributes. Therefore, the properties can be set within the respective elements. The network element contains two sub-elements of TSN and Message specifications. The TSN network speed and switch parameters are configured within the TSN network specification. Each port of a switch may have different configuration, which needs its own modeling element. Within the message element, several attributes, such as signal class, routing information, and timing constraints are specified. In order to classify the signal classes and network specification, we follow the IEEE/IEC 60802 TSN Profile for Industrial Automation, which identifies several message classes including synchronized critical, non-synchronized critical, and non-critical classes.

2.4 Conclusion and Future Work

Utilizing time-sensitive networking (TSN) to unify the network protocols within an automation system is a promising approach as it can be used for a backbone network connecting IT and OT domains and supporting their requirements. The OPC UA modeling approach can be used to facilitate the data exchange between various devices and systems in automation. However, the OPC UA configuration needs specific expertise and skill level as the model is very vast and includes many sophisticated data models. In order to simplify the utilization of OPC UA over TSN in automation systems, in this paper, we proposed a conceptual framework for the development of these systems. The proposed framework leverages the principles of model-based development and

component-based software engineering to develop simplified models of automation systems utilizing TSN for backbone network communication. These models can then be automatically translated to the OPC UA modeling format, thereby offloading the burden of understanding low-level details of OPC UA by the system developers.

Acknowledgments The work in this paper is supported by the Swedish Governmental Agency for Innovation Systems (VINNOVA) via the PANORAMA, DESTINE, PROVIDENT and INTERCONNECT projects, and the Swedish Knowledge Foundation via the FIESTA, HERO and DPAC projects. We thank our industrial partners, especially Arcticus Systems, Volvo CE and HIAB.

References

1. A. Bucaioni, P. Pelliccione, Technical architectures for automotive systems, in *2020 IEEE International Conference on Software Architecture (ICSA)* (IEEE, Piscataway, 2020), pp. 46–57
2. T. Sauter, S. Soucek, W. Kastner, D. Dietrich, The evolution of factory and building automation. *IEEE Ind. Electron. Mag.* **5**(3), 35–48 (2011)
3. IEEE, IEEE standard for local and metropolitan area network-bridges and bridged networks. *IEEE Std 802.1Q-2018* (Revision of IEEE Std 802.1Q-2014), July 2018, pp. 1–1993
4. D. Bruckner, M.-P. Stănică, R. Blair, S. Schriegel, S. Kehrer, M. Seewald, T. Sauter, An introduction to OPC UA TSN for industrial communication systems. *Proc. IEEE* **107**(6), 1121–1131 (2019)
5. M. Ashjaei, L. Lo Bello, M. Daneshalab, G. Patti, S. Saponara, S. Mubeen, Time-sensitive networking in automotive embedded systems: state of the art and research opportunities. *J. Syst. Archit.* **117**, 102137 (2021)
6. M. Ashjaei, M. Sjödin, S. Mubeen, A novel frame preemption model in TSN networks. *J. Syst. Archit.* **116**, 102037 (2021)
7. M. DammStefan, H. Leitner, W. Mahnke, *OPC Unified Architecture* (Springer, Berlin, 2009)
8. J. Pfrommer, A. Ebner, S. Ravikumar, B. Karunakaran, Open source OPC UA pubsub over TSN for realtime industrial communication, in *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1 (2018), pp. 1087–1090
9. G.T. Heineman, W.T. Council, *Component-Based Software Engineering, Putting the Pieces Together*, vol. 5 (Addison-Wesley, Boston, 2001)
10. D.C. Schmidt, Model-driven engineering. *Comput. IEEE Comput. Soc.* **39**(2), 25 (2006)
11. A. Bucaioni, S. Mubeen, J. Lundbäck, K.-L. Lundbäck, J. Mäki-Turja, M. Sjödin, From modeling to deployment of component-based vehicular distributed real-time systems, in *2014 11th International Conference on Information Technology: New Generations* (IEEE, Piscataway, 2014), pp. 649–654
12. A. Bucaioni, S. Mubeen, F. Ciccozzi, A. Cicchetti, M. Sjödin, Modelling multi-criticality vehicular software systems: evolution of an industrial component model. *Softw. Syst. Model.* **19**(5), 1283–1302 (2020)
13. S. Mubeen, J. Mäki-Turja, M. Sjödin, Support for end-to-end response-time and delay analysis in the industrial tool suite: issues, experiences and a case study. *Comput. Sci. Inf. Syst. J.* **10**, 453–482 (2013)

14. S. Mubeen, H.B. Lawson, J. Lundbäck, M. Gålnander, K. Lundbäck, Provisioning of predictable embedded software in the vehicle industry: the rubus approach, in *2017 IEEE/ACM 4th International Workshop on Software Engineering Research and Industrial Practice (SER IP)* (2017)
15. M. Ashjaei, S. Mubeen, J. Lundbäck, M. Gålnander, K. Lundbäck, T. Nolte, Modeling and timing analysis of vehicle functions distributed over switched ethernet, in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, Oct 2017, pp. 8419–8424
16. S. Mubeen, M. Ashjaei, M. Sjödin, Holistic modeling of time sensitive networking in component-based vehicular embedded systems, in *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (2019), pp. 131–139
17. R. Henia, A. Hamann, M. Jersak, R. Racu, K. Richter, R. Ernst, System level performance analysis - the symTA/S approach. *IEE Proc. Comput. Digit. Tech.* **152**(2), 148–166 (2005)
18. IEC 61131-3, International Electrotechnical Commission Standard, 2003
19. B. Houtan, A. Bergström, M. Ashjaei, M. Daneshtalab, M. Sjödin, S. Mubeen, An automated configuration framework for TSN networks, in *22nd IEEE International Conference on Industrial Technology (ICIT'21)*, March 2021. <http://www.es.mdh.se/publications/6117->
20. L.L. Bello, M. Ashjaei, G. Patti, M. Behnam, Schedulability analysis of time-sensitive networks with scheduled traffic and preemption support. *J. Parallel Distrib. Comput.* **144**, 153–171, June 2020. <http://www.es.mdh.se/publications/5835->
21. D.B. Mateu, M. Ashjaei, A. Papadopoulos, J. Proenza, T. Nolte, Letra: mapping legacy ethernet-based traffic into TSN traffic classes, in *26th IEEE International Conference on Emerging Technologies and Factory Automation*, Sept 2021. <http://www.es.mdh.se/publications/6243->
22. C. Eymüller, J. Hanke, A. Hoffmann, M. Kugelmann, W. Reif, Real-time capable OPC-UA programs over TSN for distributed industrial control, in *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1 (2020), pp. 278–285
23. S. Mubeen, P. Nikolaidis, A. Didic, H. Pei-Breivold, K. Sandström, M. Behnam, Delay mitigation in offloaded cloud controllers in industrial IOT. *IEEE Access* **5**, 4418–4430 (2017)

Virtual Reality Multiplayer Interaction and Medical Patient Handoff Training and Assessment

Christopher Lewis, Daniel Enriquez, Lucas Calabrese, Yifan Zhang, Steven J. Anbro, Ramona A. Houmanfar, Laura H. Crosswell, Michelle J. Rebaleati, Luka A. Starmer, and Frederick C. Harris, Jr.

Abstract

Virtual worlds have the potential to mirror many aspects of real life. Immersive virtual worlds constructed through the use of Virtual Reality (VR) are useful in simulating the technology, equipment, and practices of many different fields. In the medical field, VR can be heavily relied upon to circumvent a wide variety of tools, human resources, and other objects that may be limited or difficult to procure at any given time. As a result, the goal of this research was to develop a low-stakes virtual environment (VE) in which medical students could practice developing skills necessary to their profession. As such, this environment needed to mirror, as closely as possible, an environment the medical students would see frequently during their practice. The result of this work is an application for use in patient handoffs, a situation where patient care is transferred from one medical professional to another. In order to achieve this, this work created a multiplayer VR environment with an immersive virtual world simulating standardized patient rooms and standard mediums

of communication and interaction between users. While doing so, a framework was developed as the need for VR multiplayer, VR with voice communication, and a VR interaction system was seen to be needed for future VR multiplayer applications. This framework can be used to construct more applications for communication fueled environments, like the patient handoff.

Keywords

Virtual reality · Multiplayer · Framework · Software · Simulation · Eye tracking · Photogrammetry

C. Lewis · D. Enriquez · L. Calabrese · Y. Zhang · F. C. Harris, Jr. (✉)
Computer Science and Engineering, University of Nevada, Reno,
Reno, NV, USA
e-mail: christopher_je1@nevada.unr.edu; denriquez@nevada.unr.edu;
lcalabrese2018@gmail.com; yfzhang@nevada.unr.edu;
Fred.Harris@cse.unr.edu

S. J. Anbro · R. A. Houmanfar
Behavior Analysis Program, University of Nevada, Reno, Reno, NV,
USA
e-mail: stevenanbro@nevada.unr.edu; ramonah@unr.edu

L. H. Crosswell
Reynolds School of Journalism, University of Nevada, Reno, Reno,
NV, USA
e-mail: lcrosswell@unr.edu

M. J. Rebaleati · L. A. Starmer
@Reality, University of Nevada, Reno, Reno, NV, USA
e-mail: mrebaleati@unr.edu; lstarmer@unr.edu

3.1 Introduction

The medical field, and those who work in it, can be considered the backbone of a modern society. This is evident in normal life, but overly obvious in a global pandemic. A task often taking place in any hospital around the world is the patient handoff. The patient handoff is the transfer of patient care from one medical professional to another. Patient handoffs happen at a very high frequency, often occurring multiple times per patient, per day. It is estimated that the most serious adverse events in hospitals are caused by communication failures, mainly to do with patient handoff errors [1]. These errors can cause very large problems to both patients and medical professionals. Through Virtual Reality (VR), this work introduces a tool to streamline and standardize patient handoff training, a step forward in addressing these errors.

VR is rising in popularity throughout the world in many different sectors. Major sectors innovating with VR are the education and medical sectors. In fact, there has been major growth in the use and innovation of VR for training medical professionals and for healthcare in general. VR provides these sectors, and many other sectors, with engaging and

immersive educational scenarios that help reinforce taught behaviors in a kinesthetic way. These kinesthetic learning methods help to engage the user, which cements the lesson into their memories by providing multiple learning styles for the user to learn from [2].

3.2 Background and Related Work

Patient handoffs happen very regularly across any hospital, making the handoff one of the most pivotal and important actions medical personnel should learn to do effectively. Yet, handoffs have the potential to cause a large amount of harm due to their somewhat regular missteps and miscommunications [3]. This highlights the need for greater care and plentiful innovation in the education and training of medical professionals, particularly in the skill of patient handoff. The SBAR is an effective innovation in the communication and standardization of handoffs. It has been shown that the SBAR has had an incredibly positive impact in the areas it affects [4].

With the improvement in the structure of the handoff that SBAR creates, innovations have also taken place around the education or implementation of handoffs. One such innovation is with the user study this application was built for [5].

3.2.1 Medical Training Simulators

Innovations in medical training can have wide reaching impacts for the overall medical community. One such innovation is within the use of haptics for medical training simulators. Coles, Meglan, and John in [6] cover the use of haptics for training in simulators through various scenarios, such as: palpation, needle insertion, laparoscopy, endoscopy, and arthroscopy. Overall, the authors found that surgical simulations incorporating haptic feedback provides a richer training experience, compared to the simulations that don't.

VR based medical training simulators have also come to fruition. One such medical training simulator is for periodontal training [7]. The authors put haptic feedback into a VR simulator to allow the user to fully experience working on teeth. This is ultimately a safer approach since it doesn't involve a patient or real teeth. The authors also found, through various studies and expert interviews, that their work provided a realism that could serve as a, "useful instruction tool with high teaching potential" [7].

3.2.2 VR in Medicine

VR is seeing more prevalent use in the medical industry as an alternative for training applications for surgeries. New technologies, such as the simulations made by ArthroSim

[8] and ORamaVR [9], are some of the most innovative VR surgical simulators.

ArthroSim [8] is a VR knee and shoulder arthroscopy simulator. ArthroSim uses a dedicated machine for high precision surgery training. It uses a dual-monitor display that reflects the training simulator task and data, as well as the simulated imagery of a camera used for these surgeries. This machine focuses on precision due to its dedicated machinery and its use of haptic feedback in training. ArthroSim also focuses on realism, where the precise use of haptic feedback would be similar to physically performing surgery on a person. The simulated images of the camera also use accurate anatomy to reflect where the user is currently working. This technology provides a much safer and efficient solution to training surgeons, opposed to letting surgeons learn on mannequins or on patients.

OramaVR [9] is a general purpose VR medical training simulator with multiplayer capabilities. One of the largest benefits of using this system is its reliance on mass-produced VR head mounted displays (HMD). Not only are these HMDs much cheaper than a dedicated machine, but they can generally be used interchangeably with other varieties of HMDs. This allows the application to be used by a large variety of users and to be flexible for new technology and innovations within VR technology.

3.2.3 Multiplayer VR Simulators

Multiplayer VR is a relatively new technology. Although some applications, mostly video games, use multiplayer VR as a form of entertainment, not much research has been done surrounding it and its intrinsic problems. One such intrinsic problem with VR that has been studied is room-scale multiplayer VR. Sra in [10] addresses the use of a Galvanic Vestibular Stimulation system to allow the user to avoid collisions with their environment while not breaking the user's immersion. This technology looks incredibly promising for the future of VR for commercial use, as not every person has a large room to navigate in.

Multiplayer VR can also have very interesting and useful applications. One such application is in the world of construction. Du, Shi, Mei, Quarles, and Yan in [11] detail the use of multiplayer VR for building walkthroughs. A building walkthrough is a tour of a building, often before it is built, through the use of simulations, models, and CGI images. Building walkthroughs using multiplayer VR allowed this project's stakeholders to view a 3D virtual model of a building that they otherwise couldn't experience aside from inside a less interactive medium. It is also important to note that this application also allowed the users to verbally interact with each other. Due to the immersive nature of VR applications, the users of this program would also be able to experience

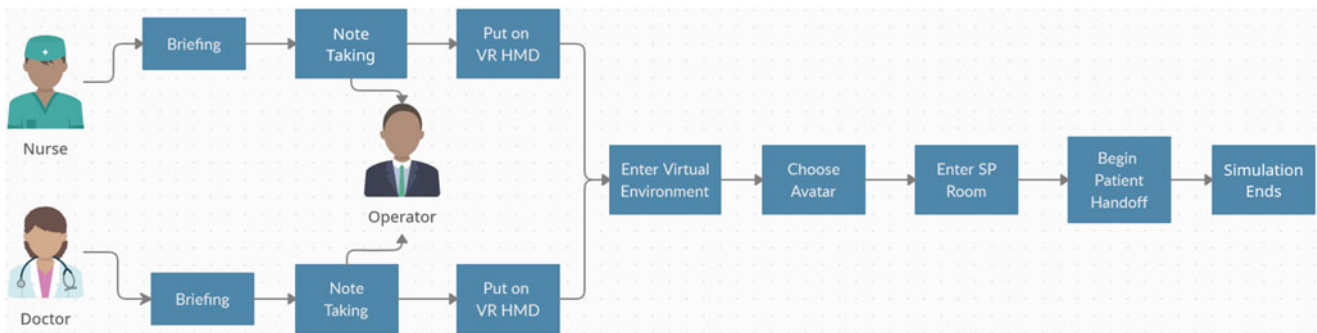


Fig. 3.1 The execution order of the application in use during the user study

the building in the exact same scale as their avatars, causing a more robust and informative walkthrough.

3.3 Implementation

3.3.1 Process of Execution

As described in Fig. 3.1, before the program is started two participants were briefed separately as to the current condition of a unique simulated patient. Each participant wrote down any information they chose about the patient to help convey the information needed during the patient handoff. Each of these files are given to the operators of the program to be uploaded into the program.

Once the participants have been properly worn the HMD, the simulation will start. Both participants are then put into the VR world together. They each start in the same room as seen in Fig. 3.2. This room allows the participants to choose their avatar and appearance freely. Once either participant is ready, they can point and press their remotes at a sign that will teleport them into the standardized patient room as seen in Fig. 3.3.

When both participants had teleported to the standardized patient room, they could start the patient handoff for each patient. Each participant had the notes they made about the patient displayed on their left hands in the form of a clipboard. This was done so that they could glance at it and help themselves complete their patient handoff using the written information, like in a physical working environment.

Finally, when both participants completed their patient handoffs, the simulation ended and they took off their HMDs.

3.3.2 Eye Tracking

For this work the Vive Pro Eye was used for eye tracking through the interface of HTC's SRanipal SDK [12]. The observer and the recording could observe timestamps and location data of when and where the user was looking throughout the application. This data could be used to determine where



Fig. 3.2 A participant configuring their avatar in the starting room



Fig. 3.3 The standardized patient room put into VR via photogrammetry

in their handoff the users were engaging with the other user, looking at the patient as seen in Fig. 3.4, or getting distracted by the environment.

3.3.3 Recording

To monitor user experience in this application, this work recorded the movements, communications, and interactions using RockVR video capture [13]. This video capturing is recording the exact screen that the observer views, including eye tracking lines, and audio from both participants. The



Fig. 3.4 The standardized patient room put from the perspective of the observer with eye tracking visible



Fig. 3.5 The overseer station used in the user study



Fig. 3.6 A participant engaged in the user study with an operator in the background

physical room during the user study [5] was also being recorded using an in-room camera that connected to an overseer station as seen in Fig. 3.5. Lastly, some participants were recorded in a more close-up manner, as seen in Fig. 3.6.

3.4 Room Creation

The VR patient room was a very accurate approximation of a standardized patient room. This room was created to have the

participants in an immersive space that is, very likely, familiar to them. To generate this room, this project needed to do three tasks accurately. First, the appearance of the room needed to emulate a standard patient room. Second, the room had to feel approximately the same size as it would in real life. Third, the appearance of the objects, the equipment and the dummy needed to appear as similar to real life as possible, with training equipment mirroring standard tools that are used in medical practice. Both the first and second tasks were done using a technique called photogrammetry, while the last task needed to be done, largely, manually.

3.4.1 Photogrammetry

Photogrammetry, broadly, is the technique of finding the correct size of virtual objects that correlates with the size and spacing of the physical objects they are associated with. The entire room, and the full body mannequin, are both generated using photogrammetry for 3D modeling. These models were exceedingly large in terms of the storage space they required on the computers. These models were so large, in fact, that Unity [14] struggled to render them. Thus, we had to lower the polygon count of the objects. To do this, every wall was replaced with larger flat polygons, rather than a large amount of small polygons. Other parts of the room were then cleaned to remove gaps and other wrong textures, all while replacing these features with larger polygons that would take up less rendering time. This removes some of the finer resolution of this technique, but ultimately doesn't change the appearance of the room to any significant degree.

3.4.2 Room Objects

A few of the room objects were not available to have photogrammetry done to them. This includes the computer monitor, the cart of medical equipment, and the IV bag and trolley. All of these room objects were instead created using computer-generated imagery (CGI). Each of these elements had to be placed and sized directly emulating the room's photogrammetry. This created a small amount of scaling issues, as the objects couldn't be exactly the correct scale as we had no real life object to compare to the virtual room. All room objects can be seen in Fig. 3.7.

3.5 Interaction

3.5.1 Avatars

The participants had the ability to change the avatar of their own personal character. This allows them to change their own

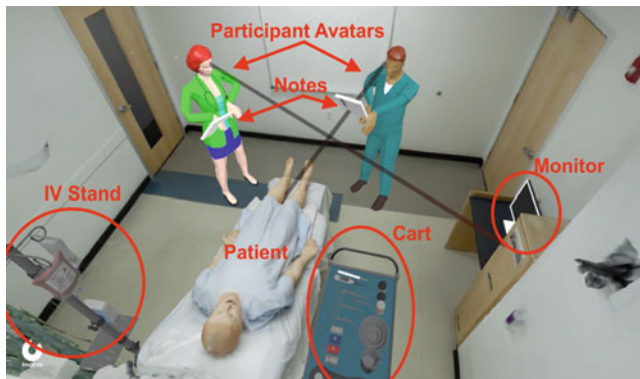


Fig. 3.7 The room objects that were made using CGI and the patient mannequin

virtual appearance to their liking to either fit their role or fit their own personal character. The participant can change to either wear a medical smock or to wear scrubs to fit their role of a doctor or a nurse. These outfits had a variety of different options associated with them as seen in Fig. 3.2. They could also change their skin color, gender, and hair color accordingly to whatever they deemed fit.

3.5.2 Voice and Audio

The participants are required to communicate verbally to each other in order to perform the patient handoff. This includes both producing audio and receiving audio throughout the application. This was accomplished using the built-in microphone and speakers found in the HMD. As a result, this application included the ability to communicate with one another via voice chat. Due to this voice chat integration, both participants can execute the patient handoff in almost exactly the same ways they would in the real world.

3.5.3 Clipboard

Each of the participants were briefed before entering the virtual world on a particular scenario of a patient's current status. Each of the participants were then instructed to write down everything they wished to convey during the patient handoff. These notes were then transferred into the application and displayed on a clipboard in their left hand to emulate a typical medical setting. These notes were only available to be seen by the person who wrote them.

3.6 Framework

3.6.1 Interactables

In many instances of multiplayer VR, some form of visible interaction can occur between objects and/or users that all

users should be able to see. In this framework, interactable objects in the shape of spheres were made that could be picked up and thrown that the other player would be able to see move and interact with the environment. If a player were to pick up the object by moving to the object, putting their controller into the object, then pressing the trigger, the object would then mirror the movements of their controller. Upon release of the trigger, the object would be thrown in the direction and speed that mimics the direction and speed of the controller.

These types of interactions require data transfers to be broadcast between all of the users involved in that instance of VR. This is to ensure the concurrency and accuracy of these interactions on all user instances. To accomplish this in the multiplayer framework, a client to server system was implemented. A client to server system has a server that communicates between the players and gets the inputs necessary for program functions such as movement and interaction updates and it shares it to all users in that instance of VR. The server also stores important information such as synchronized objects and the active scene. As a result of this server model, if a user were to grab an object, the object would exist on the server's independent storage of the object. This means that if the user were to move the object, the user needs to communicate with the server that they have grabbed the object and have moved it. The server would then broadcast to the user holding the object and all other users that the object is being moved. This handshake is very slow, and the movement of the object would appear choppy and inaccurate. However, to counteract this problem, our framework has the ownership of the object swap from the server to the player in contact with the object. This means the server is no longer keeping track of the object, but is getting frequent updates by the user holding the object as to the position, rotation, and velocity of the object. This means that there would be no latency in moving the object for the user holding the object. When the interactable is then released by the user, the server takes control of it again, causing the updates to only be broadcast from the server and thus lowering latency between the users.

It is important to note that while the user study only allowed two users into the VE at any given time, this framework allows for many users to collaborate. The varying factors for precisely how many users are: the number of assets used in the game, the internet speed of the users and server, the power of the user's computer (largely GPU based), and the number of assets set up for networking/interaction.

3.6.2 Voice

Voice communications across users are synchronized to the other player. This means that if one player speaks the other player is able to hear the audio. In this work, the medium of voice communication was done through Photon PUN Voice

chat [15], with the microphones being connected to the HMD as well as the audio from the other person coming through the speakers of the HMD. Voices were also being played and recorded on the observer's computer.

3.7 Conclusions

VR is seeing more implementation in training and education than ever before, particularly inside the field of medicine. With medicine being one of the most important parts of civilization, it is critical to improve the training of common sources of error and miscommunications, like the patient handoff. The work presented is an application where the standardization of patient handoff training could be effective in achieving reduced errors and reduced miscommunication.

This work demonstrated a framework for user communication and user interaction in multiuser VR environments. It also created an application specific to typical medical environments and training in a particular medical scenario. This application was tailored to a standardized patient room and attempted to mirror common interactions and points of communication for evaluation of communication such as voice recording, video recording, and eye tracking. As a result of this application, a framework was developed to streamline the process of these multiplayer interactions that could allow for future development for VR multiplayer interaction applications.

3.8 Future Work

3.8.1 VR Cross-compatibility

As a result of this application being made to be compatible with the HTC Vive Pro Eye's eye tracking software [16], the application doesn't work with other HMDs. This means that the program is not compatible with other popular VR hardware, such as: Oculus Rift/Quest [17], Vive Wave [18], mobile VR, etc. A future iteration on this work could allow for the inclusion of different mediums of VR. The removal of a hardware requirement would allow accessibility to people who own a different VR headset. This would, however, disable eye tracking for the HMDs that couldn't support it.

3.8.2 Avatar Eyeball Movement

As a result of eye tracking being a prominent feature of this application, it could be beneficial to share the eye movement of the other participant. A situation in which this could be useful is to experiment with the level of eye contact occurring in the hand off. Eye contact is crucial for flawless commu-

nication to occur, so allowing other participants to see what another participant is looking at could be very impactful.

3.8.3 Avatar Lip Tracking

For communication to occur smoothly in VR, increased immersion is very important. The more immersion the participants experience, the more seriously they would take the simulation, which would cause communication to become more accurate. If a participant is able to see another participant's mouth moving, it might increase the immersion of that participant. An example of technology that could be used is to implement Vive's facial tracking technology and synchronize facial expressions to the other player over the network [19].

3.8.4 Object Interaction

In medical environments, it could be useful to have the medical equipment synchronized to other players. This would allow a user to show an application of specific equipment on an object to other users, how to use a stethoscope for example. It would also be more realistic to see objects, that other people are using, move and interact similar to the real world. This may improve the immersion of the users which may improve the effectiveness of the training.

3.8.5 Notepad Improvements

The virtual notepad that is being held by the participant may see some utility to share notes if these notes were synchronized across different users. This means that it may be useful for participants to share their notes/objectives to the other users. Allowing the user to dynamically create their own notes is also an important consideration, as creating your own notes is an important function for effective communication. Current note taking technology in VR is limited, as it is difficult to precisely write things down without a physical pen-shaped object to write with. This is an issue with the precision and maneuverability of the standard VR controllers. It is also difficult to type in VR due to the fact that the user is generally standing up or moving around the environment. Standard controllers have issues with precision and speed for typing in VR on a virtual keyboard.

Acknowledgments We would like to acknowledge the UNR School of Medicine and the Orvis School of Nursing for their collaboration and for having their students participate.

This material is based upon work supported by the National Science Foundation under grant numbers 2019609 and IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. P.S.Q. Healthcare, Patient handoffs: the gap where mistakes are made. *Patient Saf. Monit. J.* Nov 2017. Last accessed 6 Oct 2021. <http://www.hcpro.com/QPS-330353-234/Patient-handoffs-The-gap-where-mistakes-are-made.html?sessionGUID=edbe08b4-46e3-0e62-79c3-b0d7966ab079&webSyncID=3298c37b-1099-3ae8-b921-25094b55dc8c>
2. R.H. Ibrahim, D.-A. Hussein, Assessment of visual, auditory, and kinesthetic learning style among undergraduate nursing students. *Int. J. Adv. Nurs. Stud.* **5**(1), 1–4 (2015). <https://doi.org/10.14419/ijans.v5i1.5124>
3. B.T. Kitch, J.B. Cooper, W.M. Zapol, M.M. Hutter, J. Marder, A. Karson, E.G. Campbell, Handoffs causing patient harm: a survey of medical and surgical house staff. *Jt. Comm. J. Qual. Patient Saf.* **34**(10), 563–570d (2008). <https://www.sciencedirect.com/science/article/pii/S1553725008340719>
4. M. Müller, J. Jürgens, M. Redaelli, K. Klingberg, W.E. Hautz, S. Stock, Impact of the communication and patient hand-off tool SBAR on patient safety: a systematic review. *BMJ Open* **8**(8) (2018). <https://bmjopen.bmj.com/content/8/8/e022202>
5. S.J. Anbro, A.J. Szarko, R.A. Houmanfar, A.M. Maraccini, L.H. Crosswell, F.C. Harris, M. Rebaleati, L. Starmer, Using virtual simulations to assess situational awareness and communication in medical and nursing education: a technical feasibility study. *J. Organ. Behav. Manag.* **40**(1–2), 129–139 (2020). <https://doi.org/10.1080/01608061.2020.1746474>
6. T.R. Coles, D. Meglan, N.W. John, The role of haptics in medical training simulators: a survey of the state of the art. *IEEE Trans. Haptic* **4**(1), 51–66 (2011)
7. C. Luciano, P. Banerjee, T. DeFanti, Haptics-based virtual reality periodontal training simulator. *Virtual Reality* **13**(2), 69–85 (2009). <https://doi.org/10.1007/s10055-009-0112-7>
8. Toltech - arthrosim arthroscopy simulator. Last accessed 15 Oct 2021. <https://www.toltech.net/medical-simulators/products/arthrosim-arthroscopy-simulator>
9. Oramavr. Last accessed 15 Oct 2021. <https://oramavr.com/>
10. M. Sra, Asymmetric design approach and collision avoidance techniques for room-scale multiplayer virtual reality, in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology, ser. UIST '16 Adjunct* (Association for Computing Machinery, New York, NY, 2016), pp. 29–32. <https://doi.org/10.1145/2984751.2984788>
11. J. Du, Y. Shi, C. Mei, J. Quarles, W. Yan, Communication by interaction: a multiplayer VR environment for building walkthroughs, in *Construction Research Congress 2016* (2016), pp. 2281–2290. <https://ascelibrary.org/doi/abs/10.1061/9780784479827.227>
12. Vive eye and facial tracking SDK. Last accessed 8 Oct 2021. <https://developer.vive.com/resources/vive-sense/sdk/vive-eye-and-facial-tracking-sdk/>
13. RockVR video capture. Last accessed 8 Oct 2021. <https://assetstore.unity.com/packages/tools/video/video-capture-75653>
14. U. Technologies, Unity real-time development platform | 3d, 2d vr & ar engine. Last accessed 10 Jan 2022. <https://unity.com/>
15. Voice for pun2. Last accessed 8 Oct 2021. <https://doc.photonengine.com/en-US/voice/current/getting-started/voice-for-pun>
16. VIVE Pro Eye specs. Last accessed 8 Oct 2021. <https://www.vive.com/us/product/vive-pro-eye/specs/>
17. Oculus | vr headsets, games, & equipment. Last accessed 8 Oct 2021. <https://www.oculus.com/>
18. Vive wave | htc vive. Last accessed 8 Oct 2021. <https://developer.vive.com/us/wave/>
19. Vive facial tracker. Last accessed 8 Oct 2021. <https://www.vive.com/us/accessory/facial-tracker/>

A Tool for Syntactic Dependency Analysis on the Web Stack

Manit Singh Kalsi, Kevin A. Gary, Vasu Gupta, and Suddhasvatta Das

Abstract

One of the most common errors developers make is to provide incorrect string identifiers across the HTML5-JavaScript-CSS3 stack. The existing literature shows that a significant percentage of defects observed in real-world codebases belong to this category. Existing work focuses on semantic static analysis, while this paper attempts to tackle challenges that can be solved using syntactic static analysis. While semantic state analysis is more powerful, it creates a greater computational burden on tool processing while simple static analysis may be computed faster, allowing for better integration in inline syntax-highlighting marker in a user interface or a quick pass through large codebases. This paper proposes a tool for quickly identifying defects at the time of injection due to dependencies between HTML5, JavaScript, and CSS3, specifically in syntactic errors in string identifiers. The proposed solution reduces the delta (time) between defect injection and discovery with the use of a dedicated just-in-time syntactic string identifier resolution tool. The solution focuses on modeling the nature of syntactic dependencies across the stack, and providing a tool that helps developers discover such dependencies. This tool was validated against a set of real-world codebases to analyze the significance of these defects.

Keywords

Web development · Software engineering · Programming languages · Static analysis · Software tools

M. S. Kalsi · K. A. Gary (✉) · V. Gupta · S. Das
School of Computing & Augmented Intelligence, The Ira A. Fulton
Schools of Engineering, Arizona State University, Tempe, AZ, USA
e-mail: manitsingh.kalsi@asu.edu; kgary@asu.edu;
vgupta31@asu.edu; sdas76@asu.edu

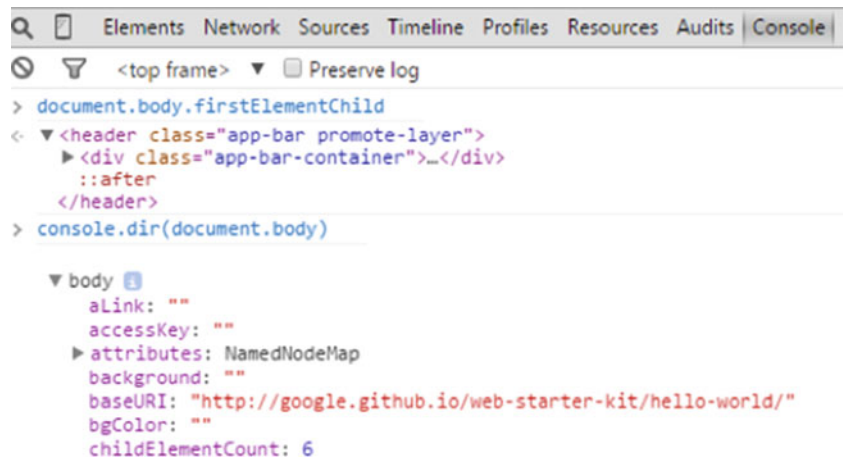
4.1 Introduction

The modern web developer repeatedly deals with syntactic dependencies in the Document Object Model (DOM) and render trees of the browser, and lacks the tool support to identify and resolve the dependencies rapidly. While developer tools embedded in modern browsers such as Chrome and Firefox allow for run-time troubleshooting, such cross-file syntax checking is rare in modern tools. More complex static analysis tools exist that probe for security defects primarily in Javascript do not work across the stack, and tend to be complex to use (meaning they are more appropriate for post-test than during development). The existence of a simple development tool focused on resolving such cross-references during development would reduce cycle time by avoiding the repeated edit→save→load in browser development troubleshooting cycle.

In most scenarios, these types of defects are not caught even during testing (unit and integration). Currently, the most common method used by developers to test such dependencies is to run the code in the browser, see the results and look for any error message in the console or any undesirable behavior in their application. The most common tools for such purpose are browser developer tools (see Fig. 4.1). There are different components of these tools that help the developer map the dependencies and find the defects if any. In most cases, any JavaScript related defect is directly caught by looking at the “console” view. But CSS related defects can only be found through a visual inspection of either the functionality of the module or through a code inspection. This round trip is not only an overhead adding to the development time and cost, but is also very error-prone.

The objective of this work is to present a tool for straightforward static analysis between dependencies in the web stack enabling fast processing while still assisting developers in catching a high percentage of injected defects.

Fig. 4.1 Console tab of Google Chrome Developer tools



4.2 Background

The web was built with the purpose of document sharing [1], with the server generating HTML pages with almost no JavaScript or CSS. This means that most of the content was static and the server was the main source of all the client side documents. The minimal JavaScript and CSS ensured very little interaction between the (DOM) and the scripts and stylesheets. As the web evolved, this trend began to change. The client side is no more a “thin-client”, instead, a lot of functionality is client driven. Static HTML pages generated by a server gave way to dynamic web pages with a lot of user interaction, runtime DOM manipulation, and client-server interaction. Mesbah et al. [2] have very aptly described modern day web application as “. . . stateful asynchronous client/server communication, and client-side runtime manipulation of the DOM tree . . .”.

As a web developer, this shift towards the front end stack means that the developer has to write code in HTML5, CSS3, and JavaScript. These are three different languages with their own characteristics; HTML5 is a markup language that is used to create the structure of the web pages; CSS3 is style-sheet language that is used to provide presentation to the HTML5 document; JavaScript is a dynamically typed interpreted language that provides dynamic nature to the otherwise static HTML5 documents. JavaScript gives the developer an ability to manipulate the underlying page structure (the DOM) at runtime. While writing a front-end web application, a developer has to keep a track of the DOM in the HTML5, the associated JavaScript, and the associated CSS3 stylesheets as well. Because of this interplay, there are a lot of syntactic dependencies created. Figure 4.2 can help better illustrate the concept of syntactic dependencies.

Given the nature of these languages and the dependencies between them, modern day web applications are prone to have a lot of defects. And as more and more behavior is moving to the front-end, errors in this technology stack are

no longer cosmetic faults but significant defects that impact the correctness of the application. The developer has to keep track of how the DOM accesses the JavaScript and vice-versa, how the DOM accesses the CSS3 stylesheet and vice-versa, how the JavaScript accesses the CSS3 stylesheet and vice-versa. Figure 4.2 illustrates the concept at a very small scale. For example, the button in `index.html` has an `onclick` event that is handled by the `getSomeData()` method in the `myScript.js` file. Similarly, the `myScript.js` file attempts to manipulate the content of the DOM by accessing the `serverResponse` DOM element by its ID. The div that contains the button uses a style `myCssClass` as defined in the `theme.css` file. So even in about 20 lines of code, it is very easy to see the nature of these dependencies. And the developer can easily make an error resolving these dependencies. These errors can range from typographical errors to non-existent constructs like ID and functions, etc. It becomes very difficult for the developer to keep track of such dependencies as the size of the codebase increases. It is essential to note that in all of the modern web applications, these three languages work in parallel, making them prone to dependency related defects. Ocariza et al. [4] found that most of such defects (specifically JavaScript) are injected by the programmers in the code itself.

“A defect is an instance in which a requirement is not satisfied.” [5] The dependencies across the HTML5-JavaScript-CSS3 stack make it highly prone to defects committed by developers. One of the most common errors developers make is to provide incorrect string identifiers across the HTML5-JavaScript-CSS3 stack. The existing literature [4–7] shows that a significant percentage of defects observed in real-world code bases belong to this category. The literature [8] also shows that 80% of the defects are caused by 20% of the modules. In this case, the main causes of defects generated due to dependencies in the HTML5-JavaScript-CSS3 stack can be traced back to the DOM. In fact, an empirical study of client-side JavaScript bugs has shown that 65% of the bugs are DOM related [4]. In another study, the authors

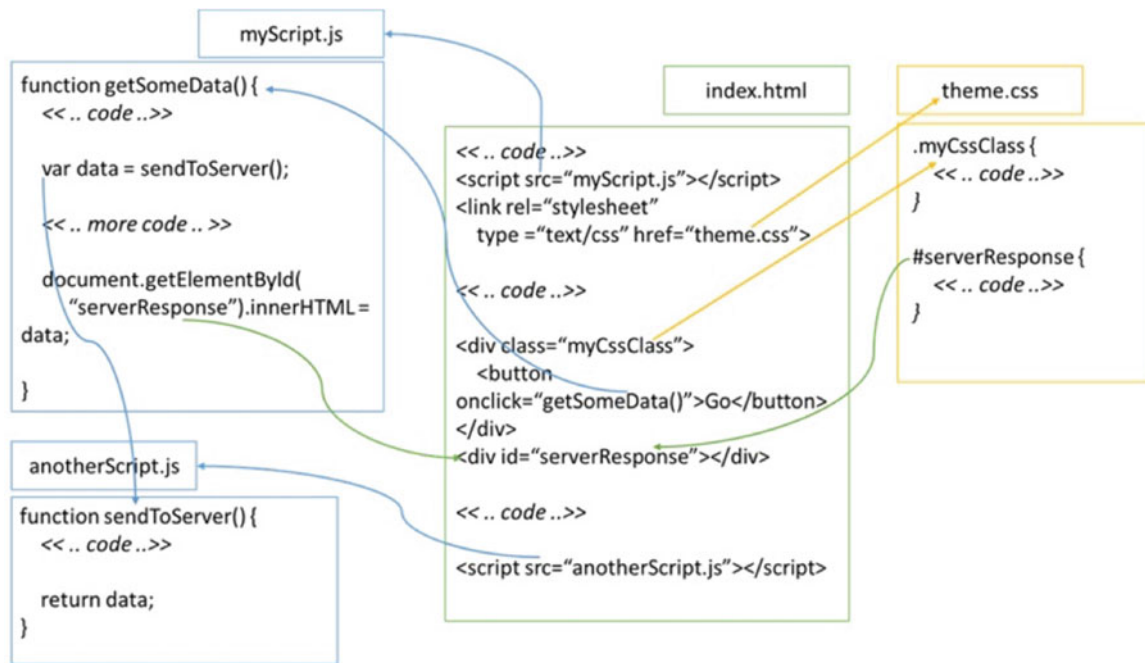


Fig. 4.2 Syntax dependencies in the HTML5-JS-CSS3 stack [3]

have observed that DOM manipulation is one of the most common usages of JavaScript in modern web application and have recommended static analysis tool designers to consider the DOM as a component in their tools [6]. This is a clear indication that most of the defects are caused by interaction between the DOM and JavaScript. Although no similar studies have been found for DOM and CSS interactions, it is easy to extend these results and expect a similar behavior for DOM and CSS interactions.

When a developer introduces a defect into code, this activity is termed as *defect injection*. When the defect is found by either the same developer or some other developer/user, this activity is termed as *defect discovery*. We contend that the existing web developer toolset and practices make this delta between defect injection and defect discovery a relatively long time. This is because of the round-trip between the IDE/text editor to the browser's dev tools to inspect the console and/or the behavior of the application to determine the existence of a defect. Even when changes are made, this round trip is a mandatory step to determine the status of the defect. This necessary evil adds to the delta between defect injection and defect discovery, thereby affecting developer productivity. Researchers recommend reducing this delta to as minimum as possible: "In order to eliminate defects from the product it is necessary to address their prevention, or detection and resolution as soon as possible after their injection during development and maintenance." [5, p. 746]. It is also recommended that such defects be found and fixed before delivery to avoid cost: "Finding and fixing a software problem after delivery is often 100 times more expensive

than finding and fixing it during the requirements and design phase." [8, p. 135]. This paper focuses on defect injection and defect discovery as an in-phase activity rather than across phases. This is due the nature of front end web development. This micro-optimization will help save many seconds per developer and when aggregated over the entire development team it will amount to a lot of valuable time.

4.3 Research Questions

The research questions that this paper targets:

- RQ1*: Are syntax errors in string identifiers referencing DOM elements in the HTML5-JavaScript-CSS3 stack significant?
- RQ2*: What is the delta time/cost between defect injection and discovery in HTML5-JavaScript-CSS3 style applications?
- RQ3*: Does a dedicated just-in-time syntactic string identifier resolution tool significantly reduce delta time/cost from RQ2 for a significant portion of real problems (RQ1)?

First, we consider RQ1 in light of the concepts discussed in previous sections. Developers often make the mistake of using incorrect string identifiers in the HTML5-JavaScript-CSS3 stack. For example, assume an "id" attribute declared for an HTML5 element is `foobar`. Developers may make the mistake of accessing it as `fooBar` or `FooBar`. As the code grows larger, with new changes being added, it's difficult to keep track of such identifiers and these errors become significant. RQ1 attempts to identify the significance

of such errors, defined as the severity of the defect on system behavior.

There are different development environments and toolchains used by developers for front-end applications which impact the delta between defect injection and defect discovery. The typical developer uses Chrome or Firefox Developer tools to test changes to code. The time it takes to round-trip test DOM-related development in these tools is considered the standard for defect injection and defect discovery. RQ2 attempts to define the typical delta distribution.

RQ3 attempts to reduce the delta observed in RQ2 by providing the developers with a dedicated just-in-time syntactic string identifier resolution tool which will help the developer to discover the defect quicker than traditional methods in practice today.

4.4 Tool Design

The HTML5-JavaScript-CSS3 stack of front-end development is tightly coupled. To successfully render a web application on a browser, these three languages have to be parsed and loaded correctly. It is important for the developer to understand these dependencies and tackle any defects generated because of them. Based on our prior work [3] we catalogued the types of dependencies in this stack as shown in Table 4.1.

Johnson et al. [9] found in a study that developers are reluctant to use static analysis tools to find bugs because of many reasons. One of the prominent reasons was the inability of static analysis tools to provide understandable results. Developers say that it is difficult to make sense out of the error messages given by static analysis tool. The authors also quoted some responses from developers, including: “it’s one thing to give an error message, it’s another thing to give a useful error message.”([9], p. 677) and “I find that the information they provide is not very useful, so I tend to ignore them.” ([9], p. 677). This shows that the developers need to be provided with descriptive and proper error messages that will help them understand the error and also help them understand how to fix the error. Based on the dependency model, we designed an error ontology that helps developers understand the error and the dependency better. The error

ontology design helps to achieve this by building on the dependency model and providing more useful information about the dependencies and the errors.

Figure 4.3 shows a classification based on the dependency model and other errors that have been taken into consideration. This classification also helps us understand dependency types that do not generate any errors; namely unused dependencies.

“Syntax” errors are represented by `ParseError`. “Unused Ref ” refers to all those dependency constructs that are declared but are not used, and do not generate any errors. “File Not Found” categorizes all the dependencies between the HTML5 document and external files. All other classifications are extensions to the dependencies discussed in the dependency model above. Next, we use this classification to generate an ontology used to describe the errors and warnings as generated by the tool as shown in Fig. 4.4.

4.5 Tool Implementation

The main focus of the implementation was to develop a static analysis tool that can generate a symbol table to manage the dependencies discussed above. It is important to note that developing an algorithm to parse HTML5, CSS3 and JavaScript was not the main focus of this work, instead to parse the codebase, popular libraries and engines were used. Different parsers were used for the different languages in the stack, with the tool tracking dependencies between them.

The HTML5 files in a web application are the door to the entire codebase. This tool starts the analysis by building a list of HTML5 files in a given directory and moves from there. This way non-dependent files in the project are ignored. The entire code for a given application is analyzed and a list of dependencies is generated, compared, and the results are computed. Furthermore, metadata associated with each dependency is tracked in order to help the developer find and fix the defect as fast as possible. The metadata includes source file, line number, column number and dependency type. The secondary features of the tool include: providing verbosity in terms of output shown, exporting results in JavaScript Object Notation (JSON) and plain text format, an HTML viewer for displaying the JSON results, rule-based analysis,

Table 4.1 Initial Dependency Model

To→From	HTML5	JavaScript (JS)	CSS3
HTML5	No dependencies identified.	1. Links from HTML5 to JS files 2. Event listeners in HTML5 file.	Class attribute in HTML5 elements.
JavaScript	Through Document object	1. Function calls within a function. 2. Global variables.	JavaScript adding CSS3 class to DOM if it exists in a CSS3 file included in the webpage.
CSS3	No dependencies identified.	No dependencies identified.	Other CSS3 selectors like id, tag name

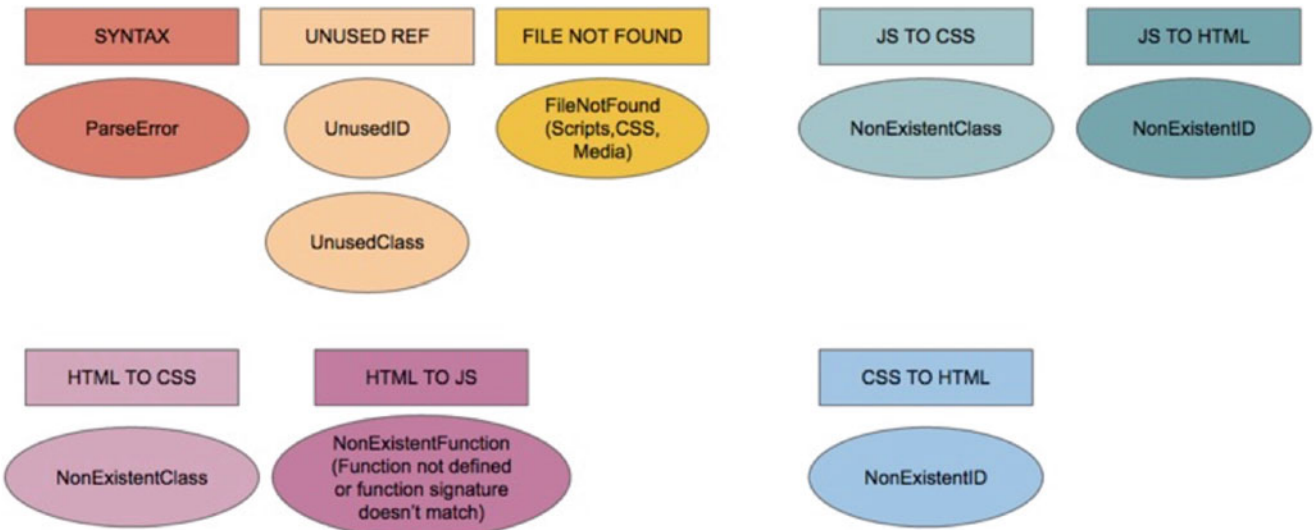


Fig. 4.3 Error classification

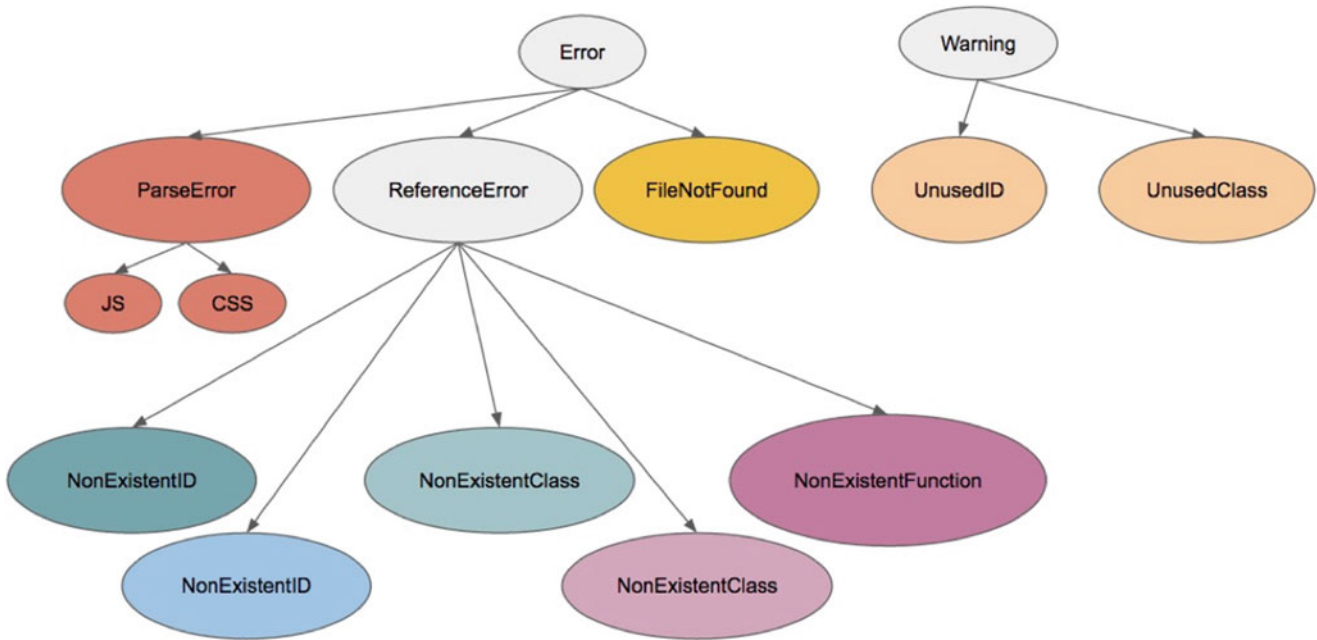


Fig. 4.4 Error ontology

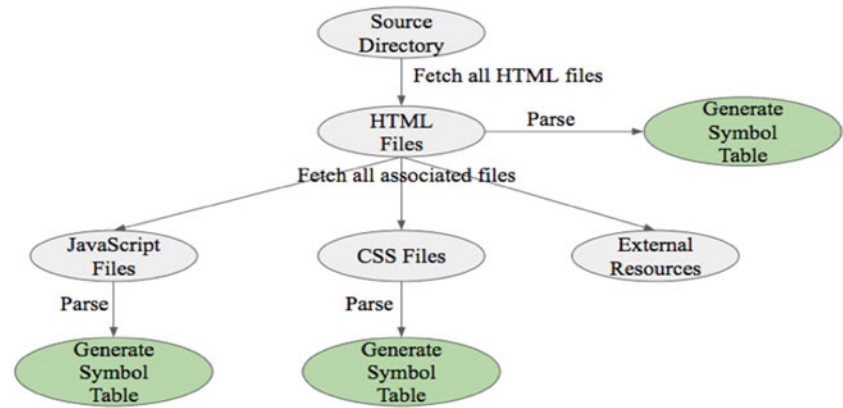
recommendation for fixes and integrated plugin development (Fig. 4.5).

For parsing HTML5, we used the Jsoup parser, a popular Java library for parsing HTML. It supports all the latest HTML5 tags, implements the WHATWG HTML5 specification, and parses the HTML5 into a Document Object Model (DOM). One major drawback is that it does not keep track of line numbers, so as a result, a custom module based on Java’s Matcher engine was made to keep track of the line numbers and column numbers. The following data is extracted per HTML5 file:

- All “id” attributes used
- All “class” attributes used
- Links to all associated CSS3 style sheets
- Links to all associated media assets (images, audio, etc.)
- Links to all associated JavaScript files
- All methods referenced through event handlers

CSS3 parsing was achieved using CSSParser, a Java library to parse CSS3 files. It takes the CSS3 file text as input and generates a Document Object Model Level 2 Style tree. It also provides the ability to choose different internal parsers as per the developers need. For the purpose of this project,

Fig. 4.5 Tool parsing and dependency construction flow



we use the SAC (Simple API for CSS) Parser for CSS3 since the focus is on modern web applications which primarily use CSS3. All CSS3 files associated with each HTML5 file are processed using this parser. Once the HTML5 parsing phase is complete and all valid associated CSS3 files are identified, each file is passed on to the CSS3 parser. For each CSS3 file, the following data is extracted:

- List of referenced IDs
- List of defined classes

Once this data is obtained, the static analyzer uses it to analyze dependencies based on class and id references.

For parsing JavaScript, we chose to use the Nashorn engine, a JavaScript engine by Oracle pre-bundled with Java 8 and above. It has a very nice API which provides a method to extract an Abstract Syntax Tree (AST) from the given JavaScript source code. The AST is provided as a JavaScript Object Notation (JSON) data structure which can be easily parsed to extract the information that is required. All JavaScript files associated with each HTML5 file are processed using this parser. Once the HTML5 parsing phase is complete and all valid associated JS files are identified, each file is passed on to the JS parser. For each JS file, the following data is extracted:

- List of referenced IDs
- List of referenced classes
- Method signatures of all methods defined

All of the above information is stored for each file along with the location of the respective file.

Once the data from all associated files is gathered, it is compared with dependencies as and when they arise. For example, while identifying the list of IDs in a JavaScript file, those IDs are simultaneously compared to existing IDs in the associated HTML5 file. If an error is found, it is immediately stored in a Results object and retrieved later for display.

Similar with CSS file processing. When a dependency is found that does not contribute to a runtime error, it is stored in the Results object as a warning. Based on the verbosity level, the user can choose to view the warning details.

4.6 Tool Presentation

Most static analysis tools come in two flavors; a standalone command-line tool and a plugin that is installed on a preferable Integrated Development Environment (IDE). Both flavors serve different purposes. The standalone command-line tool can be integrated into the build toolchain or the continuous integration tool. The integrated plugin, on the other hand, can serve the purpose of an interactive utility within the IDE that helps developers tackle defects while they code. A similar approach was taken for this tool as well.

The command-line tool supports a set of flags to control output filtering and verbosity, and can export results to plain text and JavaScript Object Notation (JSON) format. The JSON format can be helpful in exporting the results and analyzing them externally, or if someone wants to expose the tool as a service-based component rather than directly using the tool. The tool also supports an HTML Viewer that parses the generated JSON results and displays them as an HTML page.

Rule-based analysis flags help the user to filter out the results and view them one category at a time. The categories were presented as part of the error ontology in the previous section. Even though output results are filtered as per the flag values, the internal computation is the same, as we do not want the analysis to disregard any dependency and miscalculate errors. The supported values for the rules flag:

- *ParseError*: filter results to show parse errors encountered while parsing either HTML5, CSS3 or JavaScript.
- *ReferenceError*: filter results to show reference errors. Reference errors can be of various types: nonexistent class, nonexistent id, and nonexistent function.

- *FileNotFound*: filter results to show file not found errors. Note: this also checks for remote files.
- *Warnings*: filter results to show only warnings. Warnings are those dependencies that do not cause runtime errors.

We also wanted the tool to be able to suggest fixes in case of reference errors generated due to nonexistent class and nonexistent ID. These errors arise because of a referenced class or ID does not exist in either the CSS3 or HTML5 file. Under the hypothesis that these errors are caused because of typographical errors in string identifiers, we use string matching algorithms to find the most similar string and suggest a fix to the user. The underlying algorithm used for string matching is based on string distance similarity score [10]. This string matching algorithm starts by comparing each character for the two given strings. For every character that matches, one point is given, and for every other match thereafter two bonus points are given. Thus, a higher score would indicate higher similarity. The string with the highest similarity is suggested as a fix for that particular defect. To illustrate with a very simple example, consider two strings “foobar” and “fooBar”. The user might accidentally type the first string as a class or ID. When this algorithm runs, it would give the highest score to “fooBar” as compared with other strings. Thus, we can see how this simple suggestion might help the user fix a defect quickly.

A plugin was integrated with Eclipse to provide ease of use inside an IDE. Eclipse provides a Plug-in Development Environment (PDE) that helps developers to create plugins for Eclipse. These plugins are developed as a Rich Client Application (RCP). For the purpose of this research work, the plugin was made with a simple view based layout where the output of the analysis is shown in a tableview with separate rows and columns for each part of the result. The developer can also quickly jump to the location of the defect by double clicking on a particular row.

4.7 Validation

To address the research question (Sect. 4.3), we tested the tool against real world codebases. To test against real world codebases, the main task was to find repositories that did not use any JavaScript frameworks like jQuery, AngularJS, React, etc. GitHub was used as the source to find such repositories. GitHub also has “Issues” functionality which helped in analyzing if dependency defects were triaged by the codebase owners. Once the code was locally cloned from the GitHub repository, we ran the tool against each one of them to find the number of dependency defects. Further, we

analyzed the open issues to find out whether these defects were triaged or not. The main inclusion criterion was to find codebases that did not use any JavaScript frameworks. Two of the three repositories used for this experiment were found in other literature as well. The Internet Explorer (IE) Test Suite was used by Jensen et al. [11] and the Google Octane Suite was used by Andreasen et al. [12] for their validations (Table 4.2).

The defects discovered by our tool were not the same as those listed in the issue trackers of the repositories, yet are relevant in modern front end web development. On an average, 10.13% of the total defects were dependency-related defects. This is a significant amount considering that the codebases were real world codebases. Another observation was that from the total 96 defects found, 65 of them were of the type *HTML5toCSS3* dependency defects and 26 of them were *CSS3toHTML5* dependency defects demonstrating that these dependencies may not be easy to discover. The remaining defects: 2 external file dependency defects and 3 CSS3 Parse errors. Further, many warnings were found showing that dead code exists in these codebases.

The main limitation of this experiment is that the set of codebases may not be an appropriate sample of the actual modern front end web development codebases, as they do not use any frameworks like jQuery, AngularJS, React, etc. This research currently supports only plain JavaScript. But the findings can be extended to codebases with frameworks as the nature of the dependencies stays the same regardless of the use of frameworks and libraries. A second limitation is the small sample size of real-world codebases. Again, this in part due to the lack of pervasive testbeds with vanilla Javascript. But running the tool against a very small subset of real world codebases shows that such defects are significant.

The most interesting observation from the results of this experiment is that none of these defects were triaged and reported in the issue trackers of these repositories. This shows that these defects might not be easy to triage when contributions to the codebase are made by multiple developers. *HTML5toCSS3* and *CSS3toHTML5* dependency defects were a major chunk of the total defects which shows that these dependencies may not be easy to triage. Across multiple developers and across various iterations of the code, many such defects might get injected and never be caught or triaged. Further, a number of warnings show that dead code exists in these codebases. Dead code analysis was not the focus of this research, but the existence of such warnings shows there are unmet dependencies in the code that do not contribute to functionality. Such dependencies are difficult to manage when the codebases are large, change with iterations of the product, and are contributed to by various developers.

Table 4.2 Performance on real world codebases

Repo Name	Tested file/Example	Errors	Warnings	Issues in GH	% of defects
TodoMVC	VanillaJS	2	22	3	40
IE Test Suite	@supports sample	1	0	24	4
	audiomixer	4	22	24	14.28
	blobbuilder	7	45	24	22.58
	chalkboard	1	7	24	4
	chess	17	40	24	41.46
	coloringbook	2	396	24	7.69
	compatinspector	0	0	24	0
	css3filters	11	11	24	31.42
	css3mediaqueries	1	12	24	4
	editingpasteimage	4	8	24	14.28
	eme	2	14	24	7.69
	familysearch	0	0	24	0
	fishbowl	2	20	24	7.69
	html5forms	1	65	24	4
	mandelbrot	1	15	24	4
	math	0	8	24	0
	mazesolver	5	16	24	17.24
	microphone	4	34	24	14.28
	musiclounge	0	3	24	0
	particleacceleration	2	3	24	7.69
	photocapture	0	16	24	0
	picture	1	0	24	4
	readingview	4	67	24	14.28
	setimmediatesorting	0	14	24	0
	spellchecking	0	2	24	0
	sudoku	2	101	24	7.69
	svgradientbackgroundmaker	5	36	24	17.24
	toucheffects	2	15	24	7.69
	typedarrays	3	18	24	11.11
	usersselect	2	17	24	7.69
	videofformatsupport	1	19	24	4
	webaudiotuner	6	17	24	20
	webdriver	1	10	24	4
Google Octane	Entire Suite(single html)	2	408	17	10.52

4.8 Conclusions

Modern front end developers work with large-scale front-end web applications. The size of the codebase makes it a challenge for the developers to keep a track of the syntactic dependencies across the HTML5-JavaScript-CSS3 stack. A developer has to resort to using the developer tools within a browser and mentally keep a track of dependencies. A manual inspection of the developer tool and the running web application is prone to human error and as a result of which a significant number of defects due to syntactic dependencies go unnoticed. The time and effort saved per developer with the help of such a tool is significant. When aggregated over an entire development team, it can help save a lot of time

and effort. Such an in-phase micro optimization of effort in modern day software engineering would help in drastically improving developer efficiency and developer productivity.

There are many directions in which this research can be extended. Codebases considered for this research did not include JavaScript/CSS3 frameworks or libraries [13, 14]. Regardless, the concept of the dependencies presented is of framework agnostic, and can be extended in that direction. Modern approaches are rapidly moving toward a more component-oriented binding of the stack, coupling elements of the DOM, Javascript, and CSS into a packaged bundle that may then be analyzed as a unit [15]. We earlier mention that the JSON results provide a hook by which to create a service that can be dynamically invoked from many different IDEs, so developers would not be bound to Eclipse or our

command-line interface. This work can be further extended to analyze the presence of dead code and how it impacts the loading time of those web applications. Finally, the tool can be further improved from a design perspective to be real-time (e.g. syntax highlighting within a code editor) as opposed to the current implementation of a just-in-time tool (must be specifically selected to run).

References

1. T. Berners-Lee, Information management: A proposal. Unpublished manuscript (1989)
2. A. Mesbah, A. van Deursen, Invariant-based automatic testing of AJAX user interfaces, in *IEEE 31st International Conference on Software Engineering* (2009), pp. 210–220
3. K.A. Gary, V. Gupta, M.S. Kalsi, A tool for teaching web dependency analysis, in *Information Systems Education Conference* (2020)
4. F. Ocariza, K. Bajaj, K. Pattabiraman, A. Mesbah, An empirical study of client-side JavaScript bugs, in *Empirical Software Engineering and Measurement, ACM/IEEE International Symposium* (2013), pp. 55–64
5. M.E. Fagan, Advances in software inspections, in *Pioneers and Their Contributions to Software Engineering*, ed. by M. Broy, E. Denert, (Springer, Berlin/Heidelberg, 2001), pp. 335–360
6. F.S. Ocariza Jr, K. Pattabiraman, B. Zorn, JavaScript errors in the wild: An empirical study, in *Software Reliability Engineering, IEEE 22nd International Symposium* (2011), pp. 100–109
7. F.S. Ocariza Jr, K. Pattabiraman, A. Mesbah, Vejovis: Suggesting fixes for JavaScript faults, in *Proceedings of the 36th International Conference on Software Engineering* (2014), pp. 837–847
8. B. Boehm, V.R. Basili, Top 10 list [software development]. *Computer* **34**(1), 135–137 (2001)
9. B. Johnson, Y. Song, E. Murphy-Hill, R. Bowdidge, Why don't software developers use static analysis tools to find bugs?, in *Software Engineering, 35th International Conference* (2013), pp. 672–681
10. W.H. Gomaa, A.A. Fahmy, A survey of text similarity approaches. *Int. J. Comput. Appl.* **68**(13), 13–18 (2013)
11. S.H. Jensen, M. Madsen, A. Møller, Modeling the HTML DOM and browser API in static analysis of JavaScript web applications, in *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European Conference on Foundations of Software Engineering* (2011), pp. 59–69
12. E. Andreasen, A. Møller, Determinacy in static analysis for jQuery. *ACM SIGPLAN Notices* **49**, 17–31 (2014)
13. M. Madsen, B. Livshits, M. Fanning, Practical static analysis of JavaScript applications in the presence of frameworks and libraries, in *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering* (2013)
14. A. Karakochev, G. Zhang, Static analysis of large-scale JavaScript front end, in *International Conference on Web Engineering* (Springer, Cham, 2021)
15. T. Sotiropoulos, B. Livshits, Static analysis for asynchronous JavaScript programs. arXiv preprint arXiv:1901.03575 (2019)

Using Software for Computational Fluid Dynamics and Molecular Dynamics

Jeena Shetti, Stefan Pickl, Doina Bein, and Marian Sorin Nistor

Abstract

The paper surveys the current state-of-the-art at promising US companies dedicated to drug discovery using computational methods. Computational fluid dynamics and molecular dynamics are only two methods that can be adapted to drug discovery using computational methods. Drug discovery is promising with the improvements in computer power and algorithms and is likely to play a more critical role.

Keywords

Biotechnology · Drug discovery · Drug testing · Drug research · Computational method · Computational fluid dynamics · Molecular dynamics · Free energy · Neuro-science medicine · Simulations

molecular dynamics (MD) and how the simulations work. The comparison of MD simulations and Computational Fluid Dynamics (CFD) are two technological ways for this to work. Another topic that plays a significant role is free energy and the challenges that come with that. Computational methods are a faster and cost-effective way to make drug discovery more efficient. At the rate that the world is advancing, the need for MDs and CFDs will grow higher. Our future entails more and more developments that will benefit our society.

The paper is organized as follows. In Sect. 5.2, we discuss the concepts of molecular dynamics and computational fluid dynamics. Section 5.3 includes a logical approach to CFDs along with a modern-day approach. Another relative topic is free energy which is discussed in Sect. 5.4. Finally, concluding remarks are presented in Sect. 5.5.

5.1 Introduction

We are living in a generation where everyone is surrounded by technology. The advances in this field have been progressing at a rate faster than ever. Using computational methods for drug discovery and drug testing is something that can change the future of medicine. With the improvements in computer power and algorithms, drug discovery is promising and will likely play a more important role. We focused on

J. Shetti · D. Bein (✉)
Department of Computer Science, California State University,
Fullerton, Fullerton, CA, USA
e-mail: jeenashetti@csu.fullerton.edu; dbein@fullerton.edu

S. Pickl · M. S. Nistor
Department of Computer Science, Universität der Bundeswehr
München, Neubiberg, Germany
e-mail: stefan.pickl@unibw.de; sorin.nistor@unibw.de

5.2 Comparison of Molecular Dynamics Methods

MD simulations are a very interesting topic. The first MD simulations were used in the late 1950s, and scientists used them to study proteins. Over time, they have become more advanced and become more popular and visible to the biotechnology community. They are used today to study a variety of molecules such as proteins, carbohydrates, or nucleic acids and can predict how every molecule/atom will move over time. When it comes to drug discovery specifically, many scientists use the research of proteins [1]. They have a wide range of resolutions and help catch processes of conformational change, protein folding, and revealing positions of atoms at a high resolution [2]. Often, they are not used alone. Many scientists combine MD simulations with other medical techniques such as x-ray, cryo-EM, etc.. The increasing attention towards MD

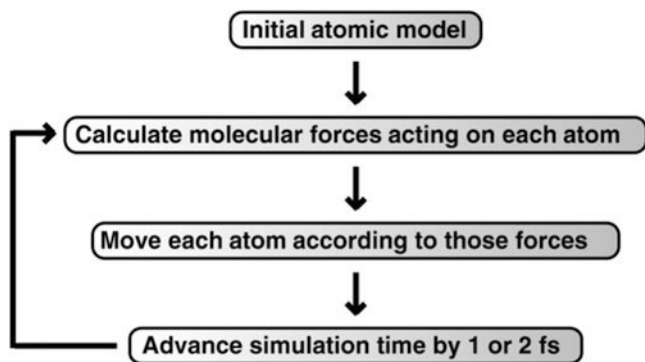


Fig. 5.1 Diagram of how molecular dynamic simulations are performed. (Image taken from [5]).

simulations has been more revolved around neuroscience medicine. Specific membrane proteins require research done by experimental methods [2]. This causes an increase in challenges that have risen year after year.

The article of [3] mentioned how “the principles of statistical mechanics allow quantitative estimates of important thermodynamic observables to be computed from MD trajectories”. As we mentioned earlier, protein is the main event when it comes to drug discovery. The complex PL, which is an association between protein and ligand ($P + L \rightleftharpoons PL$), has an equilibrium that is defined by [3]:

$$K_a = \frac{[PL]_{eq}}{[P]_{eq}[L]_{eq}}$$

What we found interesting is that the reciprocal of this equation is the equilibrium constant for the opposite reaction. Thus, that ratio becomes the ligand concentration for which an equal probability of bounded and unbounded protein is achieved. This way is much more used in pharmaceutical research than the equilibrium equation above.

When working with computational methods to assist in drug research, there are a few different ways technology can be used. Two methods that we found interesting were MD simulations and CFD. MD simulations are a more modern technology that is used to gain a better understanding of macromolecular structure-to-function relationships. They play an essential role in specifically protein motion, where even a singular image of protein conformation can tell a lot about its dynamics. It can predict key properties like the biocatalysts used, temperature, pH, or nutrient concentration [4]. Figure 5.1 is a depiction of how MD simulations are performed. After the computer model of the protein is made, we calculate the forces on the system atoms. Then, the atoms

are moved, and the process ends when the simulation time has been advanced.

5.3 Computational Fluid Dynamics

Cost reductions with CFD are one of the key reasons they are more favorable. In terms of doing physical experiments, tests, and getting engineering data, research can become expensive. CFDs are relatively cheap, and as technology advances, they are bound to become even more affordable [6]. The scientists will have the freedom to simulate different conditions. CFDs provide the ability to simulate any physical condition. This means that if any specific case studies need to be done, they can control the physical process. Another benefit is that they are a lot more effective and quicker than standard engineering data tools [7]. CFDs are used in many industries like aerospace, chemical manufacturing, meteorology, medical research, et cetera. The market for it is growing at a faster rate every year. The global market size was estimated at \$2393.08 million in 2020 and is expected to reach \$2567.73 million by the end of 2021. There is speculation that the annual growth rate will consistently be at 7.63% for the next several years [8].

A company based in Durham, North Carolina, called Resolved Analytics, used CFD to create a level of realism in digital prototypes. They have found that it cuts the development time in half and removes the need for experimental prototypes throughout the years. Similarly, companies in the pharmaceutical industry have found that using CFDs has become the most efficient research method. Products are manufactured in different forms, such as capsules, tablets, liquids, ointments, et cetera. They all have a unique method for their creation, but most of the time, it includes milling, drying, pressing, et cetera. These steps can be consolidated by just using fluid dynamic simulations. Aside from the cost, CFDs are also a great way to reduce the risk of failure.

CD-Adapco was a multinational computer software company that created applications that are used within biotechnology, their most popular being computational fluid dynamic products. In 2016, it was acquired by Siemens Digital Industries. Throughout the years, the industry and demand for their products have increased and become more popular. They believe that “engineering simulation provides the most reliable flow of information into the design process” [9]. This inevitably drives more biotech innovation and reduces product development costs. This calls for faster designs and results. CD-Adapco also created products for heat transfer, particle dynamics, computational solid mechanics, et cetera. They strive towards this success and believe that using computational methods is how the future generation of scientists will research.

5.4 Free Energy

A concept that goes together with MD simulations and CFDs is free energy. A mathematician in the 1870s, Josiah Willard Gibbs, described free energy as “the total energy of a system that is available to perform useful work. It is expressed in kilojoules (kJ) and is also called ‘available energy’” [10]. Many thermodynamic systems use this concept to measure the maximum work done by the system at constant pressure and temperature. Figure 5.2 shows the thermodynamic cycle that is used for different ways to calculate free energy.

The ΔG_{bind} is “generally impractical to run a MD simulation long enough to capture an entire binding event” [5]. Next, $\Delta G_{\text{protein}}$ is the change that occurs when the bounded ligand is annihilated. And finally, ΔG_{water} is the change that occurs when unbounded ligands are annihilated [5].

Free energy calculations must be very reliable and fast in order to show the interactions between macromolecules accurately. This holds a great deal of importance in drug discovery because of the research that has been done. It has shown the strength of interactions, macromolecule conformations, drug binding, and transport properties [2]. Therefore, the calculations made must have a level of accuracy higher than other experimental methods. Many scientists find that they can get the link between equilibrium constants and thermodynamics from the free-energy change equation:

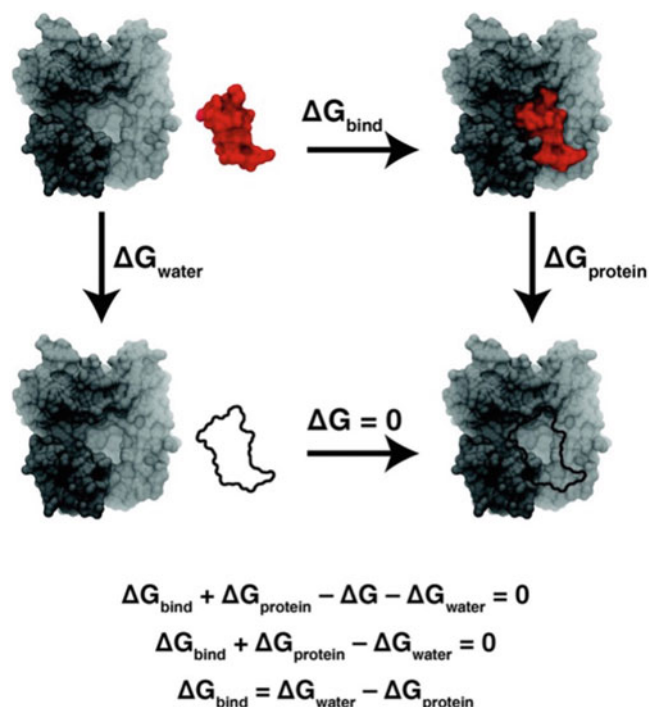


Fig. 5.2 Thermodynamic cycle. (Image taken from [5])

$$\Delta G_b^\circ = -k_B T \ln(K_a C^\circ) = k_B T \ln\left(\frac{K_d}{C^\circ}\right)$$

These calculations are not trivial, and “free energy is the thermodynamic observable whose estimation suffers the most from sampling limitations” [3]. It has been more difficult in the past few years because of insufficient sampling and inadequate force fields. Both factors are nearly impossible to test independently of each other.

5.5 Conclusion and Future Work

At the rate that technology is advancing, there is no doubt that more drug discovery methods with computational methods will become more common. CFD and MD simulations are only two of the methods that can be adapted. As companies and scientists see the benefits of them, we believe there will be a higher demand. The future of biotechnology is promising because of resources like these. Before researching this topic, we were unsure whether we would find enough information to write about.

Nevertheless, to our surprise, we read through many articles that encourage the use of computational methods. It will be very interesting to see when big companies start adapting these methods and making it a common way of research. There is still much development to be made, but we will be more advanced at this rate than ever.

Acknowledgment This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700.

References

1. W.L. Jorgensen, The many roles of computation in drug discovery. *Science* **303**, 1813–1818 (2004)
2. D.W. Borhani, D.E. Shaw, The future of molecular dynamics simulations in drug discovery. *J. Comput. Aided Mol. Design* **26**(1), 15–26 (2011)
3. M. De Vivo, M. Masetti, G. Bottegoni, A. Cavalli, Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* **59**(9), 4035–4061 (2016)
4. A. Meyers, Using computer simulations to create safer, more effective medications, Stanford University School of Engineering, 3 May 2018. [Online]. Available: <https://engineering.stanford.edu/magazine/article/using-computer-simulations-create-safer-more-effective-medications>. Accessed 15 Nov 2021
5. J.D. Durrant, J.A. McCammon, Molecular dynamics simulations and drug discovery. *BMC Biol.* **9**(1), 1–9 (2011)
6. [InternacionalWeb.com](http://www.pretechnologies.com), Advantages of computational fluid dynamics, PRE Technologies. [Online]. Available: <http://www.pretechnologies.com/services/computational-fluid-dynamics/advantage>. Accessed 16 Nov 2021
7. A. Karanjkar, Design, development and optimization using computational fluid dynamics, PharmTech, 15 Nov 2020. [Online]. Available: <https://www.pharmtech.com/view/design-development->

- and-optimization-using-computational-fluid-dynamics. Accessed 16 Nov 2021
8. Global Computational Fluid Dynamics Market Research Report (2020 to 2026) – By deployment, application and region – [researchandmarkets.com](https://www.researchandmarkets.com), Business Wire, 30 July 2021. [Online]. Available: <https://www.businesswire.com/news/home/20210730005496/en/Global-Computational-Fluid-Dynamics-Market-Research-Report-2020-to-2026%2D%2D-by-Deployment-Application-and-Region%2D%2D-ResearchAndMarkets.com>. Accessed 16 Nov 2021
 9. Siemens to acquire simulation software supplier CD-ADAPCO: Siemens Software, Siemens Digital Industries Software, 25 Jan 2016. [Online]. Available: <https://www.plm.automation.siemens.com/global/en/our-story/newsroom/siemens-press-release/43811>. Accessed 16 Nov 2021
 10. S. Cheriyaedath, Gibbs free energy and metabolism, News, 26 Feb 2019. [Online]. Available: <https://www.news-medical.net/life-sciences/Gibbs-Free-Energy-and-Metabolism.aspx>. Accessed 16 Nov 2021

Blended Modeling Applied to the Portable Test and Stimulus Standard

Muhammad Waseem Anwar, Malvina Latifaj, and Federico Ciccozzi

Abstract

Blended modeling is an emerging trend in Model-Driven Engineering for complex systems. It enables the modeling of diverse system-related aspects through multiple editing notations seamlessly, interchangeably, and collaboratively. Blended modeling is expected to significantly improve productivity and user-experience for multiple stakeholders. Case-specific solutions providing blended modeling, to a certain extent, for domain specific languages have been provided in the last few years. Nevertheless, a generic and language-agnostic full-fledged blended modeling framework has not been proposed yet.

In this paper, we propose a comprehensive and generic blended modeling framework prototype that provides automated mechanism to generate graphical and textual notations from a given domain-specific modeling language. Moreover, it offers a flexible editor to get expert's feedback on the mapping between graphical and textual notations. The proposed prototype is validated through a proof-of-concept on the Portable test and Stimulus Standard use-case. Our initial results indicate that the proposed framework is capable of being applied in different application scenarios and dealing with multiple domain-specific modeling standards.

Keywords

Model-driven engineering · Systems engineering · Modeling editor · PSS · Meta-modeling · Blended modeling · Domain specific languages · Design verification · Embedded systems · EBNF

M. W. Anwar (✉) · M. Latifaj · F. Ciccozzi
 School of Innovation, Design and Engineering, Malardalen University,
 Vasteras, Sweden
 e-mail: muhhammad.waseem.anwar@mdh.se; malvina.latifaj@mdh.se;
federico.ciccozzi@mdh.se

6.1 Introduction

Model-Driven Engineering (MDE) is a prominent engineering paradigm for complex systems. By shifting the bulk of engineering activities to a higher abstraction level than code-based approaches and by stressing the importance of automation, MDE simplifies and expedites most engineering activities. MDE facilitates system specification at high abstraction levels through models that, via model transformation processes, are used to automatically run analysis, validation, and produce implementation code [1].

Domain-Specific Modeling Languages (DSMLs) and tools traditionally focus on one specific editing notation (such as text, diagrams, tables or forms). This limits human communication, especially across stakeholders with varying roles and expertise. A notation that is well-understood by one stakeholder role may not be understood by others. Besides the limits to communication, choosing one particular kind of notation has the drawback of limiting the pool of available tools to develop and manipulate models that may be needed.

For systems with heterogeneous components and entailing different domain-specific aspects and different types of stakeholders, mutual exclusion is too restrictive and voids many of the MDE benefits. Therefore, DSMLs and tools need to enable different stakeholders to work on overlapping parts of the models using different modeling notations (e.g., graphical and textual). The seamless and collaborative modeling using multiple notations, called blended modeling, was introduced by Ciccozzi et al. [2] as blending of graphical and textual notations with seamless synchronization, which has been shown to be particularly beneficial for boosting collaboration, communication and modeling time [3].

To produce a blended modeling environment, four major steps need to be carried out. In the first step, starting from a DSML, graphical and textual notations are defined or generated. In the second step, a complete mapping between

graphical and textual concepts is carried out. In the third step, automated synchronization mechanisms based on the mapping are established to attain seamless switching between graphical and textual notations. Eventually, all pieces are embedded into a modeling environment to provide blended modeling. Existing blended modeling solutions are either specific to certain DSMLs (e.g., Ericsson Hive [4], HCL RTist [5]) or focus on specific aspects of blended modeling only (e.g., seamless synchronization [3]). Furthermore, existing studies do not cover important blended modeling features such as the automated generation of (graphical and textual) notations, or mechanisms to get domain expert's feedback on the mapping between notations.

This paper describes our first step towards a comprehensive and generic framework for the generation of blended graphical–textual modeling environments from virtually any DSML. The framework provides a prototypical solution for the aforementioned four steps: (1) it provides automated mechanisms for the generation of graphical and textual notations starting from a DSML definition (i.e. meta-model), (2) it provides a specific editor for describing mappings between graphical and textual notations, (3) it provides automatic synchronization between graphical and textual notations, and (4) it offers an automated mechanism to generate the blended modeling infrastructure. The viability of the proposed framework is established through a proof-of-concept by leveraging the Portable test and Stimulus Standard¹ (PSS) language. The PSS is an emerging system verification standard that allows the specification of test intents through a C++-like Domain Specific Language (DSL). We utilized the proposed framework to produce a blended modeling environment for the specification of test intents through the PSS textual DSL and an ad hoc graphical notation. The initial results are encouraging and show the suitability of the proposed framework to be used for multiple DS(M)Ls.

The remainder of this paper is organized as follows. Section 6.2 sets the background and scope of this research. An overview of our solution is described in Sect. 6.3, while the implementation details and the proof-of-concept through the PSS use case are described in Sect. 6.4. The core aspects of the proposed framework together with a set of future directions are discussed in Sect. 6.5. Eventually, Sect. 6.6 concludes the paper.

6.2 Background and Related Work

The aim of this section is to contextualize our research work and describes its scope. Particularly, the existing literature on blended modeling and the need for the proposed framework

are discussed in Sect. 6.2.1. Furthermore, a justification of the selection of the PSS use case is given in Sect. 6.2.2.

6.2.1 Blended Modeling

Ciccozzi et al. [2] formally introduced the concept of blended modeling as:

Blended modeling is the activity of interacting seamlessly with a single model (i.e., abstract syntax) through multiple notations (i.e., concrete syntaxes, allowing a certain degree of temporary inconsistencies.)

The notion of blended modeling may seem similar or even overlapping with multi-view modeling [6] that is based on the paradigm of viewpoint/view/model as formalized in the ISO/IEC 42010 standard. For example, Wimmer and Kramler [7] proposed a conceptual mapping between the textual languages EBNF and corresponding meta-models to achieve multi-view modeling features. In another study [8], Scheidgen proposed a multi-view modeling approach using a parsing tree technique to provide textual editors for Graphical Modeling Framework (GMF) based editors. Although these studies provide supporting features for blended modeling, some essential characteristics of blending, such as seamless synchronization between graphical and textual editors, are missing.

While multi-view modeling aims at defining viewpoints/views, blended modeling aims at providing a powerful multi-notation characterization that may be used to define viewpoints/views. Maro et al. [4] proposed a blended modeling methodology for Ericsson Hive – a proprietary UMLbased DSML. Particularly, authors proposed an approach to automatically generate an Ecore DSML from the Hive profile, exploited the features of Xtext to provide textual modeling for Hive, and achieved synchronization between graphical and textual notations. However, the synchronization is based on two separate resources (i.e., UML and Xtext) and rely on fixed bi-directional transformations. Addazi et al. [3] proposed a blended modeling approach for UML profiles where synchronization between graphical and textual notations is achieved through one single underlying resource (i.e., UML), thus minimizing the need for complex synchronization transformations. Recently, Latifaj et al. [5] proposed blended modeling support for UML-RT state machines through graphical and textual notations in the HCL RTist environment.²

To summarize, existing approaches either focus on techniques orthogonal to blended modeling, e.g. multi-view modeling, or provide partial blended modeling solutions for specific languages. Moreover, existing studies do not target

¹<https://www.accellera.org/downloads/standards/portable-stimulus>

²<https://www.hcltech.com/brochures/software/hcl-rtist>

certain core blended modeling aspects, such as mechanism to generate both graphical and textual notations automatically. Furthermore, mechanism to record domain expert’s feedback during the notations mapping process to ensure a semantically accurate synchronization is missing. To conclude, there is a dire need for a unified, automated and customizable blended modeling solution that is not dependent on a single language, but rather applicable to multiple languages and application domains. As a first step towards such a solution, a comprehensive blended modeling framework prototype is proposed in the forthcoming sections.

6.2.2 Selection of PSS Use case

The main rationale behind PSS is to specify the test scenarios once with simplicity and reuse them for different applications. In terms of reusability, simplicity and integration testing, PSS has certain advantages over existing verification methods like Universal Verification Methodology (UVM) and formal methods [9]. Here, we chose PSS as a use-case for concept-proving our solutions for the following two reasons. First, there is no full-fledged modeling environment for PSS available at the moment (e.g., existing tools provide only visualization of test scenarios through UML activity diagram after the specification of test intents). In this context, the automatic generation of PSS graphical notations through our framework is actually more challenging as compared to other standards, like e.g. EASTADL, where a formal meta-model specification exists. Secondly, the provision of blended modeling features for PSS would significantly simplify the specification of test intents. Consequently, companies could swiftly enhance their processes around PSS with our framework without any major effort.

6.3 Blended Modeling Framework: An Overview

The architecture of the proposed framework is based on three main components depicted in Fig. 6.1 and described in the following.

Generation of Graphical and Textual Syntaxes This component is responsible to automatically generate the required graphical and textual notations from the input metamodels and textual DSL/grammar, respectively. Particularly, in the “Graphical Notations” component, we developed a model-to-text transformation in Aceleo and Java for the generation of graphical notations. Furthermore, the “Textual Notations” component is implemented in Java and it provides a user interface to effectively generate textual notations from given DSL/grammar samples. The generated graphical and textual notations are saved in XML format and serve as an input for the Mapping & Synchronization component.

Mapping & Synchronization This component performs mapping and synchronization activities by utilizing the input graphical and textual notations described in XML. Particularly, in the first step, a mapping between input graphical and textual notations is created. For this, the feedback of domain experts is important to ensure the syntactical and especially semantic correctness of the mapping of modeling concepts across notations. For this, a flexible mapping editor is developed using Java and the WindowBuilder library. The mapping information is saved in XML format and is used by the Synchronizer component as input for the implementation of an Extended Backus–Naur form (EBNF) grammar. This grammar provides then rules for seamless synchronization between graphical and textual notations. The implementation of EBNF grammars in Synchronizer is carried out through the JAVACC platform [10]. The output is an executable

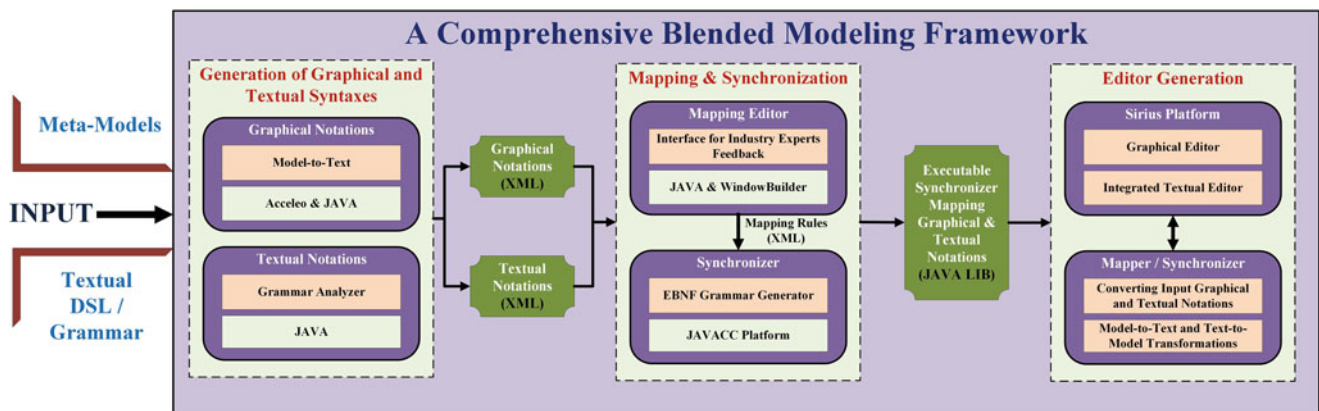


Fig. 6.1 Blended modeling framework architecture

Synchronizer, packaged as a Java library, which is used in the modeling editors to achieve seamless synchronization and switching across graphical and textual notations.

Editor Generation This component is responsible to generate the actual editors for graphical and textual notations. Here, the features of the Sirius platform [11] are utilized to provide both graphical and textual editors in the blended modeling environment. We provide model-to-text and text-to-model transformations, implemented in Java and Aceleo, to propagate model changes across notations. Note that the seamless switching between graphical and textual notations is achieved with the help of the Synchronizer.

6.4 Implementation and PSS use case

In this section, the implementation details of each component are briefly explained. The proposed framework can be downloaded as an Eclipse plugin for further exploration.³ Moreover, a demo of the framework in action is available as well.⁴ Furthermore, the application of proposed framework is demonstrated through PSS use case.

6.4.1 Input Format

The proposed framework requires a textual DSL/grammar (plain text) and a meta-model (in Ecore [12]) for the generation of textual and graphical notations respectively. In our use case, PSS carries along a C++-like DSL for the specification of test intents. The samples of PSS DSL are available in the PSS specifications and can be given as an input to our framework for the generation of textual notations for PSS. On the other hand, a meta-model of PSS is not available to the best of our knowledge. Therefore, we propose a basic PSS metamodel in EMF [12] as shown in Fig. 6.2. The main PSS concepts like actions, objects, resources etc. and their semantics are included in the meta-model. This meta-model is given as an input to our framework for the generation of graphical notations.

6.4.2 Generation of Graphical and Textual Notations

With this component the user selects a DSL/grammar file as an input through the “Browse” button and starts the generation of the required tokens (textual notation) using the

“Parse” Button. The generated grammar tokens can be saved in an XML file, which is further utilized in the mapping process. This module is capable of generating textual notations for different types of DSLs/grammars. For demonstration purposes, a grammar sample from the PSS DSL is considered, where an action is specified through input and output buffer types. The respective textual notations (tokens) are generated and subsequently saved as shown in Fig. 6.3. For the generation of graphical notations, a generic component is developed via model-to-text transformations in Aceleo and Java. Particularly, it takes a meta-model as an input and generates the corresponding graphical notation and semantic information in XML format. For demonstration, the PSS metamodel (Fig. 6.2) is given as an input and the generated XML is shown in Fig. 6.4.

6.4.3 Mapping & Synchronization

The domain expert’s feedback is crucial in the mapping process. For this purpose, a mapping editor is developed as shown in Fig. 6.5. In terms of reusability and portability, the mapping editor is flexible and can be used for any pair of graphical and textual notations. Importantly, all the interface components are generated dynamically as shown in Fig. 6.5. Particularly, it displays the graphical and textual notations that are generated by the first component. Furthermore, it displays the repository of symbols to be associated with the graphical notation via XML (containing addresses and IDs of symbols) and, therefore, other symbols specific to the domain can be added to the mapping editor. Additionally, it provides AND/OR operators to define complex mappings.

In the figure we show the mapping between the PSS graphical and textual notations, as well as the association of graphical symbols to the PSS concepts. More specifically, the symbol with Id = 1 and Name = Action is associated with the graphical action concept. The assigned symbol will be available in the blended modeling editor for the modeling of graphical action. On the other side, the textual syntax for action is specified as: action name {}. Subsequently, this mapping between graphical and textual action can be added to the queue (grid), as shown in Fig. 6.5. Similarly, the mapping between graphical and textual notations for other concepts like buffer etc. is performed and saved in XML.

The Synchronizer component is responsible to generate an EBNF grammar by utilizing the mapping XML. This step is essential for seamless synchronization and switching between graphical and textual. The JAVACC platform is used to implement this component. Few EBNF mapping grammar rules for the PSS action and buffer concepts are given here for demonstration purposes:

³Framework available for download at: <https://github.com/blended-modeling/PSS>

⁴Demo available at: https://play.mdh.se/media/t/0_4t63df9w

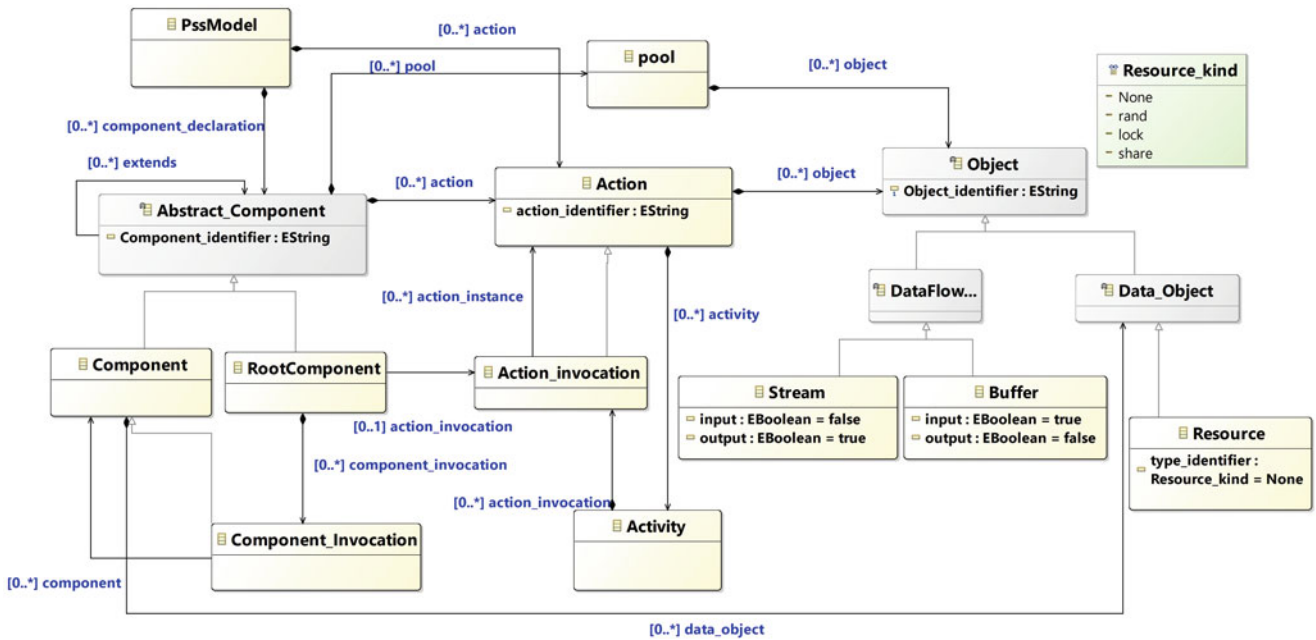


Fig. 6.2 PSS meta-model

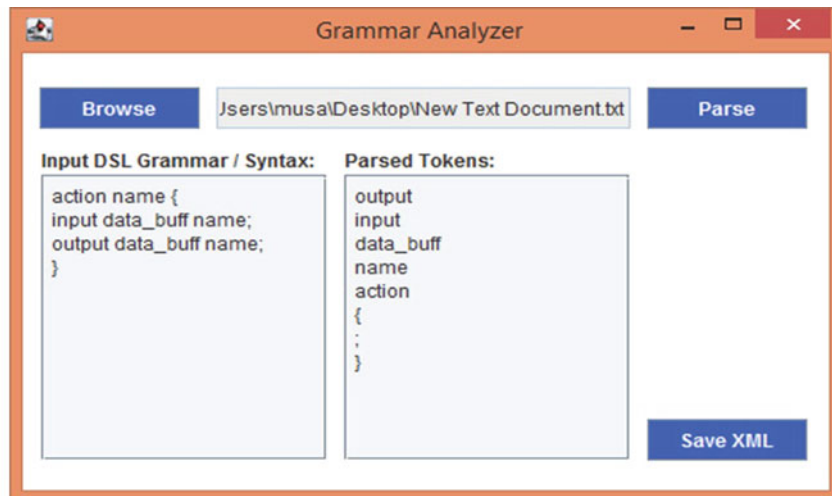


Fig. 6.3 Grammar analyzer and textual tokens

- Rule 1:** <Action> ::= <Graphical-Action> | <Textual-Action>
- Rule 2:** <Graphical-Action> ::= <Name><Symbol> | <Name> <Symbol><Relationship>(<Graphical-Data Buffer>)*
- Rule 3:** <Textual-Action> ::= action <Name> { } | action <Name> { (<Textual-Data Buffer>)* }
- Rule 4:** <Data Buffer> ::= <Graphical-Data Buffer> | <Textual-Data Buffer>
- Rule 5:** <Graphical-Data Buffer> ::= <Type><Name> <Symbol>
- Rule 6:** <Textual-Data Buffer> ::= <Type> data buff <Name>

Rule 7: <Relationship> ::= Containment <Symbol>

Rule 8: <Name> ::= ([a-z][A-Z][0-9])*

Rule 9: <Symbol> ::= ([a-z]/[/][\][A-Z][0-9])*

Rule 10: <Type> ::= input | output

For better understanding, consider a simple PSS DSL example where an action having two buffers is defined as:

```
action mem2mem {
input data buff src Buffer;
output data buff dst Buffer; }
```

On the basis of aforementioned EBNF rules, a parsing tree of given example is shown in Fig. 6.6 to achieve seamless synchronization and switching between graphical and textual

Node	Content
▷ [e] concept	
▷ [e] concept	
▲ [e] concept	
[e] name	Action
[e] attributes	
[e] operations	
▲ [e] references	
[e] name	action_instance
[e] lowerBound	0
[e] upperBound	-1
[e] containingClass	Action_invocation
[e] connectedClass	Action
▷ [e] concept	
▷ [e] concept	
▷ [e] concept	
▷ [e] concept	
▲ [e] concept	
[e] name	Buffer
[e] attributes	
[e] operations	

Fig. 6.4 Generated XML

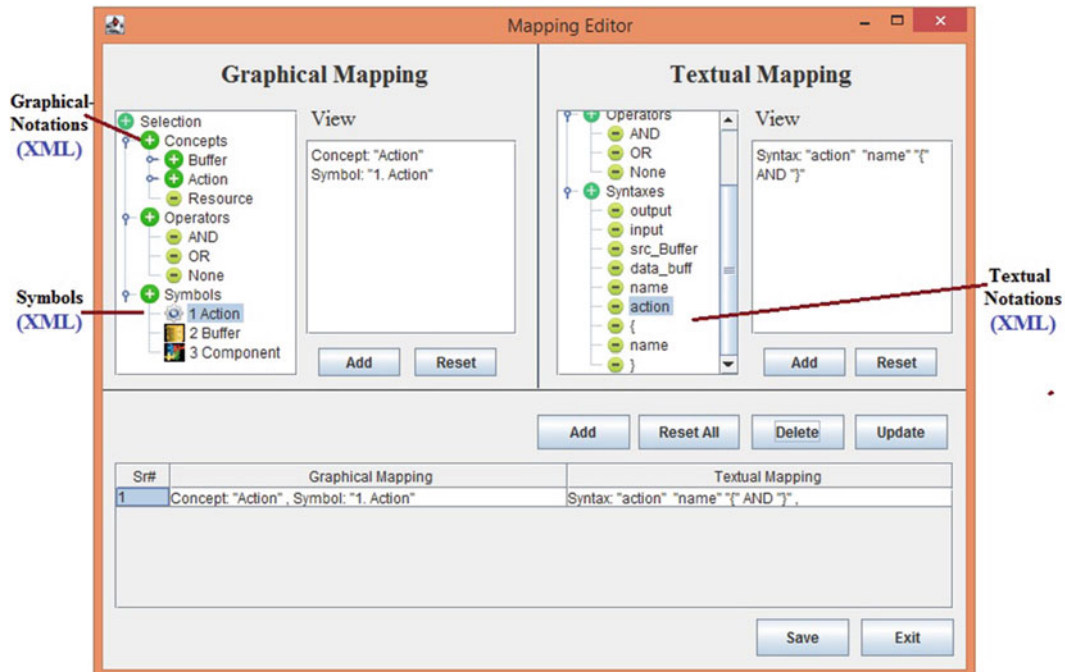


Fig. 6.5 Mapping editor

notations. Please note that terminal symbols are displayed in orange boxes

6.4.4 Editor Generation

The Sirius [11] platform is used to generate the blended modeling editor. The architecture of the editor generation

component is shown in Fig. 6.7. In the first step, a graphical editor is generated in Sirius by utilizing the PSS meta-model. In the second step, a textual view is incorporated in the graphical editor for the specification of textual PSS. The blended modeling editor supports seamless synchronization and switching between graphical and textual notations using Java services, model-to-text and text-to-model transformations, and the Synchronizer component. Particularly, Java services are used to communicate with the Synchronizer,

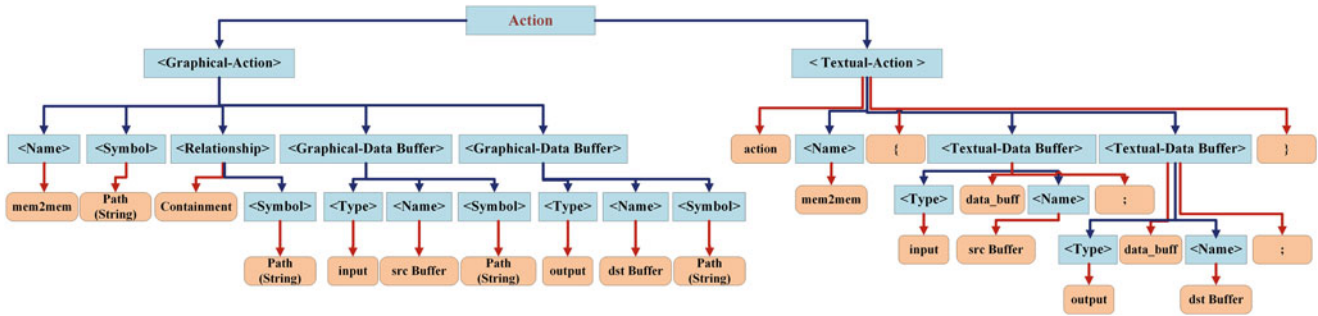


Fig. 6.6 Parsing tree of the PSS use case

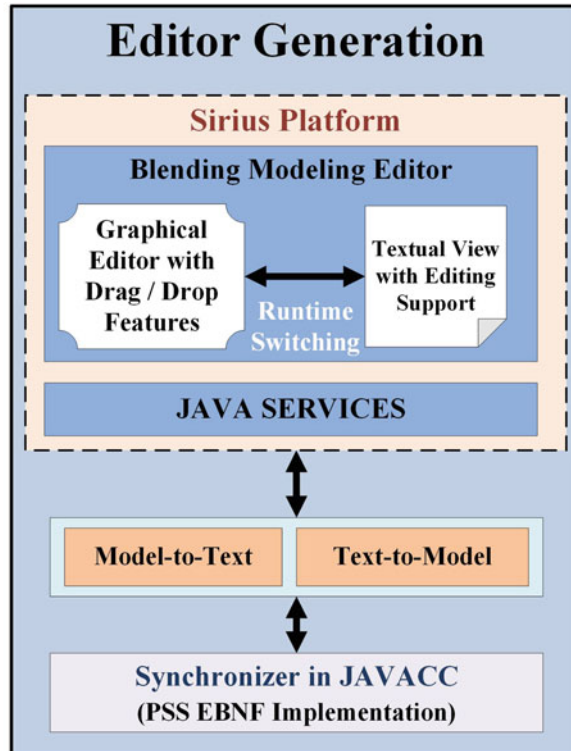


Fig. 6.7 Editor generation component

while model-to-text and text-to-model transformations are intermediate components to perform suitable propagation of model changes between graphical and textual notations and display them in the blended modeling editor accordingly for seamless switching. The generated PSS blended modeling editor offers drag/drop functionality for different PSS graphical elements (provided in a palette) with suitable symbols that were chosen during the mapping process.

6.5 Discussion and Future Perspectives

The proposed blended modeling framework prototype provides a generic and automated methodology for the generation of graphical and textual notations. Additionally, it pro-

vides a flexible editor to consider domain expert’s feedback for the mapping of graphical and textual notations. However, activities like generation of synchronization mechanisms and editors are currently not fully automated. For example, the implementation of the mapping EBNF grammar requires some manual intervention from the user.

Similarly, in the editor generation process, the model-to-text and text-to-model transformations were implemented manually and specifically for the PSS use case. In this regard, we are currently working on different industrial use cases and we expect to minimize manual interventions by the use of Higher Order Transformations (HOTs).

Scalability is an important issue in industrial contexts. Currently, the time required for seamless switching from across notations for the presented DMA example with the

proposed framework is approx. 400 milliseconds, on Dell Inspiron Core i5 5th Generation Laptop with 8 GB RAM. We are working on more complex examples to credibly analyze the time required for seamless switching and thereby the scalability of our solutions. Furthermore, we are exploring different platforms like Xtext/Sirius integration to implement Synchronizer and editor generation components.

The proposed framework requires both meta-models and DSL grammar/syntax samples to effectively generate graphical and textual notations respectively. In this context, it can be argued that the existing standards usually do not provide both meta-models and DSL grammars simultaneously and therefore, a broader applicability of the proposed framework is questionable. To address this issue, we have developed a complete solution to automatically generate Ecore meta-models from restricted natural language. This allows to develop the required meta-models without going into the technical hitches. The details regarding the automatic generation of meta-models are beyond the scope of this paper.

6.6 Conclusion and Future work

This paper presents a comprehensive and open-source framework to achieve blended modeling environment for different design and verification standards. Particularly, it provides an automated mechanism for the generation of graphical and textual notations from the given meta-models and DSL grammars respectively. Moreover, it offers a mapping editor to entail domain expert's feedback for the accurate mapping between graphical and textual notations. Furthermore, it allows seamless switching between notations with the help of EBNF grammars. Finally, it supports the development of desired blended modeling environment through the Sirius Platform. The viability of the proposed framework is shown through the PSS use case. Particularly, a complete blended modeling environment is developed for the specification of test intents through the PSS textual DSL and a PSS graphical notation, interchangeably. The initial results are encouraging

and show that the proposed framework can be applied on multiple domain-specific modeling standards to achieve the desired blended modeling.

Acknowledgement This work was supported by Vinnova through the ITEA3 BUMBLE project (rn. 18006) and the Knowledge Foundation through the HERO and MoDev projects.

References

1. C. Verbruggen, M. Snoeck, Model-driven engineering: A state of affairs and research agenda. *Enterprise, business- process and information systems modeling* (2021), pp. 335–349
2. F. Ciccozzi, M. Tichy, H. Vangheluwe, D. Weyns, Blended modelling – What, why and how, in *Proceedings of MODELS companion* (IEEE, 2019), pp. 425–430
3. L. Addazi, F. Ciccozzi, Blended graphical and textual modelling for uml profiles: A proof-of-concept implementation and experiment. *J. Syst. Softw.* **175**. Elsevier (2021)
4. S. Maro, J.-P. Steghofer, A. Anjorin, M. Tichy, L. Gelin, On integrating graphical and textual editors for a uml profile based domain specific language: an industrial experience, in *Proceedings of SLE* (2015), pp. 1–12
5. M. Latifaj, F. Ciccozzi, M. Mohlin, E. Posse, Towards automated support for blended modelling of uml-rt embedded software architectures, in *Proceedings of ECSA* (2021)
6. A. Cicchetti, F. Ciccozzi, A. Pierantonio, Multi-view approaches for software and system modelling: A systematic literature review. *Softw. Syst. Model.* **18**(6), 3207–3233 (2019)
7. M. Wimmer, G. Kramler, Bridging grammarware and modelware, in *Proceedings of MoDELS* (Springer, 2005), pp. 159–168
8. M. Scheidgen, Textual modelling embedded into graphical modelling, in *Proceedings of ECMFA* (Springer, 2008), pp. 153–168
9. G. Bhatnagar, D. Brownell, Portable Stimulus vs Formal vs UVM A Comparative Analysis of Verification Methodologies Throughout the Life of an IP Block (DVCon, San Jose, 2018)
10. V. Kodaganallur, Incorporating language processing into java applications: A javacc tutorial. *IEEE Softw.* **21**(4), 70–77 (2004)
11. V. Viyovic, M. Maksimovic, B. Perisic, Sirius: A rapid development of DSM graphical editor, in *Proceedings of INES* (IEEE, 2014), pp. 233–238
12. D. Steinberg, F. Budinsky, E. Merks, M. Paternostro, *EMF: Eclipse Modeling Framework* (Pearson Education, Upper Saddle River, 2008)

An Evaluation Framework for Modeling Languages Supporting Predictable Vehicular Software Systems

Enxhi Ferko, Igli Jasharllari, Alessio Bucaioni, Mohammad Ashjaei, and Saad Mubeen

Abstract

Handling the software complexity of modern vehicular systems has become very challenging due to their non-centralized nature and real-time requirements that they impose. Among many software development paradigms for these systems, model-based development excels for several reasons including its ability to verify timing predictability of software architectures of these systems using pre-runtime timing analysis techniques. In this work, we propose a comprehensive framework that captures the timing related information needed for the modeling languages to facilitate these timing analyses. We validate the applicability of the framework by comparing two modeling languages and their respective tool-chains, Rubus-ICE and APP4MC, that are used for software development in the vehicle industry. Based on our results, both modeling languages support the design and analysis of vehicle software, but with different. Both modeling languages support time-, event- and data-driven activation of software components and modeling of single- and multi-rate transactions. Amalthea targets applications on single nodes with multi-core architectures while RCM focuses on single-core single-node and distributed embedded systems with ongoing work for supporting single-node multi-core architectures. In comparison to Amalthea, RCM provides a generic message model which can easily be re-modeled according to protocol-specific properties.

Keywords

Model-based software development · Model-based design · Model-based verification · Modelling languages · Rubus Component Model · Amalthea · Automotive software · Real-time systems · Timing analysis · Timing verification

7.1 Introduction

The software in modern vehicles consists of millions of lines of code that run on several tens of Electronic Control Units (ECUs) [1, 2]. According to a recent study from Jaguar Land Rover [3], the size of vehicle software will soon reach 1 billion lines of code. To cope with the increasing size and complexity of the software during its development, researchers and practitioners have successfully adopted emerging software development paradigms, including model-based software development [4]. Model-based software development revolves around meta(models) [5, 6] and model transformations [7]. Among numerous benefits of this paradigm is its ability to support model-based timing verification of the software architectures. This is crucial in the vehicle industry as the systems' developers are required to verify the *timing predictability* of these systems. A system is considered timing predictable if it is possible to prove at the design time that all the specified timing requirements are satisfied [8, 9].

In recent years, several modeling languages and methodologies that are supported by the model-based timing analysis have been developed, e.g., the Rubus Component Model (RCM) [10] and Amalthea [11]. However, these languages are characterized by different approaches towards the modeling of timing information (timing properties, requirements and constraints). As a result, the timing information that a

E. Ferko (✉) · I. Jasharllari · A. Bucaioni · M. Ashjaei · S. Mubeen
Mälardalen University, Västerås, Sweden
e-mail: enxhi.ferko@mdu.se; igli.jasharllari@mdu.se;
alessio.bucaioni@mdu.se; mohammad.ashjaei@mdu.se;
saad.mubeen@mdu.se

timing analysis tool requires as an input may be partially or fully extractable from the software architectures that are modeled with these languages. Consequently, timing analysis of the software architectures that are modeled with the languages that have limited expressiveness for timing information cannot be supported without making over-estimated assumptions.

To date, few researches have focused on characterizing and comparing modeling languages with respect to timing perspective as most of them focus on other extra-functional properties and requirements. In this context, it would be beneficial to both researchers and practitioners to have means for classifying and categorizing modeling languages with respect to modeling timing information and analyses they enable. Such a classification can not only establish a basis for comparing the timing modeling capabilities of the languages, but also identify which of the modeling languages would be more appropriate for a specific vehicular application. We address the following research problem: *How can we categorize and compare modeling languages based on a theoretical framework considering their timing expressiveness?* The main contribution of this work is a comprehensive framework for characterising modeling languages supporting timing analysis of software architectures of vehicular systems. To show the applicability of the proposed framework, we apply it to categorize two industrial modelling languages widely used in the vehicular domain, namely the RCM [10] and Amalthea [11] and their respective tool-chains Rubus-ICE [12] and APP4MC [13].

7.2 Background and Related Work

7.2.1 Rubus Component Model (RCM)

Rubus is a collection methods and tools for model-based software development of real-time embedded systems. RCM, core of the Rubus approach, is a modelling language. It is developed by Arcticus Systems¹ in co-operation with partners from academia and industry. RCM has been used in the vehicular domain for more than 25 years for the development of in-vehicle control functionalities [10]. The Rubus-ICE tool suite consist of modeling and analysis tools, code generators as well as run-time infrastructure for the modeled applications. Within Rubus-ICE, the applications are developed and described graphically as inter-connected components, called Software Circuits (SWCs). The SWCs are characterized by run-to-completion semantics meaning that, upon triggering, an SWC reads the input data, processes it according to the internal behavioral logic and provides an output. Figure 7.1

illustrates a multi-rate data chain in RCM which consists of 3 SWCs.

7.2.2 Amalthea Model and APP4MC

APP4MC is an open-source European project under Eclipse Public License whose objective is to provide a tool-chain for the engineering of real-time multi-core embedded systems [13]. It originates from the Amalthea and Amalthea4Public ITEA2 research projects [14]. It is still under development as it aims to cover all development cycles of software starting from design to verification and validation of these systems. Amalthea can be seen as an extension to the AUTOSAR [15] standard from various aspects and is envisioned to be a standard for multi-core real-time embedded systems in the vehicular domain [16]. To date, it provides the design and run-time infrastructure of real-time embedded system applications. Systems are designed based on the Amalthea model [11] where the basic hierarchical element is called *Runnable*. A Runnable encapsulates the basic functions. They might be grouped in clusters and mapped to a task which is activated by *Stimuli*. Systems are designed as interconnected components, but yet not graphically supported. Figure 7.2 illustrates how this is concretely done in Amalthea.

7.2.3 Related Work

The research conducted in [4] and [17] presents an overview of the current modeling languages and their respective tool support in the vehicular domain. A comprehensive description regarding the timing modeling and timing analysis is given. The modeling languages and tools are grouped based on the four levels of abstractions for vehicular embedded systems development. In comparison to these works, we aim to develop a comparative evaluation framework for the languages based on their expressiveness to model timing information on the software architectures and timing analyses they enable. A framework for comparing real-time modeling languages based on different aspects such extension capabilities and tool support is presented in [18]. Despite we believe that the properties in [18] are very important when choosing a modeling language, in this work we aim to develop a comprehensive framework with in-depth investigation of their capabilities for describing the timing behavior of vehicular software systems.

Another comparison framework, developed in [19], takes into consideration mainly the functional, but also non-functional aspects of modeling languages. It identifies the aspects of the software architecture that should be modeled by an architectural language. An extension of this framework

¹Arcticus Systems—<https://www.arcticus-systems.com/>.

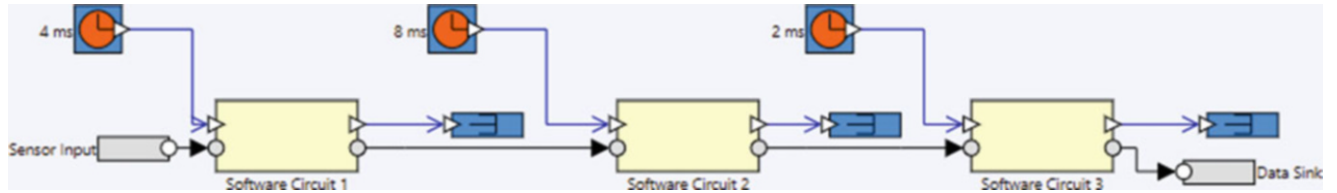


Fig. 7.1 Multi-rate data chain modeling in RCM

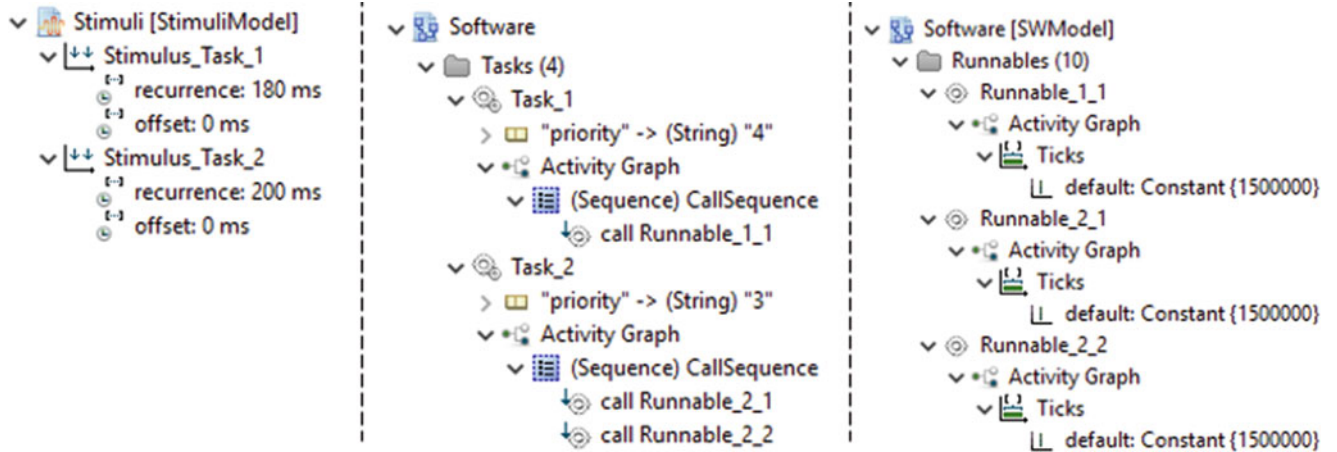


Fig. 7.2 Runnables and Tasks modeled with Amalthea

is presented in [20], which incorporates several other features of hardware architectures and non-functional characteristics such as timing and dependability for safety-critical systems. Even though the framework is extended, modeling of timing information is not the primary focus of the work in [20]. Whereas, we focus on filling this gap by providing the minimal criteria for a modeling language to describe the timing behavior of a vehicular software system and enable its timing analysis.

A survey of component-based modeling languages for embedded systems and their classification framework is presented in [21]. The survey provides a minimal criteria to classify several domain-specific modeling languages. In comparison to this work, we aim at classifying and comparing modeling languages with respect to their expressiveness to support modelling of timing information and provision of timing analysis in the vehicular domain only. The work in [22] defined a set of features and capabilities that a modeling language should support to enable timing analysis. The scope of this work is limited to two specific modeling languages. In comparison, we present a generic classification framework.

To the best of our knowledge, this work represents the first attempt to categorise and compare modeling languages for vehicular software systems based on their timing expressiveness, which is responsible for enabling timing analysis of the software architectures with efficiency and precision.

7.3 The Classification Framework

This section presents a comprehensive framework for categorizing and comparing different modeling languages based on their expressiveness to model timing information on the vehicular software architectures. The framework is graphically illustrated in Fig. 7.3. The top-level entities in this framework include the timing model and timing verification.

7.3.1 Timing Model

The timing model is a crucial input for model-based timing analysis tools. It comprises timing properties and requirements from each node (ECU) and network in the vehicular system.

7.3.1.1 Timing Properties

The timing properties are categorized by node properties and network properties. The node properties are defined based on the model in [23], which can be represented by the following tuple.

$$\tau = \{C, T, D, P, J, O, B, R\} \quad (7.1)$$

Above, C represents the worst-case execution time (WCET) that a task takes to execute on a particular hardware without considering any interference. A task can be activated

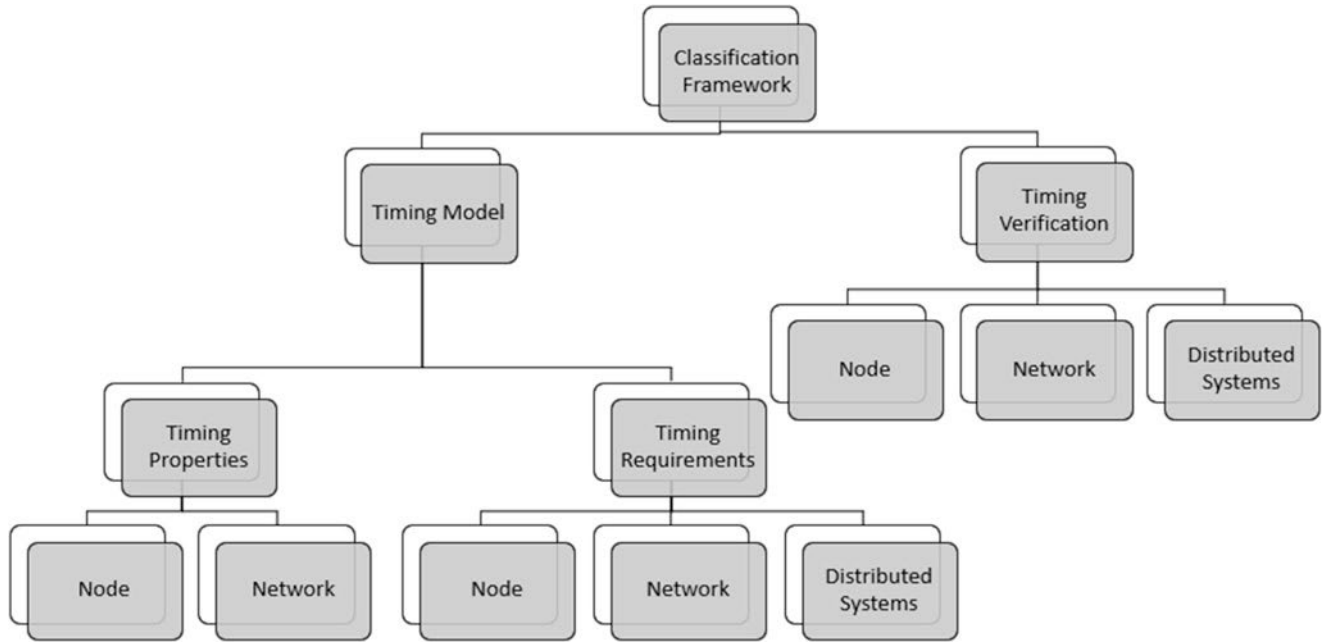


Fig. 7.3 Classification framework

periodically or sporadically with the minimum inter-arrival time denoted by T . Moreover, a deadline D is defined as the time that the task should finish its execution in order to not disrupt the service (soft deadline) or lead to a catastrophe (hard deadline). A task has a priority which indicates the importance of the task shown by P . The period is defined either manually or using a priority assignment algorithm, such as the rate monotonic priority assignment. The time difference between the earliest and latest activation time of the task is called jitter, which is shown by J . An offset O can be set for a task which indicates the time that the task is activated after its arrival. A task can be blocked by lower priority tasks due to occupied shared resources and the maximum blocking time is denoted by B . Finally, the worst-case response time of the task R is the maximum difference between the finishing and arriving time considering the interference and blocking. Besides the above timing properties, a node in a distributed embedded system can contain several transactions with multiple tasks. Therefore, additional information related to the transactions belonging to a node should be provided. In multi-core architectures, where nodes are partitioned in two or more cores, the basic properties for nodes remain the same [24,25]. However, additional run-time information is needed for example task affinity properties.

A distributed embedded system contains one or more networks. A network is usually characterized by a data transmission speed and the communication protocol. Similar to the tasks within a node, a message in a network has a set of properties, which are specified in the following tuple.

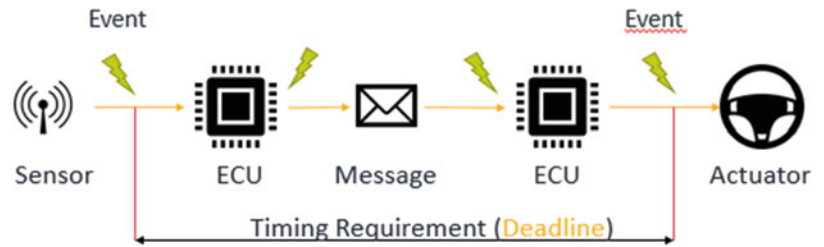
$$m = \{S, C, T, D, O, P, J, B, R\} \quad (7.2)$$

In the above tuple, a message contains a payload, which is the actual data to be transferred in the network, that is defined by S . Depending on the network speed, the message transmission time can be derived from the payload size and the message overhead, which is modeled by C . Similar to tasks, messages can be sent periodically or sporadically that is shown by T , while the deadline for the message delivery is denoted by D . A message can be activated with an offset, shown as O . The priority of the message is shown by P in the network. Moreover, the message may have a release jitter, denoted by J . The message transmission can be blocked by the lower priority message transmission that blocks the medium and the maximum blocking time is denoted by B . The worst-case response time of the message is also defined by R .

7.3.1.2 Timing Requirements

These requirements are specified on the software architectures by means of timing constraints on events or event chains [26]. Figure 7.4 exemplifies a timing requirement on the distributed functionality between the event from the sensor detecting road obstacles and the event responsible for the steering wheel actuation. Timing constraints are often specified using two attributes representing lower and upper bounds. A constraint is satisfied when the occurrences of the events happen within these limits. This framework categorizes the timing constraints into node and network constraints, which is aligned to the AUTOSAR standard [15].

Fig. 7.4 Example of a timing requirement on distributed vehicular functionality



The constraints that can apply on nodes are as follows.

- **Delay Constraint:** it constrains the time distance between the source triggering event and target response event.
- **Strong Delay Constraint:** this constraint is similar to the delay constraint with the difference that both source and target response events have equal period.
- **Order Delay:** it is a minor instance of the Strong Delay Constraint where the low attribute limit is set to zero and the high limit is set to infinite. Moreover, the occurrence of the two events are not allowed to coincide.
- **Repetition Constraint:** it constrains the activation pattern of the occurrences of a single event under the presence of jitter.
- **Repeat Constraint:** it is an instance of the Repetition Constraint where an event is not allowed to experience any Jitter.
- **Sporadic Constraint:** it constrains the activation pattern of two consecutive occurrences of a single event to be sporadic. Additionally, the time distance between the two occurrences must be equal or higher than the minimum inter-arrival time.
- **Periodic Constraint:** it is a special instance of the Sporadic Constraint where the minimum and maximum inter-arrival times are equal. This is also referred as the event's period.
- **Burst Constraint:** it is also a special instance of the Sporadic Constraint where Jitter is not allowed. Moreover, in a particular time-interval there is a fixed number of occurrences of the event and the time distance between two consecutive occurrences is higher than the minimum inter-arrival time.
- **Pattern Constraint:** it constrains the occurrences of the event to be activated following a specific pattern where fixed periodic points are pre-defined. Fixed periodic points are considered with respect to the offset.
- **Arbitrary Constraint:** it constrains the activation pattern of the occurrences of the event to be random and irregular.
- **Execution Time Constraint:** it constrains the execution time of a function where the preemptions or blocking during execution are not considered.
- **Synchronization Constraint:** it constrains a group of event occurrences in which an occurrence should be within a margin limit, known as the tolerance window.
- **Strong Synchronization Constraint:** it is similar to the Synchronization Constraint, however, the tolerance windows should not overlap.
- **Output Synchronization Constraint:** it constrains the occurrence of a group of events corresponding to a stimulus.
- **Input Synchronization Constraint:** it constrains the way how occurrences of stimuli events follow each other corresponding to a response event.
- **Reaction Constraint:** it constrains the time distance that a response event corresponding to a stimuli event must occur.
- **Age Constraint:** it is defined as the maximum data age allowed from the input to the output of a chain.

7.3.2 Timing Verification

After the system has been modeled, its timing information should be extracted in order to provide evidences on the system timing predictability. To this end, timing analysis are used to check whether or not the constraints are satisfied. We categorize the timing verification parameters as follows:

- Verification of timing constraints applied to the frequency and occurrence of a single event such as Repetition, Repeat, Period constraints.
- Verification of synchronization timing constraints applied to a group of events.
- Verification of timing constraints applied to the time distance between the occurrence of the source triggering event and the occurrence of target response event.
- Verification of timing constraints applied to event chains or distributed event chains.

7.4 Case Study

To demonstrate the applicability of the proposed framework, we use it to categorise and compare two industrial modelling languages: RCM [10] and Amalthea [11].

Table 7.1 Timing properties of element responsible for run-time behavior and communication

Timing properties					
Node	RCM	Amalthea	Network	RCM	Amalthea
WCET (C)	✓	✓	WCTT	✓	–
Priority (P)	✓	✓	Priority (P)	✓	–
Offset (O)	✓	✓	Offset (O)	✓	–
Deadline (D)	✓	✓	Deadline (D)	✓	–
Period (T)	✓	✓	Period (T)	✓	–
Jitter (J)	✓	✓	Jitter (J)	✓	–
Response	✓	✓	Response	✓	–
Time (R)			Time (R)		–
			Pattern (F)	✓	–
			Payload (S)	✓	–
			Inter-arrival	✓	–
			Time (IAT)		–

7.4.1 Evaluation Based on Modeling of Timing Properties

Both languages have been developed for supporting the modeling of timing information of automotive software. Besides, they have same level of expressiveness for modeling node-level timing properties (Table 7.1). The languages use different levels of abstraction. For instance, in RCM timing properties are linked to a software component. Whereas, in Amalthea these properties are linked to a Runnable, which is a schedulable part of a software component. Furthermore, a Runnable can be mapped to more than one task at run-time unlike RCM which supports a one-to-one mapping between software components and tasks.

In both languages, the basic components can be part of transactions (chains), which can be event- or data-triggered. The languages support single-rate and multi-rate data chains. Both languages provide a clear separation between the control and data flow [11, 12]. Both languages support modeling of multi-core applications from the perspective of proof-of-concept prototypes. However, Amalthea has more practical support for modeling multi-core vehicular software systems.

Table 7.1 summarizes the comparison among the two languages with respect to modelling of network timing properties. RCM supports the modeling of all properties that are identified in the proposed framework. Amalthea does not support modeling of network-level timing properties. This is because Amalthea focuses only on the optimization of applications developed for single-node multi-core software architectures [13]. Since Amalthea is foreseen as a complementary language to AUTOSAR to support modeling of multi-core systems, it can be argued that Amalthea may get network modeling support indirectly via AUTOSAR. RCM is capable of supporting the development of distributed vehicular

Table 7.2 Timing constraints on a single event, pair of events and chain of events

Timing requirements		
	RCM	Amalthea
<i>Single event</i>		
Repetition constraint	✓	✓
Repeat constraint	✓	✓
Sporadic constraint	✓	✓
Periodic constraint	✓	✓
Burst constraint	✓	✓
Pattern constraint	✓	✓
Arbitrary constraint	✓	✓
<i>Pair of events</i>		
Delay constraint	✓	✓
Strong delay constraint	✓	✓
Order constraint	✓	✓
<i>Set of events</i>		
Sync constraint	✓	✓
Strong sync constraint	✓	✓
Output constraint	✓	✓
Input constraint	✓	✓

software systems and makes a distinction between intra- and inter-node communication [12]. Inter-node communication is made through the use of messages and hence expressing the properties identified in Table 7.2. The identified properties are applicable to several on-board networks, e.g., Controller Area Network and TSN [27].

7.4.2 Evaluation Based on Modeling of Timing Requirements

This section performs a comparative evaluation of the two languages with respect to their expressiveness to model the timing requirements. These requirements are extracted from the TADL2 languages that provides a timing model for AUTOSAR. Since Amalthea inherits most of the timing constraints for events, event sets and event chains exactly from TADL2, it supports modeling of all the identified timing constraints as shown in Table 7.2. Similarly, RCM is also capable of modeling these timing constraint [28]. Both languages are also capable of constraining all the activation patterns of a single event occurrences. They permit the activation of events periodically, sporadically, or following a certain pattern.

In Table 7.3, we identify the timing constraints that can be specified on a chain of events and distributed chain of events. Moreover, we also identify the timing constraints, which could potentially be put on the network. RCM is capable of supporting all the timing constraints on the network as well as on the event chains and distributed event chains. On the other hand, Amalthea can model the timing constraints

Table 7.3 Timing constraints on network and event chains

Timing requirements	RCM	Amalthea
<i>Event chains</i>		
Reaction constraint	✓	✓
Age constraint	✓	✓
Input sync constraint	✓	✓
Output sync constraint	✓	✓
<i>Distributed event chains</i>		
Reaction constraint	✓	–
Age constraint	✓	–
Input sync constraint	✓	–
Output sync constraint	✓	–
<i>Network</i>		
Repetition constraint	✓	–
Periodic constraint	✓	–
Sporadic constraint	✓	–
Pattern constraint	✓	–
Strong delay constraint	✓	–
Transmission time constraint	✓	–

within a node but does not support modeling of these constraints on network messages and distributed event chains as shown in Table 7.3. However, how the languages specifies the supported timing constraints differs. For example, RCM provides two special objects for specifying the Delay and Strong Delay Constraints. Whereas, the Strong Delay Constraint in Amalthea can be derived from the Delay Constraint by configuring *Mapping Type* option to OneToOne [13, 28]. Besides timing constraints, Amalthea provides additional support for mapping of constraints to Runnables and Tasks to cores.

7.4.3 Evaluation Based on Support for Timing Verification

This subsection performs a comparative evaluation of RCM and Amalthea based on the tool support for timing verification of vehicular software architectures. The results of the evaluation are shown in Table 7.4. The timing analysis framework in Rubus-ICE is complemented by a predictable run-time environment and is supported by the Rubus real-time operating system (RTOS). The repetition constraint is proven by construction as RCM provides special clock and event objects, supported by the Rubus runtime framework, for predictable repetition rates. Similarly, the runtime framework of RCM ensures the verification of all types of single event occurrence patterns. Furthermore, the execution time constraint is verified by construction as Rubus RTOS does not allow a particular task to over-run than the specified

Table 7.4 Timing verification for events, event chains or events set

Timing verification	Rubus-ICE	APP4MC
The occurrence of a single event	✓	✓
The occurrence of a message	✓	–
Time distance between the occurrence of triggering event and response event	✓	✓
Synchronization constraints on events set	✓	✓
Constraints put on event chains	✓	✓
Constraints put on distributed event chains	✓	–

WCET. RCM supports verification of all timing constraint on messages, event chains and distributed event chains. The input synchronization constraints in RCM are enforced by the offline scheduler that ensures that multiple inputs to one or more SWCs are synchronized [28]. On the other hand, the verification of the output synchronization, delay, strong delay, age, and reaction constraints are supported by the end-to-end timing analysis framework of RCM.

On the contrary, APP4MC is an open-source development platform that aims to have open interfaces to create extensions for other tools regardless they are of the same nature or commercial [29]. Today, APP4MC covers only the design of real-time multi-core vehicular embedded systems and provides the run-time framework for the mapping of software components and runnable to tasks, tasks to cores, and schedulers to cores. The verification of the timing requirements needs to be performed by third-party tools as APP4MC lacks a tool for the validation of timing requirements. However, commercial tools do not expose many details regarding the modeling techniques or timing analysis that they enable. In contrast, Rubus-ICE, despite being a commercial tool, exposes the underlying analysis techniques and methods to the research community [12, 28, 30]. Some of the timing analysis tools supporting Amalthea models and APP4MC include Timing Architects [31], Symta/s [32], ChronSIM [33], MAST [34]. The first three tools do not clearly discuss the implemented timing analysis techniques and related models. Whereas, the MAST is an open-source academic tool that is transparent about the implemented techniques and analyses for single-core, multi-core and distributed systems. When MAST is used to analyse Amalthea models then a model transformation is needed for MAST to interpret the timing information modeled by MAST. One challenge with this model transformation is that some of the information may not be properly transformed because MAST does not include corresponding components for covering all aspects of the software that Amalthea supports. These challenges are pointed out in [35]. Consequently, the timing analysis engines supported by MAST may use some pessimistic assumptions about the missing information.

7.5 Conclusion and Future Work

In this paper, we presented a framework for capturing the timing properties and requirements that modeling languages should express to verify the timing predictability of vehicular software systems. Interpretation of timing models generated through the specification of timing properties is needed to make use of such models. Hence, the framework also targets to what extent existing tools support the interpretation and verification of timing models. We validated the applicability of the framework by considering two vehicular modeling languages, namely Amalthea and RCM.

Based on the evaluation results, we identify that both modeling languages support the design and analysis of vehicle software. However, they have different scopes. Amalthea targets applications on single nodes with multi-core architectures while RCM focuses on single-core single-node and distributed embedded systems with ongoing work for supporting single-node multi-core architectures. Both modeling languages support time-, event- and data-driven activation of software components. Furthermore, both languages support modeling of single- and multi-rate transactions. In comparison to Amalthea, RCM provides a generic message model which can easily be re-modeled according to protocol-specific properties.

Our comparative evaluation also indicates that the modeling effort and learning curve of RCM and Rubus-ICE is minimal due to the approach of abstracting low-level implementation details and explicit modeling of timing properties, e.g., jitter. Amalthea model relies on sub-models for each aspect of software architecture, which increases its understandability and usability. APP4MC (the Amalthea tool) does not offer a timing analysis engine. Hence, it relies on third party tools for the purpose of performing timing analysis of software architectures. Amalthea is characterized by a high degree of timing expressiveness regarding single-node vehicular software systems of different architectures. However, this comes with the drawback of an extensive modelling effort. This may also result in increasing the pessimism of timing analysis as assumptions need to be made when some timing properties are not available. On the other hand, Rubus-ICE has been upgraded continuously through several projects in collaboration with academia and industry. New components are added to adopt state-of-the-art research on timing analysis proposed by academia and to meet the needs of the end-users.

One possible future research direction is to extend and refine the proposed framework by considering the timing properties of software architectures of computation-demanding vehicular applications that would run on many-core and heterogeneous computing platforms. Another possible direction

is the application of the proposed framework to additional modelling languages for automotive systems such as EAST-ADL.

Acknowledgments The work in this paper is supported by the Swedish Governmental Agency for Innovation Systems (VINNOVA) via the PANORAMA, DESTINE, PROVIDENT and INTERCONNECT projects, and the Swedish Knowledge Foundation via the FIESTA, HERO and DPAC projects. We thank our industrial partners, especially Arcticus Systems, Volvo CE and HIAB.

References

1. C. Ebert, J. Favaro, Automotive software. *IEEE Softw.* **34**(3), 33–39 (2017)
2. A. Bucaioni, P. Pelliccione, Technical architectures for automotive systems, in *2020 IEEE International Conference on Software Architecture (ICSA)* (IEEE, Piscataway, 2020), pp. 46–57
3. Land Rover Newsroom. <https://media.jaguarlandrover.com/news/2019/04/jaguar-land-rover-finds-teenagers-writing-code-self-driving-future>
4. L. Lo Bello, R. Mariani, S. Mubeen, S. Saponara, Recent advances and trends in on-board embedded and networked automotive systems. *IEEE Trans. Ind. Inf.* **15**(2), 1038–1051 (2019)
5. A. Bucaioni, S. Mubeen, F. Ciccozzi, A. Cicchetti, M. Sjödin, Modelling multi-criticality vehicular software systems: evolution of an industrial component model. *Softw. Syst. Model.* **19**(5), 1283–1302 (2020)
6. A. Bucaioni, A. Cicchetti, M. Sjödin, Towards a metamodel for the rubus component model, in *ModComp@ MoDELS*. Citeseer (2014), pp. 46–56
7. R. Eramo, A. Bucaioni, Understanding bidirectional transformations with TGGs and JTL. *Electron. Commun. EASST* **57** (2013). <https://doi.org/10.14279/tuj.eceasst.57.869>
8. S. Mubeen, E. Lisova, A.V. Feljan, Timing predictability and security in safety-critical industrial cyber-physical systems: a position paper. *Appl. Sci. Spec. Issue Emerg. Paradigms Archit. Ind.* **4.0 Appl.** **10**(3125), 1–17 (2020)
9. M. Becker, D. Dasari, S. Mubeen, M. Behnam, T. Nolte, End-to-end timing analysis of cause-effect chains in automotive embedded systems. *J. Syst. Archit.* **80**, 104–113 (2017)
10. S. Mubeen, H.B. Lawson, J. Lundbäck, M. Gålnder, K. Lundbäck, Provisioning of predictable embedded software in the vehicle industry: the rubus approach, in *2017 IEEE/ACM 4th International Workshop on Software Engineering Research and Industrial Practice (SER IP)* (2017)
11. Amalthea: Deliverable: D 3.4 development of scheduling analysis and partitioning/mapping support tools. Work package 3, April 2014
12. S. Mubeen, J. Mäki-Turja, M. Sjödin, Support for end-to-end response-time and delay analysis in the industrial tool suite: issues, experiences and a case study. *Comput. Sci. Inf. Syst. J.* **10**, 453–482 (2013)
13. App4mc project. Help documentation (June 2020). <https://www.eclipse.org/app4mc/help/app4mc-0.9.8/index.html>
14. Amalthea & Amalthea4Public ITEA3 Projects. <https://itea3.org/project/success-story/amalthea-and-amalthea4public-success-story.html>
15. AUTOSAR standard V2.2.1. https://www.autosar.org/fileadmin/user_upload/standards/classic/3-0/AUTOSAR_TechnicalOverview.pdf

16. R. Höttger, U. Jahn, P. Nördemann, P. Heisig, Teaching distributed and parallel systems with app4mc, in *International Symposium on Embedded Systems and Trends in Teaching Engineering* (2016)
17. S. Mubeen, T. Nolte, On timing analysis of component-based vehicular distributed embedded systems at various abstraction levels, in *Federated Conference on Component-Based Software Engineering and Software Architecture (CompArch)* (IEEE, Piscataway, 2016), pp. 277–278
18. K. Evensen, K. Weiss, A comparison and evaluation of real-time software systems modeling languages, in *AIAA Infotech@Aerospace 2010*, April 2010
19. N. Medvidovic, R.N. Taylor, A classification and comparison framework for software architecture description languages. *IEEE Trans. Softw. Eng.* **26**(1), 70–93 (2000)
20. A. Johnsen, K. Lundqvist, Developing dependable software-intensive systems: AADL vs. EAST-ADL, in *Reliable Software Technologies - Ada-Europe* (2011)
21. I. Crnkovic, S. Sentilles, A. Vulgarakis, M.R.V. Chaudron, A classification framework for software component models. *IEEE Trans. Softw. Eng.* **37**(5), 593–615 (2011)
22. S. Anssi, S. Gérard, S. Kuntz, F. Terrier, AUTOSAR vs. MARTE for enabling timing analysis of automotive applications, in *SDL: Integrating System and Software Modeling* (2012)
23. K. Tindell, Adding time-offsets to schedulability analysis. Technical Report YCS 221, Dept. of Computer Science, University of York (1994)
24. S. Mubeen, M. Gålnander, J. Lundbäck, K.-L. Lundbäck, Extracting timing models from component-based multi-criticality vehicular embedded systems, in *15th International Conference on Information Technology : New Generations*, April 2018
25. M. Bertogna, M. Cirinei, Response-time analysis for globally scheduled symmetric multiprocessor platforms, in *Proceedings - Real-Time Systems Symposium*, Jan 2008, pp. 149–160
26. TIMMO Methodology, Version 2. TIMMO (TIMing MOdel), Deliverable 7, Oct 2009. The TIMMO Consortium. http://adt.cs.upb.de/timmo-2-use/timmo/pdf/D7_TIMMO_Methodology_Version_2_v10.pdf. Accessed 09 June 2020
27. S. Mubeen, M. Ashjaei, M. Sjödin, Holistic modeling of time sensitive networking in component-based vehicular embedded systems, in *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)* (2019), pp. 131–139
28. S. Mubeen, T. Nolte, M. Sjödin, J. Lundbäck, K.-L. Lundbäck, Supporting timing analysis of vehicular embedded systems through the refinement of timing constraints. *Int. J. Softw. Syst. Model.* 1–31 (2017)
29. R. Höttger, H. Mackamul, A. Sailer, J.-P. Steghöfer, J. Tessmer, APP4MC: application platform project for multi- and many-core systems. *Inf. Technol.* **59**(5), 243–251 (2017). <https://www.degruyter.com/view/journals/itiit/59/5/article-p243.xml>
30. M. Ashjaei, S. Mubeen, J. Lundbäck, M. Gålnander, K. Lundbäck, T. Nolte, Modeling and timing analysis of vehicle functions distributed over switched ethernet, in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, Oct 2017, pp. 8419–8424
31. Vector - Timing Architects Tool Suite. <https://www.timing-architects.com/>. Accessed 10 June 2020
32. R. Henia, A. Hamann, M. Jersak, R. Racu, K. Richter, R. Ernst, System level performance analysis - the symTA/S approach. *IEE Proc. Comput. Digit. Tech.* **152**(2), 148–166 (2005)
33. chronsIM, model-based simulation of embedded real-time systems. <https://www.inchron.com/tool-suite/chronsim/>. Accessed 10 June 2020
34. MAST-modeling and analysis suite for real time applications. <https://mast.unican.es/>. Accessed 09 June 2020
35. J.M. Rivas Concepción, J.J. Gutiérrez, J. Medina, M. Harbour, Calculating latencies in an engine management system using response time analysis with mast, in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, July 2018

A Model-Based Approach for Quality Assessment of Insulin Infusion Pump Systems

Tássio Fernandes Costa, Álvaro Sobrinho, Lenardo Chaves e Silva, Leandro Dias da Silva, and Angelo Perkusich

Abstract

Insulin infusion pumps are safety-critical systems that require the approval of regulatory agencies before commercialization to prevent hazard situations. Nowadays, many recalls are reported for insulin infusion pump systems, motivating the usage of a formal model-based approach to improve quality. However, the usage of such approaches increases costs and development time. Thus, this study aims to assist the quality assessment of such systems cost-effectively and time-efficient. We defined a coloured Petri nets model-based approach and conducted a case study on the ACCU-CHEK Spirit system to verify and validate a reference model, describing quality assessment scenarios. We also conducted an empirical evaluation of the approach with 12 modelers to verify productivity and reusability. Using the approach, 66.7% of the modelers stated no effort, while 8.3%, stated low effort, 16.7% medium effort, and 8.3% considerable effort. Given such results, we developed a web-based application to assist modelers in re-using the proposed approach. The usage of the approach can decrease development time and thus

costs, increasing confidence in quality attributes such as safety and effectiveness.

Keywords

Access/CPN · ASK-CTL · Coloured Petri Nets · Modeling · Formal specifications · Simulation · Quality assessment · Web-based application

8.1 Introduction

Formal modeling languages are relevant mathematical tools to represent the requirements of Insulin Infusion Pump Systems (IIPS) with a high level of accuracy. Coloured Petri Nets (CPN) is an example of this type of language [5], and lets designers represent complex systems using programming language, hierarchy, and timing constraints, suitable to this study's purpose. CPN suits IIPS because of these systems' existing characteristics such as concurrency, communication, and synchronization. However, handling formal techniques usually requires expert knowledge and increased costs and development time. Reusable reference models have the potential to reduce the impact of such costs.

In 2010, the FDA released the infusion pump improvement initiative [3], concerning problems such as software defects, inadequate Graphical User Interfaces (GUI), and mechanical or electrical failures. As a result, the FDA released in 2014 a guidance for industry and FDA staff to follow during the pump's life cycle [4]. The most commonly reported problems of infusion pumps include software error, human factors, broken components, battery failure, alarm failure, and over infusion and under infusion. Some of the reported problems are related to the design activity, while manufacturers describe others as unknown problems.

T. F. Costa (✉) · L. D. da Silva

Computing Institute, Federal University of Alagoas, Maceió, Alagoas, Brazil

e-mail: tfc@ic.ufal.br; leandrodias@ic.ufal.br

Á. Sobrinho

Federal University of the Agreste of Pernambuco, Garanhuns, Pernambuco, Brazil

e-mail: alvaro.alvares@ufape.edu.br

L. Chaves e Silva

Computing Department, Federal Rural University of the Semiárid, Mossoró, Rio Grande do Norte, Brazil

e-mail: lenardo@ufersa.edu.br

A. Perkusich

Electrical Engineering Department, Federal University of Campina Grande, Campina Grande, Paraíba, Brazil

e-mail: perkusic@dee.ufcg.edu.br

There still exists a large number of recalls reported by regulatory government agencies regarding IIPS. Gao et al. [2] state that from the 70 infusion pump recalls released by FDA between 2001 and 2017; software failures caused 17 recalls. Thus, regulatory agencies have increased the surveillance stringency when manufacturers submit the system under development to the certification process. Such many recalls also motivate the proposal and usage of a formal Model-Based Approach (MBA) to improve quality. However, the usage of such approaches usually increases costs and development time. To address such a problem, we present an MBA of IIPS (named MBA with CPN—*MBA/CPN*) focusing on CPN reference models as modeling artifacts to increase confidence in system behaviors and provide quality assessments. The specification of the pump system is linked to the proposed CPN model by the definition of CPN modules that represent critical parts of such a system. We describe a case study on a commercial system, i.e., the ACCU-CHEK Spirit [8], to evaluate a reference model, as part of *MBA/CPN*, by simulations and the model checking technique, describing quality assessment scenarios. We also conducted an empirical study to evaluate the *MBA/CPN*.

8.2 Model-Based Approach

Figure 8.1 illustrates an overview of the *MBA/CPN*, consisting of modeling and analysis steps: hardware and software modularization, reference model definition, reference model instantiation, safety analyses, effectiveness analyses, and abstract test generation. In the first modeling step, manufacturers elicit CPN hardware and software modules from sources of requirements. In the second step, a reference model is defined by two system refinements. The first one results in a more abstract model that does not consider real colour sets, while the second refinement generates a more detailed model considering real colour sets and time constraints. The system's refinement specification is relevant to conduct analyses using two different perspectives: safety and effectiveness. In the third modeling step, manufacturers instantiate the model refinements to analyze the system's behaviors assessing quality. In the first analysis step, the first system's refinement is used to verify safety properties, preventing the state space explosion problem. Afterward, the verified reference model can be reduced (removing hardware components and adjusting initial markings) to apply abstract test generation approaches. The same properties shall be verified again for the reduced model to ensure consistency. Finally, the second system's refinement is used to analyze safety and effectiveness properties in the second analysis step. For all steps, manufacturers document IIPS requirements using assurance cases.

We recommend using two different versions of models to reduce the state-space explosion problem, avoiding the application of more complex techniques for state-space reduction. The definition of a less expressive first refinement does not negatively impact the approach because it maintains the desired safety properties, verified using the model checking technique. We assure that the CPN model represents the functionality of a specification according to the FDA guidelines by conducting model simulations.

8.2.1 Hardware and Software Decomposition

The *MBA/CPN* requires manufacturers to develop assurance cases based on the requirements derived from sources such as similar systems, literature reviews, recalls, and guidelines. Claims, evidence, and the other elements of assurance cases have specific representations in the *MBA/CPN* using XML and the Goal-Structuring Notation (GSN) [9]. In this article, we proposed and defined the Assurance Case Exchange Standard (ACES) to assist manufacturers and regulatory agencies in specifying and exchange assurance cases during the development and certification processes. For example, the ACES includes features to enable manufacturers to carry out the system requirements traceability. The usage of ACES is the starting point for the application of *MBA/CPN*, and it remains being used during the whole development process due to its connection to the CPN reference models.

The XML specification is the basis for associating all of the remaining steps of the *MBA/CPN*. The ACES considers the main concepts of the requirements engineering process based on modular GSN. An ACES document contains, at least, the graphical notations defined in the GSN specification. Each graphical notation of the GSN elements has a representation in ACES, relating the elements to specific tags and attributes of an ACES document. For example, the evidence element contains an attribute link to enable regulatory agencies to access a design artifact provided by system manufacturers to support an argument. The usage of assurance cases based on a well-defined and independent standard platform to represent and share results obtained by manufacturers with regulatory agencies may improve the design and evaluation of systems. We provide a more detailed description of the ACES specification in the *MBA/CPN* repository [7].

8.2.2 First System Refinement

Each goal element of an ACES document that contains the `<formalDefinition>` tag relates to CPN specifications considering system refinements. The first system's refinement of the reference model of IIPS has two main modules, representing the entire system based on hardware and software

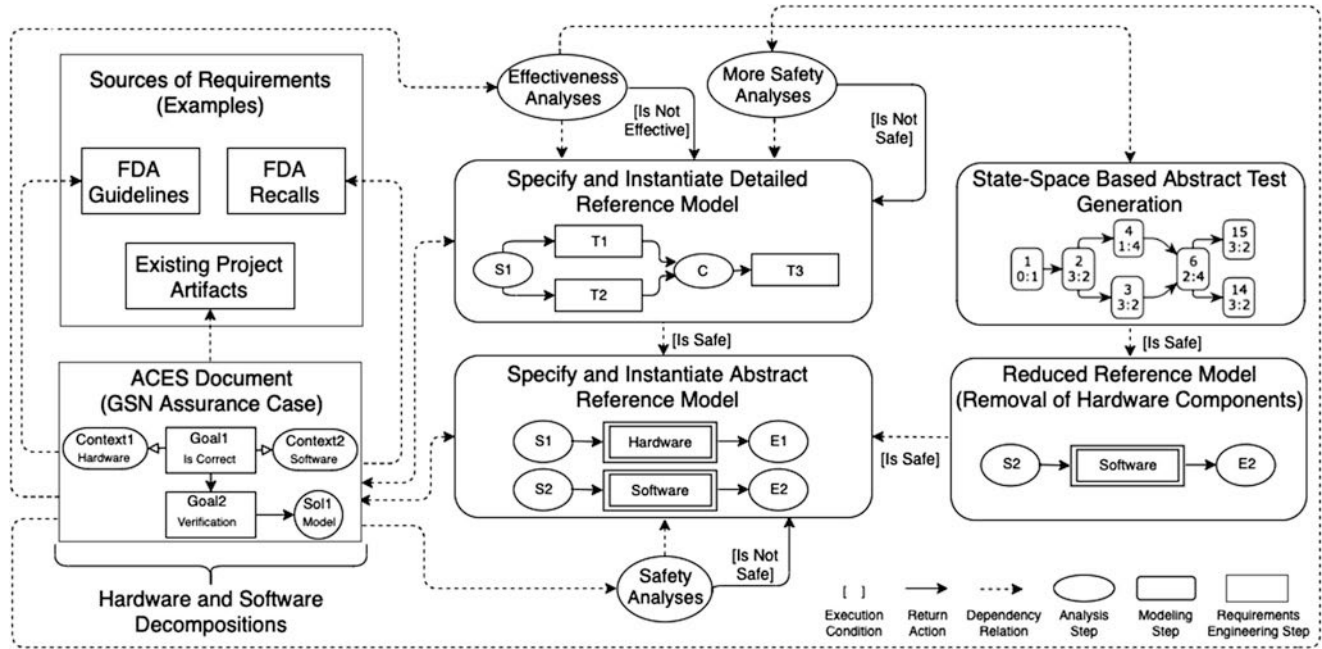


Fig. 8.1 Overview of *MBA/CPN* for quality assessment of IIPS

requirements. Thus, the modeling follows the architectural structures of module decomposition and usage.

8.2.2.1 Hardware Module

We provide the complete reference model in the *MBA/CPN* repository due to the size limitation of figures and to make it available for reuse and simulation [7]. The basic configuration of the hardware is composed of one starting button, one battery, and one cartridge. However, manufacturers can configure the number of hardware components using the modularization feature of CPN. The reference model includes this architectural redundancy tactic to help manufacturers deal with common FDA recalls caused by battery malfunctioning. Battery recalls may result in problems such as data loss, communication loss, abrupt therapy interruption, and over and under infusion [4].

8.2.2.2 Software Module

We divide the software module into three parts related to the steps of intermediate modular decomposition: battery verification, pump configuration, and insulin infusion. To start using the system, the user should configure the pump by defining the profile that guides operations: standard or personalized. The values' configuration for insulin dosages changes depending on each profile. The software sends a message to the hardware requiring the current cartridge's capacity to start configuring the pump.

For the personalized mode, it is necessary to define a set of values for different dosages daily. A text file loads the configuration dosages to define the infusion mode values as

list data structures, depending on the medical prescription for a specific clinical case. The remainder of the configuration follows the same approach as the standard mode. Besides describing the IIPS configuration step, the reference model contains standard and personalized infusion modes specifications. Finally, the model contains a submodule for the standard infusion, in which the first step comprises recording the configuration data, notifying that the pump is executing and that the cartridge is loaded.

Three types of software failures are critical to the pump's functioning due to risks to the safety of users: (i) failures that happen when the insulin dosages are being selected (no dosage has been selected), (ii) failures that happen when the insulin dosages are being selected (at least one has been selected), and (iii) failures that happen when insulin dosages are being applied. When failures occur, the system should stop running immediately, informing the user about the malfunctioning. The personalized insulin infusion mode behaves similarly to the standard mode; however, instead of the dosages' constant values, the personalized mode allows the user to define the desired dosages from a text file.

8.2.3 Second System Refinement

The second refinement of the reference model of IIPS provides an extended version of the first system refinement, including time constraints and real data type values. This refinement is composed of the same higher-level modules of

the first refinement; however, the specification improves the reference model with a more detailed representation of IIPS to comply with stringent regulatory requirements (e.g., the correct rate of infusion) with real and timed data types.

The second model refinement for the standard infusion (also available in the repository [7]) shows the administration control. A user-defined administration rate guides the administration control, and thus, considering an insulin dosage, it does not mean that the pump applies the dosage at once. The system splits the dosage up to schedule the infusion based on the administration rate, and the infusion depends on the number of insulin units allowed.

8.3 Quality Assessment Scenarios

We instantiated the reference model to represent the technical requirements of the ACCU-CHEK Spirit [8], presenting quality assessment scenarios of IIPS using model checking and simulations. Each solution element of an ACES document, defined as an artifact, links to verification and validation results generated using the IIPS model.

8.3.1 Verification of the First Refinement

The verification concerns the quality assessment of safety properties. A file called `BasalV2AC.txt` contains 24 hypothetical values of basal dosages for the personalized infusion mode. This activity consists of verifying two safety properties of IIPS that are part of the specification of the ACCU-CHEK Spirit. The model checking technique using CPN is based on the Computation Tree Logic (CTL) [9], and the functions available on the ASK/CTL library of CPN/Tools represent the specifics of the properties.

The first property defines that when the level of the cartridge is 0, the pump must be stopped ($AF \neg (cartZeroed) \vee AF(pumpStop \wedge cartEmpty)$). The ASK-CTL model checker verified that the model complied with this property. The reference model is according to the first property of IIPS for the standard and personalized modes. The second property defines that the pump cannot run when the insulin dosage is greater than the level of the cartridge ($AF \neg (cartLevel) \vee AF(pumpStop)$). The ASK-CTL model checker also verified that the model is according to the second property of IIPS. These are two examples of project artifacts associated with solution elements of the ACES document. Following the *MBA/CPN*, we reduced the reference model by removing hardware components and adjusting initial markings, verified it again to ensure conformance, and generated abstract tests to guide the validation. To illustrate the abstract test generation, we applied the MBT/CPN tool using the state space of the IIPS model, resulting in the eight abstract tests.

8.3.2 Validation of the Second Refinement

The simulation resources of the CPN/Tools software enabled the validation of the second refinement, analyzing functionalities related to the model's time constraints. Simulations are used to conduct quality assessments of safety and effectiveness properties. The format convention adopted to represent time constraints is a 1-time unit regarding 1 second, 60-time units regarding 1 minute, and 3600-time units regarding 1 hour. The simulations (more than 3000 steps) were conducted to evaluate the completeness and correctness of the model based on the eight abstract tests generated with the MBT/CPN tool. Besides, we conducted a large number of random simulations to increase confidence, considering other characteristics such as desired and undesired battery charges. Requirement's simulation results are other examples of project artifacts associated with ACES solutions.

8.4 Empirical Evaluation

8.4.1 Scoping, Modelers, and Variables

The goal-question-metric methodology [1] guided the definition of the empirical evaluation through analyzing the usage of the reference model into two aspects: (i) evaluation concerning productivity from the viewpoint of the modelers by instantiating the model; and (ii) evaluation for reusability from the viewpoint of the modelers by instantiating and extending the model. Therefore, we defined the following secondary research question (RQs): does the reference model increase modelers' productivity? (RQ1), and is the reference model reusable? (RQ2). These RQs guided the specification of the following hypotheses: productivity is not increased (H0-1), productivity is increased (HA-1), the reference model is not reusable (H0-2), and the reference model is reusable (HA-2). H0-1 and HA-1 are related to RQ1 and H0-2 and HA-2 to RQ2.

We selected 12 modelers using the convenience sampling technique and evaluated at a federal university located in Brazil. The profiles of the modelers who participated in the study relate to age, knowledge about formal methods, opinion about the training phase, resolution of the list of exercises, knowledge about CPN, and knowledge about the present work.

There are two dependent variables defined based on the research goals: modelers' productivity and reference model's reusability. There are four independent variables: the reference model, the modelers' experience, the tool support, and the environment. These variables are controlled at a fixed level, meaning that each group of modelers (i.e., control and treatment) has the same reference model, experience, tool holder, and environment.

8.4.2 Procedure and Measures

The CPN/Tools software was installed to enable the conduction of the evaluation based on four phases: first training stage, first evaluation stage, second training stage, and second evaluation stage. We prepared the modelers in the training phase to specify CPN models using the CPN/Tools. We conducted the following activities for the first training stage: (i) we asked modelers to answer a questionnaire regarding personal information, experience with formal methods, and experience with reusability; (ii) the trainer presented a short course regarding concepts and examples related to CPN, CPN/Tools, and the reference model. The trainer also presented an overview of the ACCU-CHEK Spirit; and (iii) the modelers applied the learned techniques to build simple examples assisted by the trainer.

During the first evaluation stage, the modelers instantiated the reference model to represent the commercial insulin infusion pump system ACCU-CHEK Spirit and responded to a questionnaire regarding the models' reusability attributes. We divided the modelers into control (size eight) and treatment (size four). The treatment group comprises modelers who have few experiences using CPN (i.e., less than ten hours of a CPN course), while the control group comprises modelers who finished a six-month class about CPN and, consequently, at least, six months of experience using CPN. We asked each subject to instantiate the reference model within two hours.

The next step was the second training stage, in which the reference model was explained in more detail. This step was followed by the second evaluation stage, comprising the reference model's usage to extend the specification. The following activities were conducted in the second evaluation stage: (i) each modeler implemented two new requirements within two hours: add a new battery representation and add a feature for recharging the battery; (ii) the modelers were asked to answer a questionnaire about effort and reuse.

We defined metrics to evaluate the hypotheses and assess the RQs. Thus, we addressed time to measure productivity [6], computing the time required by each group to finish the problems that we asked modelers to solve during the evaluation phase. Moreover, we measured reusability using two factors [10]: understandability and adaptability. Understandability is related to how easy the modeler recognizes the meaning of a component of the reference model and its applicability, while adaptability stands for how to ease the modeler can extend the reference model to comply with a new system's requirement.

We formalized hypotheses to conduct statistical analyses for the RQ1: the null hypothesis H_{0-1-1} , i.e., $vU > 1h$, in which vU is the meantime (in minutes) needed by the modelers to conclude instantiating the reference model; and the alternative hypothesis H_{A-1-1} , represented by $vU \leq 1h$.

One hour corresponds to six times the time spent by the researchers to instantiate, for the first time, the reference model based on the ACCU-CHEK Spirit. We also formalized the hypotheses to conduct statistical analyses related to the effort factor: the null hypothesis H_{0-1-2} , i.e., $vU > 3$, in which vU represents the mean classification of the responses for the effort questions; and the alternative hypothesis H_{A-1-2} , represented by $vU \leq 3$. We also formalized hypotheses to conduct statistical analyses related to the RQ2. Considering understandability, there is a null hypothesis H_{0-2-1} , i.e., $vU \leq 3$, in which vU represents the mean classification of the responses for the understandability questions, while the alternative hypothesis H_{A-2-1} is represented by $vU > 3$. Considering adaptability, there is a null hypothesis H_{0-2-2} , i.e., $vU \leq 3$, in which vU represents the mean classification of the responses for the adaptability questions, while the alternative hypothesis H_{A-2-2} is represented by $vU > 3$.

At the end of the experiments, we applied a questionnaire to collect metrics for the effort and reusability factors by providing a 5-point Likert scale (1) to (5). The scale interpretation concerning the metrics is based on the effort of instantiating the reference model during experiment 01: (1) represents the best result and (5) represents the worst result. For the understandability and adaptability measures, conducted during experiment 02, the scale's interpretation is (1) for the worst and (5) for the best. The modelers answered the questionnaire after extending the reference model.

8.4.3 Analysis

Figure 8.2 depicts the answers for effort and each reusability factor evaluated in experiments 01 and 02, respectively. We did not apply more complex hypotheses tests such as the *t*-test and Wilcoxon test because the hypotheses were defined to evaluate RQ1 and RQ2 only use the mean responses concerning the factors considered in this experiment. The basic descriptive statistics enabled the evaluation of the hypotheses. Considering the RQ1, we evaluated the reference model's reuse based on the analysis of the improvement of the productivity of the modelers. The 12 modelers who participated in the experiment concluded instantiating the model, and in the average case, within six minutes.

When asked to respond to the questionnaire, 66.7% of the modelers (8 of the 12 modelers) stated that the task of instantiating the reference model presented no effort while 8.3% stated low effort, 16.7% stated medium effort, and 8.3% stated considerable effort. Therefore, the hypotheses H_{0-1-1} and H_{0-1-2} were refuted, showing that the reference model presented a positive response to the RQ1 (the modelers used only 10% of the estimated time).

We evaluated the reference model's reuse to analyze the RQ2 based on the understandability and adaptability factors. Figure 8.3 presents the number of modelers stating to extend

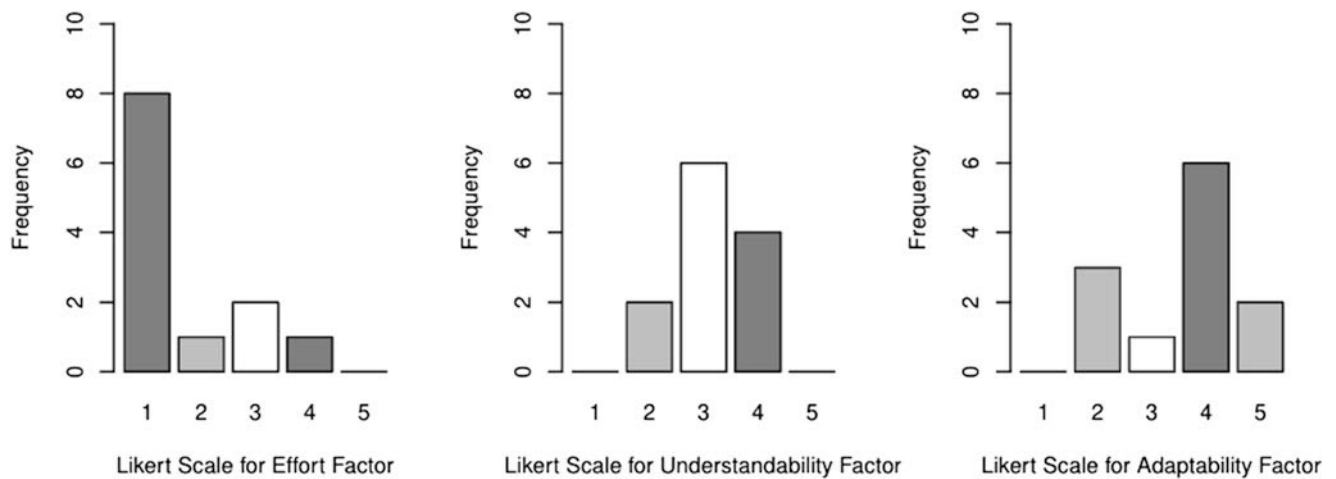


Fig. 8.2 Answers distribution in accordance with the effort and reusability

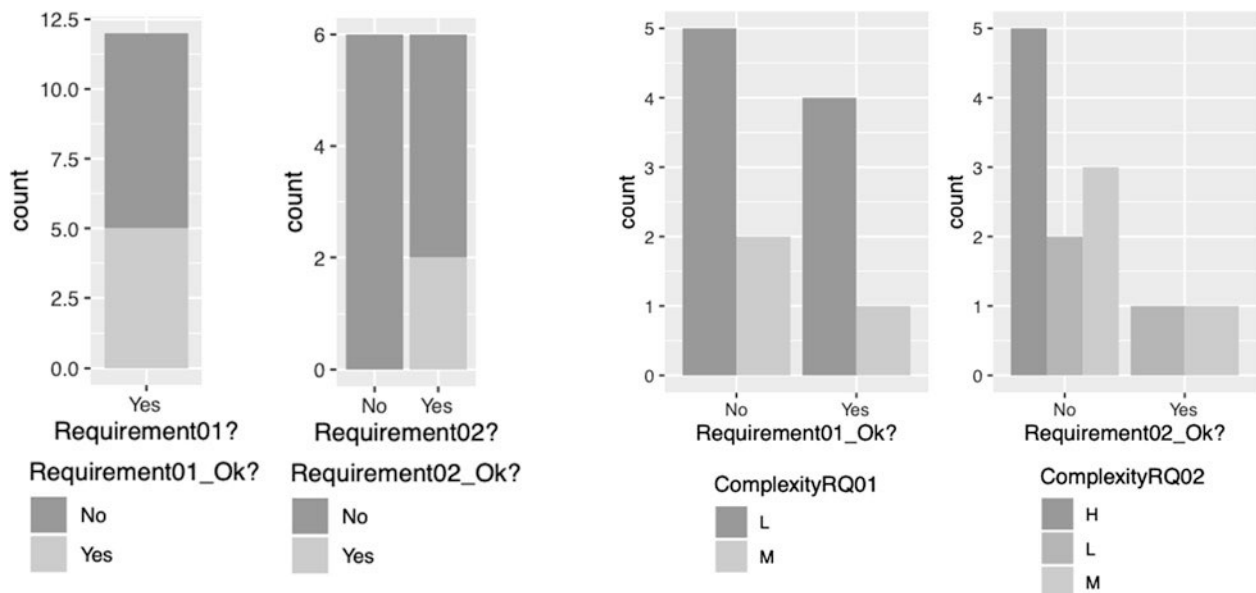


Fig. 8.3 Instantiated model vs. correctly instantiated

the model considering the requirements 01 and/or 02, also including the number of modelers who experimented with 02 correctly. All the modelers implemented requirement 01 (i.e., adding a new battery). However, only five of them implemented it successfully. The analyses of the work conducted by the seven modelers who did not implement requirement 01 successfully showed that they all failed in the same task. Six modelers attempted to implement requirement 02. However, only two of them achieved the correct answer. The results of the four modelers who did not achieve the correct implementation showed different mistakes, all related to the usage of CPN/ML.

To analyze the time spent to conduct the second experiment, we considered the group related to each modeler. Only one of the modelers successfully implemented the reference

Fig. 8.4 Answers distribution considering complexity of functionalities


model's first and second requirements within 25 and 38 minutes, respectively. The modelers scaled the first requirement as a low-complexity task while disagreed when asked about the complexity of the second requirement: the control group's modeler scaled as a medium-complex task, while the modeler of the treatment group scaled it as a low-complex task.

Analyzing understandability and adaptability, it is not possible to conclude that the reference model is fully reusable, requiring the analysis of results presented in Fig. 8.2. Most of the modelers did not fully understand each component of the reference model's roles and functioning. The high number of modules required by the complexity of IIPS negatively impacted the task of adapting the model adding new requirements. Figure 8.4 supports this claim showing modelers'

SysIIPS

Support system for using the reference model for insulin infusion pump systems

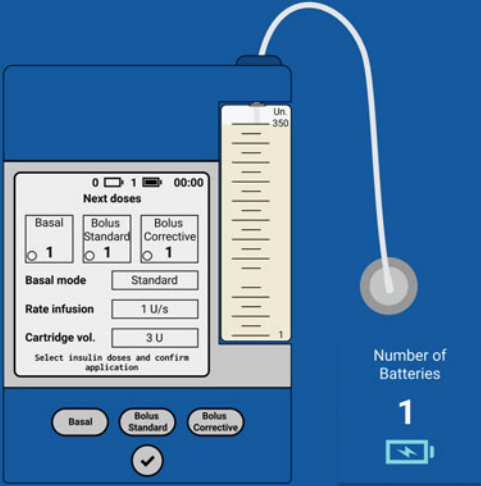
Generating evidence for certification



Explore the SysIIPS

[Objective](#) [Functionalities](#) [Benefits](#) [Model overview](#) [Current features of the model](#)

Parameter values



Basal Application Mode

Standard Mode

Personalized Mode

Parameter	Description
BOLUS	1
CBOLUS	1
CAPCART	3
UPPERDOSELIMIT	3
LOWERDOSELIMIT	1
INFUSIONLIMIT	3
BASAL	1
DADMSD	1
TADMSD	1
TADMBSD	1

Fig. 8.5 GUI sample of the web-based application

answers for the complexity of each requirement added to the reference model in the second experiment. Most of the modelers stated the requirements 01 being of low (75%) or medium (25%) complexity, while stated medium or high complexity (75%) for requirement 02.

The evaluation of $RQ2$ based on the questioning presented to the modelers showed that $H0-2-1$ was refuted due to the mean classification of the understandability factor (approximately 3.17). For $H0-2-2$, the mean classification of the adaptability factor presented approximately 3.58, refuting the null hypothesis. Thus, the reference model presented acceptable levels of understandability and adaptability.

8.5 Web-Based Application

Based on the results of the empirical evaluation, we used the ACCESS/CPN framework [11] and web services to develop a web-based application, aiming to assist modelers to re-

use our approach. The usage of ACCESS/CPN enabled us to embed the reference models in web services to conduct simulations without the GUI of the CPN/Tools software. Figure 8.5 presents a GUI sample of the web-based application. For instance, the application enables modelers to simulate the configuration of the pump, showing each simulation step (e.g., along with information related to places and transitions). The web services used to implement such application can be easily adapted to handle a different version of the reference model.

The modeler can use the system to handle the CPN model of IIPS using four features: (1) add battery; (2) remove the battery; (3) configure the insulin pump; and (4) simulate the insulin pump. Features (1) and (2) allow the user to change the structure of the CPN model by adding/removing an instance of a module that represents the battery of the insulin infusion pump. Feature (3) allows modelers to change the values of the parameters of the reference model. Finally, Feature (4) allows modelers to simulate the CPN model using the GUI, without using CPN/Tools, to improve

the understanding of the pump's behavior without the user having access to the internal structure of the model.

8.6 Conclusions and Future Work

We presented an MBA focused on CPN reference models of IIPS, aiming to assist manufacturers in assessing quality. The case study was relevant to extending the model to verify and validate two model refinements as quality assessment scenarios. We considered the formal verification of two safety properties for the first refinement using the model checking technique. We validated the second refinement using simulations to analyze the model concerning the commercial system's technical specifications. Finally, the empirical evaluation demonstrated that the MBA/CPN is reusable for the development and certification process of IIPS. As future work, we envision conducting another empirical study to evaluate the web-based application, integrated with the reference model, to analyze if understandability and adaptability are improved.

Acknowledgment This work was supported in part the CNPq and CAPES.

References

1. V.R. Basili, H.D. Rombach, The TAME project: towards improvement-oriented software environments. *IEEE Trans. Softw. Eng.* **146**, 758–773 (1988)
2. X. Gao, Q. Wen, X. Duan, W. Jin, X. Tang, L. Zhong, S. Xia, H. Feng, D. Zhong, A hazard analysis of class I recalls of infusion pumps. *JMIR Hum. Factors* **62**, 10366 (2019)
3. *Infusion Pump Improvement Initiative*. <https://bit.ly/2QAHe2V>
4. *Infusion Pumps Total Product Life Cycle*. <https://bit.ly/2EMOXrW>
5. K. Jensen, L.M. Kristensen, Colored Petri nets: a graphical language for formal modeling and validation of concurrent systems. *Commun. ACM* **58**(6), 61–70 (2015)
6. B. Kitchenham, L. Pickard, S.L. Pfleeger, Case studies for method and tool evaluation. *IEEE Softw.* **124**, 52–62 (1995)
7. *MBA/CPN Repository*. <https://bit.ly/2Qyci3k>
8. *Pump User Guide - Accu-Chek*. <https://bit.ly/2QzJMOT>
9. The Assurance Case Working Group, Goal Structuring Notation Community Standard (Version 2) (2018)
10. H. Washizaki, H. Yamamoto, Y. Fukazawa, A metrics suite for measuring reusability of software components, in *Proceedings of the 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry, Sydney, Australia* (2003), pp. 211–223
11. M. Westergaard, L.M. Kristensen, The access/CPN framework: a tool for interacting with the CPN tools simulator, in *Proceedings of the International Conference on Applications and Theory of Petri Nets* (2009)

Narrowing the Gap Between Software Engineering Teaching and Corporate Environment

Marcelo A. M. da Conceicao, Oswaldo S. C. Neto, Andre B. Baccarin, Luan H. S. Dantas, Joao P. S. Mendes, Vinicius P. Lippi, Gildarcio S. Gonçalves, Adilson M. Da Cunha, Luiz A. Vieira Dias, Johnny C. Marques, and Paulo M. Tasinaffo

Abstract

Currently, there is a gap between university education in the area of software engineering and the corporate and industrial environment. In this way, it is necessary to explore new learning models to transmit complex knowledge more quickly and interconnect these different environments. For that, the STEPES-BD was developed on the first half of 2020, in the midst of the COVID-19 Pandemic. In this project, the students have put into practice some concepts from three different undergrad and graduate courses of Computer and Electronic Engineering in the Area of Informatics of the Aeronautics Institute of Technology (*Instituto Tecnológico de Aeronáutica – ITA*) in Brazil. To approach these different worlds, it was used the Interdisciplinary Problem-Based Learning (IPBL), the Scrum Framework (SF), and also several emerging Information Technologies of public domain. This article aims to present the main results obtained by students in just one semester of 17 weeks, in a collaborative and cooperative manner, allowing them to implement a computer system prototype to help solve a real-world problem of managing health information for the COVID-19 Pandemic. In this article, it is described how the IPBL, the SF, its ceremonies, roles, and artifacts were adapted. The assessment of students experience in the STEPES-BD Project was carried out using the Tuckman Model, extracting quantitative and qualitative

results, comparing the results achieved with related works, and acquiring students' perceptions through project-end questionnaires.

Keywords

Software engineering education · Computer science education · Interdisciplinary problem-based learning · Scrum · Agile · Soft skills · Tuckman model maturity level · Scrum of scrum · Storytelling · Emerging technologies

9.1 Introduction

Rapid technological and social changes have demanded constant updates in curricula of universities and research institutes, causing the need for new educational configurations. These demands have forced who is involved in these processes to explore new learning models to transmit complex knowledge more quickly and to interconnect the academic world with the real world of companies, public, or private institutions.

The Interdisciplinary Problem-Based Learning (IPBL) approach has been providing these interconnections, through student-centered learning processes, with emphasis on different problem situations [1]. This approach has reflected the complex reality faced by undergraduate and graduate students, allowing more realistic contextualization and more intense collaborations between students to apply, even in their academic environments, concepts obtained from disciplines into real-world problems [1].

Within this context, this article describes how the IPBL and the Scrum Framework (SF) were adapted and used in just one academic semester of 17 weeks for developing a computer system prototype involving a health problem of infor-

M. A. M. da Conceicao (✉) · O. S. C. Neto · A. B. Baccarin
L. H. S. Dantas · V. P. Lippi · G. S. Gonçalves · A. M. Da Cunha
L. A. Vieira Dias · J. C. Marques · P. M. Tasinaffo
Computer Science Division, Aeronautics Institute of Technology –
ITA, São José dos Campos, Brazil
e-mail: marcelomamc@fab.mil.br

J. P. S. Mendes
State Technology Educational Center Paula Souza, São José dos
Campos, Brazil

mation management, by using some emerging technologies of public domain as Blockchain among others in the scenario of the COVID-19 Pandemic. For this, the STEPES-BD was conceived as an academic project using Specific Technological Solutions for Special Patients and Database Systems.

This interdisciplinary project brought together 25 students (84% male and 16% female) from the following undergrad (12%) and graduate (88%) courses of Computer and Electronic Engineering in the Area of Informatics at the Brazilian Aeronautics Institute of Technology: CE-245 Information Technologies; CE-240 Database Systems Project; and CE-229 Software Testing. Working collaborative and cooperatively, these students were able to develop a product as a proof of concept at the level of system of systems.

During the development of this project, the maturity level reached by each interdisciplinary Scrum Team (ST) was evaluated based upon the Tuckman Model Maturity Level (TMML) and artifacts, roles, and ceremonies were adapted and audited by the professors so that the development process could be measured in the most possible practical and natural manner.

In order to contextualize and guide the academic project, the following scenario was described as an Assigned Mission storytelling: *“For public or private organizations involved with the management of information of interest to Patients, Physicians, Hospitals, and Suppliers, which provide monitoring and decisions in the area of health knowledge, the STEPES-BD involves a computational system based on emerging technologies. Unlike other products that may exist in Universities, Research Institutes, Government Agencies, or private companies, this product was developed and tested in 17 academic weeks of the 1st Semester of 2020.*

9.2 Related Work

Francese et al. [2] report on the experience acquired in the semi-attendance Mobile Application Development Course for 55 students in 12 weeks, using PBL learning. The students were divided into 27 teams in pairs form, creating intra-team collaborations and an extra-team competitions for the best application. Teams' evaluation was performed by the professors, by the juries composed of 10 IT professional managers of national and international companies, and by the students through answers to a project-end questionnaires.

Kizaki et al. [3] report collaborative software development through PBL learning to undergraduate students, where 32 students participated and were divided into groups from 3 to 5 students, having 1 Project Manager and 1 Scrum Master per each team, as a member-of-society engineer. The evaluation of projects was performed according to the Kano Model [4].

Raibulet [5] describes the experience of collaborative activities with students of the Software Engineering course,

aiming to limit the gap between industry expectations and the preparation of undergraduate students. The project lasted 10 weeks, where 41 students were divided into groups from 3 to 5. Each team has developed its own web software project. The project evaluation was performed through questionnaires, capturing feedback from students' perception.

9.3 Tailoring Scrum and IPBL

In order to overcome the needs imposed by the social distancing due the COVID-19 Pandemic, in this project the Totally Integrated Scrum model was addressed, in which ST are multi-functional, having members distributed geographically, requiring strong extra-teams communications, as well as the adoption of Scrum of Scrums (SoS) meetings [6].

To support the interaction between the students, two levels of Scrum were created: (SoS Level #1) traditional intra-teams; and (SoS Level #2) extra-teams, involving only the Product Owner (PO) and the Teams Scrum Masters (TSMs).

For the extra-team integration and for the geographically dispersed members to be integrated, SoS meetings [7] were held where, in large projects with multiple teams like the STEPES-BD, you may create a higher level of aggregation between the PO and the TSMs, as shown in Fig. 9.1.

Furthermore, the STEPES-BD Project has applied the IPBL pedagogical concepts as its basis, integrating students from the three different disciplines, using distinct and emerging technologies for the development of a Proof of Concept (PoC). In this project, it was addressed the real-world problem of the COVID-19 Pandemic that has been afflicting the populations of the world in recent years.

As it is an academic project with a well-defined restriction in its schedule and requiring reductions in scope, its management was always carried out by using the Scrum best practices. That's why the time boxes between the ceremonies and the artifacts had to be adapted to allow students a greater opportunity to get in touch with the Scrum Agile method. For this, the STEPES-BD Project was divided into four ST representing the interest segments in the health area: Patient, Physician, Hospital, and Supplier.

Interdisciplinary Activities Although STs had the specific scopes of their well-defined segments, they had to work together, starting from the 2nd Sprint, performing interdisciplinary activities, mainly to properly address common efforts, involving integration.

Team Arrangement The choice and distribution of students to join STs were carried out in a balanced way by professors who were in charge of balancing and adjusting any differences in skills and experiences existing within STs.

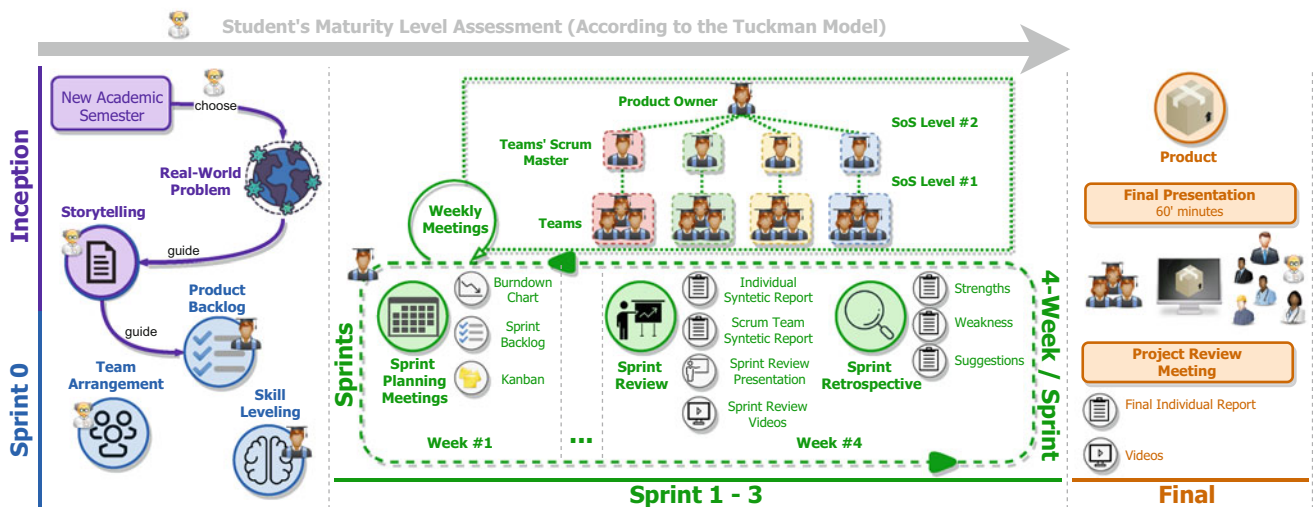


Fig. 9.1 The steps adopted in STEPES-BD project

(A) The Scrum Roles Academically Used

To enable the application of Scrum in the academic environment, 2 new backup roles had to be created (Backup of Product Owner and Backup of Team Scrum Master), to provide eventual replacements of STs key members. The selection and choice of students to play the roles of PO, PO Bkp, the four roles of TSMs, and the four roles of TSMs Bkp was carried out by professors of the disciplines based upon skills and qualifications proven in their curricula and in previous participation experiences in the development of Agile Projects, using similar structures as those quoted in [8,9].

During the STEPES-BD Project development, professors acted interested or directly involved with the Project (as Stakeholders) and constantly monitored the progress of their development, through the artifacts generated at each Sprint, as Burndown Charts, Sprint Backlogs, and Test Case Spreadsheets.

(B) The Scrum Artifacts Academically Used

The following artifacts, sorted chronologically, have also been used and adapted to meet the academic needs of the STEPES-BD Project, as shown in Fig. 9.1.

Product Backlog It was created in the beginning, updated during the development of the project, containing a dynamic and updated list of User Stories (US) that could be implemented in the project, as well as information about possible inter-teams integrations, acceptance criteria, and dependencies.

Sprint Backlogs It was generated from the Sprint Planning Meetings of each ST, containing USs selected to be developed during the Sprints. They typically involved information on USs development priorities, from scores obtained in effort planning estimates to be spent, such as Planning Poker, Integrations, USs Definition of Ready, USs Definition of Done, Acceptance Criteria, and so on.

Planning Poker Games It was generated in the Sprint Planings Ceremonies, within each ST and after the US was subdivided into Tasks. In them, the Planning Poker technique was applied, where the scores, consensually and according to the Fibonacci Sequence, were assigned, representing the necessary efforts for the complete development of each US.

Burndown Charts It was generated in the Sprint Plannings Ceremonies, within each ST and after the US was subdivided into Tasks. It allowing Stakeholders, POs, and team members to visually monitor the effort spent, inferring the existence of delays or advances in the scores made from USs.

Kanban It was generated and updated during Sprints. In this example, all USs and Tasks were registered on GitHub, right after a Sprint Planning Meeting. In this example, it is noticed that the ST used this Kanban board to hold their Weekly Meetings, indicating those responsible for their developments, and adjusting their progress as To Do, In Progress, and Done.

Individual Synthetic Reports It was prepared, updated, and presented individually by the Team Members at the end of each Sprint and during the Sprint Reviews, containing a summary of what was developed during the Sprint and individual suggestions for improving the Project.

Scrum Teams Synthetic Report It was prepared, updated, and reported by the TSMs, at the end of each Sprint, containing the information consolidated by Teams, in a similar manner to the Individual Synthetic Reports previously described.

Sprint Review Presentations It was prepared by the Team Members and presented by the TSMs, at the end of each Sprint and during the Sprint Reviews. They were created to report developments made by the ST during Sprints.

Sprint Review Videos It was prepared by the Team Members and also presented by the TSMs, at the end of each Sprint and during the Sprint Reviews. They were presented in 3-min videos, containing a summary of what was developed.

Video It was presented in an integrated manner by the TSMs, in the Final Presentation of the Project, in the form of a 5-min video, consolidating all the value deliveries.

Final Presentation It was created and presented in a collaborative manner by the TSMs, in a Final Presentation, lasting approximately 60 min, where all value deliveries were consolidated.

Final Individual Report These last artifacts were created and published individually by the students, containing: a summary of all the development carried out; the student participation in each US; and its conclusions and suggestions for the next projects.

(C) The STEPES-BD Scrum Ceremonies

To manage the work carried out during the Sprints, weekly, on Mondays, remotely and through video-conference tools, the PO and the TSMs held the **Weekly Meetings** (WM) on the second level of SoS (**SoS Level #2**), to present the project progress, as shown in Figure 1.

At these meetings, in addition to discussing relevant topics to all ST, general information and guidelines were regularly received from professors. At the first meeting of each Sprint (into SoS Level #2), US were selected to compose the Sprint Backlog for each ST. The intra-team meetings (SoS Level #1) were held weekly, one day after the extra-team Weekly Meetings (SoS Level #2), replacing the traditional Daily Meeting, both by video-conference manner.

TSMs and their developers participated in this ceremony. In the first meetings of each Sprint, called **Sprint Planning Meetings**, ST played Planning Poker, created Burndown Charts, Kanban boards were made available, and the STs Sprint Backlogs were prepared with the additional information generated in the meetings.

In the others Weekly Meetings of Sprints intra-teams, in level #1 of SoS, the STs progress and Kanban was updated,

Burndown Charts was presented, impediments were checked, and so on.

At the end of the Sprints, in the last weeks, the STs carried out the **Sprint Reviews**, where it was possible to show the USs developed and the main artifacts such as Burndown Charts.

Each ST developed a video, in addition to their technical reports, summarizing the following topics: (1) sprint's contextualization and objectives; (2) Sprint Backlog; (3) requirements; (4) US and associated Tasks; and (5) suggestions.

In the **Sprint Retrospective** were carried out through Google Forms and their results presented, containing strengths and weaknesses, in addition to suggestions for changes for the following Sprints and future works. The **Project Review Meeting**, the final results obtained were presented.

To overcome communication challenges in a completely distributed environment, the team members made massive use of online communication tools.

(D) The STEPES-BD Scrum Ceremonies

To assess the maturity levels achieved by the STs, considering the participation of its members in the ceremonies and in the uses of the artifacts, the TMML was applied.

When applying this model, it was found that, although the STs sought to practice self-management, there was a need to verify the maturity levels achieved by each of them, in order to carry out behavioral realignments and evaluations [10, 11].

Forming Already on the first Sprint, after explaining the ST objectives within the project, one can verify manifestations within the teams of the main characteristics of this stage. In this stage, team members usually start their participation with fear of how to adapt to the environment. To overcome this difficulty, acting as facilitators, professors, PO, and TSMs needed to support their members by clarifying doubts and providing technologies that could be initially used to facilitate the tasks to be performed.

Storming In any project that involves group dynamics, conflicts can arise within the teams. In the TMML, these conflicts characterize the second stage of the maturity level where each member knows the goals for that moment, but cannot clearly define their individual role or responsibility, generating conflicts that have to be resolved by leaders within the groups. In this Project, evidence of this stage only began to be noticed at the end of the first and the beginning of the second Sprint and it was, once again, the responsibility of professors, the PO, and the TSMs to support their members, to clarify doubts and provide technologies, always as facilitators of the tasks to be performed.

Norming In this third stage the Team Members already know each other and began to deal better with possible situations that could occur. In this stage, the processes, and execution of tasks became more natural. The members were already more motivated by the first results achieved. At this point, the leaders needed to actuate when necessary. In this Project, evidence of this stage only began to be noticed at the end of the second and the beginning of the third Sprint, with an ever smaller participation of the professors, the PO, and TSMs as facilitators in the project.

Performing The fourth stage is characterized by the full and self-sufficient performance of all members of a team, when they are able to autonomously develop their roles and manage themselves without the need of leaders to intervene, because each participant has a well-defined role, knows the goals and objectives to be accomplished, knows how to work collaborative and cooperatively, and understands the importance of their participation to successfully deliver values. Some evidence of this stage only began to be noticed at the end of the third Sprint, but just within two STs.

Adjourning The fifth and final stage consists of the dissolution of a group due to greater force, withdrawal or closure of a project. In this project, this stage occurred naturally, at the end of the project and at the end of the academic semester, giving rise to the beginning of the preparation of this article by those interested in documenting for the academic and scientific community the experiences acquired.

9.4 Proof of Concept

This section describes the features of the STEPES-BD Project, such as periodic deliveries developed over four Sprints.

The Sprint Zero This Sprint was dedicated to leveling the knowledge of the students, designating them to: carry out short courses through the Internet; identify and develop skills in Team Members; receive the initial guidelines for the formation of teams and definition of the Scrum roles.

In this Sprint, the initial versions of the Assigned Mission artifact was defined to make a single delivery of value at the end of the Project. In addition, the following were also elaborated: Product Backlog, involving US that could turn into future functionalities of a computational system, as well as the Project Architecture artifact. At the end of this Sprint, some students with interpersonal characteristics and appropriate techniques to be allocated to the teams and segments of the Project have already started to stand out.

The Sprint The teams chose to be conservative, which was corroborated by the only 157 points selected for the 13 USs, during the first Sprint Planning, applying the Planning Poker technique. These USs were divided into 74 tasks.

In this Sprint, according to the TMML, the ST have already managed to migrate from the Forming stage to the Storming stage. The latter, characterized by conflicts of ideas and working methods among its members [10].

For a better use of students' learning, this Sprint was chosen so as not to make extra-team integrations yet. In this way, each ST started to develop its own front ends, data models, back ends, test cases, and web servers to support the computational system. Also, each ST individually developed their functionalities to enable surveys and management from your registration data.

The Sprint 2 In this Sprint, after applying the Planning Poker technique, the STs estimated a total of 226 points for the 14 USs selected in the Sprint Backlog #2, which enabled the development of 101 Tasks, representing an increase of 44% in implemented points in relation to the previous sprint.

In this Sprint, it was found that members of the STs were already more integrated and able to work better as teams, starting the integration between teams by pairs of segments (Patient with Physician and Hospital with Supplier). However, teams still have needed some interventions from their TSMs to better standardize and organize their meetings and activities between teams showing that they were still in the Storming stage and only walking towards the Norming stage.

Considering the TMML, the STs achieved their goals and managed to migrate from Sprint 1 and from the Forming stage to the Storming stage. Due to this good performance in Sprint 1, it is believed that, perhaps due to overconfidence, the STs have decided to increase and overestimate (by another 44%, in relation to the previous Sprint) the total points of their estimated efforts to be performed in this Sprint 2.

For these reasons, it is believed that the STs have not been able to achieve a better performance nor have they migrated yet to the Norming stage in this Sprint 2.

The Sprint 3 In this Sprint, after applying the Planning Poker technique, the ST estimated a total of 256 points for 17 US selected in the Sprint Backlog #3, which led to the development of 129 Tasks, consolidating an increase of 13% in the points implemented in relation to Sprint 2 (previous).

It was verified that all STs were already in Stage 3, Norming, and out of that only 2 of the 4 STs had some signs and characteristics inherent to the Stage 4, Performing, as they already interacted with each other without the need for intermediation by the PO, professors or the TSMs. This

behavior can be seen from various alignment and integration meetings recorded in the Minutes, where members of some STs already called spontaneously and when necessary.

In this Sprint, deliveries with greater aggregate value could be observed. This was mainly due to the requirements, chosen in terms of USs to promote the development of more consolidated Tasks also aligned with the maturity levels reached. Specifically with regard to the relationship between the teams and differently from what happened in Sprint 2, only with integrations in pairs (Patient with Physician and Hospital with Supplier), in this Sprint 3, integrations started to occur between all STs simultaneously, characterizing interoperability.

Finally, another relevant factor for this project to be successfully completed was the situational awareness and the maturity levels progressively achieved by the STs over the 3 Sprints. In planning for an increase of just 13% of points in Planning Poker on Sprint 3, compared to Sprint 2, the STs have shown better accuracy over time to correctly estimate the efforts to be made for the development of USs and Tasks.

The Final Presentation It took place on June 30, 2020, in the form of a video-conference and with the live participation of 40 people, including the Rector of the Aeronautics Institute of Technology, students, professors, and Higher Education Institutions in the Paraíba Valley region, São Paulo, Brazil, among other guests interested.

A 60-min recording of the Final Project, where nine students representing the others have linked the main topics of this Project. The end of the project was marked by the dissolution of the teams, Adjourning stage [11], due to the successful conclusion of the academic semester.

9.5 Results and Discussion

This section presents the outcomes gathered from different perspectives: (1) quantitative results achieved from measurements of worked hours, completed tasks, and accomplished points; (2) students' feedback synthesis to understand the evolution of the students at the end of the project; and (3) comparison with the related works shown in Sect. 9.2.

(A) Quantitative Outcomes

In order to generate the quantitative outcomes, data from the ST #2 (Physician) were synthesized, because have presented similar performance to the other teams, in addition to better detailing of worked hours in each US.

It was possible to analyze that students in the 3rd Sprint needed fewer hours to complete the tasks. The speeds were 4.1, 3.3, and 2.7 h per point (hpp), in Sprints 1, 2, and 3,

respectively. This represented an increase in speed of 20% and 34% on Sprints 2 and 3, respectively, as compared to Sprint 1. This speed improvement was corroborated by the evolution of the teams, in which the teams became more mature and improved their management.

(B) Students' Feedback

In order to collect feedbacks from the students' perception, at the end of the project, students answered questionnaires in which the responses were synthesized into the following questions categories: **C1: What worked well?** and **C2: What can be improved?**

In Category 1 (C1), the responses were grouped into 7 topics and 43 mentions were collected: extra-teams integration (33%), union (21%), maturity (16%), meetings quality (9%), communication (9%), organization (7%), and engagement (5%). In Category 2 (C2) the responses were grouped into 10 topics and 39 mentions were collected: individual assessment (38%), communication (23%), overload (15%), and artifacts enhancement and technology specification (5% each). The other topics received only one mention each.

From these results, it can be analyzed that the main objective of the STEPES-BD Project was achieved, as it is precisely in soft skills that the main gap between academic experience and the development environment in companies and industries are.

The STEPES-BD Project achieved a good rate of students' perception in the topics of extra-team integration, union, and maturity. Otherwise, there are some improvements to do, as pointed by the students, especially in the manner of the evaluation was conducted, where the individual assessment was the main suggestion. In addition, communication was mentioned, indicating that communication processes need to be adjusted. Finally, the topic of overload was mentioned a lot. This is mainly due to the short project time and the learning curve of some technologies.

(C) Related Work Comparison

The Table 9.1 describes a comparison between the related works and the STEPES-BD Project, represented by the last record. The **Y** column displays the year of the article. The **Cit.** column displays the number of citations based on Google Scholar. The **Part.** column shows the number of students participating in the projects. The **Teams** column indicates the number of teams participating in the projects. The **Dur.** column shows the project duration, where "y" and "w" represent years and weeks, respectively. The **Type** column indicates whether the approach was composed of a single collaborative project (**S**) for all students or multiple distinct projects, one for each team (**M**). The **Mode** column shows whether the project was executed in-person (**P**), semi-attendance (**S**) or remotely (**R**). The **Eval.** column indicates how the projects

Table 9.1 Quantitative comparison with the related works

#	Y	Cit.	Part.	Teams	Dur.	Type	Mode	Eval.
[3]	14'	16	32	7	1y	M	P	K
[2]	15'	48	55	27	12w	M	S	E, F, P
[5]	18'	20	41	8	10w	M	P	F, P
S	20'	–	25	4	17w	S	R	F, P

were evaluated, where the values indicate: “E” – external members assessment; “F” – qualitative assessment by students’ feedback through questionnaires; “K” – Kano Model; “P” – artifacts evaluation by professors.

By analyzing Table 9.1, one can see the different strategies used by the 4 works. Regarding the number of students in the project, it can be seen that the project by Francese et al. [2] was broader (55 students), while in this Project had 25 students. 75% of works were carried out in just 1 academic semester: STEPES-BD Project, [2], and [5].

Only in Kizaki et al. [3], the project was longer than 1 semester and 75% of the work have used different projects for each team, differing from the STEPES-BD Project, in which all students worked cooperative and collaboratively in the same project. This can be observed by the reduced quantity of the STEPES-BD Project teams (4), against 27, 7, and 8 of other works, [2, 3, 5], respectively. Regarding the students’ interaction manner, an exclusively remote modality was adopted, due to the COVID-19 Pandemic.

In other works, the in-person modality (50%) [3, 5], and semi-attendance (25%) [2] were used. For projects’ evaluation, feedback from students through questionnaires and artifacts assessment by teachers (75%) were reported as the main form of evaluation: STEPES-BD Project, [2], and [5]. Francese et al. [2] also added the evaluation by external collaborating participants. Only in Kizaki et al. [3], the evaluations of others were not used, because they followed the Kano Model.

The originality offered in the STEPES-BD can be seen as students were harmoniously integrated into a single project in interdisciplinary teams. In addition, carrying out the project in a completely remote environment may reflect the challenges of society today, in which the home office has been expanding.

Thus, the real applicability of the approach used in the STEPES-BD Project for teaching Software Engineering disciplines is evident, providing an approximation of teaching in universities and the real needs of companies and industries.

9.6 Conclusion and Future Work

This article described the development of the academic project Specific Technological Solutions for Special Patients and Database Systems, named in Portuguese STEPES-BD

Project, from an Interdisciplinary Problem-Based Learning approach (IPBL).

In it, the best practices of the Scrum management method were used, providing undergrad and graduate students at the Brazilian Aeronautics Institute of Technology (ITA) with a practical academic experience involving the development of computational systems, using emerging technologies.

At the end of the semester, the students managed to demonstrate with their Scrum Teams the successful development of a relatively complex computational system, in terms of a Proof of Concept, in a self-managed manner.

To enable the development of this Project, the Scrum framework was adapted to support integrations and interdisciplinarity in 2 levels of Scrum of Scrum, providing students an opportunity to participate in an agile project, involving Information Technologies to meet the needs of society, and public or private companies.

For the natural continuity of this Academic Research and Development Project, it is recommended to investigate other metrics for performance evaluation to compare results to be achieved with the good practices of the Scrum method.

As future work, it is suggested that IPBL applications be further explored and deepened in other ITA disciplines of Undergrad and Graduate Programs.

Acknowledgments The authors of this article would like to thank the Brazilian Air Force, the Brazilian Aeronautics Institute of Technology, the Ecosystema Negocios Digitais Ltda (WiBOO/Wibx) and the Casimiro Montenegro Filho Foundation, for the support and infrastructure offered during the development of this ALFA Project (MEC-ITA).

References

1. M. Brassler, J. Dettmers, How to enhance interdisciplinary competence – Interdisciplinary problem-based learning versus interdisciplinary project-based learning. *Interdiscip. J. Prob. Based Learn.* **11**, 9 (2017)
2. R. Francese, C. Gravino, M. Risi, G. Scanniello, G. Tortora, Using project-based-learning in a mobile application development course – An experience report. *J. Vis. Lang. Comput.* **31**, 196–205 (2015)
3. S. Kizaki, Y. Tahara, A. Ohsuga, Software development pbl focusing on communication using scrum, in *2014 IIAI 3rd International Conference on Advanced Applied Informatics* (2014), pp. 662–669
4. N. Kano, Attractive quality and must-be quality. *Hinshitsu (Qual. J. Jpn. Soc. Qual. Control)* **14**, 39–48 (1984)
5. C. Raibulet, F.A. Fontana, Collaborative and teamwork software development in an undergraduate software engineering course. *J. Syst. Softw.* **144**, 409–422 (2018)
6. J. Sutherland, A. Viktorov, J. Blount, N. Puntikov, Distributed scrum: Agile project management with outsourced development teams, in *40th annual Hawaii International Conference on System Sciences (HICSS’07)* (IEEE, 2007), pp. 274a–274a
7. J. Sutherland, Inventing and reinventing scrum in five companies. *CutterIT J.* **14**, 5–11 (2001)
8. G.S. Goncalves, R. Shigemura, P.D. da Silva, R. Santana, E. Silva, A. Dakwat, F. Miguel, P.M. Tasinaffo, A.M. da Cunha, L.A.V. Dias,

- An agile developed interdisciplinary approach for safety-critical embedded system, in *Information Technology-New Generations* (Springer, 2018), pp. 947–949
9. L.S. Siles, M.V. Santos, R.A. Rodrigues, A. Lineu Filho, J.P. Siles, R.E. Maria, J.C. Marques, L.A. Dias, A.M. da Cunha, An integrated academic system prototype using accidents and crises management as pbl, in *Information Technology-New Generations* (Springer, 2018), pp. 419–427
 10. B.W. Tuckman, Developmental sequence in small groups. *Psychol. Bull.* **63**(6), 384 (1965)
 11. B.W. Tuckman, M.A.C. Jensen, Stages of small-group development revisited. *Group Org. Stud.* **2**(4), 419–427 (1977)

API-First Design: A Survey of the State of Academia and Industry

Nicole Beaulieu, Sergiu M. Dascalu, and Emily Hand

Abstract

The evolution of distributed and cloud-based systems has led the computing community to converge on Microservice Architecture (MSA) as a preferred solution to distributed software design. Established design methodologies applied to MSA (e.g., Data-, Model-, and Domain-Driven Design) assist in decision-making about business capacity and functionality encapsulated by the microservice. An expected result of microservice design is a well-defined Application Programming Interface (API) that facilitates the access of microservice and system capabilities. However, even with the extensive documentation and defined frameworks guiding practitioners in their execution of MSA, challenges exist in defining and exposing clean APIs. Further, the industry's current focus on maximizing business capacities exposed by distributed systems emphasizes the importance of improving API design and implementation. To this end, API-First Design is emerging as a viable approach to MSA and API design. API-First principles suggest that all capabilities of an organization and its systems are exposed via an API and that the foundation of system design is the definition of clear and well-defined APIs. A significant challenge associated with API-First Design lies in the infancy of the topic and the necessity for peer-reviewed research defining guidelines for adoption and a baseline for future research. This paper seeks to move the state of the API-First Design methodology forward by exploring publications of the academic community and grey literature available on the topic. The paper concludes with a discussion about future research opportunities that may advance the understanding and adoption of API-First Design.

N. Beaulieu (✉) · S. M. Dascalu · E. Hand
College of Engineering, University of Nevada, Reno, Reno, NV, USA
e-mail: nbeaulieu@nevada.unr.edu; dascalus@cse.unr.edu;
emhand@unr.edu

Keywords

API-first design · API-driven design · Software architecture · Software design · Software engineering · Microarchitecture · Microservice · Microservice architecture · API culture · API design

10.1 Introduction

Over the last few decades, as systems have become increasingly distributed and data-intensive, Service Oriented Architecture (SOA) has become a standard for solving the technical challenges of industry enterprise software. Today, enterprise systems encapsulate business functionality, rules, and domain logic in standalone, loosely coupled, and independently deployable applications known as microservices [1–3].

As the popularity of microservices has increased in distributed system design, their benefits, limitations, and applicability have caught the attention of academia and are a topic of modern research. Studies include design methodologies that assist practitioners and researchers alike in making savvy decisions about the functionality of and boundaries between microservices that comprise a system. Such methodologies include Model-Driven [4, 5] and Domain-Driven [6] Design. Each methodology defines a set of practices intended to maximize the benefits of microservice architecture; the approaches focus on the resulting services and their encapsulated business capacities [2, 3, 7, 8]. However, attention to the interfaces (APIs) through which services interact, are accessed, and their capacities exposed for the business is often an afterthought [1, 3]. Added is the complexity of modern distributed systems, their APIs, and the variety and scale of target device requirements as shown in Fig. 10.1.

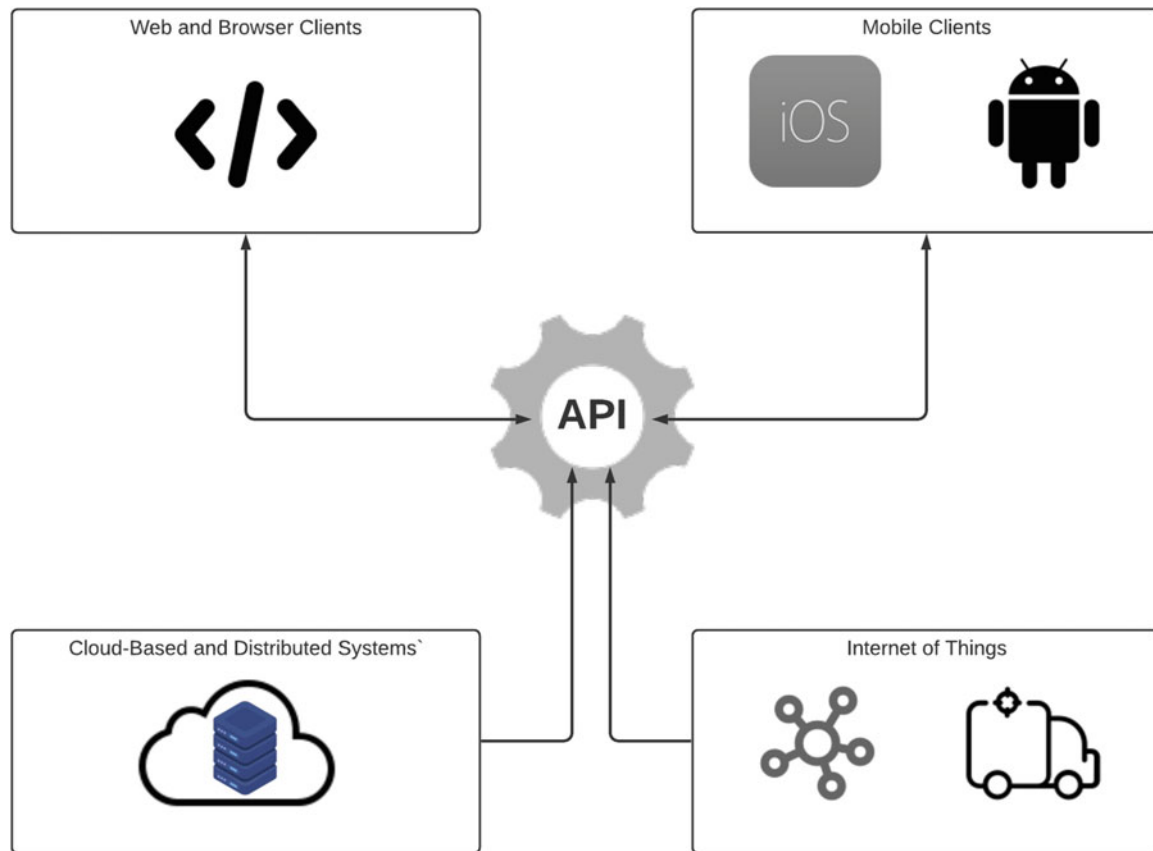


Fig. 10.1 The modern API

An emerging methodology, API-First Design, is gaining the attention of software practitioners in the industry. Its principles suggest that all business capabilities supported by a system are exposed via an API. The methodology also suggests that design focuses on the definition of clear and well-defined APIs as its foundation [7, 9]. A challenge associated with API-First Design lies in the imbalance between industry-generated grey literature (documentation that has not undergone the rigor of academic peer review or is not included in the scholarly publications traditional in academia [10]), and peer-reviewed research that may provide guidelines for adoption and a baseline for future research. Thus, given the importance of APIs in modern software solutions and the focus on API-First Design as a viable methodology for SOA and MSA, a systematic review of grey literature provides valuable context and content.

For career software professionals and practitioners of software engineering, the exploration of API-First Design presents an opportunity to explore end-to-end the possibilities created by this relatively young design methodology. This paper aims to review the existing academic research and grey literature, provide a comprehensive set of references, and identify options for potential future research of API-First Design.

This publication, in its remaining sections, is organized as follows: Section II provides a background on software architecture and methodologies; Section III overviews the state-of-the-art of API-First Design in academia and industry grey literature; Section IV outlines potential directions of future research; and Section V closes with a summary of the paper's main findings and our concluding remarks.

10.2 Background

The concept of Software Architecture emerged in the 1990s [11]. Its introduction provided lexicon and methodologies for describing software and its complexities from a high level. The concept of Service-Oriented Architecture (SOA) was introduced in 1996; its tenets include the organization and design of business functionality for distributed systems [11, 12] as well as the stability of the service Application Programming Interface (API) [11]. With the ubiquity of cloud-based, distributed, and data-intensive systems, SOA has become a standard for industry enterprise software design [1]. Specifically, systems frequently encapsulate business functionality, rules, and domain logic in loosely coupled and independently deployable applications known as microservices. A succes-

rior to SOA, Microservice Architecture (MSA) is a software architecture pattern introduced and coined in 2014 by Martin Fowler [1]. MSA more concisely defines a distributed architecture than the monolithic implementations of the SOA-era distributed enterprise systems. Microservice goals include encapsulated business capabilities, loose coupling between components, lightweight communication mechanisms, end-point intelligence, and deployment independence [1]. Many modern enterprise software systems take advantage of the benefits of MSA, including those from Netflix, Amazon, Guardian, Twitter, [2, 3], Intuit, and Figure Technologies.¹

The practices and methodologies for designing microservices are well-researched and documented. Common design approaches include Model-Driven Design (MDD) [5] and Domain-Driven Design (DDD) [6]. Introduced in approximately 2001 and an extension to MDD, MDA@[4, 13], is a software engineering methodology that encourages abstraction of the software complexities of service-based system architectures via models represented in the Unified Modeling Language (UML). The methodology encourages the separation of business logic from the platform and technical implementation concerns, allowing all to evolve independently [4, 5]. In 2003, Eric Evans coined the term Domain-Driven Design, a methodology that encourages system designers and architects to design software to reflect real-world business contexts and their users [6, 14]. While introduced before the microservice, DDD is a common design approach for microservice architectures today [8].

Even with the vast amount of documentation and research of these methodologies and with the well-defined guidelines for their adoption, challenges in identifying microservice boundary and granularity remain [2, 3, 7, 8]. Additionally, even though consequences of microservice design include an explicit interface positioned as the cornerstone of the design [15], challenges remain in defining precise published interfaces [1, 3], a critical aspect of design that is often considered late in the development cycle.

An emerging methodology for microservice design, API-First Design, is gaining popularity as demonstrated by the available grey literature authored by industry practitioners. The fundamental principles of this emerging approach are that all business capabilities are exposed via an API and that the design of the system focuses on clear and well-defined APIs as the design foundation [7]. Stated benefits of API-First Design in addition to a well-defined API include increased exposure of business capabilities, improved identification of system components, clear boundaries between frontend and backend logic and the services themselves,

improved scale, and increased developer velocity due to parallelization of effort [8, 16, 17].

Next, this paper explores the state of API-First Design publications of the academic community, provides an overview of the grey literature available on the topic, and includes a discussion about future research opportunities that may advance the understanding and adoption of API-First Design.

10.3 API-First Design

The importance of well-designed APIs is increasing as companies consider APIs as assets and commodities and therefore strive to maximize the capacities exposed via the suite of APIs [7, 18]. Even with this increased attention, challenges regarding microservice boundaries and their APIs remain [1–3, 7, 8]. Out of scope for this paper, it can be concluded that good API design is a critical aspect of modern software design and is a fundamental requirement of API-First Design. Publications targeting approaches for designing good APIs are plentiful; references include [18, 19], and [20]. With API-First Design gaining attention in academia and emerging as an industry practice for system design, the following sections provide a review of the state of API-First Design in academia and a summary of API-First-focused grey literature published by industry practitioners.

10.3.1 State of Academia

Prominent in practitioner-generated white papers and grey literature, API-First Design is a fledgling topic in academic research. Available publications generally include API-First design as a secondary theme. For example, in their 2019 paper [7], authors Wilde and Amundsen discuss the role that APIs play in the current digital landscape. They emphasize the important role of exposing capacities via an API in decentralized systems and the business for which they are designed. They further suggest that organizations with an API focus are striving to create an API Culture. In recognizing that API design is not an easy task, this paper focuses primarily on API design and management. The authors suggest that API-First Design is the preferred approach to exposing the capacities of a business via an API. The specifics of API-First Design are not discussed, leaving room to expand upon this topic in future research.

Bennett et al. discuss in their 2021 article [21] a case study conducted with Saxo Bank. The study explored test suite definition and coverage of APIs exposed via an API-First approach to service design. API-First Design is a secondary topic of the paper; the article's primary focus is test suite stability and functional coverage. The authors discover that

¹Author Nicole Beaulieu is currently the Vice President of Software Engineering for Figure's lending engineering team and was previously employed as a Senior Software Engineering Manager at Intuit.

the challenges associated with testing are often a result of a lack of initial context (IC), a concept common in Behavior-Driven Development [22]. They conclude that the API-First approach leads to improved testing capabilities and practices due to increased collaboration in test plan definition as a result of the API-focused design methodology.

In a noteworthy example of API-First research, authors Dudjak et al. present an API-First approach to designing a Backend as a Service (BaaS) platform [23]. Examples of a BaaS platform include data storage, user management, and file storage utilized by third-party application developers. The backend services, when combined by application developers, enable expedited delivery of full-featured platforms. In addition to the review of the design methodologies of BaaS components, the novelty of this research is found in the API-First approach adopted to design the system's architecture and to establish the platform's foundation. As discussed in the article, the benefits of this approach include loosely coupled services and faster time to market of products that utilize BaaS features. The authors of this paper recognize that the development process in an API-First approach begins with the definition and implementation of APIs that describe the platform; API-First Design frameworks are not included in the discussion. However, unlike the articles that reference API-First Design as a secondary topic, Dudjak et al. state the benefits of an API-First approach, including its facilitation of loose coupling between services and the representation of the services by a unique API. The goals and benefits included in this article coincide with those identified in industry and grey literature, a positive step toward closing the gap between the two sectors of the field.

While not exhaustive, the articles discussed above represent literature published by the academic community at the time of writing. The minimal number of publications discussing API-First Design indicates the novelty of the research in academia and exposes an opportunity to explore the work of our colleagues in the industry.

10.3.2 Industry and Grey Literature

In contrast to the limited research of API-First Design available in the academic setting, white papers, blogs, magazines, and video content from the industry discussing API-First design are plentiful. The extensive content provides an opportunity to understand the principles, benefits, and shortcomings of API-First Design.

In 2002, Jeff Bezos, founder and CEO of Amazon, distributed a memo to his employees that mandated the definition and use of service interfaces for technology and inter-team communication[24]. This API Mandate is an early rendition of an API-First approach to software design [25] and of building an API Culture. Fast forward nearly 20 years,

and Amazon offers a full suite of services and cloud-based products. Indicative of Amazon's evolution as an API-First technology company since the 2002 Mandate are Amazon Marketplace Webservice (MWS) [26] and Amazon Web Services (AWS) [27], API suites that facilitate the use of Amazon's marketplace and cloud-based technologies by entities external to the company itself. Another indicator of the importance of the API-as-a-commodity model is the API suite published for developer consumption by another Big 5 Tech Giant, Google [28]. Included in its API-accessed BaaS assets are Maps Platform, Google Ads, and Google Analytics [28].

A result of the digital transformation sparked by COVID-19, authors Wang and McLarty [29] discuss the shift in mental model required for business leaders to support the technological and business reinvention necessary to endure such a transformation. The authors explain that the health-care, banking, and logistics sectors may reap the benefits of API design and exposure. They also state that smaller companies may succeed amidst the more prominent companies by exposing their data via APIs. Similar to the Bezos API Mandate and the change in development and communication philosophy, the authors note that companies must learn to identify opportunities to expose business capacities via a suite of well-defined APIs [29].

In a 2020 blog post [17], author Chris Tozzi discusses the positive impacts of API-Driven development on an enterprise development strategy. Tozzi provides an invaluable summary of the benefits API-Driven [First] Design. Significant and relevant are the benefits of making the API the foundation of the architecture strategy. The approach ensures the exposure of components that may have been neglected in alternate API strategies also resulting in modular and cloud-friendly applications and services, improved CI/CD pipelines, and the assurance of compatible APIs due to the upfront focus on API development. Missing in academia, Tozzi summarizes steps that may assist in the successful adoption of API-First Design; a depiction of this process is shown in Fig. 10.2. The first steps represent an assessment of existing, new, and required APIs and the target and viable architectural patterns. The following steps focus on API development, testing, and monitoring integrated within the CI/CD pipeline. A final and essential step is the definition and adherence to a proactive API feedback loop.

Based on the results of a 2018 public survey shown in Fig. 10.3, the Lazar article [30] indicates that of modern architectural approaches to distributed system design, architects are most in favor (67%) of API-First Architectures. In the article, Lazar identifies some benefits of API-First Design that may contribute to the surveyed architectural preference. The decoupling of a system's frontend and backend components is included due to the well-defined API and clean dependencies. Furthermore, Lazar discusses API-

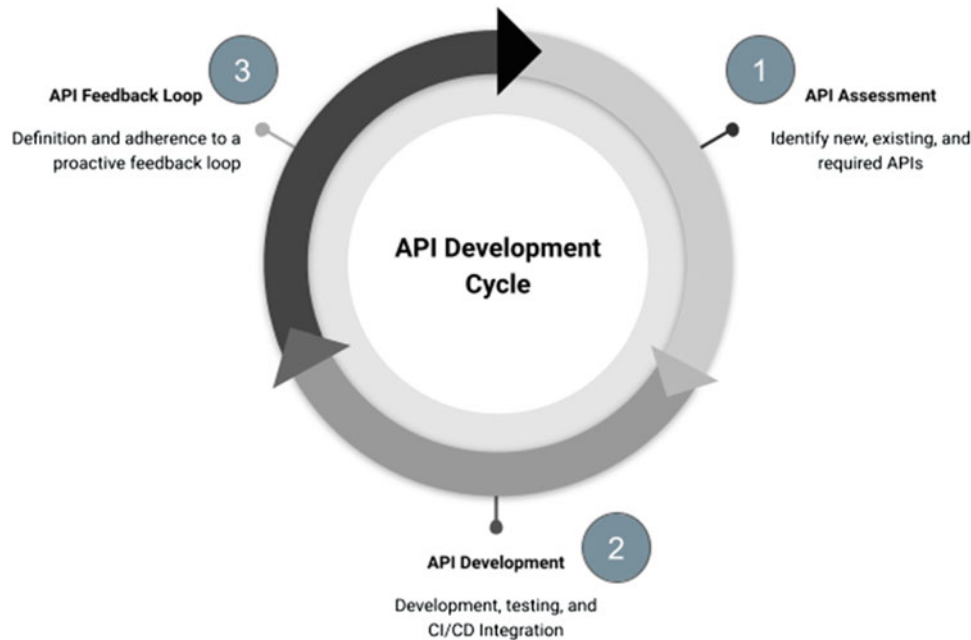


Fig. 10.2 API development cycle

First Design and its pluggable architecture. The modularity allows developers to focus on business logic and complexity rather than the application structure. Additionally, the naturally decoupled services enabled by the well-defined API increases opportunities for parallel development and an improved CI/CD pipeline [30].

Auth0 is a leader in user authentication service and provides a comprehensive API enabling integration with their services and platforms. In [11], based on experience at Auth0, Gontovnikas provides a high-level definition of API-First design, stating that the approach places the API at the center of software design, thus enabling services and software components to integrate easily. API-First Design discourages a code-first approach that leaves API attention to the end of a development cycle. Gontovnikas states that the benefits of the API-First Design approach include the ease with which software can be integrated and that the API is considered a valuable company asset. Both positively impact a company's bottom line by improving time to market, customer experience, and increasing the adaptability of implemented software [11].

The anticipation of API-First Design extends beyond the high-tech industry and audience. For example, Huerta et al. discussed the positive impact that this design approach has on the news integration and knowledge distribution ecosystems [31]. The article highlights the benefits of API-First Design and the abstraction of implementation details behind a well-defined interface designed early in the development cycle. The authors suggest that this approach supports open innovation and open API concepts also backed by the tech industry [32, 33].

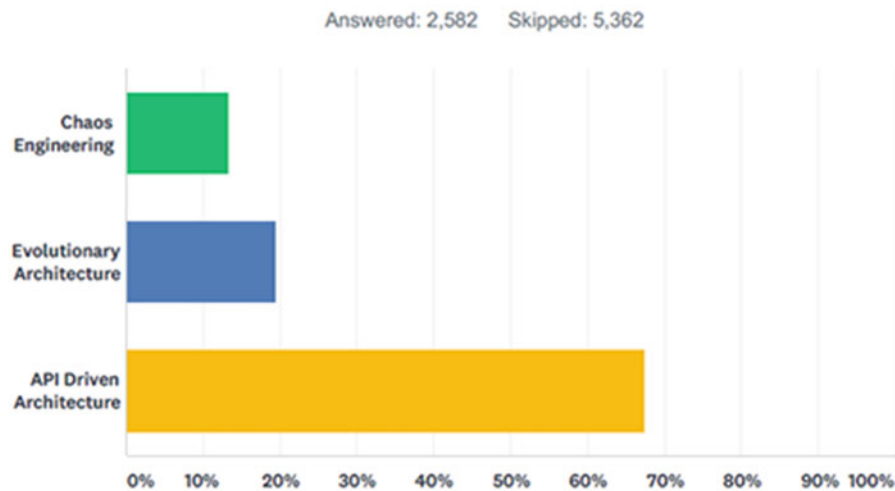
10.4 Future Research

The evolution of distributed and cloud-based technologies and the system architectures that support service-based products is leading a shift in business mindset that considers APIs an asset and commodity. In supporting this shift, software industry professionals are converging on the importance of quality APIs defined in an API-First Design approach [34]. While an imbalance in the number of publications exists between industry and academia, API-First Design is an emerging topic in the research setting.

The infancy of API-First Design presents an opportunity to explore end-to-end the possibilities presented by this design methodology. A viable first step is a comprehensive survey of the grey literature available on the topic. Additional opportunities for study include the creation of a standard definition and the development of a common language and taxonomy that assist the software engineering community in discussing API-First Design. Opportunities also exist in leveraging the adoption strategies discussed in grey publications such as [17] and [30] to design a framework through which researchers and practitioners may adopt API-First Design.

Research opportunities also exist in exploring the applicability of API-First Design. For example, model-based software generation is a well-researched topic that focuses on creating server-side code from a UML model. However, opportunities exist to automate UI/UX design and frontend code based on an API-First approach. This effort may be a natural extension to work performed by Rivero et al. in the

Q19 What new architectural approach are you most excited about?



ANSWER CHOICES	RESPONSES	
Chaos Engineering	13.22%	200
Evolutionary Architecture	19.43%	294
API Driven Architecture	67.35%	1,019
TOTAL		1,513

Fig. 10.3 Modern architectures survey, packt skill up 2018 [30]

MockAPI research [16]. A potential extension of Bennett et al. [21], additional research opportunities may exist in the examination of API-First Design and its potential benefit to the quality of fragile touch-points at the boundary of a microservice. The results may contribute relevant guidelines for test automation and quality and provide quantitative consideration of the impact on quality based on the early definition and testing of API contracts in a system.

Finally, it is worth noting that tools and standards available in the software community may prove valuable in future research and exploration of an API-First Design approach. Among them are the OpenAPI Specification [33] and RESTful API Modeling Language (RAML) [35]. Using standards and tools that support API development and testing, such as Postman [36] and Swagger [32], may also increase the viability and applicability of this research to the industry.

10.5 Summary

Microservice Architecture (MSA) is a standard design approach for distributed and cloud-based systems. However, attention to the interfaces through which services interact and are accessed is often an afterthought. As is evidenced by the quantity of grey literature about the topic, API-First Design

is an emerging methodology for software architecture and design. Its fundamental tenet is that system design focuses on creating a clear and well-defined API as the foundation for the system design. The API foundation defines the integration and communication patterns of the combined microservices [23]. Stated benefits of API-First design include increased exposure of business capabilities, improved identification of services, a clear boundary between frontend and backend logic, improved scalability, and increased developer velocity due to the natural parallelization of efforts [8, 16, 17].

While API design and implementation are ubiquitous in software engineering, API-First Design as a methodology is in its infancy. Opportunities exist in creating a standard definition and vocabulary that describes the methodology, building a framework to guide its adoption, and exploring the qualitative and quantitative benefits of applying API-First Design to architectural system design.

References

1. M. Fowler, J. Lewis, *Microservices* (2014). [Online]. Available: <http://martinfowler.com/articles/microservices.html>. Accessed 17 Oct 2021
2. J. Ghofrani, D. Lübke, Challenges of microservices architecture: A survey on the state of the practice, in *ZEUS* (2018)

3. J. Soldani, D.A. Tamburri, W.-J. van den Heuvel, The pains and gains of microservices: a systematic grey literature review. *J. Syst. Softw.* **146**, 215–232 (2018)
4. O. MDA, Object management group model driven architecture (2008). Accessed 17 Oct 2021
5. I. Sommerville, *Software Engineering*, 10th edn. (Pearson, London, 2015)
6. E. Evans, *Domain-Driven Design: Tackling Complexity in the Heart of Software* (Addison-Wesley, Boston, 2004)
7. E. Wilde, M. Amundsen, The challenge of api management: API strategies for decentralized api landscapes, in *Companion Proceedings of The 2019 World Wide Web Conference* (2019). Accessed 17 Oct 2021
8. R.H. Steinegger, P. Giessler, B. Hippchen, S. Abeck, Overview of a domain-driven design approach to build microservice-based applications, in *The Third International Conference on Advances and Trends in Software Engineering* (2017)
9. C. Munns, Building API-driven microservices with amazon API gateway - AWS online tech talks (2018). [Online]. Available: <https://www.youtube.com/watch?v=xkDcBssNd1g>. Accessed 17 Oct 2021
10. H.R. Rothstein, S. Hopewell, Grey literature, in *The Handbook of Research Synthesis and Meta-Analysis*, vol. 2 (Russell Sage Foundation, New York, 2009), pp. 103–125
11. M.H. Valipour, B. Amirzafari, K.N. Maleki, N. Daneshpour, A brief survey of software architecture concepts and service oriented architecture, in *2009 2nd IEEE International Conference on Computer Science and Information Technology* (2009), pp. 34–38
12. K.B. Laskey, K. Laskey, Service oriented architecture. *WIREs Comput. Statist.* **1**(1), 101–105 (2009). [Online]. Available: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.8>
13. I. Sacevski, J. Veseli, Introduction to model driven architecture (MDA), in *Seminar Paper, University of Salzburg* (2007)
14. V. Vernon, *Domain-Driven Design Distilled* (Addison-Wesley Professional, Boston, 2016)
15. L. Weir, *Enterprise API Management: Design and Deliver Valuable Business APIs* (Packt Publishing, Birmingham, 2019)
16. J.M. Rivero, S. Heil, J. Grigera, M. Gaedke, G. Rossi, MockAPI: An agile approach supporting API-first web application development, in *ICWE*, (2013)
17. C. Tozzi, The benefits of API-driven development and how you can implement it (2021). [Online]. Available: <https://www.merotech.com/blog/api-driven-development>. Accessed 17 Oct 2021
18. J.J. Bloch, How to design a good API and why it matters, in *OOPSLA '06* (2006)
19. J. Geewax, *API Design Patterns*. Manning (2021). [Online]. Available: <https://books.google.com/books?id=XWU0EAAAQBAJ>
20. J. Ofoeda, R. Boateng, J. Effah, Application programming interface (API) research: a review of the past to inform the future. *Int. J. Enterp. Inf. Syst.* **15**, 76–95 (2019)
21. B.E. Bennett, A practical method for API testing in the context of continuous delivery and behavior driven development, in *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (2021), pp. 44–47
22. M. Alferez, F. Pastore, M. Sabetzadeh, L.C. Briand, J.-R. Riccardi, Bridging the gap between requirements modeling and behavior-driven development, in *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems (MODELS)* (2019), pp. 239–249
23. M. Dudjak, G. Martinovic, An API-first methodology for designing a microservice-based backend as a service platform. *Inf. Technol. Control.* **49**, 206–223 (2020)
24. S. Yegge, Stevey's google platforms rant (2011), <https://gist.github.com/chitchcock/1281611>, Accessed 17 Oct 2021
25. E. Wilde, Jeff bezos' API mandate: What the five rules mean and do (2021). [Online]. Available: <https://blog.axway.com/amplify-products/api-management/jeff-bezos-api-mandate>. Accessed 17 Oct 2021
26. Develop for amazon selling partners (2021), <https://developer.amazonservices.com/>. Accessed 26 Nov 2021
27. Develop for amazon selling partners (2021). <https://developer.amazonservices.com/>. Accessed 26 Nov 2021
28. Google developers. <https://developers.google.com/>. Accessed 26 Nov 2021
29. T. Wang, M. McLarty, APIs aren't just for tech companies (2021). [Online]. Available: <https://hbr.org/2021/04/apis-arent-just-for-tech-companies>. Accessed 17 Oct 2021
30. A. Lazar, 8 reasons why architects love API driven architecture (2018). <https://hub.packtpub.com/architects-love-api-driven-architecture/>. Accessed 17 Oct 2021
31. J. Huerta, C. Childs, Accelerating news integration in automatic knowledge extraction ecosystems: an API-first outlook (2015). ArXiv abs/1509.02783
32. Swagger.io, API development for everyone (2021). [Online]. Available: <https://swagger.io/>. Accessed 17 Oct 2021
33. Openapis.org, Openapi initiative (2021). [Online]. Available: <https://www.openapis.org/>. Accessed 17 Oct 2021
34. L. Murphy, M.B. Kery, O. Alliyu, A.P. Macvean, B.A. Myers, API designers in the field: Design practices and challenges for creating usable APIs, in *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (2018), pp. 249–258
35. Raml.org, The simplest way to model APIs (2021). [Online]. Available: <https://raml.org/>. Accessed 17 Oct 2021
36. Postman.com, Postman API platform (2021). [Online]. Available: <https://www.postman.com/>. Accessed 17 Oct 2021

Part II

Data Science & Engineering

A Quality Dimension Analysis of Open Government Data in Undergraduate Public Funding in Brazil

Marcelo Moreira West and Glauco de Figueiredo Carneiro

Abstract

This paper describes the conduction of a tutorial to characterize the experience with dealing with nonconformities of open government data. The goal was to analyze data quality to identify limitations and errors relative to the use of open government data from the viewpoint of researchers and users in the context of Brazilian Student Financing (FIES). We planned the tutorial in seven steps to analyze the FIES data quality dimensions and present the results of difficulties and challenges faced by the participants.

Keywords

Open government data · Data quality · Data science

11.1 Introduction

Open government data has attracted ever-increasing attention from citizens due to its relevance to understanding several aspects of life in towns and cities. Governments' bodies worldwide have provided data that is of the main interest of citizens. The literature argues that adopting best practices to provide open data to citizens has the following goals: promoting transparency, encouraging participatory governance, and identifying government accountability for decisions. Providing open government data without considering relevant quality dimensions is a significant constraint to enhancing its practical use. Incomplete, imprecise, inconsistent, and

duplicate data are just examples of nonconformities in this scenario.

Studies have described the increased use of open government data portals in different regions of the world [1]. For example, there is a representative increase in datasets published on the Australian government portal in the order of 900% between 2013 and 2015 [1]. In the same period, the New Zealand government portal reached the milestone of 3,800 datasets, United Kingdom 23,000, United States 194,000, and Canada 240,000 [1].

Open government data can contribute to the development of aware and active citizens [2]. However, the lack of data quality is a challenge to achieving this goal. For example, there is a challenge to deal with the non-standardization of data formatting, and the non-conformity in the provision of metadata [3].

The tutorial presented in this article aims to analyze data quality to identify limitations and errors regarding the use of open government data from researchers and users not originally from information technology in the context of public funding to undergraduate students. This paper is structured as follows: Sect. 11.2 presents the main concepts of data quality dimensions. Section 11.3 details the methodological procedures for planning and conducting the tutorial. Section 11.4 presents further discussion regarding the findings in the context of open government data. Finally, Sect. 11.5 discusses the challenges faced in analyzing open government data.

11.2 Data Quality Dimensions

Analyzing data quality dimensions makes it possible to classify and identify data to achieve users' goals [4]. These dimensions contribute to selecting a specific dataset by potential users, which directly impacts the analysis results. The challenges to dealing with open government data influence

M. M. West
Instituto Federal Baiano (IFBAIANO), Salvador, Bahia, Brazil
e-mail: marcelo.west@ifbaiano.edu.br

G. d. F. Carneiro (✉)
Universidade Salvador (UNIFACS), Salvador, Bahia, Brazil
e-mail: glauco.carneiro@unifacs.br

Table 11.1 Study goal according to the GQM approach

Analyze	data quality
for the purpose of	identifying limitations and errors
with respect to the	use of open government data
from the point of view of	researchers and users
in the context of	Brazilian Student Financing (FIES)

the decision-making process due to data non-conformities. We present a set of quality dimensions already discussed in the literature and their description. *Interpretability* assesses whether it is possible to interpret the meaning and properties of data sources through the provided documentation and metadata [5]. This dimension highlights the importance of data properties to the practical use of data meaning. *Comprehensibility* evaluates how easily the data is understood [6]. In this case, the better the comprehensibility, the less effort we need to understand data. *Credibility* evaluates how genuine and trustworthy the data is [6]. Otherwise, it does not correspond to reality. *Integrity* assesses whether data is missing and whether it has sufficient breadth and depth to use [7]. We can observe a relationship between credibility and integrity, where integrity directly affects the data credibility. *Accuracy* evaluates whether the data is reliable and error-free [5]. The *Concise Representation* evaluates to what extent the data is represented in a compact form [7].

The *Consistency* evaluates whether there is compatibility between different data values [8] and *completion* evaluates the number of filled attributes [9]. Much of the open government data has errors such as missing attribute values, incorrect attribute values or different representations in the same dataset [10]. The *missing attribute values* correspond to the non-completion of mandatory attributes [10]. The *incorrect values* are considered those that do not correspond to the real situation [10]. The *different representations* occur when the attribute values have different formats [10]. The *illegal values* are considered those outside the domain of valid values [10]. To reach correct conclusions, the data must have a high level of quality [10]. In the following, we describe common errors usually found in data. The first is a set of outliers that deviate from the distribution of values. Duplicate data referring to the same entity also occur as a common error. Other occurrences are values that violate integrity constraints such as nullity and uniqueness [11].

11.3 Methodology

This section describes the tools we adopted to conduct the tutorial. Then we introduce the execution steps. Table 11.1 presents the goal of the study according to the Goal Question Metric (GQM) approach [12].

To perform the activities, the participants used *Google Meet*, a service that allows individual or collective interaction

Table 11.2 Tutorial steps performed by the participants

Step	Details
1	Getting a copy of the office tools package installation file <i>Libreoffice</i> and running
2	Obtaining the compressed file with the FIES data in "RAR" format available on the platform <i>Google Classroom</i>
3	Unzip the file obtained in the previous step
4	Validation of the unzipped file, comparing its size with with the file size available on the Brazilian Open Data Portal
5	Importing FIES data into the <i>Libreoffice Calc</i> tool
6	Analysis of FIES data quality dimensions
7	Sending the questionnaire APPENDIX A answered with the aid of the Data Dictionary ANNEX I to the organizers

through [13] videoconferences. They also used the *Google Classroom*, a component of the *Google for Education* application package made up of the (*gmail*) mail manager, (*drive*) file repository and text editor, spreadsheets and presentations (*docs*) [14]. The participants analyzed data provided by the Student Financing Fund (FIES), available on the Brazilian Open Data Portal in "CSV" format.¹ The reason to use data from the second half of 2019 was that they were the most recent available data in the portal.

(A) Planning

The FIES data quality dimensions analysis followed the seven steps presented in Fig. 11.1. The first step consisted in the installation of the Libreoffice tool. The second step corresponded to the download of the FIES open data file in a compressed format available on the platform Google Classroom. In the third step, participants uncompressed the file, and in the fourth step, they validated the content of the unzipped file, comparing its size with the file size available on the Brazilian Open Data Portal. In the fifth step, participants opened the unzipped file into the Calc tool in Libreoffice. In the sixth step the participants performed the data analysis, and in the seventh step they filled in a questionnaire. Table 11.2 presents the steps of the tutorial.

Selection of Participants We selected participants based on the following criteria: not having training in the Information Technology area, having skills in the use of computers, software, internet access, having an e-mail account on the *google* platform, and using the operating system (*Windows* or *Linux*).

Dimension Selection The criteria for selecting the data quality dimensions were: dimensions related to data usage and decision making. We choose the following dimensions: concise representation, interpretability, understandability, credibility, integrity, and accuracy.

¹<http://portal.mec.gov.br>

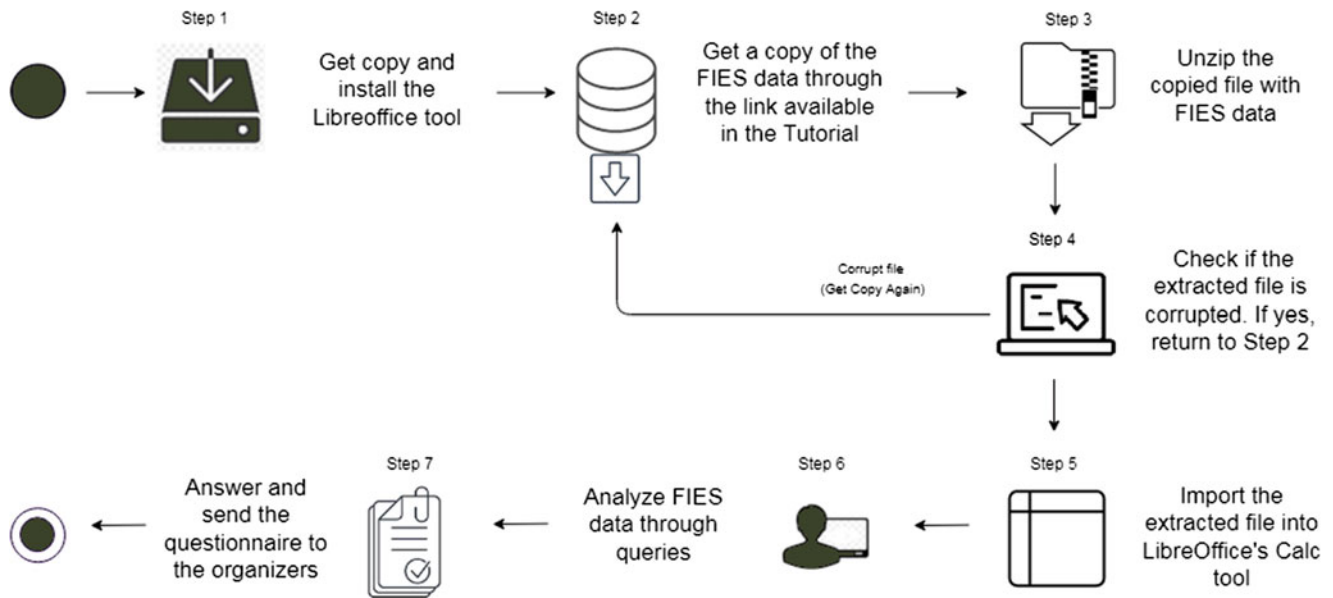


Fig. 11.1 Tutorial steps

(B) Execution

Figure 11.1 presents the steps performed by the participants. The participants took part in the tutorial in two sections. The organizers presented the proposed activities to download, install, and configure the environment in the first section. In the sequence, the participants performed the steps. The participants could ask the organizers questions regarding the activities.

Participants performed seven steps to analyze the data quality dimensions: In the first, they obtained a copy and installed the Libreoffice tool. In the second, they obtained a copy of the FIES open data in a compressed file. In the third, they unzipped the file copied in the previous step. On the fourth, they validated the extracted file.

The following tools were used by the participants: operating system Windows 10, browser Google Chrome Version 95.0.4638.54 64-bit and spreadsheet editor LibreOffice Community Version 7.2.0.4 (x86), WhatsApp Version 4.1., file unzipper Winrar Version 6.0 64 bits and mail manager.

11.4 Results and Discussion of Results

This section discusses the results of the FIES data quality analysis and also identifies opportunities for further improvements (Tables 11.3, 11.4, and 11.5).

The analysis of the quality of the FIES data revealed nonconformities as presented in Table 11.6.

Figure 11.2 presents the four stages of the analysis. The first step consisted of reviewing the literature to identify frequent errors represented in Table 11.3. The second step

aimed at identifying errors as reported in Table 11.3. Table 11.6 shows the number of errors found in the attributes of the FIES data by analyzing the responses of the participants. The evidence found were attributes filled in with spelling errors, filled in with illegal values, or not filled in. In The third step, the participants registered the errors they detected. In contrast, the fourth step aimed at proposing suggestions for improving the quality of data represented in Table 11.5.

Data analysis and error identification considered the following dimensions of data quality: interpretability, understandability, concise representation, credibility, integrity and accuracy. It is essential to highlight that the data published in open government portals must meet the principles of the selected dimensions, as they address that the users should easily interpret data. Moreover, complete, error-free, and reliable data contributes to an effective decision-making process (Table 11.7).

The organizers analyzed the answers provided by the participants during the identification of errors in the data. Participants answered the questionnaire provided in Table 11.6 supported by information provided by Table 11.8. Table 11.3 presents the list of errors found by the participants. The following errors were identified:

Missing Attribute Value They were found in the attributes Abbreviation of the Federation Unit, Name of the Municipality, Number of the National Register of the Legal Entity of the Maintainer, Name of the Corporate Name of the Maintainer, Code of the Municipality of the Maintainer and Name of the Municipality given in blank, which demonstrates non-compliance with the dimension **completeness**, according to questions 1 and 18.

Fig. 11.2 Tutorial data analysis steps



Table 11.3 Frequent quality mistakes in open data

Error identifier	Error type	Error name	Error description	Error example	Dimension of associated quality
E1	Incorrect values	Data used for a purpose other than intended	E-mail attribute filled in with a name	e-mail="Maria da Silva Souza	Consistency
E2	Incorrect value	Text type data filled with numeric type	Attribute name filled with cpf number	name="999.999.999-99"	Consistency
E3	Incorrect values	Numeric type data filled with text type	Phone attribute filled in with name	phone="Maria da Silva Souza"	Consistency
E4	Incorrect values	Date-type data filled in with a date greater than or equal to 100 years from the current date	Attribute date of birth filled with date of next century	birth date=2121-01-01	Accuracy
E5	Incorrect values	Data filled with spelling errors	County attribute filled in without accent but should be accented	county="Sao Paulo"	Accuracy
E6	Incorrect values	Data filled with more than one attribute	Attribute name filled in with name and process number	name="Maria da Silva Souza 99393346-20"	concise representation
E7	Missing attribute values	Unfilled data	Blank corporate name attribute	corporate name=""	Completeness
E8	Incorrect values	Data filled with undecipherable characters	County attribute filled with special symbols from the ASCII table	county="S - o Paulo"	Interpretability, Understandability e Credibility
E9	Illegal values	Data filled with illegal values	Attribute monthly value filled with negative value	monthly =-820,41	Consistency e Credibility

Table 11.4 Distribution of error quantity by error type

Error type	Number of	Error frequency – n(%)
Incorrect values	385949	99,297
Missing attribute value	2697	0,694
Illegal values	36	0,009
Total	388682	100

Table 11.5 Improvement suggestions

No.	Suggestion
1	Make attributes found in blank mandatory
2	Use the "Western Europe (Windows 1252/Latin1) character setting"
3	Include missing attribute descriptions in the data dictionary

Incorrect Values In the attributes Type of Contracted Surety, Sponsor's Municipality Code, Higher Education Institution Municipality Code, and Student Residence Municipality Code, data was represented in undecipherable characters and the Maintainer's Corporate Name, Education Institution Name attributes Higher (NO IES), Name of Higher Education Institution (NO IES EXT ALUNOS)

data with an incorrect accent, which demonstrates non-compliance with the dimensions: interpretability, understandability, credibility and accuracy, according to the questions 14 and 20.

Illegal Values Invalid values were found in the Monthly Value attribute, which demonstrates non-compliance with the data quality dimensions *consistency and credibility*.

Table 11.6 Distribution of error quantity by attribute

Attributes	E1	E2	E3	E4	E5	E6	E7	E8	E9	Subtotal
SG UF	0	0	0	0	0	0	351	0	0	351
NO MUNICIPIO	0	0	0	0	0	0	351	0	0	351
CO MANTENEDORA	0	0	0	0	0	0	0	0	0	0
NU CNPJ MANTENEDORA	0	0	0	0	0	0	351	0	0	351
NO RAZAO SOCIAL MANTENEDORA	0	0	0	0	375122	0	351	0	0	375473
CO MUNICIPIO MANTENEDORA	0	0	0	0	0	0	351	0	0	351
CO IES	0	0	0	0	0	0	6	0	0	6
NO IES	0	0	0	0	975	0	6	0	0	981
CO MUNICIPIO IES	0	0	0	0	0	0	366	0	0	366
CO PROCESSO	0	0	0	0	0	0	0	0	0	0
CO CONTRATO FIES	0	0	0	0	0	0	0	0	0	0
CO ADITAMENTO	0	0	0	0	0	0	0	0	0	0
CO AGENTE FINANCEIRO	0	0	0	0	0	0	0	0	0	0
NO AGENTE FINANCEIRO	0	0	0	0	0	0	0	0	0	0
SG AGENTE FINANCEIRO	0	0	0	0	0	0	0	0	0	0
NU MES	0	0	0	0	0	0	0	0	0	0
NU SEMESTRE	0	0	0	0	0	0	0	0	0	0
NU ANO	0	0	0	0	0	0	0	0	0	0
VL MENSALIDADE	0	0	0	0	0	0	0	0	36	36
CO INSCRICAO	0	0	0	0	0	0	0	0	0	0
TP FIANCA	0	0	0	0	0	0	0	0	0	0
NU ANO EXERCICIO INSC	0	0	0	0	0	0	6	0	0	6
NU PERCENTUAL PROUNI	0	0	0	0	0	0	126	0	0	126
NU PERCENT SOLICITADO FINANC	0	0	0	0	0	0	6	0	0	6
VL REPASSE	0	0	0	0	0	0	0	0	0	0
DT NASCIMENTO	0	0	0	0	0	0	0	0	0	0
ST DEFICIENCIA	0	0	0	0	0	0	0	0	0	0
ST ENSINO MEDIO ESCOLA PUBLICA	0	0	0	0	0	0	0	0	0	0
CO CIDADE	0	0	0	0	0	0	0	0	0	0
SG SEXO	0	0	0	0	0	0	0	0	0	0
DS SEXO	0	0	0	0	0	0	0	0	0	0
CO ESTADO CIVIL	0	0	0	0	0	0	123	0	0	123
DS ESTADO CIVIL	0	0	0	0	0	0	123	0	0	123
SG RACA COR	0	0	0	0	0	0	0	0	0	0
DS RACA COR	0	0	0	0	0	0	0	0	0	0
CO INSCRICAO EXT ALUNOS	0	0	0	0	0	0	0	0	0	0
CO CONTRATO FIES EXT ALUNOS	0	0	0	0	0	0	0	0	0	0
CO IES EXT ALUNOS	0	0	0	0	0	0	0	0	0	0
NO IES EXT ALUNOS	0	0	0	0	975	0	0	0	0	975
CO CAMPUS	0	0	0	0	0	0	0	0	0	0
NO CAMPUS	0	0	0	0	8877	0	0	0	0	8877
VL PERC FINANCIAMENTO	0	0	0	0	0	0	6	0	0	6
VL MENSALIDADE EXT ALUNOS	0	0	0	0	0	0	0	0	0	0
QT SEMESTRE FINANCIADO	0	0	0	0	0	0	6	0	0	6
CO ADITAMENTO EXT ALUNOS	0	0	0	0	0	0	0	0	0	0
CO CURSO	0	0	0	0	0	0	0	0	0	0
DS CURSO	0	0	0	0	0	0	0	0	0	0
CO TIPO CURSO	0	0	0	0	0	0	33	0	0	33
DS TIPO CURSO	0	0	0	0	0	0	33	0	0	33
CO PERIODICIDADE CUR	0	0	0	0	0	0	51	0	0	51
DS PERIODICIDADE CUR	0	0	0	0	0	0	51	0	0	51
VL SEMESTRE	0	0	0	0	0	0	0	0	0	0
NU ANO PROC	0	0	0	0	0	0	0	0	0	0
NU MES PROC	0	0	0	0	0	0	0	0	0	0
Total	0	0	0	0	385949	0	2697	0	36	388682

Table 11.7 Questionnaire

01	In your opinion, the registration data which contains the sponsor code equal to 17024, the FIES contract code equal to 2187276, the reference month of the payment statement equal to 8 and the reference year of the payment statement equal to 2019 are correct?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
02	You assume that the record data contains the maintainer code equal to 1534, the FIES contract code equal to 274338, the reference month of the payment statement equal to 5 and the reference year of the payment statement equal does 2012 have inconsistencies?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
03	Do you believe that it is possible to identify the names of the municipalities in the contract data number 2830492?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
04	Do you think there is no flaw in the record that contains the municipality of Canoas, the sponsor code equal to 314, the contract code 135498, the reference month of the statement equal to 1 that prevents the identification of the financial agent?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
05	Do you consider that in the municipality of Alagoinhas, sponsor code 2079, there are 3 students with disabilities?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
06	Do you consider that in the contract code equal to 2165590, registration number equal to 3966632, is it possible to identify that the candidate is indigenous?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
07	Is it correct to state that in the contract code equal to 2378896, enrollment code equal to 2945086, whose name of the higher education institution is the same as Faculdade da Amazônia Ocidental, is the course description architecture?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
08	In your opinion, the data of the institution code equal to 780, contract code equal to 2286193, maintaining code equal to 14514, high school status in public school equal to S allow to identify whether the student studied high school in the public network ?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
09	Do you consider that in the contract code equal to 2128687, enrollment code equal to 3068960, reference month of the payment statement, the data of the city code of the educational institution and the city code are redundant?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
10	Do you agree that in the municipality equal to Arapiraca, description of brown race we have three candidates registered?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
11	Have you found field(s) used with a different purpose than expected? Example of problem found: e-mail field filled in with name. If so, enter the names of the fields found that were used for a different purpose than expected:	<input type="checkbox"/> Yes	<input type="checkbox"/> No
12	Did you find alphabetic field(s) filled with numbers? Example of problem found: field name filled with cpf. If so, enter the names of the alphabetic fields that have been filled in with numbers:	<input type="checkbox"/> Yes	<input type="checkbox"/> No
13	Did you find numeric field(s) that were filled with text? Example of problem found: phone field filled in with number and name. If so, enter the names of the numeric fields that have been filled in with text:	<input type="checkbox"/> Yes	<input type="checkbox"/> No
14	Did you find field(s) filled in with spelling errors? Example of problem found: field no razao social mantenedora is incorrectly written "educacao". If so, enter the names of the fields found filled with spelling errors:	<input type="checkbox"/> Yes	<input type="checkbox"/> No
15	Did you find field(s) filled in with a date that is more than 100 years apart from the current date? Examples of problems found: field dt nascimento filled with 01/01/1900 or 01/01/2100. If so, enter the names of the fields found filled in with a date greater than 100 years from the current date:	<input type="checkbox"/> Yes	<input type="checkbox"/> No
16	Did you find records that represent the same data, but with different identifications? Example of problem encountered: name filled in one record in complete form and in another record in abbreviated form. If so, inform the number of records found that represent the same data, but with different identifications:	<input type="checkbox"/> Yes	<input type="checkbox"/> No
17	Did you find field(s) filled with more than one attribute? Example of problem encountered: name field filled with process name and number If so, enter the names of the fields found filled with more than one attribute:	<input type="checkbox"/> Yes	<input type="checkbox"/> No
18	Did you find unfilled field(s)? Example of problem found: field no razao social maintainer not filled. If so, enter the names of the fields found without filling it in:	<input type="checkbox"/> Yes	<input type="checkbox"/> No
19	Did you find field(s) with name(s) different from the names in the data dictionary? If so, answer: What fields did you find with names different from the names in the data dictionary? What do you recommend to be done to correct the names of fields that were found with names different from the names in the data dictionary? What did you do to correct the field names that were found with different names from the data dictionary names?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
20	Did you find field(s) with data with undecipherable characters? If yes, please answer: What fields did you find with data with undecipherable characters? What do you recommend to be done to fix the fields you found with data with undecipherable characters? What did you do to fix the fields found with data with undecipherable characters?	<input type="checkbox"/> Yes	<input type="checkbox"/> No

Table 11.8 Data dictionary

ANNEX I – DATA DICTIONARY	
Field	Description
SG UF	Federation Unit Acronym
NO MUNICIPIO	City Name
NU CNPJ MANTENEDORA	National Register of Legal Entity (CNPJ) of the sponsor
NO RAZAO SOCIAL MANTENEDORA	Name of the sponsor's corporate name
CO MUNICIPIO MANTENEDORA	Municipality code of the sponsor according to IBGE
CO IES	Code of the Higher Education Institution (IES) in the e-MEC electronic system
NO IES	Name of Higher Education Institution – IES
CO MUNICIPIO IES	Municipality Code of Higher Education Institution IES according to IBGE
CO PROCESSO	Reference code for the payment process in the FIES Computerized System (SisFIES)
CO CONTRATO FIES	Contract code in the FIES Computerized System (SisFIES)
CO ADITAMENTO	Identification code of the amendment to the Sis-Fies financing contract.
CO AGENTE FINANCEIRO	Code of the Financial Agent responsible for student financing: (1) Banco do Brasil (2) Caixa Economica Federal.
NO AGENTE FINANCEIRO	Name of the Financial Agent.
NU MES	Payment statement reference month.
NU SEMESTRE	Reference semester of the payment statement.
NU ANO	Payment statement reference year.
VL MENSALIDADE	Monthly fee for the financed course.
CO INCRICAO	Student registration code in the FIES Computerized System (SisFIES).
TP FIANCA	Type of surety contracted: Normal / Conventional (N); Educational Credit Operations Guarantee Fund – FGEDUC (F) and Solidarity (S).
NU ANO EXERCICIO INSC	Year of exercise referring to the student's enrollment in the FIES Computerized System (Sis-FIES).
NU PERCENTUAL PROUNI	Scholarship percentage related to the University for All Program – ProUni.
NU PERCENT SOLICITADO FINANC	Percentage of funding requested by the student.
VL REPASSE	Total amount of financing transferred to the sponsor.
DT NASCIMENTO	Student's date of birth.
ST DEFICIENCIA	The student has special needs: No (N) and Yes (Y).
ST ENSINO MEDIO ESCOLA PUBLICA	The student attended high school in a public school: No (N), Yes (S) and Partially (P).
CO CIDADE	Code of the student's residence municipality according to IBGE.
SG SEXO	Gender Acronym: (F) Female (M) Male.
DS SEXO	Description of sex.
CO ESTADO CIVIL	Marital Status Code: (1) Single (2) Married (3) Widowed (4) Separated (5) Divorced (6) Stable Union.
DS ESTADO CIVIL	Description of marital status.
SG RACA COR	Acronym Race/Color.
DS RACA COR	Description Race/Color (A) Yellow (B) White (I) Indian (N) Black (P) Brown.
CO INSCRICAO EXT ALUNOS	Student registration code in the FIES Computerized System (SisFIES).
CO CONTRATO FIES EXT ALUNOS	Contract code signed with the Student Financing Fund – FIES.
CO IES EXT ALUNOS	Code of the Higher Education Institution (IES) in the e-MEC electronic system.
NO IES EXT ALUNOS	Name of Higher Education Institution – IES.
CO CAMPUS	Campus code in the e-MEC electronic system.
NO CAMPUS	Campus name.
VL PERC FINANCIAMENTO	Percentage of funding granted to the student.
VL MENSALIDADE EXT ALUNOS	Monthly fee for the financed course.
QT SEMESTRE FINANCIADO	Number of semester financed.
CO ADITAMENTO EXT ALUNOS	Identification code of the amendment to the financing contract in the FIES Computerized System (SisFIES).
CO CURSO	Course code in the e-MEC electronic system.
DS CURSO	Course description.
VL SEMESTRE	Semester fee of the financed course

The following attributes were found without a description in the data dictionary: Course Type Code, Course Type Description, Course Frequency Code, Course Frequency Description, Process Year, Process Month and Financial Agent Acronym.

The participants found it challenging to understand and interpret the attribute that refers to the year of the payment statement and were unable to identify whether the student is indigenous through the attribute Description Race/Color, as per questions 2 and 6.

Based on the results obtained in this study, we summarize improvement suggestions in Table 11.5. The first focuses on avoiding blank data records. The second argues the

benefits of preventing indecipherable data, and the third indicates the advantages of using data to facilitate the content's understanding and interpretation.

11.5 Conclusion

At the beginning of the article, there was doubt about the limitations faced by users of data published in open portals, so it became essential to study the analysis of the dimensions of data quality to know these limitations and know how to overcome them.

Therefore, the article aimed to analyze the quality of data to identify limitations and errors about using open government data from the perspective of researchers and users in the context of Brazilian Student Financing (FIES).

It appears that the general objective was met, as effectively, with the completion of the tutorial, it was possible to find quality problems in the data and propose solutions.

The first goal to be met was the preparation of the environment to carry out the data quality analysis, in which the participants successfully achieved this goal, as they performed the installation of support software, obtained a copy of the FIES data, and imported them for the tool used in the analysis.

The second goal to be met was the analysis of the dimensions of data quality. The participants successfully reached this goal, as we conducted the study in total and the evidence tabulated.

The third goal to be met was elaborating a list of suggestions for improvements. We selected users who did not have training in the Information Technology area. The quality dimensions we focused on were: interpretability, understandability, concise representation, credibility, integrity, and accuracy-related to the use of data.

Given the proposed methodology, it is clear that the tutorial could have been carried out with other user profiles, since in this work, given the limited time, it was only possible to do it with an audience without training in exact sciences. Thus, it is recommended for future work to develop a tutorial focusing on users trained in data science so that this profile of users can consider data migration to relational and non-relational databases.

References

1. S. Sadiq, M. Indulska, Open data: Quality over quantity. *Int. J. Inf. Manage.* **37**(3), 150–154 (2017)
2. A.d.A.P. Silva, D.A.A. Monteiro, A. de Oliveira Reis, Qualidade da informação dos dados governamentais abertos: Análise do portal de dados abertos brasileiro. *Revista Gestão em Análise* **9**(1), 31–47 (2020)
3. L. Ding, V. Peristeras, M. Hausenblas, Linked open government data [guest editors' introduction]. *IEEE Intell. Syst.* **27**(3), 11–15 (2012)
4. F.G. Dutra, R.R. Barbosa, Modelos e critérios para avaliação da qualidade de fontes de informação: uma revisão sistemática de literatura. *Informação & Sociedade*. **27**(2) (2017)
5. C. Batini, B. Pernici, Data quality management and evolution of information systems, in *19th IFIP World Computer Congress – WCC 2006*, vol. 214 (2006), pp. 51–62
6. L.L. Pipino, Y.W. Lee, R.Y. Wang, Data quality assessment. *Commun. ACM* **45**(4), 211–218 (2002)
7. R. Vaziri, M. Mohsenzadeh, A questionnaire-based data quality methodology. *Int. J. Database Manage. Syst.* **4**(2), 55 (2012)
8. M. Scannapieco, T. Catarci, Data quality under a computer science perspective. *Archiv. Comput.* **2**, 1–15 (2002)
9. K.J. Reiche, E. Hofig, I. Schieferdecker, Assessment and visualization of metadata quality for open government data, in *CeDEM 2014, International Conference for E-Democracy and Open Government* (2014), pp. 335–346
10. P. Oliveira, F. Rodrigues, P. Henriques, Limpeza de dados-uma visão geral. *Data Gadgets* (2004), pp. 39–51
11. Z. Abedjan, X. Chu, D. Deng, R.C. Fernandez, I.F. Ilyas, M. Ouzzani, P. Papotti, M. Stonebraker, N. Tang, Detecting data errors: Where are we and what needs to be done? *Proc. VLDB Endow.* **9**(12), 993–1004 (2016)
12. V.R. Basili, H.D. Rombach, Towards a comprehensive framework for reuse: A reuse-enabling software evolution environment (1988)
13. D.d.F.F.A. Sant, D.V. Sant'Anna, et al., Google meet como modalidade de ensino remoto: Possibilidade de prática pedagógica. *Anais do CIET: EnPED: 2020-(Congresso Internacional de Educação e Tecnologias – Encontro de Pesquisadores em Educação a Distância)* (2020)
14. R.H.N. Diniz, J.C.F. de Almeida, B.F.L. Rodrigues, M.M.R. Marmol, E. Superior, Utilizando o google classroom como ferramenta educacional–percepções e potenciais.

Flávio de Assis Vilela and Ricardo Rodrigues Ciferri

Abstract

ETL (Extract, Transform, and Load) is an essential process required to perform data extraction from data sources, transforming the extracted data into an integrated format and loading the integrated data into a data warehouse. Moreover, the ETL process can be performed in a real-time way. This paper presents a survey of real-time ETL process applied to data warehousing environments. The related studies were collected by means of a systematic review conducted by a methodological protocol. Moreover, we grouped the studies by categories, which had not been made before in the literature. The results showed that we have a wide range of opening opportunities of original research, as there are a lot of initial ideas and in turn they have not yet been tested, validated, and compared to related work.

Keywords

Survey · ETL · Real-time · Data warehouse · Data warehousing

12.1 Introduction

Nowadays, the collecting of data that support the decision-making process has become essential for different applications domains, such as health care systems, road control

F. d. A. Vilela (✉)
Department of Computing Federal Institute of Goiás, Jataí, GO, Brazil
e-mail: flavio.vilela@ifg.edu.br

R. R. Ciferri
Department of Computing (DC), Federal University of São Carlos
(UFSCar), São Carlos, SP, Brazil
e-mail: RRC@ufscar.com

systems, and digital agriculture. These data are located in operational sources (OLTP systems). Furthermore, the operational data are available in different formats, such as relational databases, semi-structured and unstructured sources, in text files, XML or JSON files, and spreadsheets [1, 2]. The decision-making process is performed based on the data of interest gathered from the operation sources. Therefore, to provide value for the operational data and effectively aid the decision-making process, the ETL (Extract, Transform, and Load) process should be applied. This process is responsible for dealing with the operational data and performing the extraction, cleaning, processing and integration of data. Thus, the data of interest available in the operational environment are sent to the data warehousing environment and stored into a dimensional, homogeneous and integrated database named as data warehouse (DW) [1, 3, 4].

Traditional ETL process is composed and represented by a workflow of tasks. These tasks are logically split in three major steps: (1) The first one is called Extraction (E). The main goal is to extract or capture the data of interest to the decision-making process from its respective operational sources. (2) The next step is called Transformation (T), where the data cleaning, data processing and data integration is performed. (3) The third step is called Loading (L), which is responsible for loading data to the DW [5, 6]. The result of the ETL process is to make the data of interest of the operational environment available to the decision-making process, that is, storing operational data into the DW in an integrated and consistent way [5, 7].

In the traditional ETL process, the process for gathering data from the operational sources to the DW frequently takes place in a specific and predefined period of time, on a loading time window or according to the organization's business rules [1, 5, 8]. In fact, the data can be gathered once a day, once a week or even once a month. In this time, the operational and the information environments are inconsistent. Besides, when using the loading time window for gathering data, both

operational and informational environments should be offline to perform the process. To perform the ETL tasks, some Change Data Capture (CDC) techniques are applied into the ETL process to aid the extraction phase. These techniques can be a trigger, log file, delta file, snapshot or replication. In this way, the ETL process is highly dependent on operational sources.

On the other hand, a lot of applications have emerged and these applications need to gather operational data from operational sources in a real-time way, such as health care systems [9] and systems that handle data from sensors [10]. For this purpose, the ETL process has been performed in a higher frequency than traditional ETL process or in each idle time of the application so that it can simulate the near real-time approach [3, 5, 11]. In some cases, the operational data should be available in real-time into the DW to the decision-making process [12, 13]. As a consequence, the requirements for a real-time environment, that is, availability, scalability and low response time, must be ensured by the ETL process [12].

In recent years, some literature reviews were proposed in the literature which focuses on: (1) available techniques to be applied to real-time ETL process [8, 14], (2) all features of real-time ETL process into a data warehousing environment [4], and (3) some architectures to make a data warehousing with real-time ETL capabilities [15]. However, there is no work that classifies all general studies and their applied purposes. Moreover, they did not discuss how the related work deals with the requirements of a real-time environment, nor which are the techniques applied to make possible the real-time ETL process and which CDC techniques are used in the ETL process.

By solving this gap, this paper provides a literature review of the studies that deal with real-time ETL process applied to data warehousing environments. The related work is separated into two groups: the first one is related to investigations whose main goal is a general study; the second one is regarding studies that focus on applied purposes, that is, architecture, methodology, methods and applications. By the end of this paper, we show a comparison of all work and we point out which are the state of the art for the real-time ETL process.

This paper is organized as follows: Sect. 12.2 summarizes the most important studies that was proposed in the literature, Sect. 12.3 shows and discusses the ETL process and its phases, Sect. 12.3.3 discusses the methodology used to obtain the related work, Sect. 12.5 depicts the comparison among the related work and Sect. 12.6 concludes the article.

12.2 Related Work

Many investigations have been accomplished in the literature to enhance the freshness of data into the DW by applying the ETL process. In the same way, some literature reviews were proposed to highlight the main architectures to perform real-time ETL process in data warehousing environments.

Vassiliadis and Simitsis [16] proposed a literature review of traditional ETL process as well as its issues and challenges and highlighted all aspects to perform the ETL process in a near real-time ETL process. Kabiri and Chiadmi [17] showed some commercial and open source tools and some prototypes to perform ETL process, beyond to show the design of the conceptual model of the ETL process. Vassiliadis [18] proposed a literature review from the conceptual and logical point of view of the ETL process.

From the point of view of general study to perform ETL process in the real-time way, Sabry and Ali [4] showed an overview of a basic DW architecture as well as an overview of the ETL process and the main techniques to apply it into real-time data warehousing environments. However, the authors did not show architectures to perform the ETL process for real-time data warehousing environments. Bouaziz et al. [15] proposed a literature review which showed the main techniques and two other architectures to perform ETL process in a real-time way. However, the authors just showed the whole architecture and they did not group them by each phase of the ETL process. Sabtu et al. [19] showed an overview of the ETL process grouping the study on extraction, transformation and loading phases, and also they showed the real-time requirements that should be respected into a real-time ETL process. However, the authors just showed a literature review and they did not show an applied purpose to be used into a real-time ETL process.

As we can note, a lot of literature reviews have been conducted to show all related work regarding the real-time ETL process in data warehousing environments. However, the previews comparative studies did not show the following features: (1) a comparison of the related work with the characteristics of a real-time environment; (2) an indicative which was the approach of the each paper, that is, if it was an architecture, method, methodology, strategy or application; (3) they did not show which CDC technique that was used; (4) they did not show which techniques are applied to make the ETL process with real-time capabilities; (5) they did not show the state-of-the-art of each phase of ETL process. So, in this paper we present a comparison of all general studies

and applied investigations of the real-time ETL process into data warehousing environments. Finally, we will show a comparison of all related work and point out the state-of-the-art for real-time ETL process applied to data warehousing environments.

12.3 ETL Process

The ETL process is an inseparable part of the data warehousing environment. Kabiri and Chiadmi [17] says that about 70% of the whole cost of developing a data warehousing environment is spent to design and build the ETL process. This process is composed and represented by a workflow of tasks. These tasks are logically split in three major steps:

12.3.1 Extraction (E)

The first step is called Extract, which the main goal is to extract the data of interest to the decision-making process from its respective operational sources. In this step, it connects to the operational sources from the operational environment in order to gather the data of interest. To do this, this step needs to know how to connect to the operational sources in order to access the operational data, as well as it needs to deal with the heterogeneity of the whole sources of the operational environment.

12.3.2 Transformation (T)

The next step is called Transformation, where the data cleaning, data processing and data integration is performed. From this step, the whole ETL process makes value to the extracted data. In other words, this step focuses on making available accurate data, that is, the data should be correct, unambiguous, consistent and complete.

12.3.3 Loading (L)

The last step is called Loading, which is responsible for loading data to the DW [5, 6]. The result of the ETL process is to make the data of interest of the operational environment available to the decision-making process, that is, storing operational data into the DW in an integrated and consistent way [5, 7]. In this step, the data can be stored into DW in two ways: (1) bulk-load, that is, all data from operational sources are stored into DW every time the ETL process is performed; (2) incremental loading, that is, the data from operational sources are stored into the target DW periodically, in any time interval or continuously [6].

12.4 Methodology

To make it possible to find all related work for the real-time ETL process, we made a systematic review, which the main goal was to identify the architectures, strategies, techniques, methodologies, applications and general studies to allow us getting an overview of all characteristics of the real-time ETL process.

The systematic review was guided by a protocol, which means that we made use of a method to allow us to collect a wide range of related work that deal with real-time ETL process. The protocol was composed of the objectives, the research questions, the sources of search, the keywords, the selection criteria of inclusion and exclusion of a work, the idiom of the studies and finally the search strings. We used Parsifal software (<https://parsif.al/>) to support the systematic review.

For the sources of search, we searched IEEEExplore Digital Library, ACM Digital Library, Scopus, Web of Science and Google Scholar. The idiom considered for all related work was only English. The keywords were composed of the terms real-time and ETL and all of the synonyms. The search string was composed of the keywords and all of the synonyms of the terms. So, we used the following basic search string: (“RTDW” OR “Real-Time” OR “Real Time” OR “RealTime” OR “Zero-Latency” OR “Zero Latency”) AND (“ETL” OR “Extract-Transform” OR “Extract Transform” OR “Extract, Transform” OR “Extracting-Transforming” OR “Extracting Transforming” OR “Extracting, Transforming”). It is important to highlight for all sources of search, the string has a slight change in its syntax. So, we made these adjustments on the string to allow us to perform the search into each digital library. Also, we consider finding the keywords on the title or on the abstract of the articles.

12.5 Related Work of Real-Time ETL Process

In this section, we show all related work of real-time ETL process in data warehousing environments. For this purpose, we split the related work into two groups, which are general studies and for applied purposes. Table 12.1 depicts all general studies.

The knowledge of all general studies is important to be aware of all theoretical aspects that are involved in the real-time ETL process. It means that we are able to identify the concepts, characteristics and specificities about the ETL process. Moreover, these general studies can aid us to find the state-of-the-art of real-time ETL process, as they expose all content that is available until that time. Despite the fact that the general studies are important to this paper, we will show

Table 12.1 Related work which the purpose is general studies

Id	Author	Title	Year
1	Langseth, J.	Real-time data warehousing: challenges and solutions	2004
2	Vassiliadis e Simitsis	Near real time ETL	2009
3	Penzlin et al.	Current state and future challenges of real-time ETL	2009
4	Farooq e Sarwar	Real-time data warehousing for business intelligence	2009
5	Tank et al.	Speeding ETL processing in DWs using high-performance joins for changed data capture (CDC)	2010
6	Kakish e Kraft	ETL evolution for real-time data warehousing	2012
7	Revathy et al.	From DWs to streaming warehouses: a survey on the challenges for real-time data warehousing and available solutions	2013
8	Ferreira et al.	Real-time DW: a solution and evaluation	2013
9	Sabry e Ali	A survey of real-time DW and ETL	2014
10	Wibowo, A.	Problems and available solutions on the stage of extract, transform, and loading in near real-time data warehousing (a literature study)	2015
11	Gajanan Mane, N.	Near real-time data warehousing using ETL (extract, transform and load) tools	2015
12	Muddasir e Raghuv eer	Study of methods to achieve near real time ETL	2017
13	Sabtu et al.	The challenges of extract, transform and load (ETL) for data integration in near real-time environment	2017
14	Bouaziz et al.	From traditional DW to real time	2017
15	Phanikanth e Sudarsan	A big data perspective of current ETL techniques	2017
16	Chandra, H.	Analysis of change data capture method in heterogeneous data sources to support RTDW	2018
17	Muddasir e Raghuv eer	Study of meta-data enrichment methods to achieve near real time ETL	2020

the related work that focuses on applied purposes. Table 12.2 shows these works.

Table 12.2 depicts the related work that focuses on applied purposes, which the purpose can be a method, methodology, strategy, technique or architecture. To better explain all related work of applied purposes, the Table 12.3 depicts all related work that is presented in Table 12.2, making a reference of a work of the Table 12.2 by the Id column. Also, all works are classified by the following features:

1. Which is the approach of the paper. In this way, we consider: Method (M), Methodology (Meth), Application (App), Strategy (Str), Technique (Tech), Architecture (Arch), Optimization (Opt);

Table 12.2 Related work which the purpose is applied purposes

Id	Author	Title	Year
1	Bruckner et al.	Striving towards near real-time data integration for DWs	2002
2	Viana et al.	A real time data extraction, transformation and loading solution for semi-structured text files	2005
3	Naeem et al.	An event-based near real-time data integration architecture	2008
4	Chieu, T. e Zeng, L.	Real-time performance monitoring for an enterprise information management system	2008
5	Thomsen et al.	RiTE: providing on-demand data for right-time data warehousing	2008
6	Santos e Bernardino	Real-time DW loading methodology	2008
7	Shi et al.	Study on log-based change data capture and handling mechanism in real-time DW	2008
8	Majeed et al.	Efficient data streams processing in the real time DW	2010
9	Javed, M. e Nawaz, A.	Data load distribution by semi real-time DW	2010
10	Zuters, J.	Near real-time data warehousing with multi-stage trickle and flip	2011
11	Zhou et al.	An ETL strategy for real-time DW	2011
12	Santos et al.	24/7 Real-time data warehousing: a tool for continuous actionable knowledge	2011
13	YiChuan e Yao	Research of real-time DW storage strategy based on multi-level caches	2012
14	MadeSukar sa et al.	Change data capture on OLTP staging area for nearly real time DW base on database trigger	2012
15	Tank, D.	Reducing ETL load times by a new data integration approach for real-time business intelligence	2012
16	Jain et al.	Refreshing datawarehouse in near real-time	2012
17	Xue et al.	Measurement of the energy real-time DW system design and Implementation	2012
18	Jia et al.	Research on real time DW architecture	2013
19	Obali et al.	A real time DW approach for data processing	2013
20	Halenar, R.	Real Time ETL Improvement	2013
21	Freudenrei ch et al.	An on-demand ELT architecture for real-time BI	2013
22	Valêncio et al.	Real time delta extraction based on triggers to support data warehousing	2013
23	Cuzzocrea et al.	Enhancing traditional data warehousing architectures with real-time capabilities	2014
24	Lebdaoui et al.	An integration adaptation for real-time data warehousing	2014
25	Mao et al.	Dynamic mirror based real-time query contention solution for support big real-time data analysis	2014
26	Li e Mao	Real-time data ETL framework for big real-time data analysis	2015

(continued)

Table 12.2 (continued)

Id	Author	Title	Year
27	Guo et al.	A new ETL approach based on data virtualization	2015
28	Martins et al.	AScale: big/small data ETL and real-time data freshness	2016
29	Figueiras et al.	An architecture for big data processing on intelligent transportation systems. An application scenario on highway traffic flows	2017
30	Muddasir and Raghuv eer	CDC and union based near real time ETL	2017
31	Biswas et al.	Efficient incremental loading in ETL processing for real-time data integration	2019
32	Godinho et al.	ETL framework for real-time business intelligence over medical imaging repositories	2019
33	Machado et al.	DOD-ETL: distributed on-demand ETL for near real-time business intelligence	2019
34	Muddasir and Raghuv eer	A novel approach to handle huge data for refreshment anomalies in near real-time ETL	2020
35	Thulasiram and Ramaiah	Real time DW updates through extraction-transformation-loading process using change data capture method	2020

2. The work makes use of the traditional techniques of the ETL process to deal with the data into the operational environment. So, we consider: Yes or No.
3. Which phases of ETL process is approached by the work. So, we consider: Extraction (E), Transformation (T) and Loading (L);
4. Which CDC technique is applied. So, we consider: Log (Log), Trigger (Trigger), Replication (Rep), Delta Files (DF), Timestamp (TS), Snapshot (Ss) and non-mentioned (NM);
5. Which near real-time technique is applied. So, we consider: Direct Trickle Feed (DTF), Trickle and Flip (TF), Real-Time Data Cache (RTDC), Real-Time Partition (RTP), Partial Extraction (PE), Parallelism (P) and non-mentioned (NM).
6. Which real-time requirement is approached by the work. We consider: Availability (Av), Response Time (RT), Scalability (Sca) and Performance (Per).

The Table 12.3 shows the comparison of the related work presented in the Table 12.2. As we can note, the majority of the studies focused on the proposal of an architecture to real-time ETL process. Moreover, all studies make use of traditional techniques of ETL process to deal with the data into the operational environment. In other words, the extraction phase of ETL process makes use of a trigger, log file, delta file, snapshot, timestamp or replication to extract data of interest from operational sources.

Table 12.3 Comparison of all work which the purpose is general studies

Id	Analyzed characteristics					
	1	2	3	4	5	6
1	Arq	Yes	ETL	NM	NM	RT
2	Arq	Yes	ET	NM	NM	Sca
3	Arq	Yes	EL	Trigger	RTP	RT
4	Arq	Yes	EL	Trigger, Ts	RTP	RT
5	Arq	Yes	EL	NM	P	RT
6	Meth	Yes	L	NM	RTP	Per
7	M	Yes	E	Log	NM	RT
8	Arq	Yes	E	NM	NM	Per
9	Arq	Yes	E	Ts	NM	Per
10	Str	Yes	L	NM	TF, RTP	RT
11	Str	Yes	E	Log	NM	RT
12	Meth	Yes	L	NM	RTP	Per
13	Arq	Yes	EL	Trigger	RTDC, RTP	Per
14	Arq	Yes	ETL	Trigger	NM	RT, Per
15	Str	Yes	E	NM	NM	NM
16	Arq	Yes	EL	Log	RTP, PE	Per
17	App	Yes	EL	Log	TF	RT
18	Arq	Yes	EL	Log	TF, RTP	RT
19	Arq	Yes	EL	Log	TF, RTP	RT
20	Str	Yes	T	NM	NM	NM
21	Arq	Yes	ELT	NM	RTP	RT, Per
22	Arq	Yes	E	Trigger	PE	RT
23	Meth	Yes	L	NM	TF, RTP	RT
24	Meth	Yes	EL	Trigger	RTP, EP	Per
25	Arq	Yes	EL	NM	RTDC, RTP	Per
26	Arq	Yes	EL	NM	RTDC, RTP	Per
27	Arq	Yes	TEL	NM	NM	NM
28	Arq	Yes	ETL	Log	RTDC	RT, Sca
29	Meth	Yes	ET	NM	RTP	RT
30	Arq	Yes	EL	Log	DTF	RT
31	Arq	Yes	E	Ss	NM	RT
32	Arq	Yes	E	NM	NM	RT
33	Arq	Yes	ETL	Log	RTP	Av, RT, Sca
34	Arq	Yes	L	NM	RTDC	RT
35	Arq	Yes	L	NM	Rep, Ts	Per

From the point of view of ETL phases, few work deal with all ETL phases, but the majority of work deals with at least the extraction phase. Moreover, the main used CDC technique is log file and trigger. From the point of view of used real-time techniques, the main used techniques are Real-Time Data Cache and Real-Time Partition. Also, the main real-time requirement considered is Response Time and Performance. Moreover, just one work considers the Availability, Response Time and Scalability requirements.

The drawback is that the majority of the related work did not compare to each other, they just cite some studies as a related work or as a motivational example. Moreover, the studies did not validated and tested against another approach.

Another drawback is that there are newer studies that do not know the previous studies, and almost all studies did not validate experimentally the proposed approach.

From all related work, we can highlight the studies that we consider that represent the state-of-the-art for the real-time ETL process. Langseth, J. [20] and Machado et al. [21] represents the state-of-the-art for the real-time ETL process. Langseth, J. [20] developed a work in 2004 and presented some techniques to allow the ETL process to have near real-time capabilities. These techniques are applied in the majority of work that deal with real-time ETL process and in turn it is cited by almost all work. Machado et al. [21] developed an architecture to perform real-time ETL process in data warehousing environments. For this purpose, they make use of some advanced techniques to deal with data that need to be shared between heterogeneous environments. Such techniques are parallelism, buffer and publish/subscribe systems. These techniques showed great performance gain in comparison with the traditional ETL techniques.

12.6 Conclusion and Future Work

This paper presented a literature review of real-time ETL process applied to data warehousing environments. Unlike the previous surveys, this paper showed all related work that proposes a general study or were developed for an applied purpose. Also, we classify the studies by the approach of the paper, if the work consider the traditional technique of the ETL process, which phases of the ETL process is considered, which CDC technique was used, which real-time technique was used to allow perform ETL process with real-time capabilities and which real-time requirement was guaranteed. As a consequence, we pointed out the state-of-the-art studies in the literature.

From this literature review and the positive points and drawbacks, we can point out that the real-time ETL process in the data warehousing environments has a lot of open opportunities of research. Many authors have just an preliminary idea of solutions and they did not reach the complete resolution of the problem. Moreover, unfortunately the majority of the presented solutions was not tested, validated and compared to each other previous related work, which means that we have a wide range of opening opportunities to develop a series of approaches that will be methodologically correct.

References

1. R. Mukherjee, P. Kar, A comparative review of data warehousing ETL tools with new trends and industry insight, in *Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017* (2017)
2. B. Yadranjiaghdam, N. Pool, N. Tabrizi, A survey on real-time big data analytics: Applications and tools, in *Proceedings - 2016 International Conference on Computational Science and Computational Intelligence, CSCI 2016* (2017), pp. 404–409
3. H. Chandra, Analysis of change data capture method in heterogeneous data sources to support RTDW, in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)* (2018), pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/document/8510574/>
4. F. Sabry, E. Ali, A survey of real-time data warehouse and ETL. *Int. J. Sci. Eng. Res.* **5**(7) (2014). [Online]. Available: <http://www.ijser.org>
5. N.M. Muddasir, K. Raghuvver, Study of methods to achieve near real time ETL, in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)* (2017), pp. 436–441
6. R. Kimball, J. Caserta, R. Kimball, J. Caserta, *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning Conforming, and Delivering Data* (Wiley, Hoboken, 2004)
7. K. Kakish, T.A. Kraft, ETL evolution for real-time data warehousing. *Conf. Inf. Syst. Appl. Res.* **5**(2214) (2012). [Online]. Available: www.aitp-edsig.org
8. A. Wibowo, Problems and available solutions on the stage of extract, transform, and loading in near real-time data warehousing (a literature study), in *2015 International Seminar on Intelligent Technology and Its Applications, ISITIA 2015 - Proceeding* (2015), pp. 345–349
9. T. Jain, S. Rajasree, S. Saluja, Refreshing datawarehouse in near real-time. *Int. J. Comput. Appl.* **46**(18), 975–8887 (2012)
10. M. Mesiti, L. Ferrari, S. Valtolina, G. Licari, G. Galliani, M. Dao, K. Zettsu, StreamLoader: An event-driven ETL system for the on-line processing of heterogeneous sensor data, in *Advances in Database Technology - EDBT*, vol. 2016 (2016), pp. 628–631
11. X. Li, Y. Mao, Real-time data ETL framework for big real-time data analysis, in *2015 IEEE International Conference on Information and Automation, ICIA 2015 - In conjunction with 2015 IEEE International Conference on Automation and Logistics*, Lijiang (2015), pp. 1289–1294
12. A. Sabtu, N.F.M. Azmi, N.N.A. Sjarif, S.A. Ismail, O.M. Yusop, H. Sarkan, S. Chuprat, The challenges of extract, transform and loading (ETL) system implementation for near real-time environment, in *International Conference on Research and Innovation in Information Systems, ICRIS* (2017), pp. 3–7
13. K.V. Phanikanth, S.D. Sudarsan, A big data perspective of current ETL techniques, in *Proceedings - 2016 3rd International Conference on Advances in Computing, Communication and Engineering, ICACCE 2016* (2017), pp. 330–334
14. M.N. Muddasir, D. Raghuvver, Study of methods to achieve near real time ETL, in *International Conference on Current Trends in Computer, Electrical, Electronics and Communication (ICCTCEEC-2017)* (2017), pp. 436–441
15. S. Bouaziz, A. Nabli, F. Gargouri, From traditional data warehouse to real time, in *Intelligent Systems Design and Applications* (2017)
16. P. Vassiliadis, A. Simitsis, Near real time ETL, in *New Trends in Data Warehousing and Data Analysis*, vol. 3 (2009)
17. A. Kabiri, D. Chiadmi, Survey on ETL processes. *J. Theor. Appl. Inf. Technol.* **54**(2), 219–229 (2013)
18. P. Vassiliadis, A survey of extract-transform-load technology, *International Journal of Data Warehousing and Mining*, **5**, pp. 1–27.
19. A. Sabtu, N.F.M. Azmi, N.N.A. Sjarif, S.A. Ismail, O.M. Yusop, H. Sarkan, S. Chuprat, The challenges of extract, transform and load (ETL) for data integration in near real-time environment. *J. Theor. Appl. Inf. Technol.* **95**(22), 6314–6322 (2017)
20. J. Langseth, Real-time data warehousing: challenges and solutions, *DSSResources.COM* (2004)
21. G.V. Machado, Í. Cunha, A.C. Pereira, L.B. Oliveira, ‘DOD-ETL: distributed on-demand ETL for near real-time business intelligence. *J. Int. Serv. Appl.* **10**(1), 1–15 (2019)

Participatory Modeling: A New Approach to Model Graph-Oriented Databases

13

Luis A. Neumann, Enzo Seraphim, Otávio A. O. Carpinteiro,
and Edmilson M. Moreira

Abstract

This article presents a new method for modeling graph databases using the entity-relationship model. We analyze four modeling techniques available in the literature, identify the strengths and weaknesses of each one, and propose a method that minimizes the query path while maintaining the clustering base. We performed practical tests on graph bases resulting from these models. The innovation of the new proposal is the participation of the database designer who, through their knowledge of the application's business rules, interferes in the most appropriate type of mapping in the final model.

Keywords

Algorithms · Data mining · Database modelling · Databases optimization · Database systems · Graph database · Information retrieval · NoSQL · Relational databases · Storage structures

13.1 Introduction

Nowadays, companies have an increasing demand to handle large volumes of data in their business processes. To meet this demand, have been presented new database conceptions, such as the so-called NoSQL (*Not Only SQL*) databases [1], and their use has been increasing every year, according to the website Db-Engines [2] follow-up. The NoSQL databases count on horizontal scalability to manipulate large volumes

of data per second. Horizontal scalability is based on a distributed memory architecture, including replication and fragmentation of data in different servers [3].

Among several NoSQL data models, the graph-oriented model has the advantage of bringing the technical domain closer to the business domain, facilitating data modeling [4]. The motivation for this work is that many systems initially adopt the relational model but need to be switched to the graph-oriented model when the demand to manipulate a large volume of data increases. Another issue is that most of the applications deployed in NoSQL databases run parallel to data services in relational databases [5].

In these terms, several authors, such as Bordoloi and Kalita [6], Park et al. [7], Lampoltshammer and Wiegand [8], Erven [9], Lysenko et al. [10], Virgilio et al. [11], Pokorný [12], Roy-Hubara et al. [13], and Angles [14], present methods to model data in the form of graphs, from the entity-relationship model or the relational model.

Based on the above considerations, this paper presents Participatory Modeling, a method to model graph-oriented databases with the main objective of reducing the number of accesses necessary to prospect data, however, without losing the database clustering. According to the methods presented here, the results of the experiments performed on the databases modeled are presented at the end of the article.

13.2 Related Works

This section presents four methods for graph-oriented databases, from modeling performed with the entity-relationship model or from a relational model scheme to identifying edges and nodes of the graph-oriented model.

An important concept regarding graph-oriented databases is that there are only nodes and edges. There is no such thing as a table, grouping registers of the same type and same attributes. Thus, when we state that table (or entity) T1 will

L. A. Neumann (✉) · E. Seraphim · O. A. O. Carpinteiro
E. M. Moreira
Federal University of Itajubá, Itajubá-MG, Brazil
e-mail: lneumann@lna.br; seraphim@unifei.edu.br;
edmarmo@unifei.edu.br

originate node $V1$, we are saying that each record of table $T1$ will become a new type of node $V1$.

Presented by Park et al. [7], the 3NF Equivalent Graph (3EG), part of the relational model, turns entities of the entity-relationship model into nodes and relationships into edges. In its guidelines, this is similar to the solution usually proposed by developers of the Graph Database Management Systems (GDBMS), such as Neo4j [15]. In short, all tables originated from the entity will be a node, and the foreign key attributes will be edges. Tables originated from cardinality unary or binary relationships $1 : 1$ or $1 : n$ will be edges. The other relationships will give a node with edges among the new node and those corresponding to the table connected by them. In this paper, this method is called M01.

In other approaches ([6] and [16]), the authors propose modeling based on Reference Graph. Although the steps are different from the previous method, the final result is identical. The main difference between the two methods is that the first is part of the relational model, and the second is of the databases' entity-relationship model. Entities are converted into nodes. The unary relationships (cardinality $1 : 1$ or $1 : n$) are converted into self-referent edges. Finally, cardinality $1 : n$ binary relationships are transformed into node edges from n side to 1 side. All the other relationships generate a node with edges oriented to the involved entities. In this paper, this method is called M02.

Several authors, such as Lampoltshammer and Wiegand [8], Erven [9], and Lysenko et al. [10], have approached data modeling based on the simple graph concept: graph without loops and no more than one edge connecting two nodes. These concepts were combined in the Simple Graph Modeling (identified as method M03). This modeling has two phases. The first phase starts in the relational model, following all the Equivalent Graph Modeling steps. In the second phase, all nodes with more than two attributes are analyzed, and the attributes that are not either their primary key or the attribute identifier are removed from the node. These removed attributes generate a new node, being the attribute of their only property. Finally, an edge between the original node and the new node is created.

The Simple Graph Modeling (Graph RDF) produces nodes with fewer attributes (the one considered the primary key and the other that identifies the data). This modeling approximates the graph data representation to the representation method called RDF (*Resource Description Framework*), which proposes to decompose the information into the *subject*, *predicate*, and *value* [17].

The Model-driven design (in this paper called method M04) was proposed by Virgilio et al. [11] and intended to reduce the number of accesses necessary in a base traversal from the junction of two or more nodes into one. Like the other ones, the process starts from the entity-relationship

model and, in short, regroups entities divided into two or more due to the normalization rules [18].

Based on the analysis of the strengths and weaknesses of the four models presented here, we developed the method proposed in this article.

13.3 Participatory Modeling

In this section, we present the Participatory Modeling. It receives this name because it requires the database designer's participation, besides the modeling algorithms. The designer interferes in the modeling due to his knowledge of the application's business rules of the database.

The process consists of three phases (Fig. 13.1):

- **Phase 1:** the entity-relationship model (ER) is analyzed, and the entities and relationships are modified with the eventual incorporation of entities by relationships or other entities and with the conversion of relationships with a degree (D_R) higher than two into relationships with degree 2, generating a simplified entity-relationship model (ER2). The flowchart of Figs. 13.2, 13.3, and 13.4 summarizes this phase.
- **Phase 2:** the ER2 is migrated to graphs, converting entities into nodes and relationships into edges.
- **Phase 3:** at the designer's discretion, this phase is performed with the generated graph. The nodes are analyzed, and the designer can transfer identification attributes to another node to reduce the number of attributes of the original node.

The first step of *Phase 1* (Fig. 13.2) is analyzing each entity in the model to determine which can be removed and have their attributes incorporated by other entities or even incorporated by their relationship. To be eligible as an entity that can be incorporated, the entity shall:

- participate in only one relationship;
- have a low maintenance profile; that is, once the system is in place, this entity will have few additions of elements and little change from the existing ones. At this moment, the designer's participation is essential as his analysis will determine the maintenance profile of the entity. This profile represents the Degree of Maintenance of the Entity (D_{ME}), taking the value 1 for low maintenance entities or 0 for high maintenance entities.

Entities that meet these two criteria will have Entities' Incorporation Degree (D_{EI}) equal to 1; otherwise, D_{EI} will be equal to 0. The process for calculating the incorporation degree is shown in the flowchart of Fig. 13.2.

Fig. 13.1 Proposed method' flow

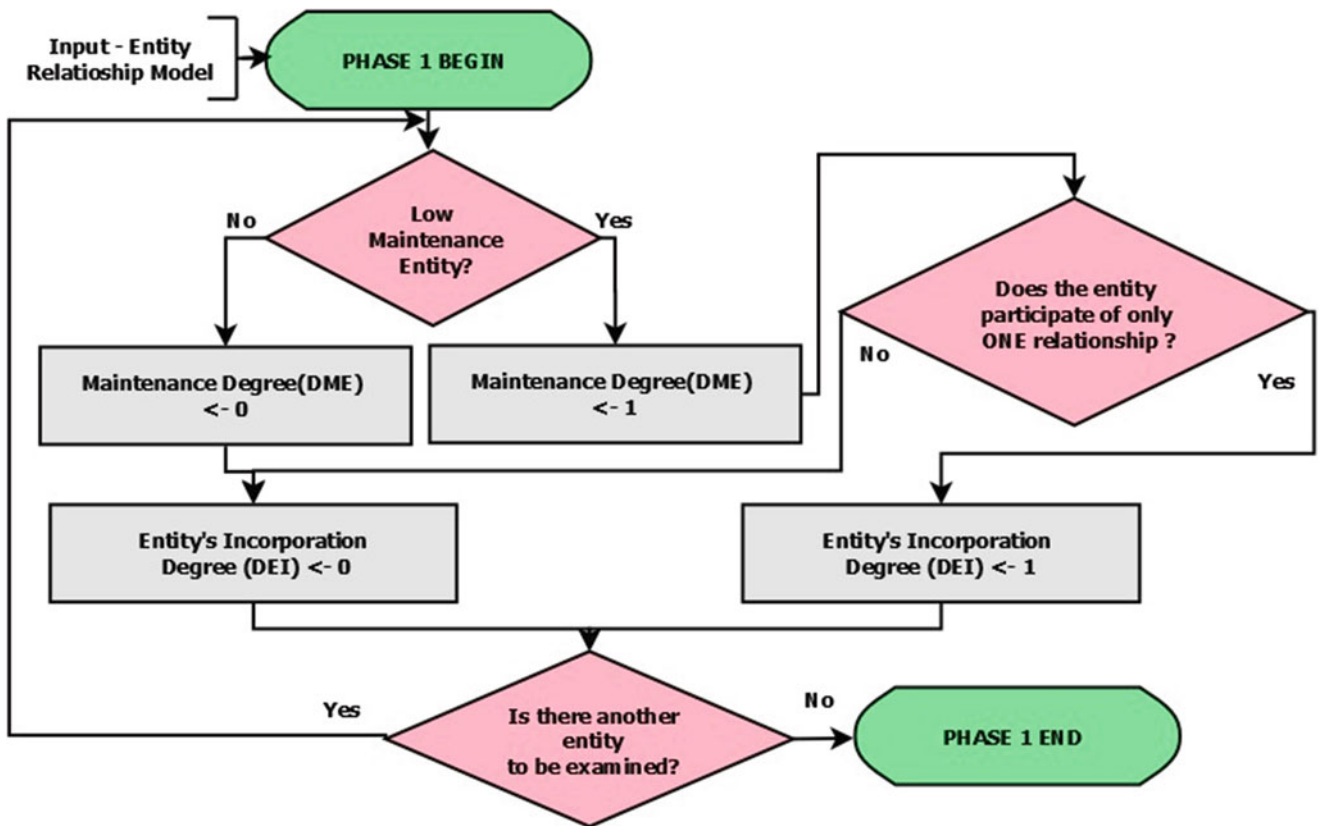
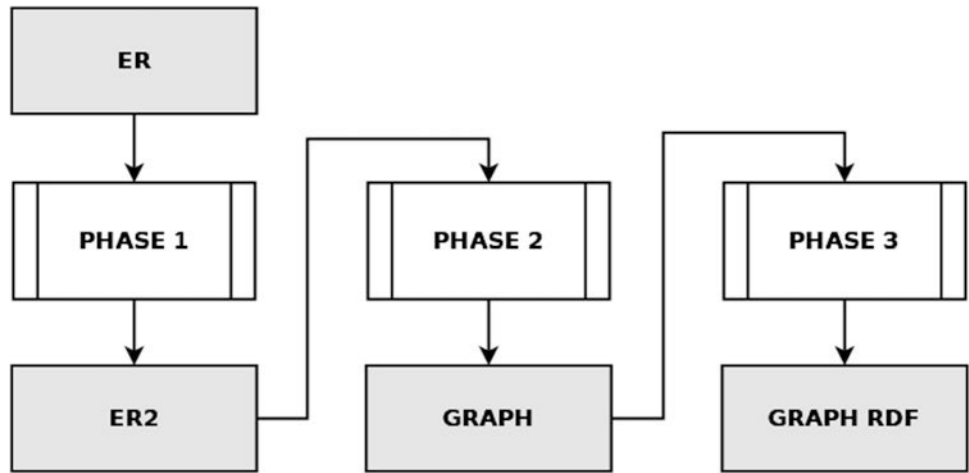


Fig. 13.2 Phase 1—step 1—calculation of the entities' incorporation degree

Phase 1, step 2 (Fig. 13.3), proceeds with analysing each relationship in the model, calculating the Relationship's Incorporation Degree (D_{RI}) sequentially and recursively:

- if the relationship has attributes, then $D_{RI} = 0$;
- else D_{RI} will also be zero (0) case the Relationship Degree (D_n) is equal to 2, and its cardinality is $m : n$ ($m > 1 \wedge n > 1$);
- otherwise, it will still be zero (0), if $D_n = 2$, the cardinality is $1 : n$, and the entity of n side is $D_{EI} = 0$;

- if none of these conditions occur, the result will be the sum of the Entity's Incorporation Degree (D_{EI}) of each entity connected by the relationship.

Then, in step 3 (Fig. 13.4), for each relationship with $D_{RI} > 0$, that is, which has an entity to be incorporated, our method will perform the following sequence:

- if $D_n > 2$, then remove from the model the entity with $D_{EI} = 1$, and join its attributes with the relationship attributes;

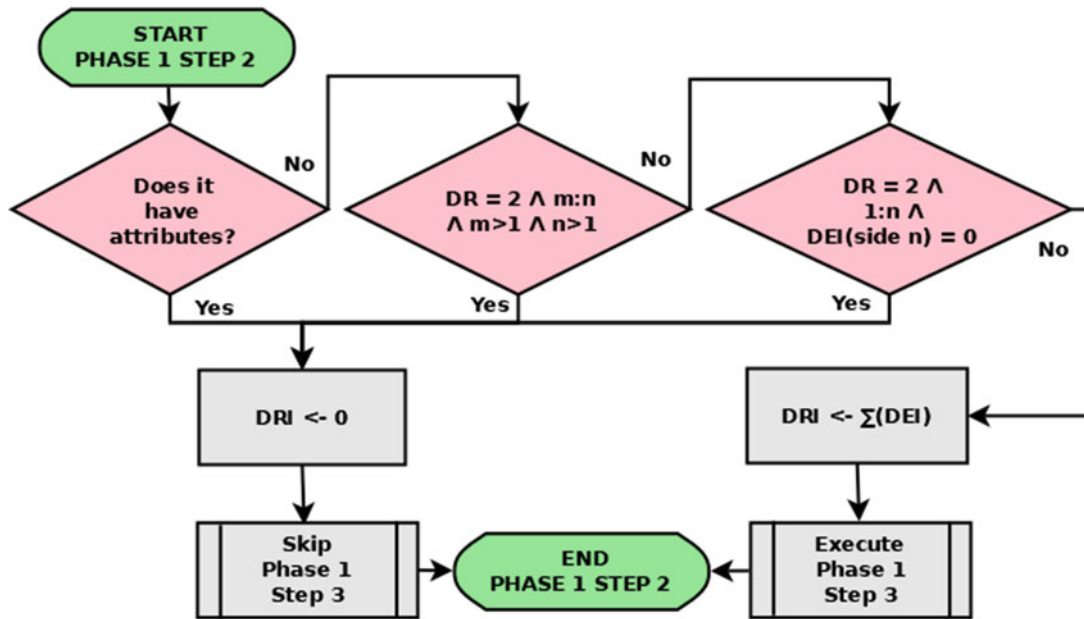


Fig. 13.3 Phase 1—step 2—calculation of the relationship's incorporation degree I

- otherwise, if $D_n = 2$, then remove the entity with $D_{EI} = 1$ from the model, and the respective relationship, incorporating its attributes in the remaining entity;
- if two or more entities that participate in the relationship have an incorporation degree equal to 1, one of them shall be elected to be incorporated into the relationship. In this situation, it is up to the designer to define which entity shall be incorporated based on his knowledge.

The procedure is performed recursively until the relationship has no more entities that can be removed from the model; that is, all relationships in the model have Relationship's Incorporation Degree equal to zero ($D_{RI} = 0$).

At the end of this process, relationships of degree greater than 2 ($D_n > 2$, involving three or more entities) will be replaced:

- by the creation of an entity with the same name as the relationship (E_R);
- for each entity involved in the relationship (E_R), a relationship of degree 2, cardinality 1 : 1, will be created between the entity E_{R1} and E_{R2} .

That done, there are no more relationships of degree greater than 2; all relationships in the model have degree 1 or 2.

Once *Phase 1* is executed, the new entity-relationship model (ER2) will include the following sets:

- original entities of the model (E_E);
- entities created from relationships (E_R) (in this phase);

- unary relationships (R_1);
- binary relationships of type 1 : 1 (R_{R11}) and 1 : n (R_{R1n});
- relationships created in this phase (R_{EE}), also binary.

The objective of *Phase 2* is to map the E-R model resulting from *Phase 1* (ER2) for the graph-oriented database model (GDBM). In the first step of *Phase 2*, for each entity ($E_E \wedge E_R$), a node (V) will be created in the MBDG with the same name and with properties, which will be the entity attributes. If the created node does not yet have a primary key property, it will be created for such purpose in the node. The node's primary key is the property used to identify it on the edges that connect it.

After defining the nodes in the first step, the second step identifies and converts the model relationships. Therefore, it will be recognized the unary relationships (R_{R11}), binary relationships of cardinality 1 : 1 or 1 : n (R_{R11} and R_{R1n}) that were not converted in the previous phase, the binary relationships of cardinality m : n (R_{mn}) of the original MER, and the binary relationships 1 : 1 created in the last phase (R_{EE}). Thus:

- unary relationships (R_1), generate bi-directional edges of the node for itself;
- binary relationships, R_{R11} , R_{R1n} , R_{mn} , and R_{EE} , generate bi-directional edges among the nodes corresponding to the entities connected by it;
- R_{R1n} relationships generate directed edges of the node corresponding to the n side entity for the node corresponding to the 1 side entity of the relationship.

Once this phase is finished, the model is converted.

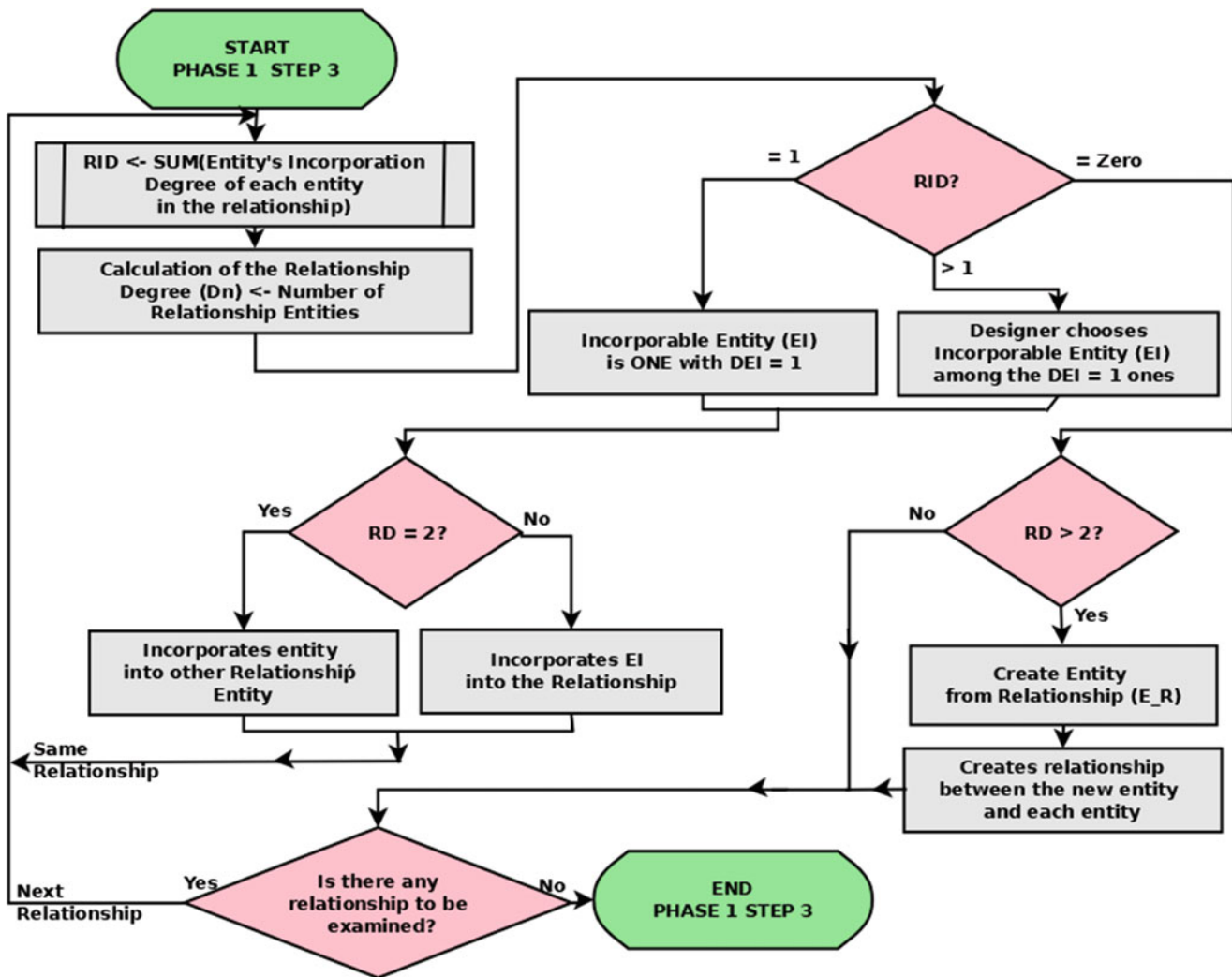


Fig. 13.4 Phase 1—step 3—calculation of the relationship's incorporation degree II

13.4 Experiments

The *Movies* database was used in this work, whose data were obtained in [19] to compare the *M01*, *M02*, *M03*, and *M04* methods and validate the method proposed herein. From these methods, we generated a relational database whose entity-relationship model can be seen in Fig. 13.5.

We implemented the proposed database in Neo4j, importing data from each entity to the nodes and edges of each method.

The first entity from *Movies* is the *Person* containing 9505 tuples. A *Person* has a *Name*, *Date* of birth, and *Sex*. This *Person* has a performance (*ACTED*) in the *Movie*. The performance was in the function (*Role*) of “Actor,” “Actress,” “Director,” “Screenwriter,” “Producer,” or “Musician” in a *Movie*, including different interpretations in the same *Movie*. This entity contains 12,509 tuples. A *Movie* has a *Title*, *Year* of release, and *Duration* in minutes. The *Movie* can be classified into different classes (*Category*), with

a *name* such as “Drama,” “Comedy,” “Documentary,” etc. Several countries can produce a *Movie*, but it can only be nominated for the award for only one *Country*, which has a *Name*. Annually, movies and persons can be nominated for different cinema awards. Each *Award* has a *Name* and different award categories. The different awards can be for the *Movie* or a *Person* due to his/her performance in the movie, but the same award cannot be attributed to a movie and a person. For example, the “Best Movie” award is granted only to movies, and the “Best Director” award is granted only to persons who acted in a movie.

There are the following relationships between these entities:

- degree 2:
 - IS1, type $m : n$, between *Movie* and *Category* entities;
 - FROM, type $1 : 1$, between *Movie* and *Country* entities;

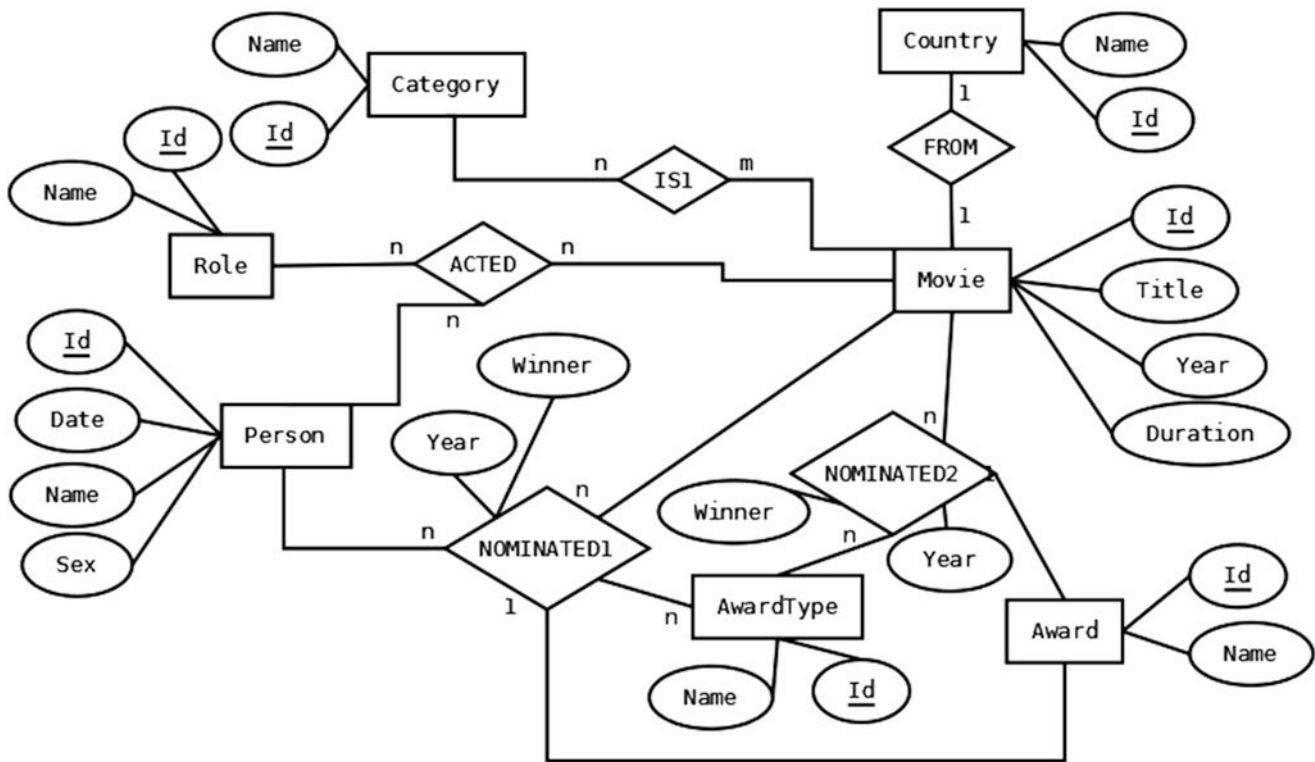


Fig. 13.5 *Movies*—entity-relationship model

- degree 3:
 - ACTED, among *Person*, *Movie* and *Role* entities;
 - NOMINATED2, among *Movie*, *Award* and *TypeAward* entities;
- degree 4:
 - NOMINATED1, among *Person*, *Movie*, *Award* and *TypeAward* entities;

13.4.1 Modeling the Movies Database According to the Participatory Modeling

As seen, this method starts with the entity-relationship model, and its first step is to calculate the degree of incorporation of the entities of the model. Table 13.1 shows the calculation of the entities' incorporation degree. The participatory role of the designer is evidenced:

- in the analysis of the *Country* entity. If the designer considers that it has a low maintenance profile, it can be incorporated into another entity; otherwise, this entity and its relationship will be maintained in the model;
- in the analysis of the *TypeAward* entity, which is related to two different entities, *NOMINATED1* and

NOMINATED2, which makes it not incorporable in the first moment. However, according to the designer's analysis, it can be considered related to only one entity since awards linked to persons are different from awards linked to movies; whoever participates in one relationship does not necessarily participate in another.

Based on the entities' incorporation degree, the relationships' incorporation degree higher than one is then calculated, the sum of the entities' incorporation degree that participates in it. Table 13.2 shows this calculation.

The *ACTED* relationship, degree 3, has an incorporation degree equal to 1, resulting from the *Role* entity. This entity shall be eliminated from the model, and its attributes incorporated by the relationship, which becomes degree 2 and has an incorporation degree equal to 0 (zero).

The *NOMINATED2* relationship, also degree 3, has an incorporation degree equal to 1, resulting from the *TypeAward* entity. Therefore, this entity shall be eliminated from the model, and its attributes incorporated by the relationship, which becomes degree 2 and has an incorporation degree equal to 0 (zero).

The *NOMINATED1* relationship, degree 4, has an incorporation degree equal to 1, resulting from the *TypeAward* entity. So, this entity shall be eliminated from the model, and its attributes incorporated by the relationship, which becomes degree 3 and has an incorporation degree equal to 0 (zero).

Table 13.1 *Movies*—method M05—phase 1—entities’ incorporation degree

Entity (E_E)	Number of Relationships it participates in	Maintenance degree		Relationships D_{RI}	D_{EI} = if $(D_{ME} + D_{RI}) = 2$ then 1 else 0
		Maintenance	D_{ME}		
Category	1	low	1	1	$1 + 1 = 2 \rightarrow 1$
Movie	6	high	0	0	$0 + 0 = 0 \rightarrow 0$
Role	1	low	1	1	$1 + 1 = 2 \rightarrow 1$
Award	2	high	0	0	$0 + 0 = 0 \rightarrow 0$
TypeAward	1(*)	low	1	1	$1 + 1 = 2 \rightarrow 1$
Person	3	high	0	0	$0 + 0 = 0 \rightarrow 0$
Country	1	low	1	1	$1 + 1 = 2 \rightarrow 1$

(*) Although it has two relationships, records participating in one relationship do not participate in another

Table 13.2 *Movies*—method M05—phase 1—relationships’ incorporation degree

Relationship	D_R	Entities							D_{RI}	
		E_{E1}	$D_{EI}(E_{E1})$	E_{E2}	$D_{EI}(E_{E2})$	E_{E3}	$D_{EI}(E_{E3})$	E_{E4}		$D_{EI}(E_{E4})$
IS1	2	Movie	0	Category	1					0(*)
FROM	2	Movie	0	Country	1					1
ACTED	3	Movie	0	Person	0	Role	1			1
NOMINATED1	4	Movie	0	Award	0	TypeAward	1	Person	0	1
NOMINATED2	3	Movie	0	Award	0	TypeAward	1			1

(*) has a value of zero because the relationship has a cardinality $m : n$
 $D_{RI} = D_{EI}(E_{E1}) + D_{EI}(E_{E2}) + D_{EI}(E_{E3}) + D_{EI}(E_{E4})$

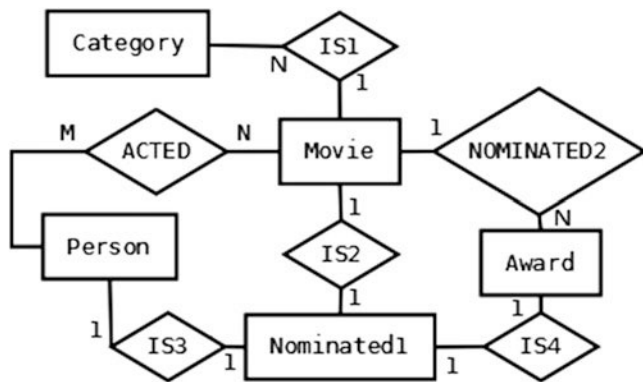


Fig. 13.6 *Movies*—method M05—end of phase 1—ER2

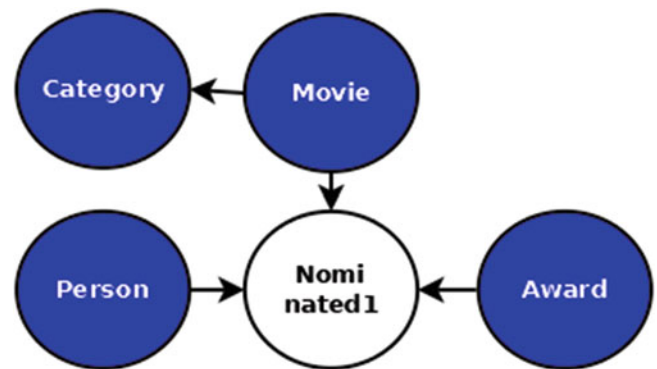


Fig. 13.7 *Movies*—method M05—phase 2—converts entities to nodes and relationships to edges

The IS1 relationship, degree 2, is of type $m : n$, so the relationships’ incorporation degree is equal to 0 and must be left as it is.

Finally, the FROM relationship, degree 2, has an incorporation degree equal to 1, resulting from the Country entity, which shall then be incorporated by the Movie entity. There are now only relationships of degree 2 and incorporation degree equal to 0 in the entity-relationship model. Thus, phase 1 is complete. This model can be seen in Fig. 13.6.

In phase 2, the first step is to convert the entities into nodes and the relationships into edges, as shown in Fig. 13.7.

After that, the process is finished, and the graph is generated. Step 3 of the method (the participatory action) provides the possibility of dividing nodes with more than two proper-

ties into two nodes, one with the node’s identifying properties and the second with the other ones.

In this database, the attribute *Duration* of the node *Movie* could be moved to a second node named *OtherMovie*, with the same key as the *Movie* node and this attribute. Likewise, the attribute *Birth* of the node *Person* could be moved to the new node (*OtherPerson*). Therefore, it is up to the database designer to decide whether this step is necessary or not. For this experiment, the nodes were not divided.

The final result of the modeling can be seen in Fig. 13.8, where there are also the final models generated by methods M01, M02, M03, and M04, whose databases were created for the comparative experiments of this article.

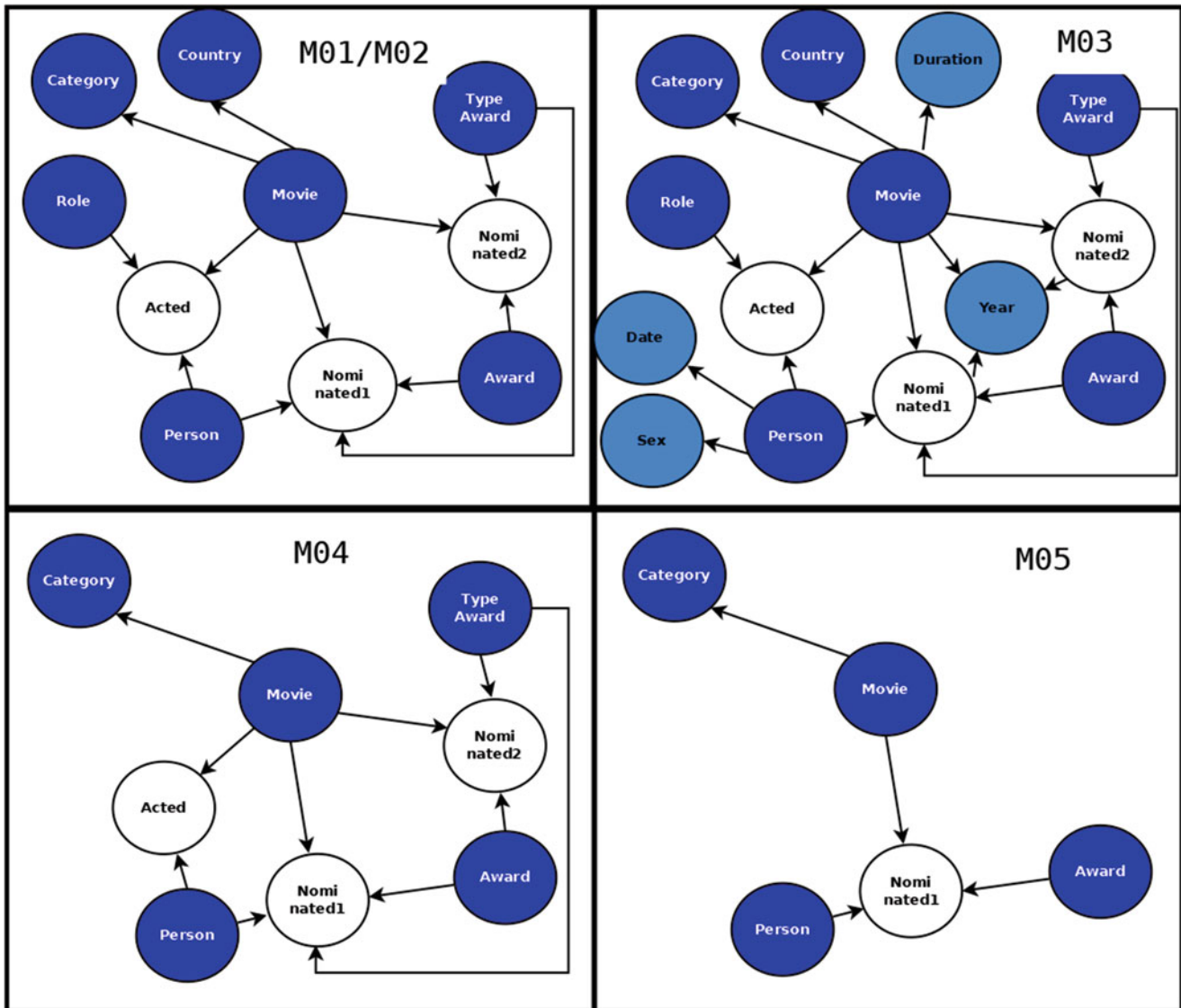


Fig. 13.8 *Movies*—final result of the five methods

We implemented four databases. We implemented the first according to *M01* and *M02* methods since the result is the same. To the second database, we implemented it according to the method *M03*. Regarding the third database, we followed the *M04* method. Finally, we implemented the fourth database according to the Participatory Modeling, referred to as the method *M05*.

Ten different queries were performed (Table 13.3) to explore the several generated models' distinct characteristics. Each query was performed twelve times, discarding the times of the first two queries, for they were considered a warm-up for each database [20]. The queries were sorted according to the Participatory Modeling method's execution time in an ascending way for comparison with the other methods.

The graphs are presented with the average execution time of each query by method (Fig. 13.9) and the number of accesses to the nodes and edges covered by the query (Fig. 13.10).

The Participatory Modeling method was faster than the other methods in almost all queries, except for the query C3 in analyzing each query's average execution times.

In part, the gain in time in the Participatory Modeling method is due to the smaller number of accesses to the nodes and the edges (Fig. 13.10) during the query, compared to other methods.

The elimination of nodes and relationships by the proposed method justifies the smaller accesses in the queries since the path in the graph is smaller. However, the result of

Table 13.3 *Movies*—queries performed

	Entities	Objective
1	Person, movie	Shortest path (Fernanda Torres, Charles Chaplin)
2	Person, movie	All shortest path (Fernanda Torres, Keanu Reeves)
3	Movie, TypeAward, nominated2	Shorter movies that won the oscar
4	Person, movie, nominated1, TypeAward	Persons with the most awards
5	Person, acting, role, movie	Persons with the most performances
6	Person, movie, nominated1, nominated2, award, category	Persons with the most nominations without awards
7	Person, movie, nominated1, nominated2, award, category	Persons with the most nominations
8	Movie, country, nominated1, nominated2, award	Persons with most acting as director
9	Movie, award, nominated1	Most nominated films that were awarded in all nominations
10	Person, acting, movie, nominated1, TypeAward	People with better nominations as best actor/atress

Fig. 13.9 *Movies*—average query execution time

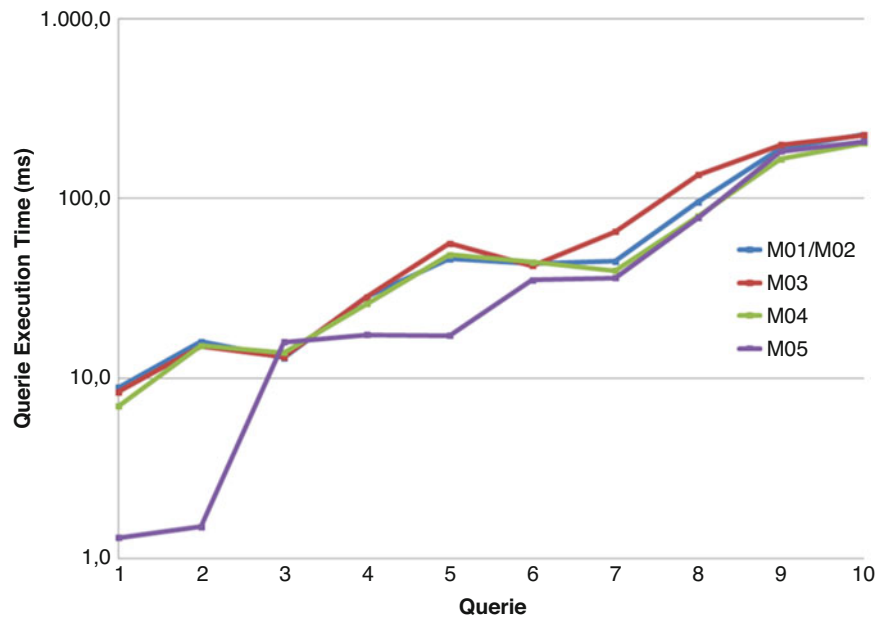
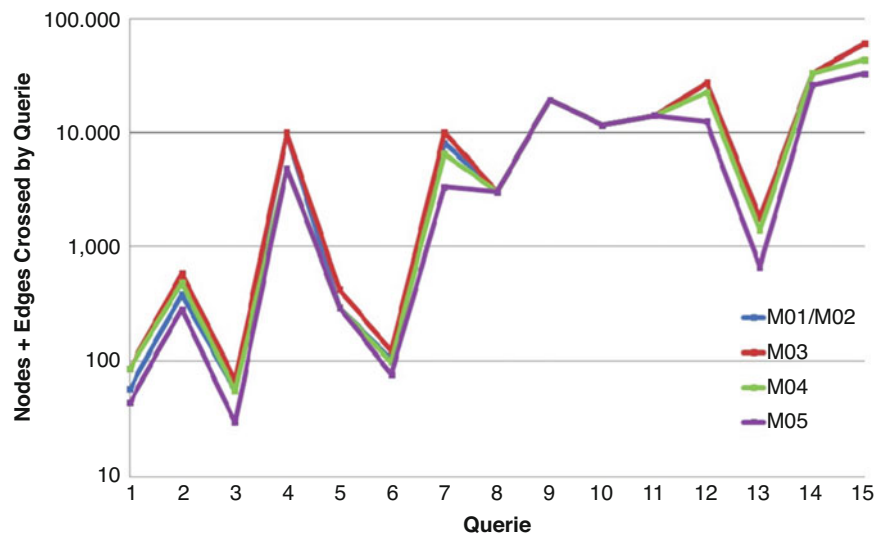


Fig. 13.10 *Movies*—nodes crossed by query



query C3, where the path is taken in the proposed method was shorter than in the others, but the time was longer, indicates that we should investigate other factors. Perhaps the results

were affected by the number of attributes or the syntactic structure of the query. We will examine these conjectures in future works.

13.5 Conclusion

This article proposes a new modeling method built from the analysis of the best characteristics of four other methods concerning the performance in retrieving information from the graph-oriented database. The main innovation of the proposed method is the conversion of relationships of degrees higher than 2, on which all other methods, like the relational model, create an intermediate node with edges between the involved entities. Besides, in the proposed method, one or more entities can be removed from the model. Its attributes are incorporated as relationship attributes without creating an intermediate node. Thus, the path to be taken when consulting the graph is reduced.

Another particularity of the proposed method is the designer's useful participation, instead of the direct automation through mathematical formulas, in the conversion, since the database designer knows the business data and, based on them, can influence the modeling process.

The conducted queries proved that the unification of nodes, with the consequent elimination of edges between them, brings a better performance in queries that return properties that would previously be distributed in different nodes. However, the unification makes performance worse if the query return belongs to the properties of only one of the nodes. What can be concluded is that none of the methods will always be more or less suitable than other ones. Therefore, the method must be chosen according to the application's characteristics that manipulate the database to be modeled.

References

1. J. Ribeiro, J. Henrique, R. Ribeiro, R. Neto, NoSQL vs. relational database: A comparative study about the generation of the most frequent N-grams, in *2017 4th International Conference on Systems and Informatics (ICSAI)* (IEEE, Piscataway, 2017), pp. 1568–1572
2. Kühne-Gasse. DB-Engines Ranking (2020). <https://db-engines.com/en/ranking>. Accessed 8 Sep 2020
3. R. Cattell, Scalable SQL and NoSQL data stores. *SIGMOD Record* **39**, 12–27, 12 (2010)
4. I. Robinson, J. Webber, E. Eifrem, *Graph Databases*. (O'Reilly Media, Sebastopol, 2013)
5. F. Ravat, J. Song, O. Teste, C. Trojahn, Efficient querying of multidimensional RDF data with aggregates: Comparing NoSQL, RDF and relational data stores. *Int. J. Inf. Manag.* **54**, 102089 (2020). Elsevier
6. S. Bordoloi, B. Kalita, ER model to an abstract mathematical model for database schema using reference graph. *Int. J. Eng. Res. Devel.* e-ISSN **6**, 51–60 (2013)
7. Y. Park, M. Shankar, B.-H. Park, J. Ghosh, Graph databases for large-scale healthcare systems: A framework for efficient data management and data services, in *2014 IEEE 30th International Conference on Data Engineering Workshops (ICDEW)* (IEEE, Piscataway, 2014), pp. 12–19
8. T.J. Lampoltshammer, S. Wiegand, Improving the computational performance of ontology-based classification using graph databases. *Remote Sens.* **7**(7), 9473–9491 (2015)
9. G.C. Galvão Van Erven. MDG-NoSQL: Modelo de Dados para Bancos NoSQL Baseados em Grafos (2015)
10. A. Lysenko, I.A. Roznovăț, M. Saqi, A. Mazein, C.J. Rawlings, C. Auffray, Representing and querying disease networks using graph databases. *BioData Mining* **9**(1), 23 (2016)
11. R. De Virgilio, A. Maccioni, R. Torlone, Model-driven design of graph databases, in *Conceptual Modeling* (Springer, Berlin, 2014), pp. 172–185
12. J. Pokorný, Conceptual and database modelling of graph databases, in *Proceedings of the 20th International Database Engineering & Applications Symposium (ACM, New York, 2016)*, pp. 370–377
13. N. Roy-Hubara, L. Rokach, B. Shapira, P. Shoval, Modeling graph database schema. *IT Professional* **19**(6), 34–43 (2017)
14. R. Angles, The property graph database model, in *AMW* (2018)
15. Neo4j (2020). <https://neo4j.com/>. Accessed 8 Sep 2020
16. S. Bordoloi, B. Kalita, Designing graph database models from existing relational databases. *Int. J. Comput. Appl.* **74**(1), pp. 25–31 (2013)
17. J.F. Sequeda, M. Arenas, D.P. Miranker, On directly mapping relational databases to RDF and OWL, in *Proceedings of the 21st international conference on World Wide Web* (ACM, New York, 2012), pp. 649–658
18. E.F. Codd, A relational model of data for large shared data banks. *Commun. ACM* **13**(6), 377–387 (1970)
19. Kaggle. Oscar nominations from 1927 to 2015 (2017). <https://www.kaggle.com/theacademy/academy-awards>. acessado em 08/setembro/2020
20. D. Gordon, Warm the cache to improve performance from cold start (2015). <https://neo4j.com/developer/kb/warm-the-cache-to-improve-performance-from-cold-start/>. Accessed 8 Sep 2020

Islam Akef Ebeid, John R. Talburt, and Md Abdus Salam Siddique

Abstract

Here we study the problem of matched record clustering in unsupervised entity resolution. We build upon a state-of-the-art probabilistic framework named the Data Washing Machine (DWM). We introduce a graph-based hierarchical 2-step record clustering method (GDWM) that first identifies large, connected components or, as we call them, soft clusters in the matched record pairs using a graph-based transitive closure. That is followed by breaking down the discovered soft clusters into more precise entity profiles in a hierarchical manner using an adapted graph-based modularity optimization method. Our approach provides several advantages over the original implementation of the DWM, mainly a significant speed-up, increased precision, and overall increased F1 scores. We demonstrate the efficacy of our approach using experiments on multiple synthetic datasets. Our results also provide some evidence of the utility of graph theory-based algorithms despite their sparsity in the literature on unsupervised entity resolution.

Keywords

Database · Data mining · Data cleaning · Data quality · Information quality · Graph theory · Graph clustering · Record linkage · Unsupervised entity resolution · Unsupervised data curation

14.1 Introduction

Entity resolution is crucial in data cleaning, curation, and integration [1]. It also refers to finding duplicate records within the same table, across various tables, or multiple databases that might refer to the same real-world entities. Traditional and supervised approaches in entity resolution rely on handcrafting rules for matching records. However, the move towards automating entity resolution for data cleaning, curation, and integration has become the goal for many organizations. Thus, unsupervised entity resolution methods have proliferated. However, unsupervised approaches suffer from higher inaccuracies than supervised ones due to inefficiencies, especially in the clustering step, where the goal is to discover unique entity profiles.

The base of any entity resolution system is string similarity or fuzzy matching. Traditional and supervised entity resolution relies heavily on human input to guide the entity matching process using predefined rules. Defining those rules depends on handcrafting simple lexical, semantic, and syntactic conditions for matching records based on attribute similarity like the OYSTER system described in [2]. On the other hand, unsupervised entity resolution approaches could be divided into probabilistic, machine learning, and graph-based methods. Some work has been done in probabilistic unsupervised entity resolution that relies on estimating statistical models [3–5]. However, probabilistic methods that rely more on natural language processing, fuzzy matching, and token frequency have proliferated more [6–8].

Probabilistic token-based methods in unsupervised entity resolution typically rely on an automated pipeline of preprocessing, blocking, matching, clustering, and canonicalization. Preprocessing refers to multiple steps that involve

I. A. Ebeid (✉) · J. R. Talburt · M. A. S. Siddique
 Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR, USA
 e-mail: iaebeid@ualr.edu; jrtalburt@ualr.edu; msiddique@ualr.edu

merging and parsing data files, tokenizing, and normalizing the unstandardized references. Blocking refers to the strategy used to mitigate the quadratic complexity of pairwise comparisons in unsupervised entity resolution. That strategy relies on quick and dirty techniques that divide the pre-processed unstandardized references into chunks or blocks, avoiding string matching across the whole dataset. Each block can be processed separately, where pairwise string similarity can be applied with less computational cost. The matched pairs from each processed block are then combined and clustered to infer the unique latent entity profiles represented in the data file. The clustering process aims to resolve any conflicts in pairwise matching and find records that indirectly match. Those conflicts typically occur due to the reliance on frequency-based blocking in unsupervised entity resolution systems [1]. In addition, many challenges riddle the unsupervised entity resolution process; the most critical are parameter tuning, finding appropriate thresholds for matching probabilities, and exact record clustering. Also, that process might need to be applied multiple times due to the low accuracy of relying on fuzzy matching alone without rules as in supervised methods.

14.1.1 Contribution

Here we study the problem of record clustering in probabilistic unsupervised entity resolution. We propose a graph-based 2-step hierarchical record clustering technique. We modify a state-of-the-art algorithm in unsupervised entity resolution named the Data Washing Machine (DWM) [1]. Integrating both techniques leverage the power and precision of graph theory methods and the speed and robustness of probabilistic token-based methods. Despite its many advantages, the DWM presents with some drawbacks and limitations. For example, the iterative approach used in the DWM relies on assessing the quality of the entity clusters using Shannon's entropy-based metric. The metric is difficult to interpret due to arbitrary threshold setting; hence it causes unnecessary iterations and higher computational costs.

More specifically, we modify the mechanism of iterating over a full, unsupervised entity resolution pipeline multiple times and replace it with a single shot graph-based approach that minimizes and localizes iterations to discover and optimize entity profiles. The optimization approach is done hierarchically and over two steps after remodeling the matched record pairs as edges in a record graph. The first step relies on iteratively discovering large connected components or soft clusters in the unweighted, undirected record graph after applying a threshold to prune the record graph using a graph-based transitive closure algorithm named CC-MR [9]. The second step relies on breaking down those large soft clusters characterized by a high recall into more precise entity

clusters also iteratively by adapting a graph-based clustering and community detection algorithm called Louvain's method [10]. The result is a significant speed-up when compared with the original implementation of the DWM. In addition to a significant increase in precision when applied to the synthetic benchmark data set described in [11]. Thus our contributions are summarized as:

- Introducing Graph-based Hierarchical Record Clustering (GDWM) by adapting Louvain's method to a graph-based transitive closure algorithm named CC-MR [9] used in the original DWM.
- Contributing to the literature and methods in graph-based unsupervised entity resolution and showing that a hybrid method, combining both probabilistic unsupervised entity resolution pipelines with graph theory approaches, provides better results than using graph-based approaches alone [12] or probabilistic token-based approaches alone [1].
- Performing various experiments on a sampled synthetic benchmark dataset [11] with known ground truths to demonstrate our approach's effectiveness, accuracy, and robustness.
- The code and datasets for our experiments are accessible at <https://bitbucket.org/oysterer/dwm-graph/src/master/GDWM18/>

The rest of the paper is organized as follows. In Sect. 14.2, we describe previous work that has been done in graph-based approaches in entity resolution. In Sect. 14.3, we describe both the algorithm and the method proposed. Section 14.4 describes our experiments on the proposed approach using a synthetic dataset in various configurations and settings. In Sect. 14.5, we discuss the results and the implications of the proposed approach. Finally, we conclude in Sect. 14.6.

14.2 Related Work

Here we review the literature on graph-based methods in entity resolution. Despite their sparsity in the literature on unsupervised entity resolution, graph-based methods and algorithms have been adapted before to the problem of entity resolution.

14.2.1 Token-Based Graph Entity Resolution

In token-based graph entity resolution, the goal is to construct a bipartite undirected graph of token nodes and record nodes and cluster the record nodes into unique entity profiles using methods such as SimRank [13]. In [14], the authors introduced a graph-based entity resolution model. The model

transformed the input data set into a graph of unique tokens where connectivity reflects the co-appearance of tokens in references. The graph was clustered using a weight-based algorithm that considered three types of vertices: exemplar, core, and support vertices. The algorithm then constructed r radius maximal subgraphs from the original token graph to discover clusters related to unique entities. Token-based methods, however, are computationally expensive and memory intensive due to the lack of an integrated blocking strategy.

14.2.2 Record-Record Similarity-Based Graph Entity Resolution

Record-record similarity graphs link structured unstandardized references in a weighted undirected graph where the nodes represent unique records. The connectivity represents the degree of similarity between individual references. That approach of constructing a record graph allows to directly utilize a whole set of graph clustering algorithms that graph theory and network science researchers have already developed. While [15] applied a graph clustering algorithm to optimize minimal cliques in the graph by approximating the NP-hard graph clique problem through pruning. Moreover, [16] developed the FAMER framework to combine multi-source data using blocking, matching, and clustering schemes. The framework modeled the merged data as a similarity record graph and then leveraged graph clustering techniques to resolve the entities.

Other work has leveraged the graph's structure instead of just the weights between records. In [17], the authors proposed three algorithms to cluster the similarity graph based on structure rather than edge weights. They argue that graph-based transitive closure, such as in [9], produces high recall but low precision because the graph's structure is not considered during clustering. They justified using maximal clique algorithms to leverage the graph's structure, which increases precision. There are also centrality and node importance-based methods where the edge weights are not considered, and node scores are propagated, such as in [18]. In addition, the authors introduced the notion of a node resistance score in a co-authorship graph to model entity similarity. Node resistance can be considered a PageRank score [19], where a random walker computes the probability of walking from the source node to the target node iteratively until convergence. Also, in [20], the authors introduced a graph-based model that linked two graph datasets by aggregating similarity scores from neighboring record nodes. Despite their accuracy, record-record similarity methods are complicated and require extensive graph theory knowledge to tune the adapted methods.

14.2.3 Hybrid Graph-Based Entity Resolution

Hybrid methods that combine token-based bipartite graphs and similarity-based record-record graphs have been investigated in [12] and [21]. The authors proposed an algorithm that combines text similarity with a graph-based algorithm. They first partition the data into a bipartite graph of record pair nodes and frequent term nodes to learn a similarity score of the record pair nodes. Then, the result was used to construct a record-record graph and used to power the CliqueRank algorithm, which runs on the blocks of records identified by the first part, known as the ITER algorithm. The probability of a matching pair of records is then updated iteratively. The authors combined two distinct methods: the random walk-based approach and the graph clustering-based approach. However, the output of their approach is matched pairs of records without introducing any clustering approach that would resolve the entity profiles.

14.3 Method

14.3.1 Framework

In this section we describe in detail the proposed framework (GDWM) as shown in Fig. 14.1.

14.3.1.1 Merged Data File

Merging data files is a crucial step in entity resolution. Data files usually come in multiple files containing unstandardized duplicate records. Those files are generally of different sizes and layouts, making it challenging for entity resolution algorithms. Merging data files is sometimes done using a single layout and sometimes by combining the files without knowing the specific layout of the metadata. A standard layout means that the data is organized in the same fields. In our case, we merge all incoming data files into one file without relying on a specific layout where each row has a unique identifier, and the unique entity profiles are unknown. More formally, a record set R consists of tuples $r_i = (u_i, f_i)$ where $u_i \in U$ is a unique identifier and $f_i \in F$ is a set of attributes with unknown headers.

14.3.1.2 Parsing

The parsing process includes preprocessing and tokenization and is central to our unsupervised entity resolution approach. Tokenization is preceded by a preprocessing step where the text from all the provided attributes and fields is combined into one string and normalized to the upper case. That is followed by removing special characters from the text. We use the standard approach to tokenizing text in English, splitting the text based on spaces between tokens. The parsed

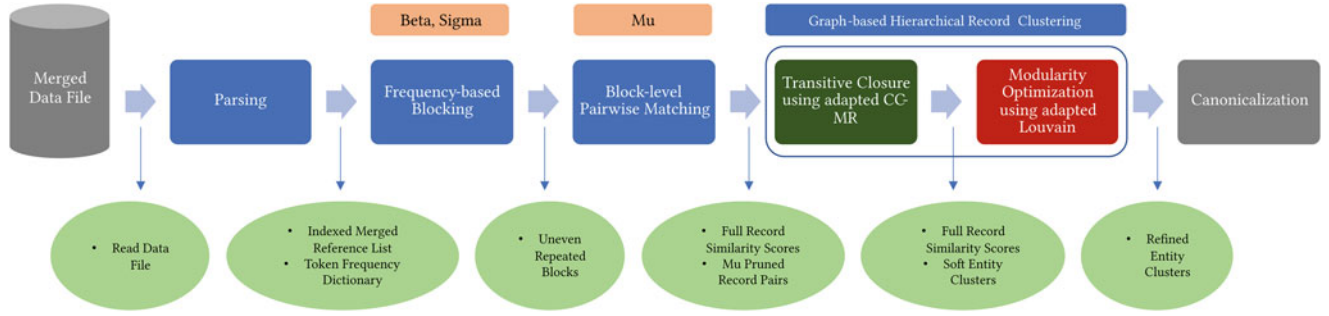


Fig. 14.1 The GDWM framework

data are then loaded into memory, and a dictionary of unique token frequencies is computed. More formally, the processed data file contain a unique distinct set of tokens T with different counts C given that t_i is a distinct unique token where $t_i \in T \subseteq f_i$ and c_i is the corresponding count of each unique token t_i where $c_i \in C$ and $C = |t_i \in T|$.

14.3.1.3 Blocking

The goal of the blocking process is to reduce the potential number of pairwise string similarity operations across a data file as much as possible. Blocking usually rely on inexpensive techniques to quickly filter out records in the data file with a very low probability of matching. Frequency-based blocking typically assumes that two records that refer to the same entity should share at least one token [1]. Other blocking techniques include methods such as Locality-Sensitive Hashing [22]. Here we adapt the frequency-based blocking technique presented in the DWM [1]. The blocking method relies on two parameters, beta β , and sigma σ . First each reference $r_i \in R$ is processed using σ to filter out tokens $t_i \in r_i$ with frequency $c_i \in C$ and $C = |t_i \in T|$ above σ . That process could be seen as a stop word removal step where tokens that appear too much are assumed to be non-discriminative and unimportant. Following for each reference token $t_i \in r_i$ with frequency $c_i \in C$ and $C = |t_i \in T|$ below β is considered a blocking token $t_{Bi} \in T$. Blocking tokens are identified for each reference in a list L where the filtered records are repeated. The list is then grouped by blocking tokens regardless of reference. Each block will include all the references where the same blocking token appeared, and the number of blocks will be equivalent to the number of unique blocking tokens in the dataset. This process is formalized in Algorithm 2.

14.3.1.4 Pairwise Matching

After blocks have been identified, the next step is pairwise matching between references in each block. Even though any string similarity metric can be used on the blocked references to identify matches, we choose to use a string similarity algorithm named the Scoring Matrix used in [1]

Algorithm 1 Blocking algorithm

Input : $C(T)$ unique token frequencies, R record set, β , σ

Result: B list of blocks

```

B ← list
for r ∈ R do
  if c_i ∈ C ∀ t_ij ∈ r | t_i ∈ T ≥ σ then
    | r ← remove t_ij from r
  end
  if c_i ∈ C ∀ t_ij ∈ r | t_i ∈ T ≤ β then
    | L ← (t_ij, r)
  end
end
L_s ← sort(L, t_i ∈ T)
T_u ← unique(t_ij ∈ L_s) for t_i ∈ T_u do
  for l_s ∈ L_s do
    if t_i = t_j ∈ l_s then
      | b_i ← r ∈ l_s
      B ← B + b_i
    end
  end
end
return B

```

and developed in [7] and tested in [23] as an adapted Monge-Elkan [24] algorithm. The Scoring Matrix method relies on lining up stop-word-filtered tokens of both references of the two records at hand as rows and columns of a matrix in order of appearance. The cell values of the matrix are the normalized Levenshtein's Edit Distance [25] between each token in the first and the second records. After computing the Levenshtein distance between each token, the algorithm then normalizes the edit distance between all the rows and columns in the matrix that are non-zero, dividing by the maximum length between the two tokens. The algorithm then loops on all the rows and columns and averages all the maximum normalized Levenshtein distances in the matrix for each row token or column token, and that becomes the final normalized similarity score between the two references. We will not formalize the Scoring Matrix algorithm here, as more details can be found in [7].

14.3.1.5 Graph-Based Hierarchical Record Clustering (GDWM)

As mentioned before, entity resolution aims to find the unique latent entity profiles in a dataset that multiple records might represent. The blocking algorithm considers two similar records having a high probability of representing the

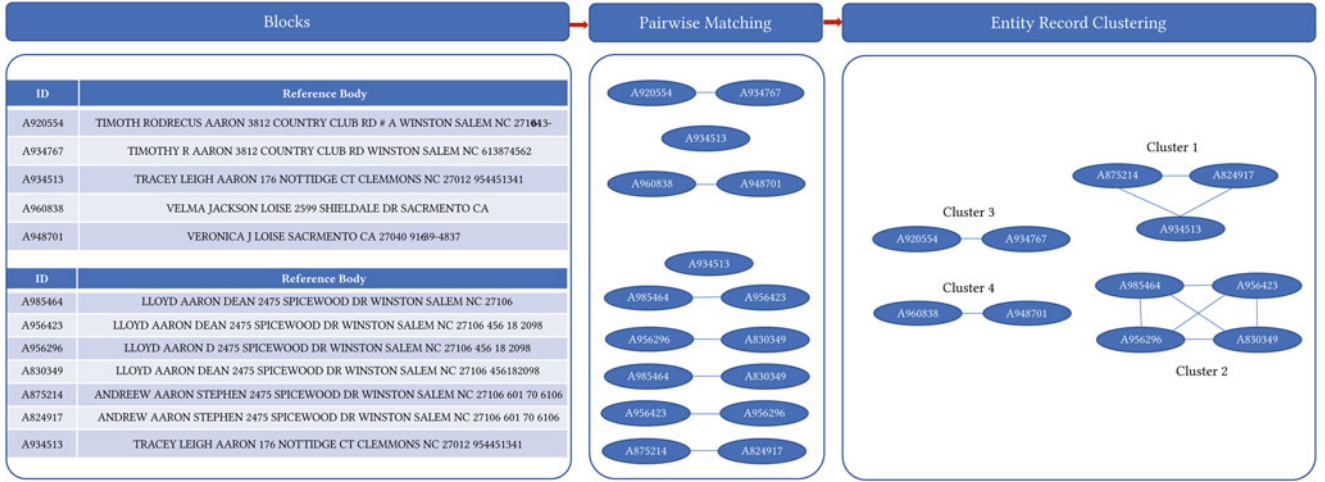


Fig. 14.2 The typical unsupervised entity resolution approach

same entity if they have a minimum of one token in common. That increases the number of records representing one entity, as shown in Fig. 14.2. As a result, a record may appear in more than one entity, making the process of clustering more expensive and the underlying graph of matched records more complex. In addition, the matching process also implies that one record might be similar in an indirect way to another record by an algebraic property known as transitivity [26]. In a simple sense, if record a matches record b and record b matches record c , then by transitivity record a matches record c . Hence transitive closure is a widely used approach to finding binary relations in ordered sets of pairs. When adapted to entity resolution, it also considers matched records as an edge list in a graph.

Hence clustering can be formalized by reformulating the problem as a graph where $G = V, E$. A vertex/node $v \in V$ represents one record $r \in R$ where $r \equiv v$ while an edge $e \in E | E \subseteq V \times V$ represents whether two records $e_i = (r_i, r_j)$ are matched and an edge weight $w_e \in W(E) : E \rightarrow \mathbb{R}$ represents the normalized similarity between the two records calculated in the pairwise matching step using the Scoring Matrix algorithm [7]. We also define a set of clusters $Q \subseteq P(V)$ where elements of Q are a subset of V and P is a partition of V . The clustered graph conventionally can be seen as a graph of subgraphs where each meta-vertex represents each subgraph of vertices or records. Hence this hierarchical relationship can be formalized as follows in Eqs. (14.1) and (14.2) and also fits in our framework as illustrated in Fig. 14.3:

$$V' = Q \quad (14.1)$$

$$E' = (Q_i, Q_j) : \exists (v_i, v_j) \mid \{v_i \in Q_i, v_j \in Q_j, (v_i, v_j) \in E\} \quad (14.2)$$

Here we view the post pairwise matching entity clustering process as a two-step process. We formulate the problem as a graph and apply a hierarchical graph clustering approach to discover the latent entity profiles. Soft clusters are discovered using the efficient graph-based transitive closure algorithm utilized in the DWM [1] and adapted from [9] called CC-MR. The input to CC-MR is a pruned unweighted undirected graph of matched pairs. Next, we convert each soft cluster into a complete undirected weighted graph and prune it from the previously computed similarity scores between matched pairs. Finally, a Modularity optimization algorithm adapted from [10] is used to hierarchically discover and refine the original soft clusters increasing the precision of the overall clustering process while eliminating the need for reiterations compared with the DWM framework. Our 2 step hierarchical method is described formally in Algorithm 2.

Transitive Closure Using Adapted CC-MR

As mentioned before, the ordered list of matched records can be considered an edge list or an adjacency matrix representing a graph of records. Following a simple graph clustering approach to finding the strongly connected components in the graph is used to perform transitive closure as the first step in our 2 step clustering method. CC-MR was introduced in [9] and used in [1] as a single-step clustering method for the DWM. CC-MR starts by creating star subgraphs from the matched pair list. That is done by simply grouping the pair list by the smallest node. Then, the algorithm iteratively checks whether its vertices are assigned to subgraph components where the center of the subgraph component is the smallest vertex in its first-order neighborhood, relying on the MapReduce framework for scalability. The DWM provides an efficient implementation of the algorithm without relying on MapReduce. The algorithm is formalized and described

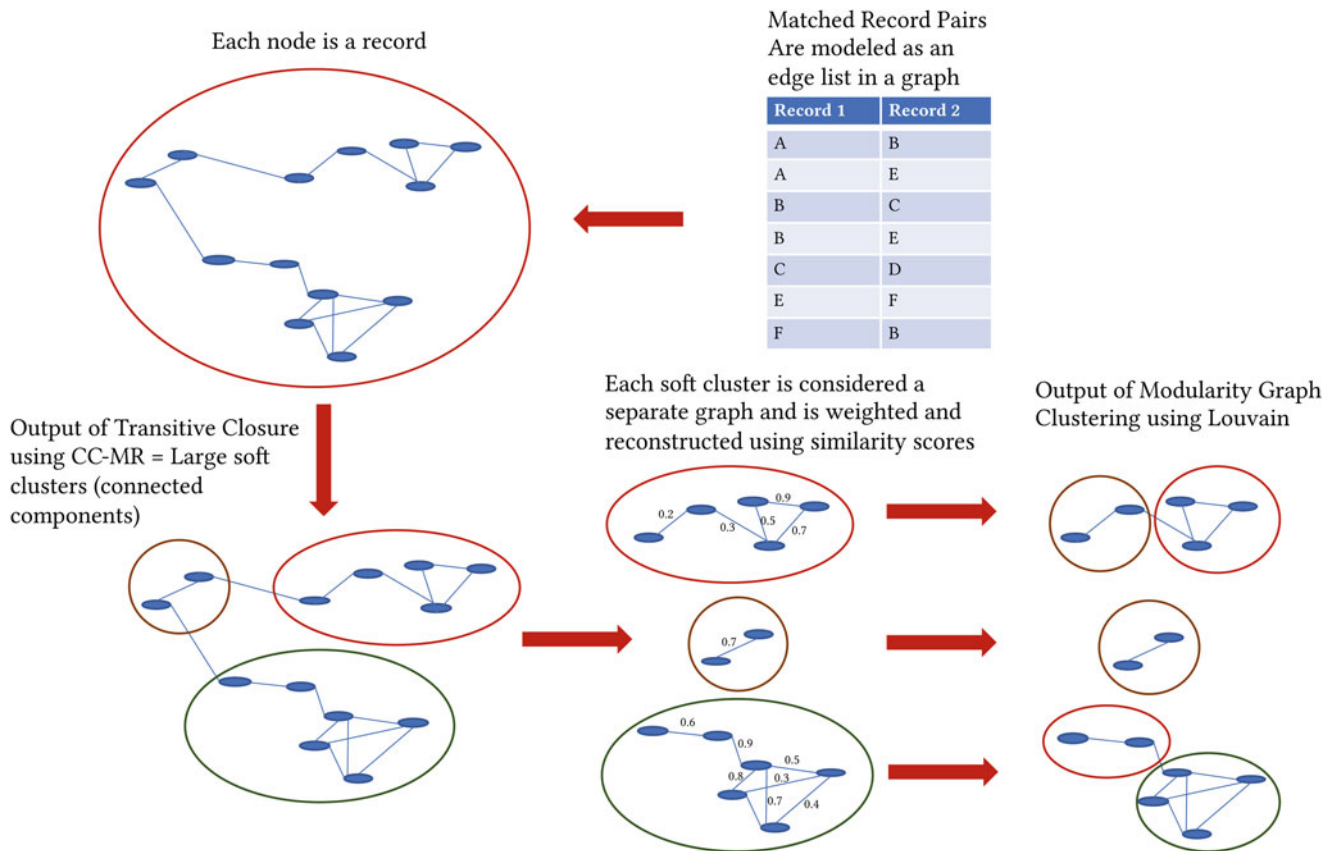


Fig. 14.3 The matched pairs with their similarity scores are considered an edge list of a graph. Then they are pruned using a threshold μ . Kolb et al. [9] is applied to discover largely connected components or soft

clusters. Blondel et al. [10] is applied to each graph, and final resolved entity clusters are achieved

thoroughly in [9]. Also, we provide a short pseudo-formal description of the implementation we used in Algorithm 4.

Modularity Optimization Using Adapted Louvain

Next, we pass the soft clusters produced by the CC-MR transitive closure in addition to the previously computed similarities between records on the block level to our Modularity optimization algorithm adapted from the Louvain method described in [10]. In the original DWM implementation after transitive closure, an evaluation step is applied where each cluster is evaluated using Shannon's entropy-based algorithm. Then, a normalized quality metric is computed for each cluster, and a threshold epsilon ϵ is applied to decide whether the cluster shall be reiterated or canonicalized. The clusters with qualities lower than ϵ are reiterated starting from blocking, where the similarity threshold parameter μ is incremented, and the lower quality clusters are further broken down. That process carries on until μ reaches one or no clusters left below the quality threshold.

Multiple issues emerge with that approach. The most salient is that trying to tune the similarity parameter μ and

the quality threshold ϵ is tricky, as shown later in Sect. 14.4.3. In addition, clusters that are good enough but their quality have been slightly below the threshold might stay in the pipeline for 2 or 3 iterations increasing the computational cost unnecessarily. It is worth mentioning that these issues are addressed in a later version of the DWM using non-graph-based approaches. Nevertheless, in the current version that we are working with, those are the issues that motivated us to redesign the entity clustering process allowing us to eliminate the need for a quality metric and reiterations. Furthermore, in our Modularity based hierarchical record graph clustering approach (GDWM), the adapted Modularity-based algorithm is a second step that involves a minimal number of iterations to optimize the Modularity metric over a record-record similarity graph, which does not require threshold setting.

We model each soft cluster coming out of the transitive closure process of the DWM as a complete graph. Then, after pruning, the graph of records is further broken down and clustered by optimizing a graph clustering quality heuristic known as Modularity [27]. Modularity is a heuristic often used in graph theory [28] to measure the strength of clusters

Algorithm 2 Graph-based hierarchical record clustering

Input : S, μ : Pairs of matched records and their computed similarity scores, μ similarity threshold
Result : Q : list of clusters indexed by the least record node in each cluster

$P \leftarrow TransitiveClosureUsingAdaptedCCMR(S, \mu)$
returns connected components as soft clusters see algorithm 3;
 $P_s \leftarrow sort(P, r \in P)$
 $q_c \leftarrow unique(r \in P_s)$

```

for  $c \in q_c$  do
  for  $p(c) \in P_s$  do
    for  $n1 \in p(c)$  do
      for  $n2 \in p(c)$  do
        if  $n1 \neq n2$  and  $S(n1, n2)$  exists or  $S(n2, n1)$  exists
        then
           $E' \leftarrow append(n1, n2, s \in S)$ 
        end
      end
    end
  end
end
 $V' \leftarrow initialize\ node\ clusters\ randomly$ 
 $V' \leftarrow AdaptedLouvain(G(V', E'))$  see Algorithm 4
 $V' \leftarrow append\ single\ nodes\ as\ separate\ clusters$ 
 $M \leftarrow ComputeModularity(V', E')$ 
if  $M \geq 0$  then
   $Q \leftarrow rename\ cluster\ IDs\ to\ least\ record\ in\ cluster\ V'$  and
  append
end
if  $M \leq 0$  then
   $Q \leftarrow assign\ records\ in\ V'$  to one cluster and append
end
end
return  $Q$ 

```

Algorithm 3 Transitive closure using adapted CC-MR

Input : S, μ : Pairs of matched records and their computed similarity scores, μ similarity threshold
Result : P : List of soft clusters as pairs indexed by the least record in the cluster
 $P \leftarrow list\ of\ soft\ clusters$ $R_p \leftarrow initialize\ list\ of\ tuples$ **for** $s \in S$ **do**

```

  if  $s_j \geq \mu$  then
     $R_p \leftarrow append(s_j)$ 
  end
end
 $R_p \leftarrow sort\ by\ first\ element\ in\ pair$  while no convergence do
  for  $(r_i, r_j) \in R_p$  do
     $P \leftarrow append\ pair\ where\ all\ record\ nodes\ belong\ to\ the\ connected\ component$ 
    with the smallest record
  end
end
return  $P$ 

```

discovered in a graph. More formally, we combine equations (14.1) and (14.2), then reformulate each cluster Q as a separate weighted undirected graph where $G' = (V', E')$. Next, the graph is pruned to include only edges available in the dictionary of matched records created in the pairwise matching step, where the edge weights are the precomputed similarity scores. The goal in this second step is to further break down the soft clusters in a hierarchical manner by optimizing through maximizing Modularity M using the adapted Louvain's method [10]. The assumption here is that clusters in a graph are often defined by the density of edges between a set of nodes. Hence, the number of connections or edges between clusters characterized by high Modularity is often very low. Modularity is defined in the equation

below as M . It measures the difference between the actual number of edges and an expected number of edges between nodes. An expected number of edges between nodes can be considered a random rewiring of the graph given the same nodes. Optimizing Modularity through maximization in a graph is a hard problem and is often tackled through various ways to reduce the number of comparisons between all nodes in a graph.

$$M = \frac{1}{2n} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2n} \right] \partial(Q_i, Q_j) \quad (14.3)$$

In the equation above, n is the number of edges in the graph, and k is the degree of a node. In addition $A_{i,j}$ is a weighted adjacency matrix constructed from E' . In addition, i and j are indices for each unique record in the file, represented as a vertex in the graph as $v' \in V'$. And $e' \in E' | E' \subseteq V' \times V'$ and $e' \equiv A_{i,j}$.

Louvain's method optimizes Modularity by efficiently measuring the difference in Modularities between different configurations of node clustering instead of rewiring the network using a simple equation and procedure [10]. The algorithm starts by randomly assigning each node in the graph to a unique random cluster. Next, a random node is picked and placed in the same cluster as its randomly chosen one of its closest neighbors. Then Modularity difference is measured between when the node was in its cluster and when the node is in the current neighbors' cluster. If the change in Modularity was positive, then the node is assigned to its new cluster with its neighboring node. If the delta Modularity is negative, the algorithm iterates and places the node in each cluster in the graph until a positive or zero deltas is achieved. If the delta Modularity was zero, the node remains in its cluster. The second phase consolidates the discovered clusters into a new graph of nodes representing whole clusters summing up all involved edge weights. Then the process reiterates on the newly reconstructed graph until no delta Modularity appears. The final clusters from this step can be canonicalized as entity profiles directly. This process is described formally in Algorithm 4.

14.3.1.6 Canonicalization

The clusters representing unique entity profiles are defined and persisted to the disk in this final step through a link index file. The link index file describes the final entity clusters of the unsupervised entity resolution framework as a list of ordered pairs where the first element of each pair represents the entity profile identifier the record is assigned to. The entity profile identifier is simply the least record identifier in the cluster. The second element of each pair represents the unique record identifier.

Algorithm 4 Adapted Louvain

Input : $G = (V', E')$: dictionary of nodes with randomly initialized clusters, list of pruned weighted edges as a tuples
Result: V' : dictionary of nodes reassigned to new clusters

```

while True do
   $M_{initial} \leftarrow \text{ComputeModularity}(V', E')$  for  $v_1 \in V'$  do
    for  $v_2 \in V'$  do
      update cluster ID of  $v_1$  to match cluster ID of  $v_2$   $\Delta M \leftarrow$  compute difference in Modularity M according to Louvain's reduced equation
       $L_M \leftarrow (v_2, \Delta M)$ 
    end
     $c_i \leftarrow \arg \max (L_M)$   $v_2 \leftarrow v_1$ 
    assign the current node to the maximum cluster update  $V'$ 
  end
   $M_{current} \leftarrow \text{ComputeModularity}(V', E')$  if  $M_{current} \geq M_{initial}$  then
     $G \leftarrow$  whole clusters become new nodes  $V'_{reconstructed}$ , the new edges weights are summation of weights between clusters  $E'_{reconstructed}$ 
  end
  else
    return  $V'$ 
  end
end

```

14.4 Experiments

14.4.1 Dataset

Our graph-based record clustering method (GDWM) is tested on a synthetic benchmark dataset. We use the synthetic dataset described in [11], a simulator-based data generator that uses probabilistic approaches to generate coherent individual data for persons that do not exist except for S3, which represents generated addresses and names of restaurants that do not exist. The data fields are names, addresses, social security numbers, credit card numbers, and phone numbers mixed in several layout configurations. Some samples are labeled as mixed layout, meaning that each row might come with a different order of those attributes and might not be delimited. The standard label means that all the rows in the data file have the same order and attributes. The generator described in [29] used a probabilistic error model to inject various errors in the previously developed simulated dataset. For example, in this excerpt of a generated data file shown in Table 14.1, the first four records are almost identical except for that record A956296 has a missing last name and the format of the phone numbers or whether they exist at all, all are errors injected and generated on purpose. In addition, the last two records are almost identical except for the first name where an intentional error was introduced. A ground truth set recording the actual clusters of the simulated records is then sampled from the generated synthetic database. Next, the corresponding references are pulled from the generated synthetic database to create various sample files with different sizes, levels of quality, and layouts. Sizes of files can vary from 50 to 20K rows, as shown in Table 14.2.

Table 14.1 An excerpt from the synthetic dataset used in the evaluation

ID	Reference body
A985464	LLOYD AARON DEAN 2475 SPICEWOOD DR WINSTON SALEM NC 27106
A956423	LLOYD AARON DEAN 2475 SPICEWOOD DR WINSTON SALEM NC 27106 456 18 2098
A956296	LLOYD AARON D 2475 SPICEWOOD DR WINSTON SALEM NC 27106 456 18 2098
A830349	LLOYD AARON DEAN 2475 SPICEWOOD DR WINSTON SALEM NC 27106 456182098
A875214	ANDREEW AARON STEPHEN 2475 SPICEWOOD DR WINSTON SALEM NC 27106 601 70 6106
A824917	ANDREW AARON STEPHEN 2475 SPICEWOOD DR WINSTON SALEM NC 27106 601 70 6106

Table 14.2 Below is a table describing the statistics of the dataset we are using. A good quality sample indicates that the sample had a limited injected number of errors, while a poor quality sample indicates that a high number of errors were injected

Sample name	Quality	Layout	Number of rows
S1	Good	Standard	50
S2	Good	Standard	100
S3	Moderate	Standard	868
S4	Good	Standard	1912
S5	Good	Standard	3004
S6	Moderate	Standard	19,998
S7	Good	Mixed	2912
S8	Poor	Standard	1000
S9	Poor	Standard	1000
S10	Poor	Mixed	2000
S11	Poor	Mixed	3999
S12	Poor	Mixed	6000
S13	Good	Mixed	2000
S14	Good	Mixed	5000
S15	Good	Mixed	10,000
S16	Poor	Mixed	2000
S17	Poor	Mixed	5000
S18	Poor	Mixed	10,000

14.4.2 Experimental Setup

Here we compare our hierarchical graph-based record clustering approach (GDWM) built using several modules from the DWM with the same version of the original implementation of the DWM. We set the blocking frequency β to 6, and the stop word threshold σ is set to 7 for all runs on all samples. Note that these values are fixed and were set based on experience with running both algorithms. They were picked by observing the execution time of each algorithm and the final F1 score, which will be discussed later. When running both algorithms on each of the 18 samples, we varied μ between 0.1 and 0.9. That gave nine runs per sample.

Table 14.3 Evaluation metrics and statistics

Statistic name	Symbol	Description
True positives	TP	Number of record pairs that appeared together in the same cluster correctly
True negatives	TN	Number of record pairs that did not appear in the same cluster correctly
False positives	FP	Number of record pairs that appeared together in the same cluster falsely
False negatives	FN	Number of record pairs that did not appear in the same cluster falsely
Precision	P	$TP / (TP + FP)$
Recall	R	$TP / (TP + FN)$
F1-score	F1	$2 \times P \cdot R / P + R$
Balanced accuracy	A	$TP(TN + FP) + TN(TP + FN) / (TP + FN)(TN + FP)$

When running the DWM, we varied μ and ϵ between 0.1 and 0.9. That gave us 81 runs per sample. Both the DWM and our GDWM are implemented in Python with the help of various libraries and packages such as NetworkX [30]. We ran the two setups on an Intel i7 Windows machine with 32 GBs of RAM. The ground truth is a list of each record and its membership cluster identifier. After canonicalization, the saved link index is grouped by the least record identifier in each profile. Thus, all records belonging to the same cluster will have the same record identifier as the first element in the link index pair. We then loop on each pair in the canonicalized link index and examine whether they belong together in the ground truth entity profile. Finally, we measure the following statistics against the ground truth for each sample run (Table 14.3).

14.4.3 Evaluation

Here we present the results after running the 18 samples on different similarity thresholds μ ; as mentioned before that the DWM ran for each μ level and each ϵ level nine times. In addition, the blocking frequency β and the stop word frequency σ are both fixed and set to 6 and 7, respectively. We then chose the ϵ value that yielded the best F1 score and lowest execution time in case of a tie.

For the GDWM, we ran on each sample 9 times, varying the similarity threshold μ from 0.1 to 0.9. We recorded both runtime statistics. For each sample run for each algorithm, we chose the lowest μ that yielded the highest F1-score, followed by the lowest execution time in case of a tie. We then averaged precision, recall, F1 scores, balanced accuracies, and execution times for both algorithms across the 18 samples. Table 14.4 demonstrates the results given each metric and statistic averaged across all 18 samples. For precision, the improvement appears to be the largest as our approach

Table 14.4 Average precision, recall, F1 scores, balanced accuracy, and execution time for each algorithm

Method	Precision	Recall	F1-score	Balanced accuracy	Time (seconds)
DWM (original)	0.7243	0.5782	0.6297	0.7689	203.1
GDWM (in this paper)	0.8478	0.6862	0.71479	0.8431	24.8

Bold values indicate that the results from running the proposed framework (GDWM) were better than the results from running the original implementation (DWM)

targets an increased precision relying on the fact that the first step of discovering large soft clusters using the graph-based transitive closure approach CC-MR plays the role of a staging step for the Modularity optimization algorithm through stabilizing the recall even at high μ levels. The next step, which involves Modularity optimization on each soft cluster using Louvain’s method, is responsible for the increased precision. That increased precision can be attributed to first modeling each set of the soft clustered records into a complete graph then weighting the graph using the previously computed similarity scores in the pairwise matching step. That approach automatically prunes the complete graph since not all the records have precomputed similarity scores. That approach also reduces the Modularity optimization done using Louvain and prevents any possible below zero final modularities that often happen when using Louvain’s methods, also known as the resolution limit problem [10]. Second, Louvain’s method is highly efficient when applied to large graphs due to its hierarchical approach in discovering clusters, and it fits nicely as a second step after transitive closure using CC-MR to provide a fully hierarchical graph clustering method. The hierarchical clustering approach also helps preserve the high recall of the entity clusters across the two steps. That is also shown in the increased balanced accuracy, which is a valuable measure for problems such as entity resolution that is usually characterized by having a class imbalance as the number of matched pairs are usually way less than the number of unmatched pairs causing a very high number of true negatives [31]. The significant speed-up achieved across all samples can also be attributed to using the reduced delta Modularity equation provided in the original Louvain’s method paper during optimization [10].

Figure 14.4 shows the interplay between μ , sample size, and F1 scores across the 18 sample run. It is worth mentioning that tuning ϵ with the similarity threshold μ can be difficult when running the original implementation of the DWM. Nevertheless, on a sample level and as shown here in Fig. 14.4, our approach, the GDWM, has outperformed the iterative approach used in the DWM, especially at μ levels where GDWM had the most significant overall F1

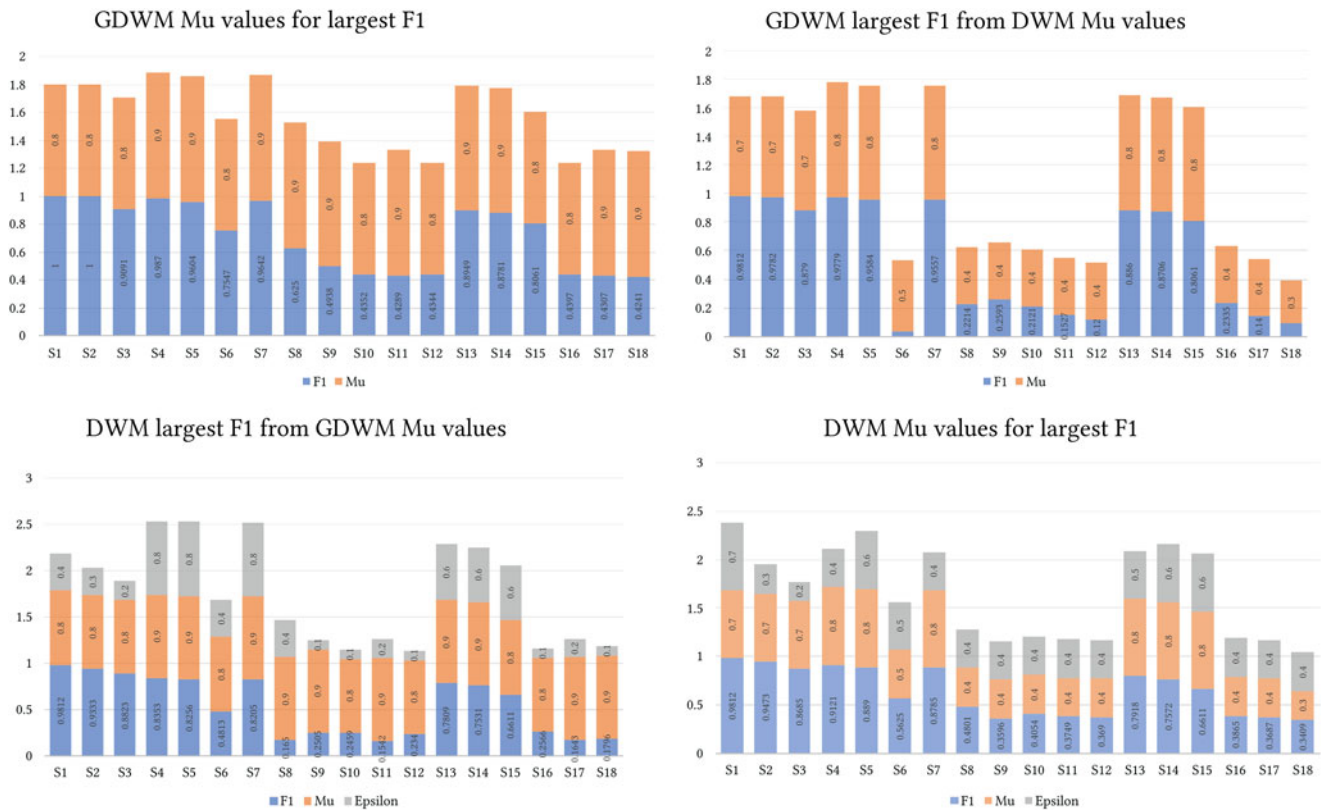


Fig. 14.4 Best F1 score for each sample with its corresponding Mu level for each algorithm

score. Those μ levels where GDWM outperformed DWM are also higher than the μ levels where DWM had the most significant F1 score on its own. However, at those levels, GDWM outperformance was not as consistent as with the original GDWM μ levels. For example, at μ level 0.3 on poor quality sample files such as S8, S9, S10, S11, S12, S16, S17, and S18, DWM slightly outperformed GDWM. That might be attributed to the existence of a mechanism to iterate and adjust quality levels ϵ in the DWM has made it perform better than GDWM in the case of poor data quality files at low levels of μ .

From the perspective of parameter sensitivity, at lower μ levels, recall appears to be more sensitive to changes than in higher μ levels. That can be seen in Fig. 14.5, showing 4 sample files with different qualities, sizes, and layouts. The figure demonstrates the sensitivity of the similarity threshold μ with the F1 score, precision, recall, and balanced accuracy. An observation is that in small data samples with good quality like S1 and S2, μ increases precision. Recall seems to be pretty stable across all μ values because it is more sensitive to blocking frequency β and block sizes in general. In large data samples such as S6 with 20K rows, there seems to be an onset for μ where μ before 0.5 does not affect the quality of the clusters produced by GDWM. While low-quality samples like S18 recall starts and continue to be very low at the β we set, affecting the overall performance and F1 scores.

14.5 Discussion

Our GDWM approach eliminates the need to reiterate over a big chunk of the remaining records that have not been well clustered, as is the case in the DWM. Our approach also eliminates the need for using a metric such as Shannon's entropy that has been utilized in the DWM to evaluate clustering quality. Instead, it uses a Modularity optimization method for graphs through an efficient implementation of Louvain that does not need threshold setting. Combining the graph-based transitive closure approach named CC-MR that has been used in the original implementation of the DWM which acts as a first step in identifying connected components or soft clusters by treating the identified matched pairs of records as an edge list of a graph with a Modularity based clustering approach that we introduced here through the implementation of Louvain's method, which acted as a second step in identifying and breaking down the original soft clusters, helped in quickly and early on breaking down any soft clusters that existed right after transitive closure hierarchically and efficiently.

Our approach, however, is not free of limitations despite its efficacy in identifying latent unique entity profiles. For example, our experiments are highly controlled and applied to a synthetic dataset on data for non-existent individuals;

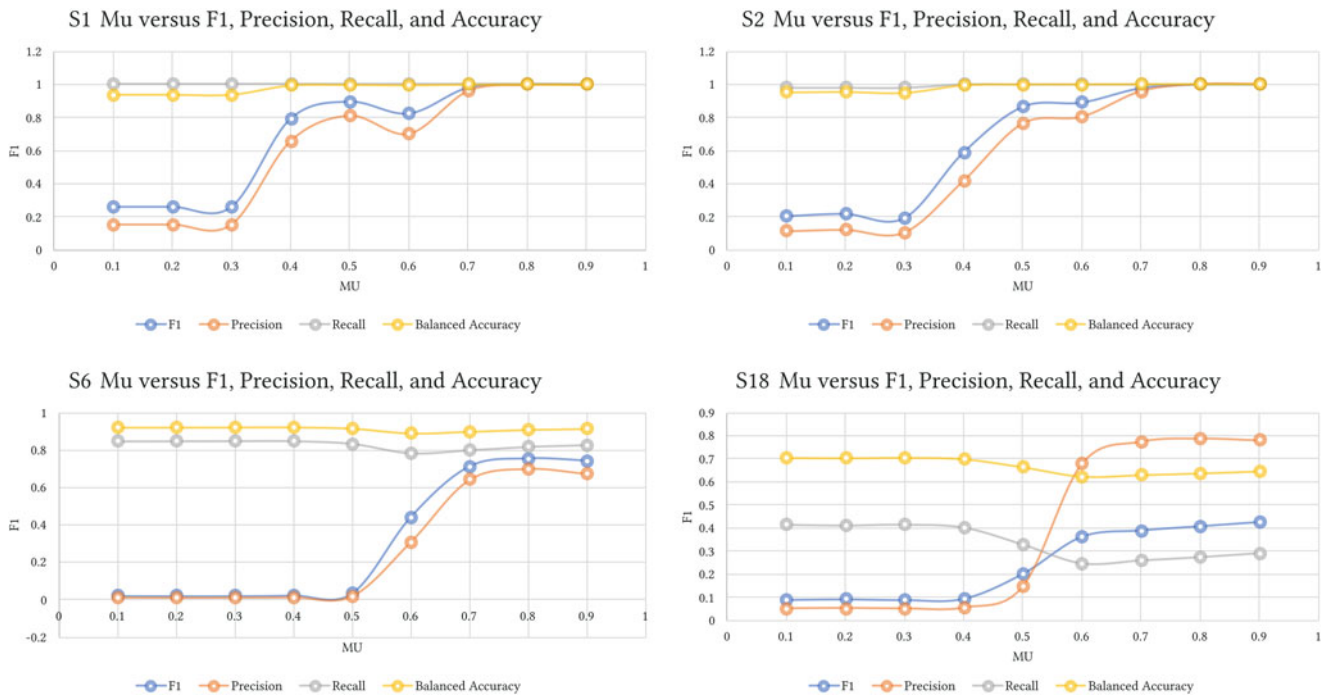


Fig. 14.5 Parameter sensitivity on four samples

despite the usefulness of the dataset, it is not diverse enough to explore the interplay between other parameters when applying our approach or any other unsupervised entity resolution approach for that matter. Therefore, we aim to diversify the benchmark datasets used in experimentation in the future. In addition, we plan to study the sensitivity of other crucial parameters like the blocking frequency β and the stop word frequency σ instead of fixing them in experiments. In addition, in the future, we aim at studying how many hierarchical levels of clusters are discovered using our approach and how that can inform the discovery of latent entity profiles using our graph method with fewer optimization iterations.

14.6 Conclusion

This paper demonstrated a graph-based hierarchical record clustering approach to unsupervised entity resolution named GDWM. Our method combines a graph-based transitive closure algorithm with an adapted Modularity-based graph clustering approach based on Louvain's method in community detection to provide a two-step, efficient hierarchical record clustering capable of producing exact F1 scores in some cases. We built on the work done by Talburt et al. [1], where the authors presented an innovative probabilistic self-assessing and iterative approach to unsupervised entity resolution named the Data Washing Machine (DWM). We integrate our GDWM with the DWM to overcome limitations, especially in self-assessment and reiterations where our ap-

proach eliminated those needs. In addition, we performed various experiments on 18 synthetic benchmark datasets with different sizes, qualities, and layouts to evaluate our adapted algorithm. Finally, we demonstrated the efficacy and advantages our clustering algorithm provides in terms of F1 scores, accuracy, and speed up of execution time compared with the original implementation of the DWM.

Acknowledgement This material is based upon work supported by the National Science Foundation under Award No. OIA-1946391.

References

1. J.R. Talburt, A.K., D. Pullen, L. Claassens, R. Wang, An Iterative, self-assessing entity resolution system: first steps toward a data washing machine. *Int. J. Adv. Comput. Sci. Appl.* **11**(12) (2020). <https://doi.org/10.14569/IJACSA.2020.0111279>
2. J.R. Talburt, Y. Zhou, A practical guide to entity resolution with OYSTER, in *Handbook of Data Quality: Research and Practice*, ed. by S. Sadiq (Springer, Berlin, 2013), pp. 235–270. https://doi.org/10.1007/978-3-642-36257-6_11
3. T.N. Herzog, F.J. Scheuren, W.E. Winkler, *Data Quality and Record Linkage Techniques*. Springer Science and Business Media (Springer, New York, 2007)
4. P. Lahiri, M.D. Larsen, Regression analysis with linked data. *J. Am. Stat. Assoc.* **100**(469), 222–230 (2005). <https://doi.org/10.1198/016214504000001277>
5. A. Tancredi, B. Liseo, A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.* **5**(2B), 1553–1585 (2011). <https://doi.org/10.1214/10-AOAS447>
6. M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, S. Fienberg, Adaptive name matching in information integration. *IEEE In-*

- tell. Syst. **18**(5), 16–23 (2003). <https://doi.org/10.1109/MIS.2003.1234765>
7. X. Li, J.R. Talburt, T. Li, Scoring matrix for unstandardized data in entity resolution, in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)* (2018), pp. 1087–1092. <https://doi.org/10.1109/CSCI46756.2018.00211>
 8. A. Alsarkhi, J.R. Talburt, A method for implementing probabilistic entity resolution. *Int. J. Adv. Comput. Sci. Appl.* **9**(11), 7–15 (2018)
 9. L. Kolb, Z. Sehili, E. Rahm, Iterative computation of connected graph components with MapReduce. *Datenbank-Spektrum* **14**(2), 107–117 (2014). <https://doi.org/10.1007/s13222-014-0154-1>
 10. V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**(10), P10008 (2008). <https://doi.org/10.1088/1742-5468/2008/10/P10008>
 11. J.R. Talburt, Y. Zhou, S.Y. Shivaiah, SOG: a synthetic occupancy generator to support entity resolution instruction and research. *ICIQ* **9**, 91–105 (2009)
 12. D. Zhang, D. Li, L. Guo, K. Tan, Unsupervised entity resolution with blocking and graph algorithms. *IEEE Trans. Knowl. Data Eng.* 1–1 (2020). <https://doi.org/10.1109/TKDE.2020.2991063>
 13. G. Jeh, J. Widom, SimRank: a measure of structural-context similarity, in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), pp. 538–543
 14. F. Wang, H. Wang, J. Li, H. Gao, Graph-based reference table construction to facilitate entity matching. *J. Syst. Softw.* **86**(6), 1679–1688 (2013). <https://doi.org/10.1016/j.jss.2013.02.026>
 15. H. Wang, J. Li, H. Gao, Efficient entity resolution based on sub-graph cohesion. *Knowl. Inf. Syst.* **46**(2), 285–314 (2016)
 16. A. Saeedi, M. Nentwig, E. Peukert, E. Rahm, Scalable matching and clustering of entities with FAMER. *Complex Syst. Inform. Model. Q.* **0**(16), Art. no. 16 (2018). <https://doi.org/10.7250/csimq.2018-16.04>
 17. U. Draisbach, P. Christen, F. Naumann, Transforming pairwise duplicates to entity clusters for high-quality duplicate detection. *J. Data Inf. Qual.* **12**(1), 3:1–3:30 (2019). <https://doi.org/10.1145/3352591>
 18. N. Kang, J.-J. Kim, B.-W. On, I. Lee, A node resistance-based probability model for resolving duplicate named entities. *Scientometrics* **124**(3), 1721–1743 (2020). <https://doi.org/10.1007/s11192-020-03585-4>
 19. L. Page, S. Brin, R. Motwani, T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web* (Stanford InfoLab, Stanford, 1999)
 20. M. Sadiq, S.I. Ali, M.B. Amin, S. Lee, A vertex matcher for entity resolution on graphs, in *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (2020), pp. 1–4. <https://doi.org/10.1109/IMCOM48794.2020.9001799>
 21. D. Zhang, L. Guo, X. He, J. Shao, S. Wu, H.T. Shen, A graph-theoretic fusion framework for unsupervised entity resolution, in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, Paris (2018), pp. 713–724. <https://doi.org/10.1109/ICDE.2018.00070>
 22. P. Malhotra, P. Agarwal, G.M. Shroff, Graph-parallel entity resolution using LSH & IMM, in *EDBT/ICDT Workshops* (2014), pp. 41–49
 23. A. Al-Sarkhi, J.R. Talburt, Estimating the parameters for linking unstandardized references with the matrix comparator. *J. Inf. Technol. Manag.* **10**(4), 12–26 (2018)
 24. A.E. Monge, C. Elkan, et al., The field matching problem: algorithms and applications, in *KDD*, vol. 2 (1996), pp. 267–270
 25. V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, in *Soviet Physics Doklady*, vol. 10, No. 8 (1966), pp. 707–710
 26. S.V. Ovchinnikov, On the transitivity property. *Fuzzy Sets Syst.* **20**(2), 241–243 (1986). [https://doi.org/10.1016/0165-0114\(86\)90080-1](https://doi.org/10.1016/0165-0114(86)90080-1)
 27. S.E. Schaeffer, Graph clustering. *Comput. Sci. Rev.* **1**(1), 27–64 (2007). <https://doi.org/10.1016/j.cosrev.2007.05.001>
 28. M.E.J. Newman, Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* **103**(23), 8577–8582 (2006). <https://doi.org/10.1073/pnas.0601602103>
 29. Y. Ye, J.R. Talburt, Generating synthetic data to support entity resolution education and research. *J. Comput. Sci. Coll.* **34**(7), 12–19 (2019)
 30. A. Hagberg, P. Swart, D.S. Chult, *Exploring Network Structure, Dynamics, and Function Using NetworkX*. Los Alamos National Lab. (LANL) (Los Alamos, NM (United States), 2008)
 31. J.P. Mower, PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinf.* **6**(1), 1–15 (2005)

Semantic-MDBScan: An Approach to Assign a Semantic Interpretation to Behavior Changes Detected in Data Stream Scenarios

Eldane Vieira Júnior, Rita Maria Silva Julia, and Elaine Ribeiro Faria

Abstract

A great variety of real-world problems can be satisfactorily solved by automatic agents that use adaptive learning techniques conceived to deal with data stream scenarios. The success of such agents depends on their ability to detect changes and on using such information to conveniently adapt their decision-making modules. Several detecting change methods have been proposed, with emphasis on the M-DBScan algorithm, which is the basis of this proposal. However, none of these methods is able to capture the meaning of the identified changes. Thus, the main contribution of this work is to propose an extended version of M-DBScan, called Semantic-MDBScan, with such ability. The proposed approach was validated through artificial datasets representing distinct scenarios. The experiments show that Semantic-MDBScan in fact achieves the intended goal.

Keywords

Data stream · Behavior change detection · Semantic assignment · Novelty detection · Classification · Clustering · Markov chain · Spatial entropy · Adaptive learning · M-DBScan

15.1 Introduction

A great variety of problems have been solved by Machine Learning (ML) techniques in different areas as entertainment, financial market, and medicine. Among these techniques, the

E. V. Júnior (✉) · R. M. S. Julia · E. R. Faria
Computer department, Federal University of Uberlândia, Uberlândia, M.G., Brazil
e-mail: eldane.vieira@ufu.br

batch learning and the adaptive learning can be highlighted [1]. In batch learning, the agents are trained through examples retrieved from databases, being inadequate for dynamic scenarios. Although, adaptive learning can cope with such problems, producing a learning model that is continuously adjusted [2].

Data streams are composed of independent and continuous sequences of events, presenting the following characteristics: the data arrive online; there is no control over data arrival order; there is no limit in the data stream's length; a processed data sample must be discarded [3, 4].

In non-stationary environments, data distribution is not constant, and this characterizes a phenomenon known as concept drift. Any change perceived in the data distribution of such environments is called a change detection [5].

For agents that act in non-stationary environments, the models are updated according to supervised or unsupervised techniques. Among the supervised ones, the updating can occur whenever there is a decrease in the model performance measure [6], or it can be updated at regular intervals [7].

Concerning both kinds of supervised techniques, it is interesting to point out that the classification of a data sample only expresses the meaning (class) of such sample, not being able to represent the nature of the data samples in a sequence.

The unsupervised techniques, similar to the first type of supervised techniques, also update the models. However, differently from those supervised techniques, the unsupervised ones retrieve information from the data themselves (and not from labeled databases). Among them, the Micro-Clustering DBScan (M-DBScan) [8] can be highlighted, since it is the basis of this work. M-DBScan is an unsupervised change detection technique based on entropy and Markov Chain (MC), but not capable to assign a meaning to a change, just like the other supervised or unsupervised data stream technique.

In this work, the authors propose an algorithm inspired by M-DBScan—named Semantic-MDBScan—which is able

to detect behavior changes in data stream scenarios and to assign a meaning (semantic) to changes. Then, Semantic-MDBScan corresponds to an extension of M-DBScan, since it allows the association of an adequate meaning to every data sequence that causes a behavior change.

The proposed approach was validated through artificial data sets representing distinct test scenarios. The results show that Semantic-MDBScan succeeded in the tasks of detecting behavior changes in data stream scenarios and of providing meanings to such changes.

This paper is organized as following: Sect. 15.2 presents the theoretical foundations; Sect. 15.3 resumes the related works; Sect. 15.4 describes the Semantic-MDBScan; Sect. 15.5 presents the experiments and the results; finally, Sect. 15.6 summarizes the conclusion and points out some future works.

15.2 Theoretical Foundations

This section resumes the main theoretical topics used in this work.

15.2.1 Change Detection in Data Stream

A data stream is a sequence of data samples that presents the following characteristics [9]: the data arrives online; there is no control over the order that the data arrives; the data stream size is unlimited; a data already processed is normally discarded, since it deals with limited resources, as memory.

In data stream scenarios can occur an event known as concept drift, where the data distribution changes over time [5]. Consider that D represents a data stream, composed by data samples d_1, d_2, d_3, \dots , which can be organized in sequences S_1, S_2, S_3, \dots , and each sequence represents a *context*, composed by samples generated by stationary distributions G_i , where $i \geq 1$ [3]. Sequences in succession may present a transition composed of data samples generated by distinct distributions, so a data sample generated by the distribution G_i can be seen as a noise for distribution G_{i+1} [3].

Change detection can deal with different types of change: abrupt or gradual. In the abrupt changes, the concept change from one to another suddenly. The gradual changes are more difficult to be identified since the beginning of a change can be seen as noise by the change detection algorithm [3], and the new concept is presented progressively, but over time the new pattern will prevail. Abrupt and gradual changes can present the aspect of a recurrent change, in which, after some time, an already known concept can reappear [5]. The experiments (Sect. 15.5) were elaborated in scenarios of abrupt and gradual change, both with recurrence of concepts.

15.2.2 M-DBScan

The algorithm M-DBScan is a technique to detect changes over the unlabeled data obtained from a stream [8]. This algorithm has an incremental clustering process, based on DenStream [10], which is a well-known technique for data stream scenarios. Succeeding the clustering process, a behavior change procedure is applied [11].

The M-DBScan clustering process is composed of two phases, online and offline. The online phase keeps a statistical summary of the data, based on micro-clusters, that maintain information as radius, center, and weight, representing the dynamics of data arrival from the stream. There are different types of micro-clusters, potential-micro-cluster (p-micro-cluster), outlier-micro-cluster (o-micro-cluster), and core-micro-cluster (c-micro-cluster), each of them presents its own characteristics, as described in [10]. The offline phase produces clusters from the result of the online phase [10].

The clusters created in the offline phase are used in a novelty detection module to point out changes over the data. It uses a Markov Chain (MC), and each state in the MC represents a group created in the offline phase. The transitions between the states in the MC are activated considering the dynamics of the data arrivals and the assignment of data to micro-clusters.

15.2.2.1 Novelty Detection

The M-DBScan algorithm uses entropy to identify dissimilarities that could be seen as a novelty. Related works as [8, 12, 13] have shown better results using the spatial entropy than the temporal entropy, so, it was chosen for this work. The spatial entropy provides the spatial distribution of the data into clusters. The entropy calculations, and the entropy update, follow the specifications presented in [8, 13].

An entropy update will occur every time that a new data sample is assigned to a micro-cluster in a state of the MC. So, the probabilities in the MC transitions will also be updated. If an MC transition is no longer activated, a decay factor will be applied over its probability value.

The detection of a novelty requires that the entropy surpasses its threshold. If it occurs, the novelty is registered in the sliding window to be used later in behavior change detection.

15.2.2.2 Behavior Change Detection

The verification of a behavior change happens every time that a novelty is identified. If a behavior change has occurred, it means that the data have presented sufficient variations represented by the novelties in the sliding window.

A behavior change detection demands the occurrence of a minimum amount of novelties in an interval of time represented by a sliding window with size k . Every novelty detected is registered in the sliding window, and a novelty can

be removed from the window if it has slid k times, using as reference the moment that this novelty was registered. Although, if a behavior change were detected, from this point on any novelty will be ignored for a period equal to k . This procedure is applied to prevent the detection of novelties related to an already known change [8].

15.3 Related Works

This section presents the state of art in the domain of analyzing data tendency in data stream scenarios according to two distinct purposes: data classification and change detection.

Concerning the data classification, the following works can be highlighted: [14–16] and [17]. In these works, the classification only represents the meaning of the sample classified, not being able to portray the sequence.

Concerning the change detection objective, according to [3], change detection techniques can be classified into two main categories: (1) the first one monitors changes through performance indicators and (2) the second category monitors the changes by analyzing data distribution on time windows.

Next, some works that fit the first category: Drift Detection Method (DDM) [4] deals with scenarios with abrupt alteration in the data, and it assumes that there is a change whenever the classification error increases. The Early Drift Detection Method (EDDM) [4] corresponds to an extension of DDM and it deals with gradual drifts. The work described in [18] introduces the Cumulative Sum Approach (CUSUM), which points out a change whenever a parameter as the classification error exceeds a limit. Page-Hinkley [19] is similar to CUSUM, but it uses the average of a Gaussian signal to define a change.

In the second category, there is the technique Adwin [20], which uses two windows that keep information from different moments, and they are statistically compared in order to indicate a change. M-DBScan proposed in [8] also belongs to the second category. In [13], the authors evaluate the performance of the algorithm M-DBScan in a real-time strategy game, which is a scenario that demands effectiveness in the task of detecting change. In [12], the authors used the M-DBScan to show that an adequate representation of the problem can improve the performance of this algorithm.

The work in [21] does not fit in any mentioned categories, but it uses a neural network to detect change regions in images and identify their semantic labels. This work does not deal with a non-stationary scenario and according to the authors each period image corresponds to a semantic label, which is different from the semantic of a change in a data sequence.

Finally, it is interesting to point out that all these related works conceived to perform change detection are not able to identify the meaning of a change, which is the main objective of the present work.

15.4 Semantic-MDBScan

The main contribution of this paper is to propose the Semantic-MDBScan: an algorithm that operates in data stream scenarios, able to detect changes in the data sequences, as well as to identify the semantic of the change.

For this, Semantic-MDBScan performs a behavior change detection process and, whenever it detects a change, it tries to assign a meaning to it. The detection of behavior changes in Semantic-MDBScan is similar to M-DBScan's, in which it demands the occurrence of a minimum number of novelties in the sliding window [8]. If Semantic-MDBScan detects a change processing a data sample, it will be classified and the result is used as the semantic of this change.

In the flowchart (Fig. 15.1), the dashed-line square delimits the original proceedings of M-DBScan. The algorithm firstly checks whether there is an available data sample in the stream (arrow #1): if so, the algorithm retrieves such sample (arrow #2) and presents it as an input to the online part of the algorithm (arrow #3). In sequence, a decay factor is applied to all micro-clusters but the one to which the last data sample was assigned (arrow #4). The decay factor is used to decrease the relevance of the attributes used to represent these micro-clusters, so they tend to disappear if they do not meet the minimum requirements of a micro-cluster. Then, the algorithm checks whether it is time to perform a rating verification, (arrow #5). If so, the algorithm converts into an o-micro-cluster every p-micro-cluster that does not fulfill anymore the constraints to be considered a p-micro-cluster [10] (arrow #6). Next, the algorithm discards all o-micro-clusters that do not satisfy anymore the constraints of an o-micro-cluster (arrow #7). Then, the offline part of the algorithm (see Sect. 15.2.2) is initialized (arrow #8), but it can be initialized if it is not time to run the rating verification (arrow #9). Once the offline part of the algorithm is concluded, the entropy value and its threshold are calculated (arrow #10 and #11, respectively). At this point, the algorithm has to check whether or not there has already been any behavior change in the processing of the current sliding window (arrow #12). If not (arrow #13), Semantic-MDBScan checks whether the entropy value is greater than its threshold. If it is (arrow #14), it characterizes a novelty occurrence, which is registered in the sliding window, if it is not (arrow #19) it will try to get a new data sample from the stream. In the following, the system checks whether the number of novelties, registered in

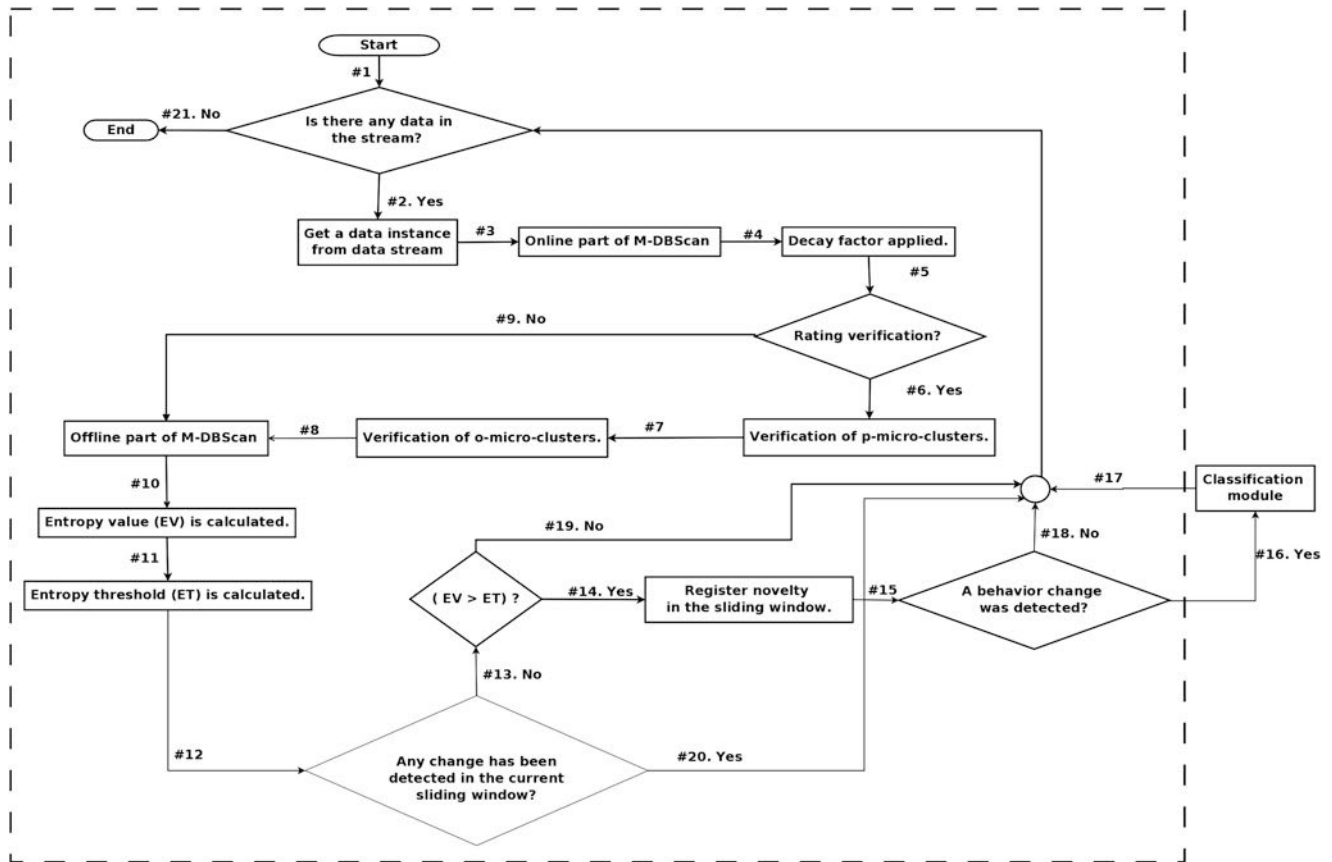


Fig. 15.1 Semantic-MDBScan flowchart

the sliding window, has reached the minimum value required to point out a change detection (arrow #15). If so (arrow #16), the current data sample is submitted to the *Classification module* and the label is used as the semantic (meaning) of the new behavior. Next, the algorithm tries to get a new data sample from the stream (arrows #17), this will also happen if no behavior change is detected (arrows #18). If there is a new data sample (arrow #2), the whole process is repeated. Otherwise, the processing of the Semantic-MDBScan is completed (arrow #21).

It is important to mention, that if a behavior change is detected, the arrows #13, #14, #15, and #16 will not be executed, for a period equal to the size of the sliding window.

15.5 Experiments and Results

The experiments aim to evaluate the ability of Semantic-MDBScan in assigning a correct semantic to behavior changes detected by the algorithm.

For comparative purposes, two classification methods will be investigated: the K-Nearest-Neighbor (KNN) [22]; and

Multilayer Perceptron (MLP) [23]. Both approaches are evaluated through seven distinct datasets. For this, seven labeled datasets—called *TrainingDataSet*—were created with the purpose of being used by KNN and also in the training of the MLPs of each dataset.

All *TrainingDataSet* and datasets were created from data generated by the framework Massive Online Analysis (MOA) [24], using the *random radial basis function generator*. The first five datasets present abrupt/recurrent changes, whereas the sixth and seventh ones present gradual/recurrent changes.

15.5.1 Test Datasets

In all artificial datasets created for the experiments, real changes occur at fixed interval of m samples, being $m = 1000$ for Datasets 2, 3, 4 and 6, and $m = 20,000$ for Datasets 1, 5 and 7. Each sample is represented by 10 attributes. The seven datasets are presented in Table 15.1.

In every *TrainingDataSet*, each behavior (class) counts on a number of samples equal to 15% of m related to the

Table 15.1 Description of datasets

Dataset	ChangeID	Timestamp	Semantic value
Dataset 1	1	20,000	class2
	2	40,000	class3
	3	60,000	class4
	4	80,000	class1
	5	100,000	class2
	6	120,000	class3
	7	140,000	class4
	8	160,000	class2
	9	180,000	class4
Dataset 2	1	1000	class2
	2	2000	class1
	3	3000	class2
Dataset 3	1	1000	class3
	2	2000	class2
	3	3000	class1
	4	4000	class3
Dataset 4	1	1000	class2
	2	2000	class3
	3	3000	class4
	4	4000	class1
Dataset 5	1	20,000	class2
	2	40,000	class3
	3	60,000	class4
	4	80,000	class1
	5	100,000	class2
	6	120,000	class3
	7	140,000	class4
	8	160,000	class2
	9	180,000	class4
Dataset 6	1	1000	class2
	2	2000	class1
	3	3000	class2
Dataset 7	1	20,000	class2
	2	40,000	class3
	3	60,000	class2

dataset associated to that *TrainingDataSet*. This value was empirically defined to assure a classification runtime of KNN that does not compromise the efficiency of the real-time processing.

Considering the creation of the datasets, one difference refers to the value of the parameter *speed*, which defines the movement of kernels throughout the arrival of data from the stream: for Dataset 1, it is equal to 500, and for all other datasets it is equal to 100.

The test datasets are described in Table 15.1. The columns of such table can be described as following: *ChangeID* enumerates the behavior change in each dataset; *Timestamp* indicates the samples when there is a change; *Semantic value* represents the semantics that prevails in that data sequence.

15.5.2 Settings in the Experiments

The experiments were executed using parameters defined empirically. The clustering parameters are: $\mu = 10$; $\beta = 0.105$; $\epsilon = 0.45$; $\lambda = 0.05$, where μ is the minimum number of points required to be a micro-cluster; ϵ is the radius limit for micro-cluster; β is the outlier threshold; and λ is the decay factor.

Spatial entropy parameters are: minimum number of novelties for a change = 2; size of sliding window = 20. Further: $\eta_s = 0.05$; $\gamma = 0.5$; $\delta = 0.02$; $\theta = 2$, where η_t is the weighting factor used to control the intensity of the probability update; γ and δ are the weighting factors used in the threshold calculation; and θ is the number of standard deviations used to define a normal distribution for entropy values.

In KNN, the value of K was defined through experimental tests in which values from 1 to 10 were evaluated for the datasets described in Table 15.1. The best results were obtained using $k = 5$, which presented 89% of correct classifications.

In order to reduce the FPs in the experiments with Datasets 6 and 7, where the changes are gradual, the following parameters were altered: firstly, the minimum amount of novelties to declare a behavior change is set to 4; secondly, the decay factor (λ) was reduced to 0.001, to slow down the forgetting mechanism of the micro-clusters, which consequently increases the stability of the MC states, reducing the occurrence of FPs.

An MLP was created for each dataset, being each one trained using data retrieved from the *TrainingDataSet* corresponding to such dataset. The architectures of these MLPs were empirically defined throughout the training: 10 neurons in the input layer; 2 hidden layers (the first with 6 neurons and the second with 3). The quantity of neurons of the output layer depends on the dataset to which the MLP will be applied, varying from 2 to 4, since it must have one output neuron for each class in the dataset (Table 15.1). The activation function of the MLPs is the hyperbolic tangent. The learning rate is 0.1, the training is concluded when the error is shorter than 0.01.

Tables 15.2 and 15.3 present the results obtained by the experiments involving all the datasets. Such results are presented using the following measures: *True Positive (TP)*: indicates the number of correct change detections with a correct semantic assignment. The delay allowed for the detection is up to 50% of m ; *False Positive (FP)*: refers to the number of incorrect change detections produced; *False Negative (FN)*: it occurs in the following situations: (1) The algorithm detects a real change, but after the maximum delay allowed (referred as *FN-D*); (2) The algorithm detects a real change, but it assigns an incorrect semantic (referred as *FN-S*); and (3) The algorithm does not detect a change (referred as *FN-B*, being a blind *FN*); *True Negative (TN)*: indicates the number of

samples that correctly was indicated as not being a behavior change. *F1*: indicates the balance of precision and recall.

15.5.3 Results

Table 15.2 presents the results of the experiments that evaluate the ability of Semantic-MDBScan with KNN to assign an adequate meaning to the changes, and Table 15.3 shows the results produced by Semantic-MDBScan using MLP.

Before analyzing the results, it is interesting to remember that the classifiers (KNN or MLP) have no influence over the change detection process, they are only used to assign a meaning to changes.

Concerning datasets 1, 2, 3, 4 and 5, with abrupt/recurrent changes, Tables 15.2 and 15.3 show that both Semantic-MDBScan versions (KNN and MLP) presented good results.

Still analyzing the results shown in these tables for datasets 1, 2, 3, 4 and 5, for both approaches they presented F1-score values that vary from 0.57 to 0.8, which are very good results. Analyzing the number of TPs in the mentioned Tables, it can be observed that the algorithm detected over 50% of the real changes. Figure 15.2 details the occurrences of TPs, FPs, and FNs mentioned in Tables 15.2 and 15.3.

Concerning Dataset 6, which presents gradual/recurrent changes, the Semantic-MDBScan, using KNN or MLP, presented an F1-score equal to 0.4 in both approaches. It is

justified by the fact that this dataset is more defiant than the abrupt/recurrent ones.

Considering both approaches and all datasets, there were only three FN-S in the experiments, all of them related to Dataset 7 (see Fig. 15.3b), being two associated to Semantic-MDBScan with KNN and one produced by Semantic-MDBScan with MLP. Such result indicates success in the task of assigning a correct meaning to changes.

Concerning the Dataset 7, the value 0.22 of F1-score obtained by Semantic-MDBScan with MLP and 0.11 obtained by Semantic-MDBScan with KNN, reflects the higher number of FP occurrences obtained in these experiments. It can be explained by the high value of m in this dataset combined with the variation of semantics in the gradual transition in the changes. The occurrence of TP, FP and FN in the experiments with Datasets 6 and 7 can be observed in details in Fig. 15.3.

In order to validate the results, the authors applied the Wilcoxon test (tool KEEL [25]). The null hypothesis used is the following one: Semantic-MDBScan with MLP does not present a better result than Semantic-MDBScan with KNN in terms of correctly assigning a semantic to the detected behavior changes. The results are R^+ equal to 17.5 and R^- equal to 10.5, with an asymptotic p-value equal to 0.498, which does not allow for the rejection of the null hypothesis.

Even being a simpler classifier method than MLP, and having a computational cost superior to the MLP, KNN presented a good performance, since few samples had to be classified, just those that caused a behavior change (for example, in a dataset with $m = 1000$, it will be 0.1% of the samples or less). Besides that, the incremental characteristic of KNN, turns it into an good classifier for dynamic scenarios, and it does not require a training process, as the MLP does.

Table 15.2 Results of the experiments: *Semantic-MDBScan with KNN*

Dataset	Measures				
	TP	FP	FN	TN	F1
Dataset 1 (9 changes)	6	6	3	199,985	0.57
Dataset 2 (3 changes)	2	0	1	3997	0.8
Dataset 3 (4 changes)	3	1	1	4995	0.75
Dataset 4 (4 changes)	3	2	1	4994	0.66
Dataset 5 (9 changes)	7	3	2	199,988	0.73
Dataset 6 (3 changes)	1	1	2	3996	0.4
Dataset 7 (3 changes)	1	13	2	79,984	0.11

Table 15.3 Results of the experiments: *Semantic-MDBScan with MLP*

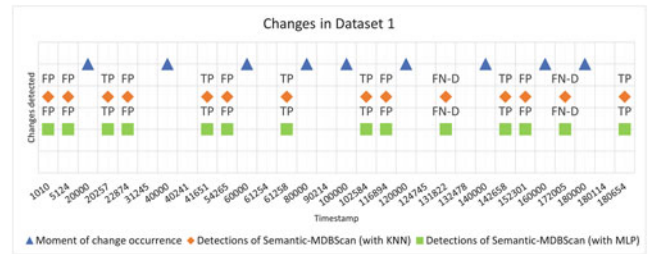
Dataset	Measures				
	TP	FP	FN	TN	F1
Dataset 1 (9 changes)	6	6	3	199,985	0.57
Dataset 2 (3 changes)	2	0	1	3997	0.8
Dataset 3 (4 changes)	3	1	1	4995	0.75
Dataset 4 (4 changes)	3	2	1	4994	0.66
Dataset 5 (9 changes)	7	3	2	199,988	0.73
Dataset 6 (3 changes)	1	1	2	3996	0.4
Dataset 7 (3 changes)	2	13	1	79,984	0.22

15.6 Conclusions and Future Works

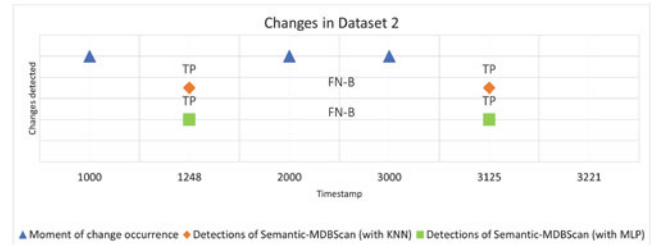
This paper presents a new algorithm based on M-DScan, named Semantic-MDBScan, whose purpose is to identify behavior changes in data stream scenarios and to assign an adequate meaning to them. Two distinct techniques to assign such meaning were investigated: MLP and KNN. The Semantic-MDBScan can be useful to build intelligent agents, since they will be capable to perceive changes and adapt their actions based on changes, turning the Semantic-MDBScan into a solution for real-world problems.

The results confirm that Semantic-MDBScan succeeded in almost all test scenarios, just presenting a lower performance in databases with gradual/recurrent changes and a long interval between the changes. The results also demonstrated that KNN is adequate to be used in order to reach the purposes pursued by this work, even being a more simple algorithm.

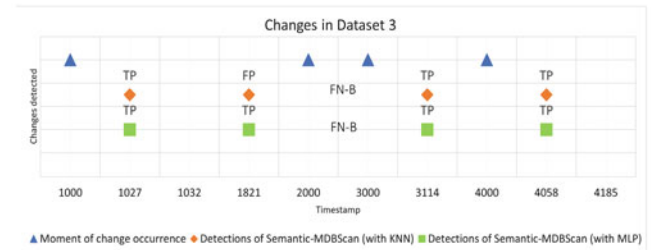
Fig. 15.2 Detections in datasets 1(a), 2(b), 3(c), 4(d) and 5(e)—abrupt/recurrent changes



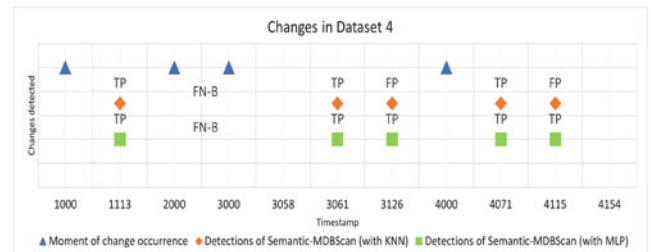
(a)



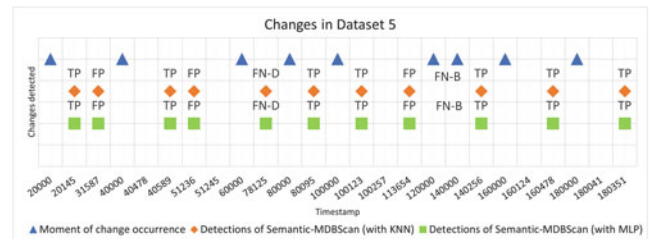
(b)



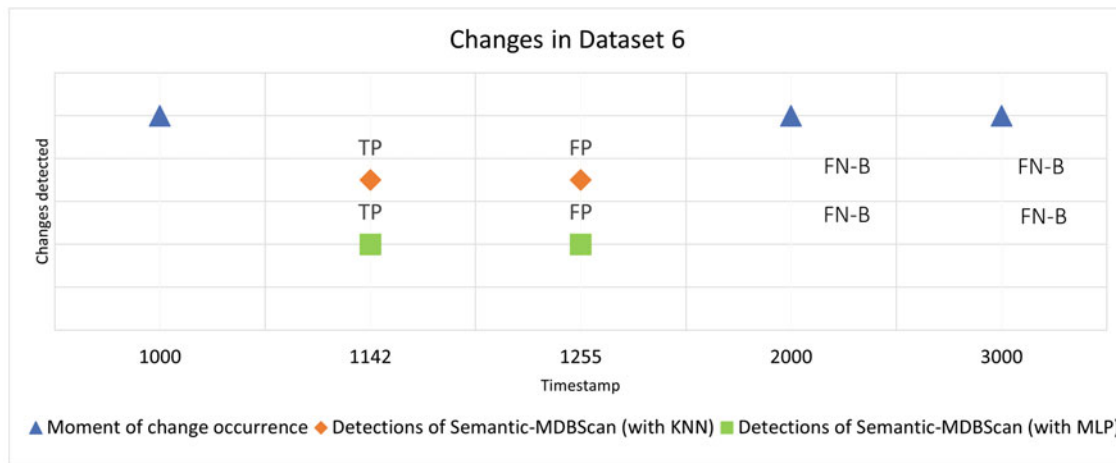
(c)



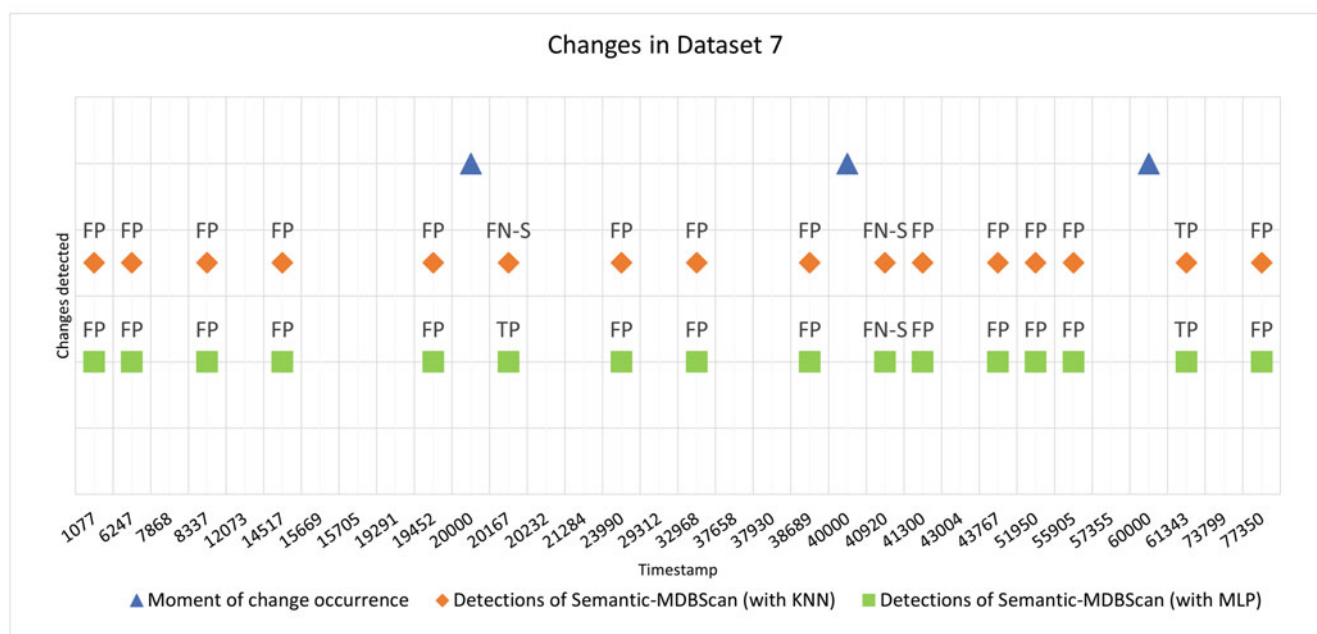
(d)



(e)



(a)



(b)

Fig. 15.3 Detections in datasets 6(a) and 7(b)—gradual/recurrent changes

In future works, the authors will use Semantic-MDBScan in games to perceive changes over the adversary's behavior, and improve the Semantic-MDBScan by better exploring the incremental aspects of KNN.

References

1. K. Faceli, A.C. Lorena, J. Gama, A.C. Carvalho, et al., *Inteligência Artificial: Uma abordagem de aprendizado de máquina* (LTC, Rio de Janeiro, Brazil, 2011)
2. A. Bifet, R. Gavaldà, Adaptive learning from evolving data streams, in *International Symposium on Intelligent Data Analysis* (Springer, Berlin, 2009)
3. J. Gama, *Knowledge Discovery from Data Streams* (CRC Press, Boca Raton, 2010)
4. B. Krawczyk, L.L. Minku, J. Gama, J. Stefanowski, M. Woźniak, Ensemble learning for data stream analysis: a survey. *Inf. Fusion* **37**, 132–156 (2017)
5. J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation. *ACM Comput. Surv.* **46**(4), 1–37 (2014)
6. A. Haque, L. Khan, M. Baron, Sand: Semi-supervised adaptive novel class detection and classification over data stream, in *THIRTIETH AAAI Conference on Artificial Intelligence* (2016)
7. N. Hatamikhah, M. Barari, M.R. Kangavari, M.A. Keyvanrad, Concept drift detection via improved deep belief network, in *Iranian Conference on Electrical Engineering (ICEE)* (IEEE, Piscataway, 2018)
8. R.M. Vallim, J.A. Andrade Filho, R.F. De Mello, A.C. De Carvalho, Online behavior change detection in computer games. *Expert Syst. Appl.* **40**(16), 6258–6265 (2013)
9. J. Gama, M.M. Gaber, *Learning from Data Streams: Processing Techniques in Sensor Networks* (Springer, Berlin, 2007)

10. F. Cao, M. Estert, W. Qian, A. Zhou, Density-based clustering over an evolving data stream with noise, in *Proceedings of the 2006 SIAM International Conference on Data Mining*. (SIAM, Philadelphia, 2006), pp. 328–339
11. R.M.M. Vallim, Mineração de fluxos contínuos de dados para jogos de computador, Ph.D. Dissertation, Universidade de São Paulo, 2013
12. E. Vieira, R.M.S. Julia, E.R. de Faria, Adapting the markov chain based algorithm M-DBScan to detect opponents' strategy changes in the dynamic scenario of a starcraft player agent, in *Proceedings of International Conference on Agents and AI* (SCITEPRESS, Setúbal, 2020)
13. E. Vieira, R.M.S. Julia, E.R. de Faria, Mining data stream to detect behavior change in a real-time strategy game, in *Proceedings of Machine Learning and Data Mining in Pattern Recognition* (ibai publishing, Leipzig, Germany, 2019)
14. L. Jalali, H. Oh, R. Moazeni, R. Jain, Human behavior analysis from smartphone data streams, in *International Workshop on Human Behavior Understanding* (Springer, Berlin, 2016), pp. 68–85
15. T. Fawcett, F. Provost, Activity monitoring: Noticing interesting changes in behavior, in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (1999)
16. F. Mota, M. Paula, I. Drummond, Combined classification models applied to people personality identification, in *ITNG 2021 18th International Conference on Information Technology-New Generations* (Springer, Berlin, 2021), pp. 457–462
17. J. Read, F. Perez-Cruz, A. Bifet, Deep learning in partially-labeled data streams, in *Proceedings of the 30th Annual ACM Symposium on Applied Computing* (ACM, New York, 2015), pp. 954–959
18. E.S. Page, Continuous inspection schemes. *Biometrika* **41**(1/2), 100–115 (1954)
19. R. Sebastiao, J. Gama, A study on change detection methods, in *Progress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence, EPIA* (2009), pp. 12–15
20. A. Bifet, R. Gavaldá, Learning from time-changing data with adaptive windowing, in *Proceedings of the 2007 SIAM International Conference on Data Mining* (SIAM, Philadelphia, 2007), pp. 443–448
21. S. Xiang, M. Wang, X. Jiang, G. Xie, Z. Zhang, P. Tang, Dual-task semantic change detection for remote sensing images using the generative change field module. *Remote Sens.* **13**, 3336 (2021)
22. S.A. Dudani, The distance-weighted k-nearest-neighbor rule. *IEEE Trans. Syst. Man Cybern.* **SMC-6**(4), 325–327 (1976)
23. S. Haykin, *Neural Networks and Learning Machines, 3/E* (Pearson Education India, London, 2010)
24. A. Bifet, G. Holmes, R. Kirkby, B. Pfahringer, MOA: massive online analysis. *J. Mach. Learn. Res.* **11**, 1601–1604 (2010). [Online]. Available: <http://portal.acm.org/citation.cfm?id=1859903>
25. J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *J. Multiple-Valued Logic Soft Comput.* **17**, 255–287 (2011)

Heber Miranda Floriano, Mario Jino, and Ferruccio de Franco Rosa

Abstract

With the massive use of e-learning and Learning Management Systems (LMS) in the education domain, the development of methods and techniques for evaluating the usability of systems is required. This is a critical and important task, as different user profiles interact with educational systems. For example, teachers and students of different ages and limitations (cognitive or physical) demand user-friendly systems. We present a systematic literature review, describing and comparing works that address the usability and accessibility assessment of e-learning and m-learning systems. The study provides a current view of the available methodological resources, in addition to pointing out gaps in the literature. This review is aimed at researchers seeking to improve the interface of educational systems.

Keywords

Assessment · Distance learning · E-learning · LMS · Survey · Usability

H. M. Floriano (✉)
UNIFACCAMP, Campo Limpo Paulista, SP, Brazil

M. Jino
FEEC/UNICAMP, Campinas, SP, Brazil
e-mail: jino@dca.fee.unicamp.br

F. de Franco Rosa
UNIFACCAMP, Campo Limpo Paulista, SP, Brazil

CTI Renato Archer, Campinas, SP, Brazil
e-mail: ferruccio.rosa@cti.gov.br

16.1 Introduction

Nowadays, with the massive use of information and communication technologies, large-scale teaching methods, such as distance learning systems, present various problems [1]. One of the main issues for many user profiles is usability, which is being neglected by educational institutions [2]. Developing usability assessment methods, which could identify possibilities for improvement in educational systems, is imperative. In this context, revealing the state of the art of usability assessment methods aimed at educational systems is necessary.

We present a systematic literature review, describing and comparing works that address the usability and accessibility assessment of e-learning and m-learning systems. The remainder of this article is organized as follows: Sect. 16.2 presents related works, i.e., other literature reviews that emerged from the keywords search; in Sect. 16.3 we describe the methodology used in the systematic review; Sect. 16.4 details the results obtained from searches in the scientific databases. Section 16.5 presents a discussion on the analyzed works; Sect. 16.6 presents our conclusions.

16.2 Related Work

Three literature reviews were identified in our study. The works were classified according to their objectives and applications. Table 16.1 (Appendix I) presents a summary of related works. A synthetic description of these works is presented below.

Herbig et al. [3], present a literature review on the user's cognitive state when using e-learning platforms.

Pereira and Rodrigues [4] present a literature review on e-learning. The study provides a critical analysis of e-learning projects, and it was divided into (i) apps from recent online

mobile stores, and (ii) standalone operating system apps. According to the authors, mobile learning (m-learning) is an extension of distance education, supported by mobile devices with wireless technologies embedded.

Hasani et al. [5] present a literature review on User-Centered Design (UCD). UCD is a method used to develop user interfaces for e-learning systems. According to the authors, the most used methods with UCD are questionnaires, interviews, high fidelity prototyping, and usability testing. Projects that used UCD with higher user involvement in various stages of development produced a design with good usability.

16.3 Review Methodology

We present a systematic literature review based on the guidelines proposed by Kitchenham [6] and on the review presented by de Mendonça et al. [7]. We aim to address the following question: “What is the state-of-the-art in usability assessment of educational systems?”.

First, exploratory research was carried out to define the study parameters, such as the search time interval, scientific databases, keywords. The search period comprises the years 2010–2021, due to the technology advances in recent years; distance education has taken gigantic proportions.

The language defined was English. Adapting the syntax to each database, the following search string was used: *((("Abstract":usability) OR ("Document Title":usability)) AND ((("Abstract":LMS) OR ("Document Title":LMS))) ((("Document Title":usability) AND ("Document Title":e-learning)) OR ((("Document Title":e-learning)))*.

We considered all articles returned from three well-known computer science databases. In the initial search, we obtained 11,645 papers, with 7716 articles from Springer Link, 69 from IEEE Xplore, and 3860 from ACM Digital Library.

The inclusion and exclusion criteria were defined by three researchers (two doctors and one expert in the field of computing), as follows: *Inclusion*: (I1) Research on the usability of e-learning systems and LMS in general; (I2) Research on the accessibility of e-learning systems; (I3) Research on the usability of mobile learning systems; and (I4) Research on the effectiveness of usability of e-learning systems versus face-to-face teaching. *Exclusion*: (E1) Articles that were written in languages other than English; (E2) Articles related to the usability of e-learning systems; (E3) Publications that are not scientific works; (E4) Abstracts that lack depth or relevant results; and (E5) Articles with more than eleven years since publication.

We submitted the collected papers to the inclusion and exclusion criteria. The evaluation considered the title, abstract, and keywords. Thirty-four articles were categorized as works adhering to the research theme and were evaluated in their

entirety according to the criteria. Two researchers analyzed the selected works and generated the final list. Three works are literature reviews related to the topic and are described in Sect. 16.2.

16.4 Results From the Review

The synthetic analysis of the selected studies is organized as follows: (i) Approaches aimed at assessing the usability of educational systems (Sect. 16.4.1); (ii) Approaches aimed at evaluating the usability of mobile learning systems (Sect. 16.4.2); (iii) Approaches aimed at assessing the accessibility of e-learning systems (Sect. 16.4.3). Table 16.2 (Appendix II) presents a synthesis of the analyzed works.

16.4.1 Approaches Aimed at Assessing the Usability of Educational Systems

Ternauciuc and Vasii [8] propose a test strategy for the e-learning platform Campus Virtual (based on Moodle).

Trilaksono and Santoso [9] proposed an instructional design aimed at kinesthetic students. According to the authors, the proposed design provided a high rate of acceptability, as well as better user learning.

Awad et al. [10] evaluated usability features (and related costs) of a commercial LMS platform (Blackboard). The level of user participation and frequency of resource usage were analyzed.

Althobaiti and Mayhew [11] performed a usability assessment of an LMS (Jusur). The authors list important factors to be considered when assessing usability, namely: content, learning and support, visual design, navigation, accessibility, interactivity, self-assessment, learning ability, and motivation.

Phongphaew and Jiamsanguanwong [12] evaluated an LMS (MyCourseVille) to identify the main interface problems. A usability assessment method was used, associated with five usability attributes, to assess student and teacher interfaces. The method is aimed at LMS developers and can be used as a guideline for evaluating interfaces.

Rasheed Hasan [13] evaluated the possibility of using Facebook as an LMS instead of Moodle. According to the authors, despite Facebook having presented some problems regarding usability, students were satisfied with its use as an LMS.

Solomon and Makara [14], seeking to understand the effectiveness of online learning environments as pedagogical tools, applied four principles to understand the nature of interactions between students and instructors, namely: (i) Perceptions of lecturers and students about the manner and quality of interactions; (ii) Interactions and cooperation

between students; (iii) Expectations that are communicated in courses that use LMS, as perceived by instructors and students; and (iv) Manners how the use of LMS tools structure and influence the time of teachers and students inside and outside the classroom.

Kataoka et al. [1] present a study on language teaching, proposing the effective use of LMS. According to the authors, with the use of information and communication technologies, large-scale language teaching methods present several problems. R. Desai et al. [15] performed comparative evaluations of LMS tools. According to the authors, LMS represents an advance in learning techniques, providing different approaches and teaching techniques.

Valerio and Naranjo-Zeledón [16] have described fundamental usability features of LMSs. The characteristics were obtained through the System Usability Scale (SUS) questionnaire, which was implemented in the Moodle and Sakai LMSs.

U. Desai et al. [17] evaluated aspects of learning and collaborative problem solving by students. According to the authors, efficient tools are needed to model and evaluate goal-oriented discussion forums built from active student collaborations.

Alyami and Alagab [18] evaluated the impact of different learning strategies in virtual environments regarding learning satisfaction. Al-Omar [19] assessed the usability of an LMS using data mining techniques (sentiment analysis and clustering) and the SUS questionnaire.

Alvarado et al. [20] applied usability criteria on web and mobile platforms aiming to help students and teachers to reduce problems that could arise, such as difficulty in tests, communication issues, lack of engagement, among others. Little [21] presents the main concerns and functionality problems in the development of LMS and VLE systems.

Olivé et al. [22] presented a software framework aimed at identifying students at risk of dropping out of a course. Data from previous versions of the course were used for training the models. The objective is to facilitate the evaluation and use of prediction models in LMS.

Zareravasan and Ashrafi [23] and Soledad Fabito et al. [24] present factors that influence students' intention to use LMS. Fernando et al. [25] present characteristics that affect the quality of e-learning systems.

Mtebe and Kissaka [2] propose heuristics to evaluate LMSs. Penha and Correia [26] performed heuristic evaluations of usability in LMSs. Medina-Flores and Morales-Gamboa [27] used general heuristic evaluation criteria to develop an approach for evaluating the usability of LMSs by experts. Lehong et al. [28] propose a set of usability guidelines.

Rodríguez Morales et al. [29] present a usability study supported by the inquiry method in conjunction with the questionnaire technique, measuring the following indicators:

satisfaction, learning, operation, attractiveness, content, and communication. The best results were obtained in the learning and content indicators; with lower scores, those regarding the operation indicator, specifically accessibility and availability. De Freitas et al. [30] present an approach for assessing the impact of gamification on university education, focusing on how data is presented to students.

Ugras and Sener [31] evaluated the usability of an online educational platform, focusing on children's needs. A qualitative usability test was conducted using a multi-method approach.

16.4.2 Approaches Aimed at Evaluating the Usability of Mobile Learning Systems

Kumar and Gounda [32] performed a heuristic evaluation on mobile learning applications (m-learning). According to the authors, the interfaces are not effective because they are generic and do not consider access through mobile devices. The use of heuristics could help detect more usability issues in mobile learning apps.

Hasan [33] evaluated usability issues in desktop and mobile interfaces from the students' perspective. Seventeen usability problems were identified in the interfaces of an LMS (Moodle).

16.4.3 Approaches Aimed at Assessing the Accessibility of E-Learning Systems

In the context of this work, we consider the term accessibility as a characteristic of usability (broader concept).

Lee and Lee [34] evaluated the usability of educational applications for smartphones from the blind users perspective. According to the authors, even when the accessibility requirements were met, the usability level for the applications was low for people with vision disabilities.

Melton and Fenwick [35] have developed a user interface that uses voice commands for learning management (Amazon Alexa for Moodle). According to the authors, the integration of Alexa and Moodle features provided people with disabilities access to LMS information in an adequate and faster manner.

16.5 Discussion

Investigating the factors that influence the students' intention to continue using a LMS is important. The usefulness is a strong indication of intended use [23]. Social influence, information quality, and good university management practices also influence the intended use [24]. The quality of e-learning

systems is influenced by system quality, service quality, information quality, instructor competence, user characteristics, individual impact [25].

What can we do to increase user interest in LMSs? According to Rodriguez Morales et al. [29], we need to implement various design improvements to achieve a higher degree of usability. According to de Freitas et al. [30], a quick feedback to students brings various benefits, such as increased motivation, involvement, satisfaction, and improved performance.

The use of LMSs is becoming common; however, institutions have not performed usability tests before adopting these systems [2]. Lehong et al. [28] reported a difficulty of interaction faced by students to access knowledge resources, due to generally working in isolation. According to Ugras and Sener [31], there is a lack of usability studies focused on the children’s needs. The authors argue that in terms of information architecture, we should seek to improve the usability of LMS platforms for children.

16.6 Conclusion

We present a systematic literature review on usability assessment of educational systems (e.g., LMS), including m-learning and accessibility issues. The approach used in the literature review allowed the verification and analysis of usability problems. Thirty-three works were selected from 11,645 articles initially retrieved, following inclusion and exclusion criteria. The articles were evaluated, classified, and organized into tables for effortless consultation and comparison. This work is meant to be useful for researchers seeking to develop methods and techniques for assessing usability in LMSs.

A.1 Appendices

A.1.1 Appendix I

Table A.1 Summary of related work

Reference	Application domain			Contribution			
	A	B	C	1	2	3	4
Herbig et al. [3]				X			X
Pereira and Rodrigues [4]		X			X		X
Hasani et al. [5]				X			X
<i>Our Study</i>	X	X	X			X	X

Application Domain: (A) E-learning; (B) M-learning; (C) Accessibility. *Contribution:* (1) Process; (2) Method; (3) Conceptualization; (4) Literature Review

A.1.2 Appendix II

Table A.2 Summary of analyzed work

Reference	Application domain			Contribution						
	A	B	C	1	2	3	4	5	6	7
Solomon and Makara [14]	X					X				
Little [21]	X					X				
Pereira and Rodrigues [4]	X									
Alyami and Alagab [18]	X									
Medina-Flores and Morales-Gamboa [27]	X				X					X
Ternauciuc and Vasii [8]	X				X					
R. Desai et al. [15]	X				X					
Ugras and Sener [31]	X				X					
Mtebe and Kissaka [2]	X					X				
Alvarado et al. [20]	X						X			
Althobaiti and Mayhew [11]	X									X
Trilaksono and Santoso [9]	X				X					
de Freitas et al. [30]	X				X					
Al-Omar [19]	X									
Olivé et al. [22]	X						X			
Phongphaew and Jiamsanguanwong [12]	X									X
Hasan [33]	X									
Kataoka et al. [1]	X				X					
Lee and Lee [34]		X	X		X					
Penha and Correia [26]	X									
Rodriguez Morales et al. [29]	X				X					
Lehong et al. [28]	X									
Zarerasavan and Ashrafi [23]	X				X					
Melton and Fenwick [35]			X						X	X
Awad et al. [10]	X				X					
Kumar and Goundar [32]		X				X				
Fernando et al. [25]	X				X					
Valerio and Naranjo-Zeledón [16]	X									
Rasheed Hasan [13]		X			X					
Soledad Fabito et al. [24]	X					X				
U. Desai et al. [17]	X									

Application Domain: (A) E-learning; (B) M-learning; (C) Accessibility. *Contribution:* (1) Process; (2) Method; (3) Conceptualization; (4) Framework; (5) Prototype; (6) Appliance; (7) Methodology

References

1. Y. Kataoka, A.H. Thamrin, J. Murai, K. Kataoka, Effective use of learning management system for large-scale Japanese language education, in *Proceedings of the 10th International Conference on Education Technology and Computers*, 2018, pp. 49–56. <https://doi.org/10.1145/3290511.3290564>
2. J.S. Mtebe, M.M. Kissaka, Heuristics for evaluating usability of learning management systems in Africa, in *2015 IST-Africa Conference*, 2015, pp. 1–13. <https://doi.org/10.1109/ISTAFRICA.2015.7190521>
3. N. Herbig, P. Schuck, A. Krüger, *User Acceptance of Cognition-Aware e-Learning: An Online Survey*, 2019. <https://doi.org/10.1145/3365610.3365615>
4. O.R.E. Pereira, J.J.P.C. Rodrigues, Survey and analysis of current mobile learning applications and technologies. *ACM Comput. Surv.* **46**(2), 27:1 (2013). <https://doi.org/10.1145/2543581.2543594>
5. L.M. Hasani, D.I. Sensuse, R. R. Suryono, User-centered design of e-learning user interfaces: A survey of the practices, in *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, 2020, pp. 1–7. <https://doi.org/10.1109/IC2IE50715.2020.9274623>
6. B. Kitchenham, Procedures for performing systematic reviews. *Keele UK Keele Univ* **33**(TR/SE-0401), 28 (2004) 10.1.1.122.3308
7. R.R. de Mendonça, F. de Franco Rosa, R. Bonacin, Um estudo sobre análise, representação e detecção de intenções de criminosos em postagens em mídia social, *www/internet 2019*, p. 27
8. A. Ternauciuc, R. Vasiu, Testing usability in Moodle: When and How to do it, in *2015 IEEE 13th International Symposium on Intelligent Systems and Informatics (SISY)*, 2015, pp. 263–268. <https://doi.org/10.1109/SISY.2015.7325391>
9. K. Trilaksono, H.B. Santoso, Moodle based learning management system development for kinesthetic learning style, in *2017 7th World Engineering Education Forum (WEEF)*, 2017, pp. 602–606. <https://doi.org/10.1109/WEEF.2017.8467180>
10. M. Awad, K. Salameh, E.L. Leiss, Evaluating learning management system usage at a small university, in *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, 2019, pp. 98–102. <https://doi.org/10.1145/3325917.3325929>
11. M.M. Althobaiti, P. Mayhew, Assessing the usability of learning management system: User experience study, in *E-Learning, E-Education, and Online Training*, 2016, pp. 9–18.
12. N. Phongphaew, A. Jiamsanguanwong, Usability evaluation on learning management system, in *Advances in Usability and User Experience*, 2018, pp. 39–48
13. L. Rasheed Hasan, Is it possible to use facebook instead of moodle Learning management systems (LMS) to support the learning process?,” in *2020 The 4th International Conference on E-Society, E-Education and E-Technology*, 2020, pp. 47–51. <https://doi.org/10.1145/3421682.3421695>
14. T.C. Solomon, K. Makara, How does LMS use affect instructional time?, in *Proceedings of the 9th International Conference of the Learning Sciences – Volume 2*, 2010, pp. 137–138
15. R. Desai, V.K. Ajay, K. Kumar, Sumangali, OSLMS: Open source softwares to E Learning – A comparative study, in *2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2015, pp. 33–37. <https://doi.org/10.1109/ICCICCT.2015.7475244>
16. C.L. Valerio, L. Naranjo-Zeledón, Usability of open source LMS platforms in academia: Benchmarking from the active learning approach, in *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 2020, pp. 1–6. <https://doi.org/10.23919/CISTI49556.2020.9141037>
17. U. Desai, V. Ramasamy, J. Kiper, Evaluation of student collaboration on canvas LMS using educational data mining techniques, in *Proceedings of the 2021 ACM Southeast Conference*, 2021, pp. 55–62. <https://doi.org/10.1145/3409334.3452042>
18. S.M. Alyami, A.M. Alagab, The difference in learning strategies in virtual learning environment and their effect on academic achievement and learning satisfaction for distance teaching training program students, in *2013 Fourth International Conference on e-Learning “Best Practices in Management, Design and Development of e-Courses: Standards of Excellence and Creativity,”* 2013, pp. 102–112. <https://doi.org/10.1109/ECONF.2013.40>
19. K. Al-Omar, Evaluating the usability and learnability of the ‘Blackboard’ LMS using SUS and data mining, in *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, 2018, pp. 386–390. <https://doi.org/10.1109/ICCMC.2018.8488038>
20. J.V. Alvarado, A.F. Alfaro, M.C. Rivas, C.G. Rodríguez, Collaborative logical framework: An e-learning assesment tool in.LRN platform, in *2016 XI Latin American Conference on Learning Objects and Technology (LACLO)*, 2016, pp. 1–9. <https://doi.org/10.1109/LACLO.2016.7751748>
21. B. Little, Concerns with learning-management systems and virtual learning environments. *ELearn* **2010**(7) (2010). <https://doi.org/10.1145/1833513.1837142>
22. D.M. Olivé, D.Q. Huynh, M. Reynolds, M. Dougiamas, D. Wiese, *A Supervised Learning Framework for Learning Management Systems*, 2018. <https://doi.org/10.1145/3279996.3280014>
23. A. Zareravasan and A. Ashrafi, Influencing factors on students’ continuance intention to use learning management system (LMS), in *Proceedings of the 9th International Conference on Information Communication and Management*, 2019, pp. 165–169. <https://doi.org/10.1145/3357419.3357429>
24. B. Soledad Fabito, R. L. Rodriguez, A.O. Trillanes, J.I.G. Lira, D.Z. Estocada, P.M.Q. Sta Ana, Investigating the factors influencing the use of a learning management system (LMS): An extended information system success model (ISSM), in *2020 The 4th International Conference on E-Society, E-Education and E-Technology*, 2020, pp. 42–46. <https://doi.org/10.1145/3421682.3421687>
25. E. Fernando, S. Titan, Meyliana, Factors influence the success of E-learning systems for distance learning at the university, in *2020 International Conference on Information Management and Technology (ICIMTech)*, 2020, pp. 294–299. <https://doi.org/10.1109/ICIMTech50083.2020.9211163>
26. M. Penha, W.F.M. Correia, Usability recommendations for a learning management systems (LMS) – A case study with the LMS of IFPE, in *Advances in Usability, User Experience and Assistive Technology*, 2019, pp. 451–460
27. R. Medina-Flores, R. Morales-Gamboa, Usability evaluation by experts of a learning management system. *IEEE Rev. Iberoam. Technol. del Aprendiz.* **10**(4), 197–203 (2015). <https://doi.org/10.1109/RITA.2015.2486298>
28. S. Lehong, J. van Biljon, I. Sanders, Open-distance electronic learning environments: Supervisors’ views on usability, in *2019 Conference on Information Communications Technology and Society (ICTAS)*, 2019, pp. 1–7. <https://doi.org/10.1109/ICTAS.2019.8703605>
29. G. Rodriguez Morales, P. Torres-Carrion, J. Pérez, L. Peñafiel, Improving the design of virtual learning environments from a usability study, in *Information and Communication Technologies of Ecuador (TIC.EC)*, 2019, pp. 100–115
30. S. de Freitas et al., How to use gamified dashboards and learning analytics for providing immediate student feedback and performance tracking in higher education, in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 429–434. <https://doi.org/10.1145/3041021.3054175>

31. T. Ugras and O. Sener, A usability study with children on an online educational platform, in *Design, User Experience, and Usability: Interactive Experience Design*, 2015, pp. 228–239
32. B.A. Kumar, M.S. Goundar, Usability heuristics for mobile learning applications, *www/internet 2019*, p. 24.
33. L. Hasan, Usability problems on desktop and mobile interfaces of the moodle learning management system (LMS), in *Proceedings of the 2018 International Conference on E-Business and Applications*, 2018, pp. 69–73. <https://doi.org/10.1145/3194188.3194192>
34. Y. Lee, J. Lee, A checklist for assessing blind users' usability of educational smartphone applications. <https://doi.org/10.1007/s10209-017-0585-1>, p. 24.
35. M. Melton, J. Fenwick, Alexa skill voice interface for the moodle learning management system. *J. Comput. Sci. Coll.* **35**(4), 26–35 (2019)

Part III

Cybersecurity

Khandaker Abir Rahman, Avishek Mukherjee, and Kristina Mullen

Abstract

A method of alternate user authentication that relies on sensory data from a smartwatch has been explored in this paper. This attempt to beef up the authentication security was made by taking the user-defined hand gesture into account while wearing a smartwatch. Eventually, the preset hand gesture would work similar way to the password-based authentication scheme. In our experiment, we recorded the 3D coordinate values measured by the accelerometer and gyroscope over a set of gestures. We experimented with 50 gesture samples comprising of five different gesture patterns and ten repeated samples for each pattern. We developed an Android WearOS smartwatch app for sensor data collection, implemented our method of sensor data processing, and performed a series of experiments to demonstrate the potential of this method to achieve high accuracy.

Keywords

User authentication · Smartwatch dynamics · Gesture recognition · Gesture-based authentication · Motion-based authentication · Biometrics · Sensors · Gyroscope · Accelerometer · Wear OS

17.1 Introduction and Background

With the increased popularity of devices containing motion sensors that can log user movement and interaction, various modalities including keystroke dynamics (KD) [1–3], mouse

dynamics [4, 5], touchscreen interaction [6], and device movement [7] have enabled the viability of using behavioral biometrics for user authentication. In recent years, research in gesture recognition methods, have been an active area of interest among security researchers. Some of the earlier work in this area have focused on various security aspects. More recently, user authentication using gesture-based recognition methods have been one such application that has seen varying levels of success. For instance, in [8], a Leap Motion Controller was used to evaluate the usability of various forms of mid-air gestures for authentication. It was concluded by this research that the easier-to-use gestures have weaker security and that users tend to forget their gestures.

With the widespread adoption of smart devices, researchers have used the data from multiple on-device sensors for various security purposes, including one-time authentication and continuous user verification. For instance, the research in [7] explored the movement pattern-based authentication using smartphones and found viability in using these movement patterns to thwart common attacks. These studies have helped pave the way for the concept of using smartwatch gestures to achieve user authentication.

Initial studies using smartwatch sensors were not pertaining to user authentication. The work in [9], for example, explored the capability of smartwatch devices to detect arm, hand, and finger gestures correctly. It showed the viability in smartwatch motion sensor data to detect unique motion. Similarly, [10] sought to test the validity of fine-motor gestures. The research concluded that five fine-motor gestures chosen by the authors could be recognized with high accuracy supporting the potential for using these gestures in future gesture-based recognition systems. Following the viability of gesture recognition, [11] found that specific gestures could be segmented from continuous movement data, allowing for the use of these gestures as device controllers.

The research efforts for gesture-based user authentication using wearables can be separated into two categories - those

K. A. Rahman (✉) · A. Mukherjee · K. Mullen
 Department of Computer Science and Information Systems, Saginaw
 Valley State University, University Center, MI, USA
 e-mail: krahman@svsu.edu; amukher1@svsu.edu; kmullen@svsu.edu

that involve predetermined gestures, that is gestures that are decided and set for each user before testing, versus those that use user-created gestures, or gestures that are chosen by the user. The work in [12–14] fall into the former category. Even with high accuracy, this approach suffers from one major limitation: the system generated predetermined gestures would be hard to memorize, be easily forgotten, and therefore, lacks usability. On the other hand, the work in [15–17] fall into the second category, that of user-created gestures. In [15], researchers collected motion sensor data from individuals wearing a smart wrist-worn device while they signed their name to validate and distinguish genuine signatures from falsified signatures. Meanwhile, [16, 17] recorded sensor data from a wrist-worn smart device's and combine the user's free range of motion while typing their username and password with keystroke dynamics to aid in the authentication of users. However, none of these approaches in [15–17] use hand gestures as a primary mode of user authentication rather, they work as an auxiliary layer of security and triggers when the primary authentication method (signature, username-password, etc.) have been compromised.

There has been some research [18, 19], where smartwatch hand-gestures have been considered as the primary mode for user authentication and is closely related to our work. The work in [18], however, is a preliminary study for an authentication system designed exclusively for the smartwatch itself. To be more specific, the method outlined by the authors do not support using the smartwatch to authenticate users on a connected device (computer login, smartphone unlocking, etc.). In contrast, our method can be used as an authentication scheme for any supported device paired with a smartwatch. Similarly, [19] designed a motion gesture authentication scheme using Dynamic Time Warping for authentication. The scope of this research was limited to a small set of users who had created their own gestures for use in their experiment. Therefore, the scope for using other pattern matching methods is considerable. In addition, large-scale data analysis would be required to establish the performance of the system.

The smartwatch hand gesture recognition method explored in this paper is a standalone user authentication system. In addition, the method developed is intended to authenticate a user not only on the smartwatch itself but also on a connected device (by unlocking the smartphone screen or logging in to a computer, for example) via Bluetooth or Wi-Fi connectivity. We developed a data collection application for the smartwatch, collected 50 user samples collected and implemented our own signal processing methods to identify different user patterns, and finally showed the efficacy over a range of experiments.

The authentication scheme implemented here is also appealing from a security point of view because it possesses inherent resiliency to spoof attacks as authentication samples can be collected without necessarily looking or interacting with the device or account itself.

The rest of this paper is organized as follows. Section 17.2 shows the data collection procedures and characteristics of the data. We describe our data processing and experimental setups in Sect. 17.3. Experimental results are shown in Sect. 17.4, and we conclude in Sect. 17.5.

17.2 Data Collection

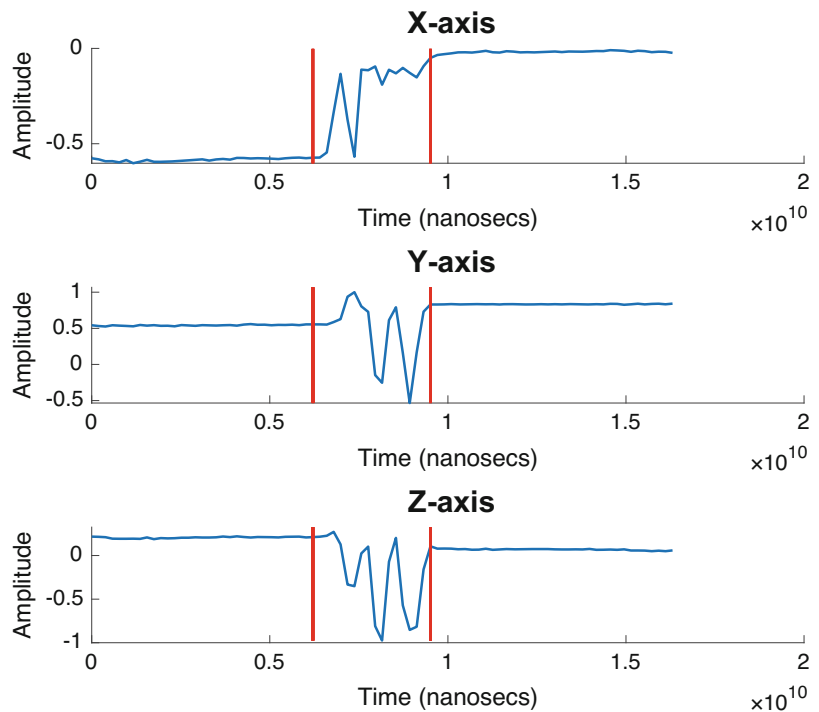
We used the Android platform to develop a simple application to run smartwatches on the WearOS operating system. The app was built for devices running Android SDK 30 (Android 8.0) or higher. The app consists of a simple interface that shows the current accelerometer and gyroscope values and a button to start-stop recording. Internally, we used Android's Sensor Manager class to access the raw data from the accelerometer and gyroscope. The app was designed to capture the raw data as fast as the hardware allows access to both the sensors. For both the accelerometer and gyroscope, the sampling rate was typically around 5Hz. Finally, the measured data for each experiment were stored locally in the smartwatch as comma-separated text files and transferred automatically to a common Wi-Fi connected Windows 10 workstation over the Android Debug Bridge (ADB) tool. This auto-transfer feature of gesture samples enables user authentication in a myriad number of scenarios from user login to computer account to a web account, for instance. Hence, the smartwatch acts more like a physical authentication key.

The smartwatch used for the data collection was Fossil Gen 4 Carlyle. This smartwatch runs on Android WearOS and was paired with an Android smartphone. For the experiments, five user-created gesture patterns have been developed. Patterns were randomly chosen, keeping them short and easy to remember. Each of these patterns was taken from a sitting position, with the smartwatch wore on the dominant hand. Table 17.1 provides descriptions for each pattern that was created. For each of these patterns, ten samples were recorded over the course of four days, with three samples taken on day one and day two each and two samples taken on day three and day four each. Over the course of four days, a total of 50 samples were recorded across all patterns. Each sample began with a calibration period of around six seconds, with gesture movement immediately following, and then a resting period of around six seconds after the gesture was completed.

Table 17.1 Pattern descriptions

Pattern number	Pattern description
1	Movement from the left shoulder down to the right hip, with two finger snaps at the end
2	Forearm movement with the beginning position referred to as center- center, up, center, left, center, down, center, up, center right, center.
3	Arm movement in the shape of a 5-point star
4	Arm movement with the beginning referred to as start, and the first position referred to as center- start, push arm forward into the center position, up, downwards and to the left, center, right, downwards and to the left, center, create a small circle ending the circle back in the center position, back to start.
5	Arm movement beginning by moving the arm up and back down to the starting position twice. When the arm returns to the starting position for the second time, move the wrist to the left and then to the right, passing over the starting position. Return to the left and repeat the movement. When the wrist reaches the right for the second time again, return to the left, and then move the entire arm upwards in a hair tucking behind the ear motion. Finish by following the motion through back to the starting position.

Fig. 17.1 The outcome of the alignment method to identify the start and end of the movement



17.3 Data Processing and Experimental Setups

17.3.1 Pre-processing of Data

As mentioned in the previous section, the movement data was collected over several days to add some randomness to the movement. A by-product of this collection process was that there was a timing offset for most of the signals collected. To compensate for this, a stationary period of six seconds was recorded at the start and end of the recorded movement. When comparing two signals, the stationary period above was used to synchronize the signals before computing the squared error between them. The alignment was done using a threshold heuristic based on standard deviation from the first

three seconds of recorded sensor data on both sensors. The starting position of the movement was identified by setting the threshold at least ten times the threshold found in the first three seconds (i.e., stationary period). A similar strategy was implemented to identify the end of the recorded movement by looking at the last three seconds of sensor data. Finally, the process was repeated across all three axes data reported by each sensor, and the final starting and ending timestamps were determined by looking at the mean values reported by every axes. In addition, the signal amplitude was normalized based on the largest amplitude measured across all axes.

An example of the alignment result can be seen in Fig. 17.1, where the amplitude of the signal measured on all three axes on the accelerometer is shown. The vertical barriers indicate the subset of the signal data that is clipped and considered during the pre-processing stage. It can be seen

Fig. 17.2 The visual similarity in recorded accelerometer signals for two samples from the same pattern

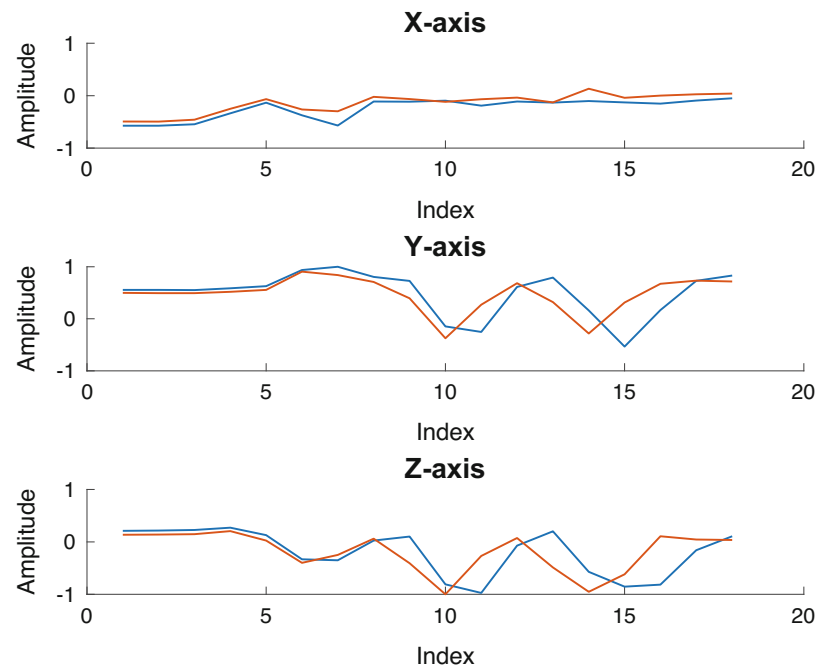
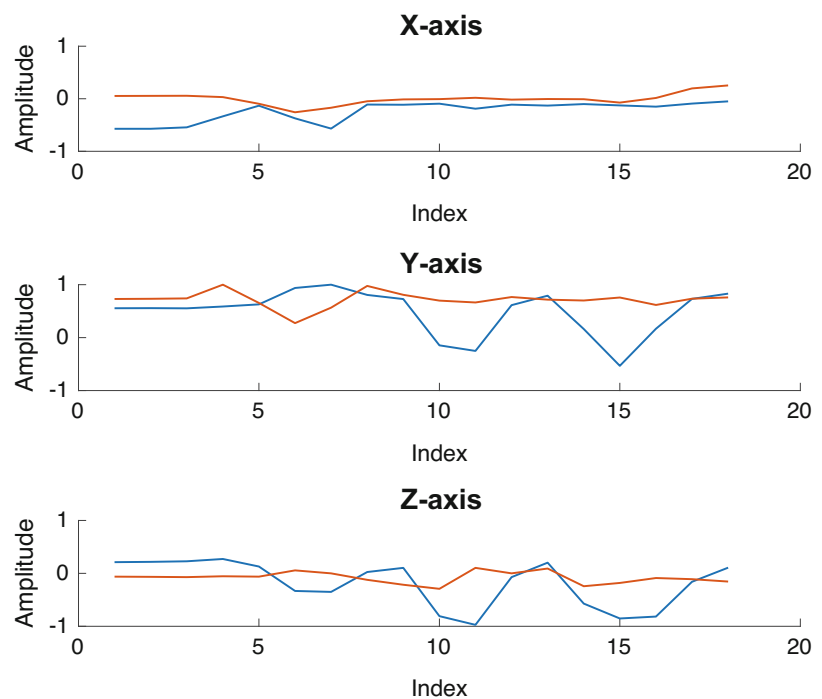


Fig. 17.3 Visual dissimilarity in recorded accelerometer signals for samples from two different pattern



that the alignment method described above works very well to identify the start and end times of the movement.

17.3.2 Error Computation

To compare the reference and the authentication/test samples, the difference between the two sensor signals was computed. First, the length of signals was considered, and the longer

of the two signals was truncated to match the length of the shorter signal. Then the pair-wise squared error between each corresponding amplitude was computed for all axes, and finally, the total difference reported by the accelerometer sensor for comparison was computed. Likewise, an identical process can be used to represent the total difference based on the gyroscope data.

As an example, Figs. 17.2 and 17.3 show the computational error when using the accelerometer data across two

sets of signal pairs. Figure 17.2 shows the aligned signals on each of the three axes when comparing a pair of signals that belong to the same pattern. Naturally, after correct alignment, the signal data follows very closely to one another, with a very low error value. On the other hand, Fig. 17.3 shows the aligned signals when looking at signal data from different patterns.

17.4 Experimental Results

Our method was evaluated based on two sets of experimental setups – (1) intra-pattern comparison and (2) inter-pattern comparison. In the case of intra-pattern comparison, for each pattern set, a comparison matrix was computed between sample R_i and sample R_j , where $i =$ samples 1, 2, and $j =$ randomly chosen three samples from samples 3 to 10. In other words, the first and second measurements for each pattern were compared against the remaining measurements within the same pattern set. This resulted in a total of 30 comparisons. In the case of inter-pattern comparisons, to compare the performance when measuring different pattern sets, a total of 20 random comparisons were generated. Each pattern was compared against at least one other different pattern to evaluate the performance of our system. Therefore, a total of 50 comparisons were generated using 50 samples. Figure 17.4 shows the distribution of errors on both the accelerometer and gyroscope across all axes.

It can be very clearly seen that the relative frequency of lower errors is much higher when computed on pairs be-

longing to the same pattern, whereas the distribution of error values for inter-pattern comparisons are relatively higher, meaning a high level of dissimilarity successfully detected. It can be concluded that our current pre-processing and computational methods work well to distinguish between patterns belonging to different datasets.

17.4.1 Additional Synchronization

While our current system already produces very good results, the algorithm can be further optimized by improving the alignment accuracy of the signals. Therefore, as a final step, we used the native alignment algorithm called `alignsignals` [20] implemented in MATLAB to further synchronize the two signals before comparing them. The `alignsignals` function uses an autocorrelation algorithm to detect if two signals are time-shifted copies of each other. The idea is that for signals belonging to the same pattern, the function should do a very good job at further aligning the signals, whereas, for signals belonging to different pattern sets, the alignment would fail to find a good synchronization since the signals do not have any correlation with each other.

The results of applying the alignment function were very encouraging. Figures 17.5 and 17.6 below shows the improvement `alignsignals` brings when comparing signals belonging to the same pattern. Figure 17.5 is based on the original alignment, while Fig. 17.6 shows the signal data comparison after both pairs have been additionally pre-processed using the above function. The improvement in synchronization is quite significant in the latter case. At the same time, it can

Fig. 17.4 Error rates of the accelerometer (left) and gyroscope (right) over three axes

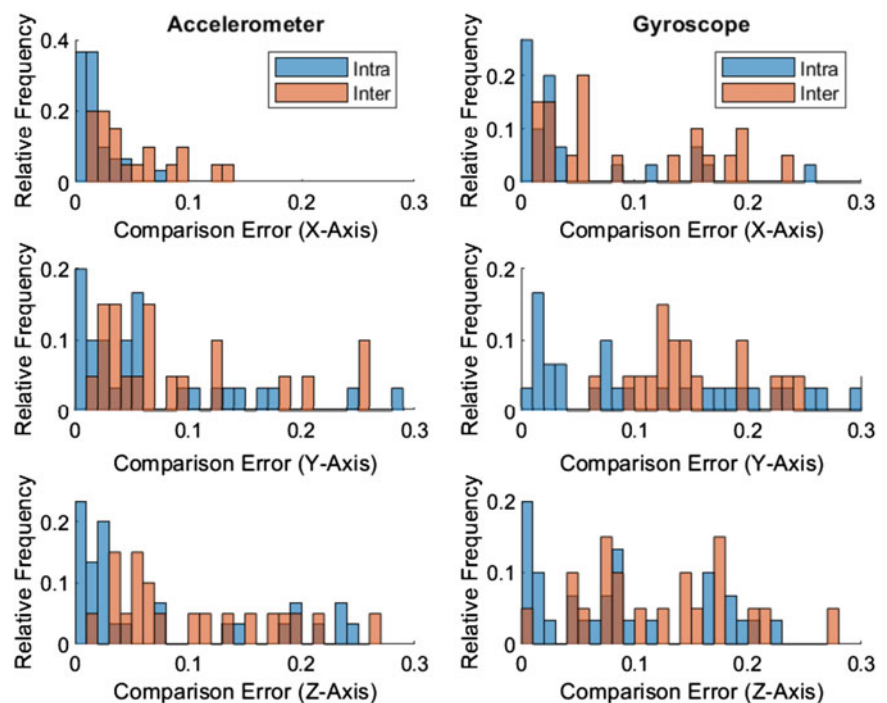


Fig. 17.5 Signal data comparison after applying the original alignment for the same pattern

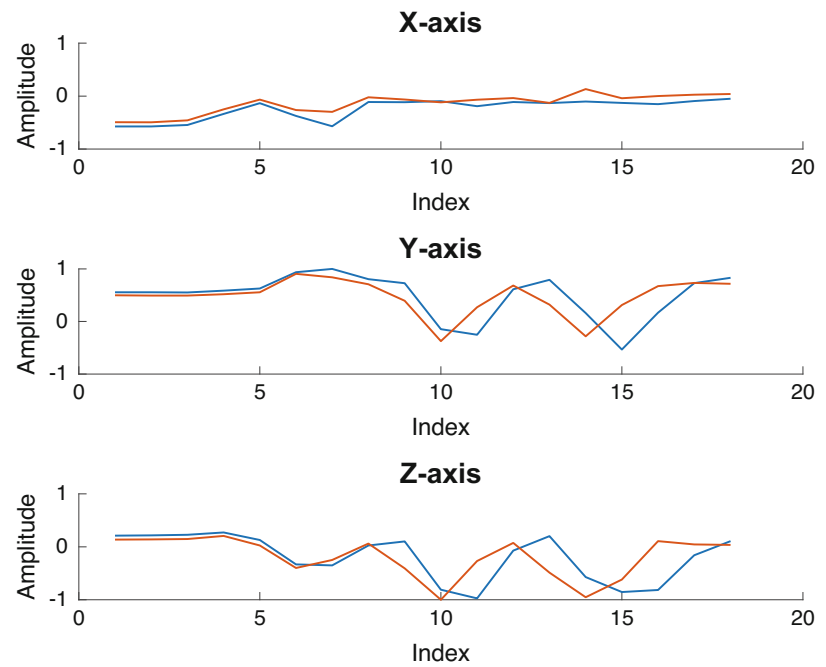
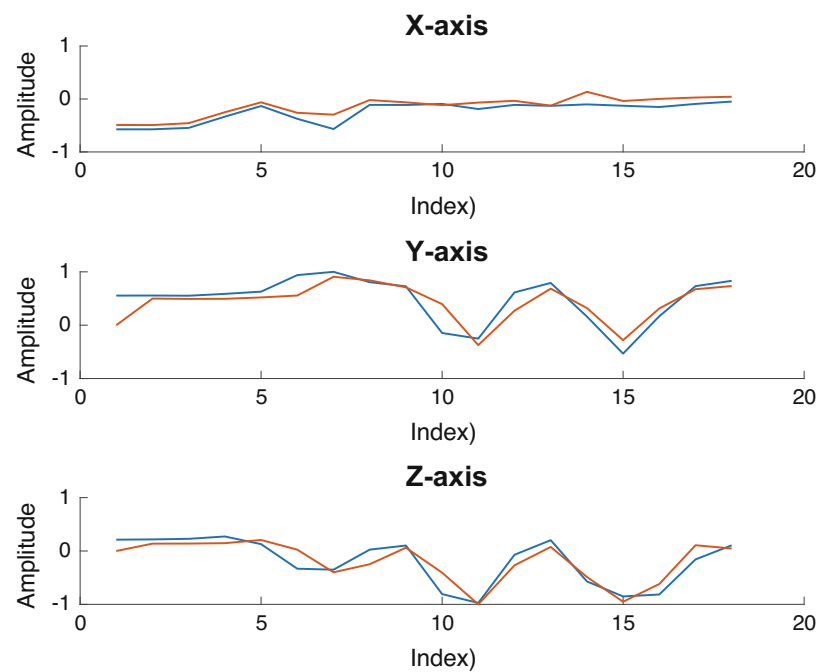


Fig. 17.6 Signal data comparison after applying the native alignment algorithm for the same pattern



also be seen from the figures Figs. 17.7 and 17.8 that signals belonging to different pattern sets are mostly unaffected by the alignsignals function, and hence, it preserves the level of dissimilarity at a higher level.

17.4.2 Highlights of the Outcomes

The experimental setup was re-evaluated using the additional synchronization method discussed above and summarized in

Fig. 17.9. As expected, there was a noticeable improvement in the overall distribution for the intra-pattern and inter-pattern comparisons compared to the errors shown in Fig. 17.4.

In summary, our current method works very well and can be combined with a threshold to establish a secure authentication scheme. Additional synchronization methods such as those described above improve the overall accuracy at the cost of higher computational complexity. Ideally, the accuracy should improve even further over time, as more

Fig. 17.7 Signal data comparison after applying the original alignment for two different patterns

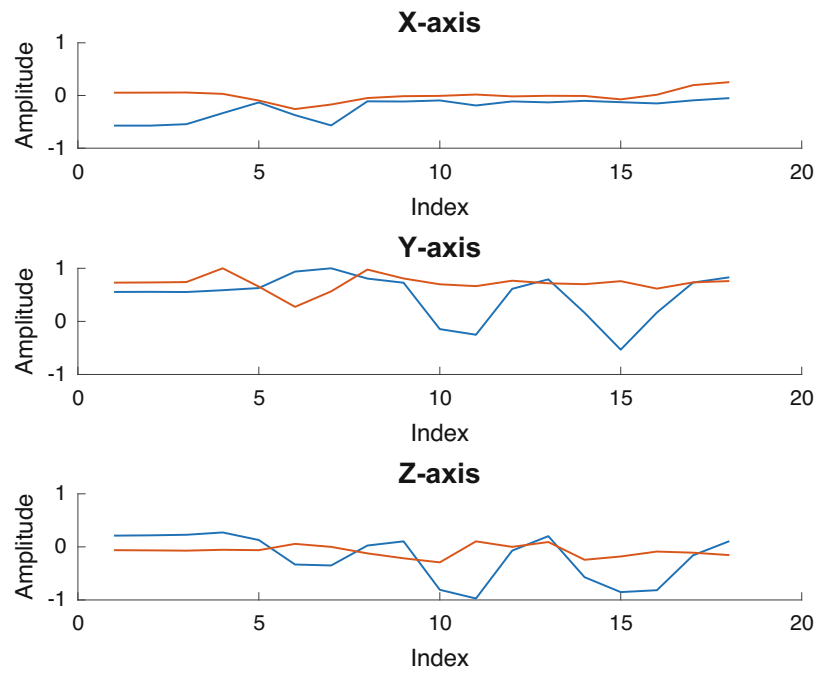
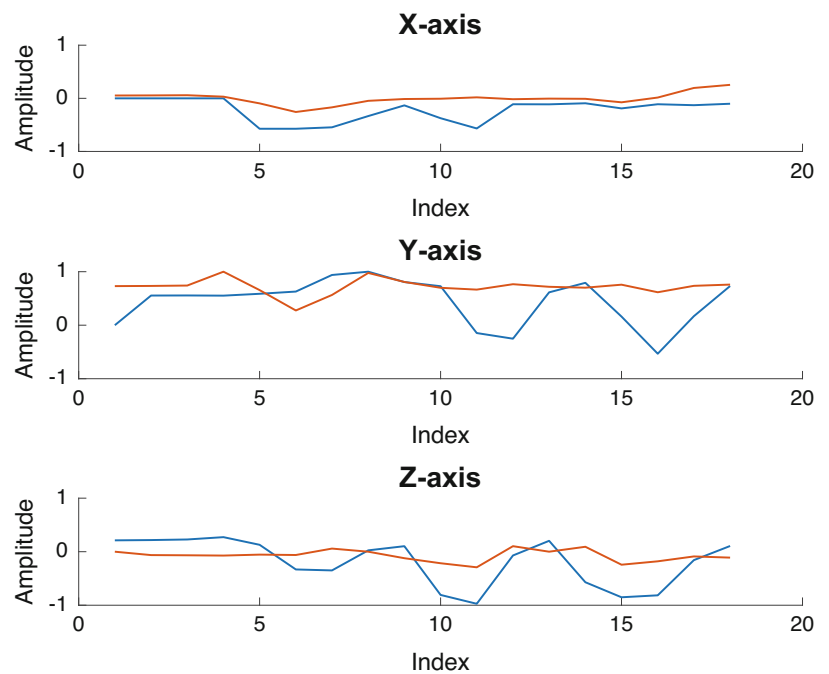


Fig. 17.8 Signal data comparison after applying the native alignment algorithm for two different patterns



measurements are recorded for each pattern, which can be logged from daily use.

Overall, our system performed very well in distinguishing signal patterns from different gestures as well as identifying patterns from the same gesture. The accuracy of the system was measured by computing the sum of the squared error between corresponding timestamps across any two patterns across all axes and sensors. Figure 17.10 below shows a comparison between the combined average error on all axes

when comparing signals belonging to the same (Intra) and different (Inter) patterns.

It can be seen that the overall error measured by our system on both sensors remains close to 0.05 in more than 80% of cases when comparing signals within the same pattern. On the other hand, more than 90% of inter-pattern comparisons successfully resulted in a higher combined error of 0.2 (or above) across both sensors. Thus, our system can be used to establish a clear threshold between different patterns, thereby providing a reliable authentication scheme.

Fig. 17.9 Error rates of the accelerometer (left) and gyroscope (right) over three axes applying the alignment function

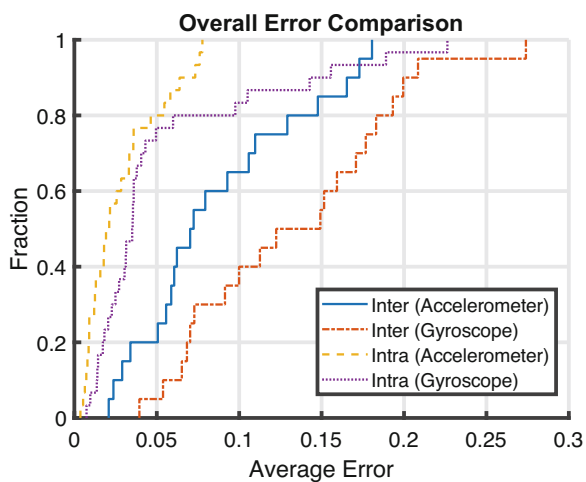
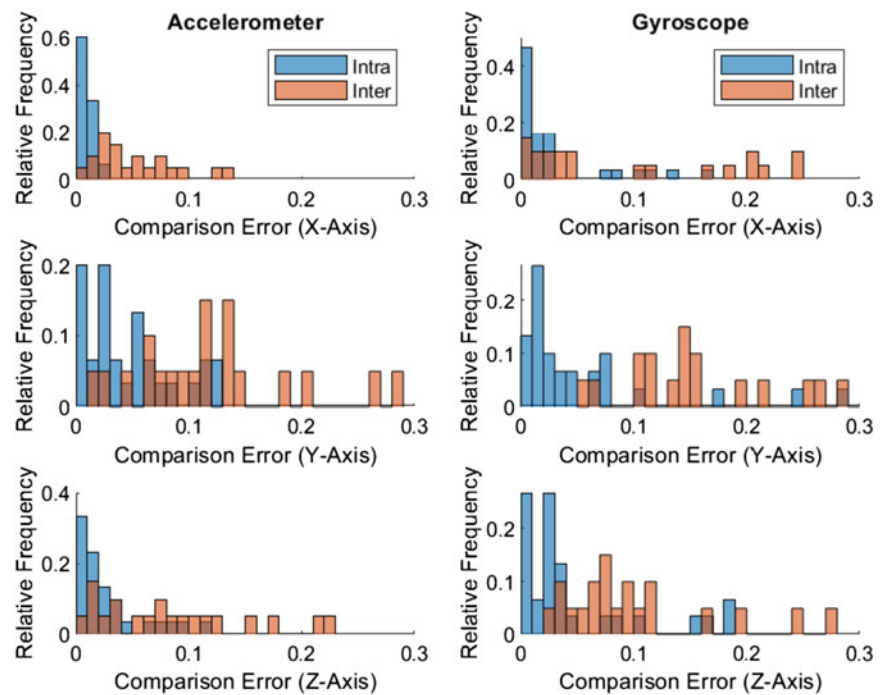


Fig. 17.10 Overall error for inter and intra-pattern comparisons

It is also noticeable that the accelerometer in general, performs better by giving off lower error rates compared to the gyroscope errors for intra-pattern comparisons. On the other hand, gyroscope performs better by giving off higher error rates for inter-pattern comparisons. Likewise, the Y and Z axes overall perform better than X-axis, especially when it comes to inter-pattern comparisons. Therefore, applying more weights on the better-performing sensors and/or axes towards decision making (accept/reject the user) would be an interesting dimension for future exploration.

17.5 Conclusion and Future Works

This work explores hand gestures as a mode for user authentication using the smart wearable. We analyzed 50 samples for five hand gesture patterns that had signal data from two sensors and three axes each. We performed a series of inter and intra pattern comparisons to demonstrate the efficacy. Our method achieved an error rate close to 0.05 for around 80% of the intra-pattern comparisons and also successfully generated higher error rates above 0.2 for more than 90% of the inter-pattern comparisons. We believe the scope for improving the accuracy (i.e., decreasing the error for intra-pattern comparison and increasing for inter-pattern comparison) is still at large. Therefore, for future works, we can regard the following for immediate improvements of our methodology – (1) apply weighted fusion methods to combine accelerometer and gyroscope comparisons together, (2) generate higher-level features such as total duration, the distance between peaks, and (3) consider sophisticated pattern matching algorithms such as Principal Component Analysis, Support Vector Machine, and Convolution Neural Network.

Acknowledgment Research reported in this publication was supported in part by funding provided by the National Aeronautics and Space Administration (NASA) under award number 80NSSC20M0124.

References

1. D. Gunetti, C. Picardi, Keystroke analysis of free text. *ACM Trans. Inf. Syst. Secur.* **8**(3), 312–347 (2005)
2. K. A. Rahman, D. Neupane, A. Zaiter, M. Hossain, Web user authentication using chosen word keystroke dynamics in *18th IEEE International Conference on Machine Learning and Applications*, USA, 2019
3. F. Monrose, A. Rubin, Authentication via keystroke dynamics, in *4th Conference on Computer and Communication Security*, USA, 1997
4. K. A. Rahman, R. Moormann, D. Dierich, M. Hossain, Continuous user verification via mouse activities, in *8th International Conference on Multimedia Communication, Services and Security*, Poland, 2015
5. C. Shen, Z. Cai, X. Guan, Y. Du, R. Maxion, User authentication through mouse dynamics. *IEEE Trans. Inf. Forensics Secur.* **8**(1), 16–30 (2013)
6. K. A. Rahman, J. Maes, How discernible user impromptu behavior is when unlocking touch screen? in *International Conference on Computer and Information Technology*, Bangladesh, 2017
7. K. A. Rahman, D. Tubbs, M. Hossain, Movement pattern based authentication for smart mobile devices in *17th International Conference on Machine Learning and Applications*, 2018, pp. 1054–1058
8. X. Xu, J. Li, Y. Lian, X. Xiao, S. Qu, X. Wang, Which one to go: Sec. and usability evaluation of mid-air gestures. *CoRR abs/1811.10168* (2018)
9. C. Xu, P. Pathak, P. Mohapatra, Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch, in *16th International Workshop on Mobile Computing Systems and Applications*, USA, 2015, pp. 9–14
10. H. Wen, J. Rojas, A. Dey, Serendipity: Finger gesture recognition using an off-the-shelf smartwatch, in *CHI Conference on Human Factors in Computing Systems*, USA, 2016, pp. 3847–3851
11. P. Zhu, H. Zhou, S. Cao, P. Yang, S. Xue, Control with gestures: A hand gesture recognition system using off-the-shelf smartwatch, in *4th International Conference on Big Data Computing and Communications*, Chicago, IL, USA, 2018, pp. 72–77
12. J. Yang, Y. Li, M. Xie, MotionAuth: Motion-based authentication for wrist worn smart devices, in *International Conference on Pervasive Computing and Communication*, USA, 2015, pp. 550–555
13. X. Yu, Z. Zhou, M. Xu, X. You, X. Li, ThumbUp: Identification and authentication by smartwatch using simple hand gestures, in *International Conference on Pervasive Computing and Communication*, USA, 2020, pp. 1–10
14. Y. Zhao, R. Gao, H. Tu, Smartwatch user authentication based on the arm-raising gesture. *Interact. Comput.* **32**(5–6), 569–580 (2020)
15. G. Li, H. Sato, Handwritten signature authentication using smartwatch motion sensors, in *44th Annual Computers, Software, and Applications Conference*, Spain, 2020, pp. 1589–1596
16. B. Chang, X. Liu, Y. Li, P. Wang, W.T. Zhu, Employing smartwatch for enhanced password authentication, in *Wireless Algorithms, Systems, and Applications*, Lecture Notes in Computer Science, vol. 10251, (Springer, 2017)
17. B. Chang, Y. Li, Q. Wang, W.T. Zhu, R.H. Deng, Making a good thing better: Enhancing password/PIN-based user authentication with smartwatch. *Cybersecurity* **1**, 7 (2018)
18. A. Lewis, Y. Li, M. Xie, Real time motion-based authentication for smartwatch, in *International Conference on Communication and Network Security*, 2016, pp. 380–381
19. Y. Li, M. Xie, Understanding secure and usable gestures for real-time motion based authentication, in *INFOCOM 2018*, pp. 13–20
20. Native Alignment Algorithm in MATLAB. URL: <https://www.mathworks.com/help/signal/ref/alignsignals.html>

Nir Drucker  and Shay Gueron 

Abstract

The AES was standardized in 2001 by NIST and has become the de facto block cipher used today. AES is a block cipher with a block size of 128 bits and is based on the proposal by Rijmen and Daemen, named “Rijndael”. The Rijndael proposal includes a definition for a block cipher with 256 bits block size (and a 256-bits key), which we call here Rijndael256. This variant has not been standardized. This paper describes software optimization methods for fast computations of Rijndael256 on modern x86-64 platforms equipped with AES-NI and with vector AES-NI instructions. We explore several implementation methods and report a speed record for Rijndael256 at 0.27 cycles per byte.

Keywords

Rijndael · Rijndael256 · AES · x86-64 · Software optimization · Wide block ciphers · AES new instructions · Vector instructions · Encryption throughput · Encryption

18.1 Introduction

Advanced Encryption Standard (AES) is the most ubiquitous symmetric cipher, and is used in many applications and scenarios. A prominent indication for its significance is the fast growing volume of AES-encrypted online data, which

N. Drucker
IBM Research, Haifa, Israel

S. Gueron (✉)
University of Haifa, Haifa, Israel

Amazon Web Services Inc., Seattle, WA, USA

is strongly supported by the industry (e.g., Intel’s AES-NI processor instructions [1, 2]).

Recently, Intel added [3] new vectorized capabilities to the existing AES-NI instructions, namely VAESENC, VAESENCLAST, VAESDEC, and VAESDECLAST (VAES* for short). These instructions are intended to further speed up the performance of AES software, to a new theoretical throughput of 0.16 C/B. Software developers can use these instructions to speed up AES modes such as AES-Counter (CTR) and AES-CBC, as well as more elaborated schemes such as AES-GCM and AES-GCM-SIV [4] (a nonce misuse resistant AEAD).

By definition, the block size of AES is 128 bits. Any 128-bit block cipher is a keyed permutation of $\{0, 1\}^{128}$. Therefore, regardless of the key length, ciphertext blocks can be distinguished (with high probability) from random strings, at the “birthday bound” of 2^{64} blocks (in reality, way below that bound, when considering some security margins). Technically, the PRP-PRF distinguishing advantage is $\frac{1}{2}q^2/2^{128}$ when viewing q encryption samples. This limits the number of encrypted blocks allowed with a given key. If we wish to keep the distinguishing advantage upper bounded by, say, 2^{-32} , the constraint $\frac{1}{2}q^2/2^{128} \leq 2^{-32}$ translates to re-keying after $\sim 2^{48}$ blocks. While 2^{48} blocks ($=2^{52}$ bytes) represent a huge amount of data, it is conceivable that current and near-future applications, especially at the cloud-scale, could benefit from lifting the required key rotation after 2^{52} bytes. This motivates the exploration of a standardized 256-bit block cipher as a useful cryptographic primitive. Indeed, with 256 bits block size (permutation of $\{0, 1\}^{256}$) the distinguishing advantage of $\frac{1}{2}q^2/2^{256}$ poses no real constraints on the allowable number of processed blocks for any imaginable real scenario, even when the adversary advantage must remain below 2^{-32} .

We point out that AES is based on the Rijndael proposal [5] that also defines a block cipher with 256 bits block size (and a 256-bits key), namely Rijndael256. However, only the

128-bit block size was standardized by the National Institute for Standards and Technology (NIST). Thus, Rijndael256 is a straightforward candidate to consider as a standardized 256-bit block cipher, provided that security analysis supports it as it supports AES (such analysis is still required, see for example [6]). Note that code that executes Rijndael256 can use the AES-NI, as shown in [2][Figure 30. “Using the AES instructions to compute a 256-bit block size RINJDAEL round”]. In 2016, Gueron and Mouha [7] reported that the throughput of code that uses AES-NI and executes pipelined Rijndael256 (encryption) are 1.54 cycles per byte (C/B hereafter). By comparison, the throughput of AES on the same processor is 0.65 C/B.

With this in mind, we focus here on software optimizations for reaching the best achievable Rijndael256 performance on modern processors.

Our Contribution We demonstrate for the first time the use of the new vector AES-NI extension to speed up the Rijndael256 block cipher in Electronic CodeBook (ECB) and CTR modes. We then measure our implementation on three different CPUs with different micro-architectures and analyze the results. Our results show that on modern CPUs the throughput of Rijndael256 is 0.27 C/B, which is comparable to the performance of the standard 128-bit block size AES.

Organization The paper is organized as follows. Section 18.2 provides some background notation. Section 18.3 details our implementations. We report our experiments setup in Sect. 18.5 and performance results in Sect. 18.6. We conclude with Sect. 18.7.

18.2 Background and Notation

18.2.1 AES

AES is defined in [8]. We describe it briefly, focusing on its encryption procedure (decryption is described in [8]). AES encryption takes a plaintext block of 128 bits and a key of size 128/192/256 bits as input and produces a 128-bit ciphertext. The 128/192/256 bits key is expanded into 10/12/14 round keys, respectively. The plaintext is whitened by XOR-ring with the (first 128 bits) of the key, and then goes through 39/47/55 back-to-back transformations that can be organized as 9/11/13 identical AES rounds and an additional final round. The j th AES round, $j=1, \dots, 9/11/13$, is the sequence of transformations

$$X = \text{SubBytes}(\text{ShiftRows}(S))$$

$$\text{AESRound} = \text{MixColumns}(X) \oplus \text{RoundKey}[j]$$

operating on the 128-bits state S , where $\text{RoundKey}[j]$ is the j th round key. The last round is the sequence

$$X = \text{SubBytes}(\text{ShiftRows}(S))$$

$$\text{AESLastRound} = X \oplus \text{RoundKey}[j]$$

where $j = 10/12/14$.

18.2.2 AES-NI

In 2009, Intel introduced six new instructions (AES-NI) to support AES computations. These instructions were and are defined in Gueron’s [1, 2] with samples and coding analysis. AESENC, AESENCLAST support encryption; AESDEC and AESDECLAST are building-blocks suitable for decryption (using the Equivalent Inverse Cipher); AESIMC and AESKEYGENASSIST support the Key Expansion. Since their introduction, the AES-NI extension has been added to most commodity processors (including AMD and ARM). Their use for fast AES encryption as well as other usages as well has become ubiquitous.

18.2.3 Vector AES-NI

Recently, Intel has added a vectorized version of the AES-NI (VAES*), as illustrated in Algorithm 1 (see [9] for details). These instructions can perform one round of AES encryption/decryption on $KL = 1/2/4$ 128-bit operands (two qwords), having both register-memory and register-register variant (we use only the latter here). The inputs are two source operands, which are 128/256/512-bit registers (named xmm, ymm, zmm, respectively), that (presumably) represent the round key and the state (plaintext/ciphertext). The special case $KL = 1$ using xmm registers degenerates to the current version of AES*. The vectorized instructions allow, theoretically, to quadruple the throughput of AES computations (compared to the xmm implementation), although the speed up may be lower on real processors, due to the execution-ports collisions when carrying out some needed shuffles and XOR operations.

18.2.4 Rijndael256

Rijndael256 [5] is a block cipher that receives as inputs a 256-bit plaintext block and a 256-bit key. It treats the plaintext as a 4×8 matrix, where the 16 plaintext bytes p_0, \dots, p_{15} are mapped to the matrix elements $a_{0,0}, a_{0,1}, \dots, a_{1,0}, \dots, a_{4,4}$ (see Fig. 18.3 for the matrix illustration).

Algorithm 1 VAES* instructions [3, 9]

```

Inputs: SRC1, SRC2 (wide registers)
Outputs: DST (a wide register)
1: procedure VAES*(SRC1, SRC2)
2:   for  $i := 0$  to  $KL - 1$  do
3:      $j = 128i$ 
4:     RoundKey[127 : 0] = SRC2[j + 127 : j]
5:     T[127 : 0] = (Inv)ShiftRows(SRC1[j + 127 : j])
6:     T[127 : 0] = (Inv)SubBytes(T[127 : 0])
7:     T[127 : 0] = (Inv)MixColumns(T[127 : 0])
                                     ▷ Only on VAESENC/VAESDEC.
8:     DST[j + 127 : j] = T[127 : 0]  $\oplus$  RoundKey[127 : 0]
9:   return DST

```

The algorithm starts by expanding the key in a process called key scheduling to generate an array of 14 256-bit round keys ks . Figure 18.1 provides our key schedule implementation in C-style pseudocode. In a typical (software) application, the key is expanded once and then applied to some input stream of multiple blocks. Therefore, its effect on the overall process is relatively small.

The Rijndael256 encryption and decryption flows are similar, and for brevity we only discuss the Rijndael256 encryption (see Algorithm 2).

The Rijndael256 rounds are similar to the AES rounds and also use the *ShiftRows*, *SubBytes*, and *MixColumns* transformations. The difference between the AES and the Rijndael256 round lies in the bytes shuffling functionality, *ShiftRows*, where the two last rows of a given state are shifted by 3 and 4 columns, instead of by 2 and 3 columns as in AES. Details are provided, below, in the implementation Sect. 18.3.

Algorithm 2 Rijndael256 block cipher [5]

```

Inputs: state A 256-bit plaintext block, ks the key schedule
Outputs: A 256-bit ciphertext block
1: procedure Rijndael256(in, ks)
2:    $state = in \oplus ks[0]$ 
3:   for  $i = 1$ ;  $i < N$ ;  $i++$  do
4:      $state = Round(state, ks[i])$ 
5:   return LastRound(state, ks[14])

6: procedure Round(state, key)
7:    $state = ShiftRows(SubBytes(State))$ 
8:   return MixColumns(state)  $\oplus$  key

9: procedure LastRound(state, key)
10:   $state = ShiftRows(SubBytes(State))$ 
11:  return  $state \oplus key$ 

```

18.3 Our Implementations

Our implementation extends Gueron’s implementation of [2] and [1] that showed how to leverage AES-NI for Rijndael256. We describe our Rijndael256 key expansion [5][Section 3.6] implementation in Fig. 18.1

The encryption round implementation of [1, 2] is based on AESENC and the Intel® AVX® architecture that includes 16 128-bit registers (*xmm*). This means that the Rijndael256 state is split across two *xmm* registers as described in Fig. 18.3 Panel (a). Applying two AESENC on these registers and in particular the *ShiftRows* transformation leads to the permutation in Panel (c). By contrast, the Rijndael256 *ShiftRows* transformation performs the permutation described in panel (d). To this end, the implementation of [2] performs a pre-permutation on the registers (panel (b)) before invoking the AESENC transformation. This is achieved through using the VPBLENDVB and PSHUFB instructions as in Fig. 18.2. Here, registers *xmm1* and *xmm2* hold the two halves of the Rijndael256 state (originally the input plaintext to encrypt), *xmm6* and *xmm7* hold the left half and right half of Rijndael256 round key, respectively. The updated (post-round) state is written into registers *xmm1* and *xmm2* and registers *xmm8* and *xmm5* hold the shuffling mask (0x03020d0c0f0e0908b0a050407060100) and the blend masks used for the pre-permutation (Fig. 18.3).

On processors that have the AVX512 extension, we can store two 32-byte states in one wide 512-bit register (*zmm*). Subsequently, we can use the VPERMB instruction to replace the two serial operations VPBLENDVB and PSHUFB with one instruction and by that reduce the register dependency. For that, we use a mask with the following eight quad-words a_0, \dots, a_7 .

$$\begin{aligned}
 a_0 &= 0x1b1a050417161100 \\
 a_1 &= 0x13120d0c1f0e0908 \\
 a_2 &= 0x0b0a151407060110 \\
 a_3 &= 0x03021d1c0f1e1918 \\
 a_4 &= 0x3b3a252437363120 \\
 a_5 &= 0x33322d2c3f2e2928 \\
 a_6 &= 0x2b2a353427262130 \\
 a_7 &= 0x23223d3c2f3e3938
 \end{aligned}$$

18.4 Rijndael256: The Best Possible Achievable Throughput

The AES algorithm (with a 128-bit key) consists of 10 AES rounds. Thus, the best AES throughput on a processor with AESENC on a single port, and with throughput 1, is $\tilde{10}/16 = 0.625$ C/B. By contrast, Rijndael256 consists of a 32-byte state and 14 rounds, where every round involves two AESENC operations (every instruction only operates on half of the state), and some extra shuffling. Consequently, the best possible throughput is at best $\tilde{2} * 14/32 = 0.875$. Better throughput is achievable when the platforms have AESENC on two execution ports. Note that the latency of the instruction only determines the number of independent blocks that we need to process in order to fill the pipeline.

Fig. 18.1 Rijndael256 key schedule in “C-style pseudocode”. Here, SLL and SHUF32 are short for `_mm_slli_si128` and `_mm_shuffle_epi32` C intrinsics, respectively

```

1 round1(in1, in2, idx, mask_ff)
2   __m128i t1 = AESKEYGEN(k, idx)
3
4   t1 = mask_ff ? SHUF32(t1, 0xff) :
5                 SHUF32(t1, 0xaa)
6
7   __m128i t2 = SLL(in1, 0x4)
8   in1 ^= t2
9   t2 = SLL(t2, 0x4)
10  in1 ^= t2
11  t2 = SLL(t2, 0x4)
12  in1 ^= t2 ^ t1
13  return *in1
14
15 intrlv_round(out[2], in[2], idx)
16   out[0] = round1(in[0], in[1], idx, 1)
17   out[1] = round1(in[1], in[0], 0, 0)
18
19 rijndael256_key_expansion(outKS, inKey)
20   vals[14] = {0x01, 0x02, 0x04, 0x08, 0x10,
21              0x20, 0x40, 0x80, 0x1b, 0x36,
22              0x6c, 0xd8, 0xab, 0x4d}
23
24   __m128i t[2] = {inKey[15:0], inKey[31:16]}
25
26   outKS[0] = t[0]
27   outKS[1] = t[1]
28
29   for(size_t i = 0; i < 14 ; i++) {
30     intrlv_round(outKS[2*(i+1)], t, vals[i])
31   }

```

Fig. 18.2 Assembly code snippet from [2] that demonstrates the use of AES-NI for computing a Rijndael256 round

```

1 vpbldv %xmm2, %xmm1, %xmm5, %xmm3
2 vpbldv %xmm1, %xmm2, %xmm5, %xmm4
3 pshufb %xmm8, %xmm3
4 pshufb %xmm8, %xmm4
5 aesenc %xmm6, %xmm3
6 aesenc %xmm7, %xmm4

```

Table 18.1 shows the latency and throughput of AESENC on different x86-64 processors [10].

Remark 18.1 (Counter Incrementation in CTR Mode) The (NIST specification) standard CTR mode is expressed in Big-Endian convention. Therefore, on Little-Endian architecture (including x86-64 and ARM), the 4 bytes of the counter need to be swapped, incremented, and swapped back. Gueron and Krasnov [11] showed an optimization that alleviates this overhead for CTR mode. Their optimization was contributed to OpenSSL and is part of the CTR implementation thereof. For simplicity, in our studies, we encoded the counter using Little-Endian convention, where counters are incremented by simply using the 32-bit Single Instruction Multiple Data (SIMD) addition.

18.5 Experimental Setup

We carried out the performance experiments on three different platforms:

- **ICL:** Dell XPS 13 7390 2-in-1 with the 10th Intel[®] Core[™] Generation (Micro architecture Codename “Ice Lake”[ICL]) Intel[®] Core[™] i7-1065G7 CPU 1.30 GHz. This platform has 16 GB RAM, 48K L1d and 32K L1i cache, 512K L2 cache, and 8MiB L3 cache. We turned off the Intel[®] Turbo Boost Technology (to work with a fixed frequency).
- **SKL:** Lenovo P1 Gen 3 (type 20TH, 20TJ) Laptop (ThinkPad)—Type 20TJ with the 8th Intel[®] Core[™] Generation (Micro architecture Codename “SkyLake”[SKL]) Intel[®] Core[™] i9-10885H CPU 2.40 GHz. This platform has 64 GB RAM, 32K L1d and 32K L1i cache, 128K L2 cache, and 2MiB L3 cache. On this platform, we could not turn off the Intel[®] Turbo Boost Technology (through the BIOS). Instead, we turned off the Hyper Threading capability and the power management feature of the CPU.
- **IVB:** Lenovo T430s Laptop (Thinkpad) with the 3rd Intel[®] Core[™] Generation (Micro architecture Codename “Ivy Bridge”[IVB]) Intel[®] Core[™] i7-3520M CPU 2.90 GHz. This platform has 8 GB RAM, 64K L1d and

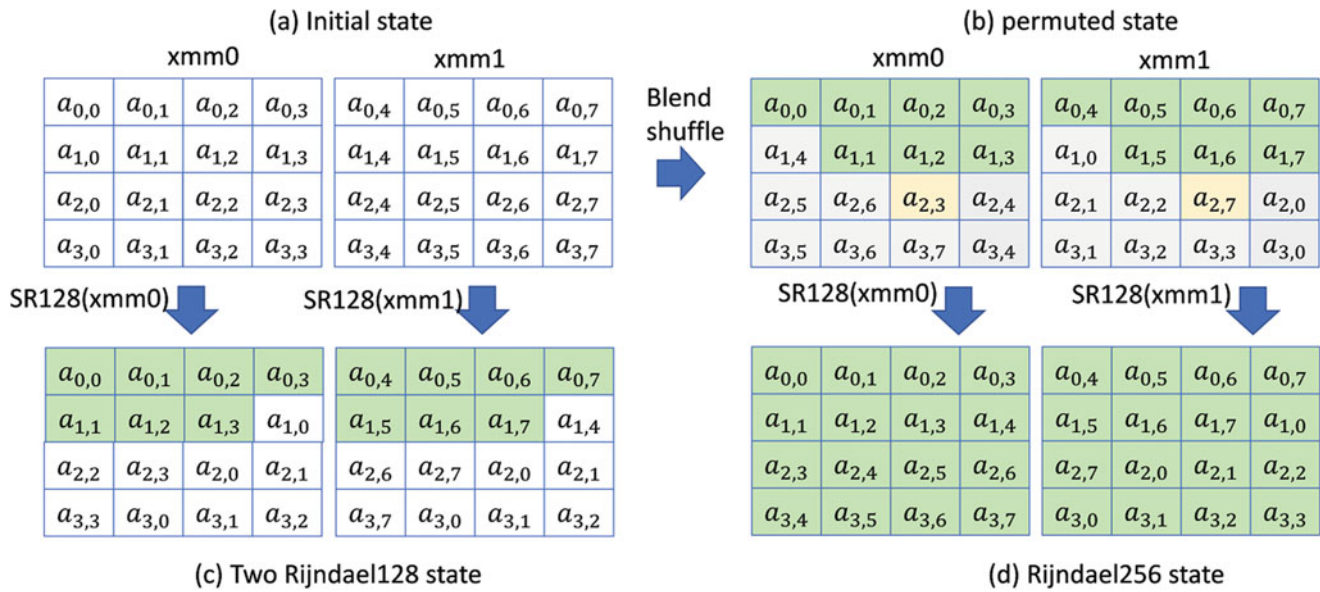


Fig. 18.3 The Rijndael256 state split to two registers $xmm0$, $xmm1$. Panel (a) is the initial state at the beginning of every round. Panel (c) show the results of permuting the two registers by applying the AES *ShiftRows* operation $SR128()$ on the two registers. Panel (d) is the

desired Rijndael256 state. Green and white cells indicate correct and incorrect permuted cells. Panel (b) shows the pre-permutation required to go from (a) to (d). Gray cells were cross lanes between the registers

Table 18.1 The latency and throughput of AESENC and AESENCLAST on different processors [10]. Note the progress through the successive processor generations

Architecture	Latency	Throughput
Ivy bridge [IVB]	8	1
Haswell [HSW]	7	1
Broadwell [BDW]	7	1
Skylake [SKL]	4	1
IceLake [ICL]	4	0.5

64K L1i cache, 256K L2 cache, and 1MiB L3 cache. We turned off the Intel[®] Turbo Boost Technology.

18.5.1 Our Optimized Code

We wrote our code in C, using x86-64 C intrinsics. The implementation uses the AVX512 and VECTOR-AES instructions. The code was compiled with both GCC (version 10) and Clang (version 12) compilers, in 64-bit mode, and with the “O3” optimization level (i.e., -O3 compilation flag), and run on a Linux OS (Ubuntu 20.04.2 LTS).

For the study, we compared several Rijndael256 implementations, encrypting data in ECB and in CTR modes. The explored code options are:

- `enc`—code based on [2] that uses AES-NI intrinsics.

- `enc_1N`—code based on [2] that uses AES-NI intrinsics, where we use the GCC/Clang unroll loops attribute (i.e., `_Pragma("GCC unroll N")` and `_Pragma("clang loop unroll_count(N)")` for GCC and Clang, respectively) with an unroll parameter N .
- `enc_xN`—encryption function that processes N (independent) blocks in parallel. The code is based on [2], where we interleave independent operations in order to get perfectly full pipelining. This is done by duplicating every original-code-row N times.
- `vaes`—Code that uses the VAES* instructions, and processes two blocks at a time.
- `vaesxN`—Code that uses the VAES* instructions and processes $2N$ (independent) blocks in parallel. The code interleaves independent operations as above.

18.5.2 Measurements Methodology

The performance measurements are reported here in cycles per byte (C/B), and the executables’ performance is profiled by using the RDTSCP instruction (per single core). Lower cycles count indicates faster execution (i.e., better performance). We used the following measurement methodology in our experiments. Every measured function was isolated, run 100 times (warm-up), followed by 1000 iterations that were measured and averaged. To minimize the effect of background tasks running on the system, we repeated every experiment 10 times and recorded the minimal result.

18.6 Performance Results

The performance results are shown in Fig. 18.4 and Tables 18.2, 18.3. on ICL, our new vectorized implementation is 5.8× faster than the non-vectorized implementation. In addition, we see that the loop unrolling attribute does not affect the final throughput. This indicates that the tested compilers cannot (automatically) fully pipeline the code. The SKL

processor has AESENC on only a single execution port. Here, we report throughput of 0.7 C/B. This is even better than our 0.875 C/B estimation, and our explanation for that is that we could not turn off the Intel Turbo feature on that platform (and hence received some extra performance boost). These results give us the performance for systems that are configured in this way.

Fig. 18.4 The performance (in C/B; lower is better) of Rijndael256 in ECB and CTR modes, measured on Ice Lake processor. The results show different code implementations (see the text for details), compiled under GCC-10 and Clang-12. Different compilers achieve different performance results on the same code. The best performance of 0.27 C/B is achieved by using vector AES, pipelining 8 blocks, and compiling with GCC-10

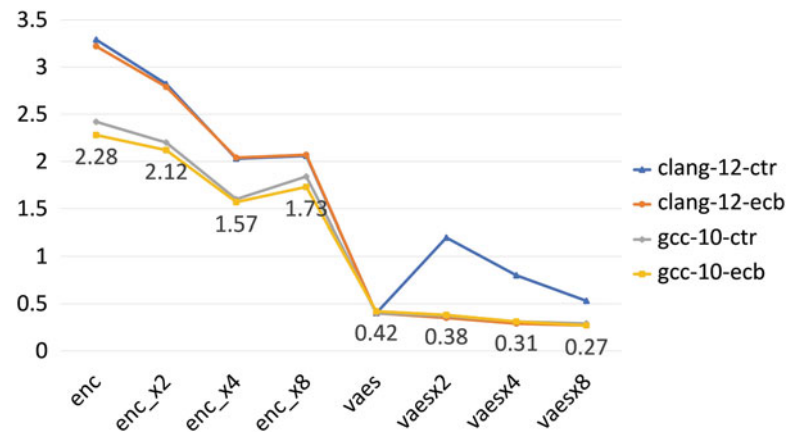


Table 18.2 The performance (in C/B; lower is better) of Rijndael256 in ECB and CTR modes, measured on ICL processor. The table provides detailed information from Fig. 18.4. The boldface numbers show the best

performance per column. The results show that aggressive parallelization with CTR mode brings the performance close to the ECB mode that is the fastest possible (though only “theoretical”) mode

Method	ECB mode			CTR mode			CTR-vs-ECB
	gcc-10	clang-12	gcc-clang speedup	gcc-10	clang-12	gcc-clang speedup	
enc	2.28	3.22	1.41	2.42	3.29	1.36	1.06
enc_12	2.26	3.2	1.42	2.42	3.28	1.36	1.07
enc_14	2.24	3.21	1.43	2.41	3.28	1.36	1.08
enc_18	2.24	3.21	1.43	2.41	3.3	1.37	1.08
enc_x2	2.12	2.79	1.32	2.2	2.82	1.28	1.04
enc_x4	1.57	2.04	1.30	1.6	2.03	1.27	1.02
enc_x8	1.73	2.07	1.20	1.84	2.06	1.12	1.06
vaes	0.42	0.4	0.95	0.4	0.4	1.00	0.95
vaesx2	0.38	0.35	0.92	0.37	1.2	3.24	0.97
vaesx4	0.31	0.29	0.94	0.31	0.8	2.58	1.00
vaesx8	0.27	0.27	1.00	0.29	0.53	1.83	1.07

Table 18.3 The performance (in C/B; lower is better) of Rijndael256 in ECB and CTR modes, measured on SKL (left) and IVB (right) processors. The boldface numbers show the best performance per column

Method	ECB mode		CTR mode		Method	ECB mode		CTR mode	
	gcc-10	clang-12	gcc-10	clang-12		gcc-10	clang-12	gcc-10	clang-12
enc	1.1	1.21	1.1	1.23	enc	4.41	4.12	4.38	4.18
enc_12	1.09	1.2	1.12	1.24	enc_12	4.35	4.08	4.27	4.21
enc_14	1.06	1.21	1.09	1.26	enc_14	4.34	4.07	4.24	4.13
enc_18	1.05	1.21	1.12	1.23	enc_18	4.31	4.06	4.25	4.09
enc_x2	0.87	0.92	0.85	0.95	enc_x2	3.36	2.96	3.43	2.97
enc_x4	0.7	0.86	0.78	0.87	enc_x4	2.62	2.46	2.67	2.33
enc_x8	0.91	0.91	0.92	0.92	enc_x8	2.39	2.23	2.59	2.23

18.7 Conclusion

As anticipated, we see that code optimizations need to take the specific processor characteristics into account. In our case, the performance is dominated by the AESENC instruction latency, its throughput, and the level of parallelization offered by the architecture (AVX, AVX512). Table 18.1 shows the information per the different processors. Other factors that affect the performance are the counter increments, shuffles, blends, XOR operations and memory access. Our code is optimized (per architecture) for these considerations as well. It is interesting to see that the compilers (GCC & Clang) behave differently when requested to automatically pipeline vectorized code (for leveraging VAESENC). In our studies, we found that GCC outperforms Clang in most cases.

In the end, we were able to write optimized code that executes Rijndael256 at the rate of 0.27 C/B (on ICL), 0.27 C/B (on SKL), and 2.23 C/B (on IVB). The speed record of 0.27 C/B on the latest processor generation (ICL) indicates that the 256-bit block cipher can be implemented in software in an extremely efficient way.

It is interesting to compare the performance for the 256-bit block cipher Rijndael256 to that of the standard AES (with 128-bit blocks). We take one case to estimate the cost of doubling the block size.

We focus on parallelizable modes, CTR in particular, non-vectorized implementation pipelining 8 blocks: AES-CTR with a 256-bit key performs at 0.58 C/B, respectively on ICL, and we compare this to AES-Rijndael256 (with a 256-bit key) that performs at 1.84 C/B on the same processor, showing a $\sim 3.3\times$ factor. The wide SIMD architectures alleviate the relative overheads of the shuffling and blending and reduce the relative performance impact of the doubled block size.

We made our code available as an open-source sample under the Apache-2.0 license and uploaded it to the GitHub repository <https://github.com/shay-gueron/Rijndael256>.

Acknowledgments This research was partly supported by: NSF-BSF Grant 2018640; The Israel Science Foundation (grant No. 3380/19); The Center for Cyber Law and Policy at the University of Haifa, in conjunction with the Israel National Cyber Bureau in the Prime Minister's Office.

References

1. S. Gueron, Intel® Advanced Encryption Standard (AES) New Instructions Set Rev. 3.01. *Intel Software Network* (2010)
2. S. Gueron, Intel's new aes instructions for enhanced performance and security, in ed. by O. Dunkelman, *Fast Software Encryption* (Springer, Berlin, 2009), pp. 51–66
3. Intel. Intel® 64 and IA-32 architectures software developer's manual, June (2021)
4. S. Gueron, A. Langley, Y. Lindell, AES-GCM-SIV: Nonce Misuse-Resistant Authenticated Encryption. RFC 8452 (2019)
5. J. Daemen, V. Rijmen, The Design of Rijndael, AES – The Advanced Encryption Standard (2002)
6. Y. Liu, Y. Shi, D. Gu, B. Dai, F. Zhao, W. Li, Z. Liu, Z. Zeng, Improved impossible differential cryptanalysis of large-block Rijndael. *Sci. China Inf. Sci.* **62**(3), 32101 (2018)
7. S. Gueron, N. Mouha, Simpira v2: A family of efficient permutations using the AES round function, in *ASIACRYPT (1)*. Lecture Notes in Computer Science, vol. 10031 (Springer, Berlin, 2016), pp. 95–125
8. NIST. FIPS pub 197: Advanced encryption standard (AES) (2001)
9. N. Drucker, S. Gueron, V. Krasnov, Making AES great again: The forthcoming vectorized AES instruction, in ed. by S. Latifi, *16th International Conference on Information Technology-New Generations (ITNG 2019)* (Springer International Publishing, Cham, 2019), pp. 37–41
10. Intel. Intel® intrinsics guide, Oct (2021)
11. S. Gueron, V. Krasnov, The fragility of AES-GCM authentication algorithm, in *ITNG* (IEEE Computer Society, Washington, 2014), pp. 333–337

Ping Wang

Abstract

Cybersecurity ethics has emerged as a new and increasingly significant area for research and education. New and complex ethical dilemmas and conflicts in values and judgments arise as new cybersecurity technologies and policies are constantly explored and implemented to defend our cyberspace for a safe and secure living and work environment. As various cyber threats, attacks, and risks pose increasing challenges to the diverse and interconnected world we live in, there is an increasing demand for quality cybersecurity education to prepare and produce qualified and ethically competent professionals to address the cybersecurity challenges. The national Centers of Academic Excellence in Cyber Defense Education (CAE-CDE) designation by the U.S. National Security Agency and Department of Homeland Security (NSA/DHS) is a high-quality program that promotes excellence in cybersecurity education for developing qualified cyber talent. Strong curriculum and courses supported by regular mentoring are essential to successful preparation of cybersecurity professionals. This research paper proposes a new credit course in cybersecurity ethics supported by an adopted comprehensive model of mentoring for an undergraduate cybersecurity education program at a CAE-CDE designated university in the United States. The curriculum proposal will present the rationale, course description, mappings of learning outcomes and topics to the CAE-CDE knowledge unit, suggested methods of assessment, and mentoring activities. The goal of this research is to contribute a new course design with ethical mentoring to

enrich and enhance national and international cybersecurity curriculum and education.

Keywords

Cybersecurity · Ethics · Curriculum · CAE-CDE · Knowledge unit · Mentoring

19.1 Introduction

We work and live in a digitally connected world with increasing and more complex cyber threats, attacks, and risks. Accordingly, there has been a substantial shortage of and increasing workforce demand for well-educated and qualified cybersecurity professionals to defend our cyber space and critical assets [1, 2]. A recent cybersecurity workforce survey study shows that the shortage of cybersecurity professionals is nearly three million across the world and about half a million in North America and the majority of organizations reported concerns of risks of cybersecurity attacks due to insufficient cybersecurity staff [3]. In addition, the cybersecurity domain is faced with increasingly complex ethical dilemmas and conflicts of values and judgement, such as security versus privacy, security versus accountability, security versus social equality and fairness, security restrictions versus civil liberties, etc. [4]. There has been a lack of awareness of and inadequate guidance on cybersecurity ethics in both research and practitioner communities [5].

Higher education providers are expected to prepare and produce qualified cybersecurity professionals to meet the workforce demand and ethical challenges in the cybersecurity field. However, a realistic challenge for cybersecurity

P. Wang (✉)
Robert Morris University, Pittsburgh, PA, USA
e-mail: wangp@rmu.edu

education is to ensure that the graduates are qualified for the expectations of cybersecurity jobs in terms of knowledge, skills, and abilities (KSAs). The U.S. National Initiative for Cybersecurity Education (NICE) published the NICE Cybersecurity Workforce Framework (NCWF) and recent revisions to provide a standard taxonomy and lexicon for specific cybersecurity work categories, job roles and tasks for the cybersecurity industry and corresponding expectations of knowledge, skills, and abilities (KSAs). NCWF serves as a guidance for employers, education and training providers, and public and private industry sectors to define cybersecurity work categories, job roles, and their required professional skills [6–8]. The NICE framework includes the knowledge of ethics related to cybersecurity and privacy as an essential professional qualification for all cybersecurity work roles [6].

A credible and rigorous quality assurance mechanism is needed to evaluate and maintain the program quality of cybersecurity education and training providers. The national Centers of Academic Excellence in Cyber Defense Education (CAE-CDE) designation program in the United States jointly sponsored by the National Security Agency (NSA) and Department of Homeland Security (DHS) has been the most comprehensive and reputable national standard for evaluating, certifying, and maintaining the quality of cybersecurity education programs. The CAE-CDE program has established specific and measurable criteria for program evaluation and assessment of cybersecurity knowledge units. The knowledge units (KUs), including the Cybersecurity Ethics KU, are a key component of the CAE-CDE designation program quality standard that applies to all cybersecurity degree programs for designation [9]. The CAE-CDE KUs and associated learning outcomes and topics are valuable reference guide for developing strong cybersecurity curriculum and courses.

Effective implementation and delivery of a strong cybersecurity program will benefit from supporting mentoring activities. Prior research has indicated that mentoring is an important factor for student academic and professional success in cybersecurity education [1, 10, 11]. A comprehensive mentoring model that includes academic mentoring, career guidance, extra-curricular mentoring, and ethical guidance is found to have contributed to student academic and professional success in cybersecurity student development [1, 12].

The following sections of this article will review relevant theoretical background on cybersecurity ethics education and mentoring, present a cybersecurity ethics curriculum proposal using the case study approach, describe and discuss the cybersecurity ethics course design and curriculum mappings, and provide conclusions and suggest future directions for follow-up research in this area.

19.2 Background

Ethics is a set of values for making judgements and decisions within a certain group, such as a cultural, social, or professional group, on acceptable lifestyle or business conduct. Ethical values may agree or conflict with other values and developments. Cybersecurity aims to protect and secure certain valued properties and assets in the digital space. However, cybersecurity technologies and policies may cause conflicts with certain ethical values in both private and public sectors.

Ethical issues are wide-ranging in the private business domain. The most frequently discussed ethical issues in business cybersecurity research literature include privacy, data protection, trust, control, accessibility, confidentiality, ethical codes of conduct, data integrity, informed consent, transparency, availability, and accountability [13]. However, additional research literature is needed to provide thorough analysis of businesses' ethical responsibilities to stakeholders in the contexts of cybersecurity incidents, such as a ransomware attack or a business data breach [13].

Ethical dilemmas and conflicts of values may also occur in the public domain cybersecurity, such as in securing critical infrastructures as part of national security. A potential conflict of values is between the privileged access or backdoors of government agencies to ICT services may jeopardize the essential values of privacy and freedom for a democratic society; For example, the use of artificial intelligence (AI) as a defensive and preventive cybersecurity technology in law enforcement, national security, and counter-terrorism may lead to faster detection, identification, and response to threats, but it may expose civilians to the threat of authoritarian government abuse of the technology in the cyber domain [14].

Cybersecurity professional organizations have adopted code of ethics to emphasize the importance of ethics in cybersecurity professional practice. For example, the International Information System Security Certification Consortium, or (ISC)², has adopted the following four mandatory Code of Ethics Canons as high-level guidance for (ISC)² certified professionals:

- Protect society, the common good, necessary public trust and confidence, and the infrastructure.
- Act honorably, honestly, justly, responsibly, and legally.
- Provide diligent and competent service to principals.
- Advance and protect the profession [15].

ISACA (Information Systems Audit and Control Association) has adopted comparable code of professional ethics for its members and certification holders [16].

Due to the ethical challenges for computing and cybersecurity technology, research in computing and cybersecurity education has also recognized the growing importance of introducing and integrating ethics education in cybersecurity curricula to train students to be aware of the ethical dilemmas and capable of making ethical judgment and decisions [17–19]. Ethical commitment to one’s organization and national security is one of the six key personal traits needed for professional success in the future cybersecurity workforce [20].

Cybersecurity education is multidisciplinary involving computing science and technology, management, communication, critical thinking, problem-solving, and analytical and decision skills [1]. Ethics component should be included and integrated into a cybersecurity curriculum for ethical judgement and decision making. The United States Naval Academy has adopted a comprehensive and interdisciplinary approach to cybersecurity education that successfully blends technical courses such as programming and networking with non-technical courses such as policy, economics, law, and ethics [21].

To complement ethics course work in the cybersecurity curriculum, ethical mentoring and guidance should be used to support the development of student ethical competency. An effective mentoring relationship between faculty mentors and student mentees will guide the student mentees toward success. Student success is defined as both academic success in completing the academic program and professional success in gaining readiness for the cybersecurity profession. The comprehensive mentoring model in Fig. 19.1 below, which includes ethical guidance, serves to help students to achieve both academic success and professional success.

The four mentoring components are interrelated and work together to contribute to students’ overall success in cybersecurity education and professional preparation. Mentoring in ethics may involve the following topics and activities:

- Ethical professional behavior
- Seminar sessions on the importance of authorization in penetration testing
- Data security and privacy laws
- Compliance and reporting requirements
- Non-disclosure agreement
- Security clearance requirements
- Conflict of interest disclosures
- Copyright protection and plagiarism
- Ethics in cybersecurity research

Mentors should also guide student mentees and make them aware of important code of ethics for major cybersecurity professional groups and societies, such as the mandatory Code of Ethics for professionals certified by (ISC)² and ISACA [15, 16].

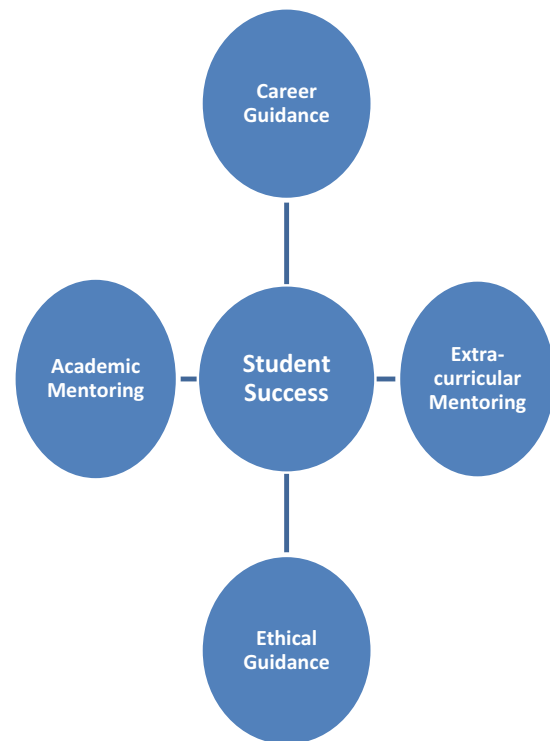


Fig. 19.1 Comprehensive mentoring model for cybersecurity education [1]

19.3 Curriculum Proposal

This research presents a sample cybersecurity curriculum proposal using the case study methodology. The case study is based on the curriculum of a Bachelor of Science (BS) degree program in cybersecurity at a CAE-CDE designated university in the northeast of the United States. The 123-credit degree program is also accredited by ABET-CAC and already has well-rounded technical and non-technical courses for a 4-year sequence, but the curriculum does not have a course in cybersecurity ethics yet. This research is to propose a 3-credit senior level seminar course in cybersecurity ethics.

19.3.1 Rationale

The course in cybersecurity ethics is necessary for graduates of the program to be aware of the complex ethical dilemmas and be able to make sound ethical judgment and decisions in the cybersecurity profession. This is in support of the ABET accredited cybersecurity program learning outcome in ethics: Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.

19.3.2 Course Description

This is a senior seminar that addresses important ethical issues, challenges, frameworks, and best practices in cybersecurity. The course work includes substantial discussions of readings and case studies to develop strong competency in ethical standards and judgment for cybersecurity decisions and behavior for service in public and private sectors.

19.3.3 Course Learning Outcomes

The course learning outcomes are mapped to the outcomes for the Cybersecurity Ethics (CSE) knowledge unit for the CAE-CDE designation, which expects students to be able to: (1) Explain how ethical foundations are applied to situations arising from the interconnected world; (2) Examine diverse ethical dilemmas; and (3) Describe the role of cybersecurity in supporting and encouraging ethics, as well as where cybersecurity practices can cause ethical conflicts [9].

19.3.4 Course Topics

The course topics are mapped to the suggested topics in the CSE KU for CAE-CDE programs, which include the following [9]:

- Ethical codes and frameworks
- Ethics and cyberspace
- Ethical issues, such as property, rights of others, availability, respect and principles of community, resource use, allocation, and abuse, and censorship
- Ethics-based decision tools
- Cybersecurity and social responsibility

19.3.5 Assessment Methods

Assessment of student learning is essential to measure students' achievement in knowledge, skills and abilities and necessary for continuous improvement of curriculum and course design. Assessment activities for the cybersecurity ethics course may utilize the following activities and assignments for evaluation:

- Case studies
- Class discussions
- Individual essay assignments
- Individual research assignment
- Team project and presentation
- Midterm and final exam

These activities will measure student knowledge and skills and abilities to analyze, evaluate and make judgement on the cybersecurity ethics topics covered in the course.

19.4 Ethical Mentoring

To support student learning in the proposed course in cybersecurity ethics, faculty mentoring of students is necessary, which is based on the Ethical Guidance component in the comprehensive mentoring model [1].

In addition to the formal coursework in Cybersecurity Ethics done in the cybersecurity program, students will receive continuous ethical training and guidance during their junior, senior, and graduate years from faculty mentors and government and business partnerships and volunteers to develop a strong ethical character and behavior for public or corporate service. The group and individual ethical guidance will include presentations and discussions led by faculty and government and corporate cybersecurity professionals on various ethical issues such as cybersecurity professional codes of conduct, privacy protection, conflict of interest, security clearances, and specific ethics policies and expectations for federal, state and local government services.

19.5 Conclusions

There is increasing demand for qualified cybersecurity workforce to defense our cyberspace from various threats and attacks. Ethical issues are increasingly significant in the cybersecurity field with complex dilemmas and challenges for cybersecurity professionals. Cybersecurity education programs are expected prepare and produce high-quality and ethical cybersecurity professionals. This research presents a case-based curriculum proposal including a cybersecurity ethics course supported by ethical mentoring for student success. The curriculum and course design is mapped to the current Cybersecurity Ethics Knowledge Unit for the U.S. national CAE-CDE designation and ABET-CAC cybersecurity program accreditation. The supplemental ethical mentoring is based on the comprehensive mentoring model from recent research in cybersecurity education.

This study is preliminary on the topic of curriculum design for cybersecurity ethics education. The proposed course design is limited to the outline of the course description, learning outcomes, topics, and assessment. More data will be available after implementations of the proposed curriculum design. Future research directions on this topic may include the impact of the coursework on student knowledge, skills and abilities on cybersecurity ethics and the effectiveness of the instructional and mentoring activities.

Acknowledgement This research was supported by a grant award from the United States Department of Defense Cyber Scholarship Program (Grant No. H98230-20-1-0352).

References

1. P. Wang, R. Sbeit, A comprehensive mentoring model for cybersecurity education, in *Advances in Intelligent Systems and Computing*, ed. by S. Latifi, (Springer Nature Switzerland AG, 2020), pp. 17–23
2. U.S. Department of Labor BLS (Bureau of Labor Statistics), (2021). Retrieved from <https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm>
3. (ISC)², *Cybersecurity Professionals Focus on Developing New Skills as Workforce Gap Widens: (ISC)² Cybersecurity Workforce Study 2018*, (2018). Retrieved from <https://www.isc2.org/research>
4. M. Christen, B. Godijn, M. Loi, The ethics of cybersecurity, in *The International Library of Ethics, Law and Technology*, vol 21, (SpringerOpen, 2020)
5. K. Macnish, J. van der Ham, Ethics in cybersecurity research and practice. *Technol. Soc.* **63**(101382), 1–10 (2020)
6. NICE (National Initiative for Cybersecurity Education), *NICE Cybersecurity Workforce Framework (SP800-181)*, (2017). Retrieved from <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-181.pdf>
7. NICE (National Initiative for Cybersecurity Education), NIST (National Institute of Standards and Technology), *NICE Cybersecurity Workforce Framework (SP800-181 Revision 1)*, (2020). Retrieved from <https://doi.org/https://doi.org/10.6028/NIST.SP.800-181r1>
8. P. Wang, M. Dawson, K.L. Williams, Improving cyber defense education through national standard alignment: Case studies. *Int. J. Hyperconnect. Internet Things* **2**(1), 12–28 (2018)
9. NIETP (National IA Education & Training Programs), *2020 CAE Cyber Defense (CAE-CD) Knowledge Units*, (2021). Retrieved from <https://public.cyber.mil/ncae-c/documents-library/>
10. K. Dean, *What everybody ought to know about mentoring in InfoSec*. AT&T Cybersecurity, (2019). Retrieved from <https://www.alienvault.com/blogs/security-essentials/what-everybody-ought-to-know-about-mentoring-in-infosec>
11. E. Dallaway, How to mentor an information security professional. *Infosecurity Magazine*, (2016). Retrieved from <https://www.infosecurity-magazine.com/blogs/how-to-mentor-an-infosec-pro>
12. P. Wang, Cybersecurity student talent recruitment and development: A case study. *Issues Inf. Syst.* **22**(2), 210–222 (2021)
13. G. Morgan, B. Godijn, A care-based stakeholder approach to ethics of cybersecurity in business, in *The Ethics of Cybersecurity. The International Library of Ethics, Law and Technology*, vol 21, ed. by M. Christen, B. Godijn, M. Loi, (SpringerOpen, 2020), pp. 119–138
14. E. Vigano, M. Loi, E. Yaghmaei, Cybersecurity of critical infrastructure, in *The Ethics of Cybersecurity. The International Library of Ethics, Law and Technology*, vol 21, ed. by M. Christen, B. Godijn, M. Loi, (SpringerOpen, 2020), pp. 157–178
15. (ISC)², *(ISC)² Code of Ethics*, (2021). Retrieved from <https://www.isc2.org/Ethics>
16. ISACA, *Code of Professional Ethics*, (2021). Retrieved from <https://www.isaca.org/credentialing/code-of-professional-ethics>
17. L. Adaryukova, O. Bychkov, M. Kateryna, A. Skyrda, The introduction of ethics into cybersecurity curricula, in *Proceedings of the 16th International Conference on ICT in Education, Research and Industrial Applications, Kharkiv, Ukraine, October 6–10, 2020*. 25–36, (2020)
18. J. Blanken-Webb, I. Palmer, S.-E. Deshaies, N.C. Burbules, R.H. Campbell, M. Bashir, A case study-based cybersecurity ethics curriculum, in *Workshop on Advances in Security Education (ASE 18)*, (USENIX) Association, Baltimore, MD, (2018)
19. Grosz et al., Embedded ethics: Integrating ethics across CS education. *Commun. ACM* **62**(8), 54–61 (2019)
20. J. Dawson, R. Thomson, The future of cybersecurity workforce: Going beyond technical skills for cyber performance. *Front. Psychol.* **9**(744), 1–12 (2018)
21. T. Emmersen, J.M. Hatfield, J. Kosseff, S.R. Orr, The USNA’s interdisciplinary approach to cybersecurity education, in *Computer*, (IEEE Computer Society, 2019 March), pp. 48–57

Performance Evaluation of Online Website Safeguarding Tools Against Phishing Attacks; a Comparative Assessment

20

Rama Al-Share, Fatima Abu-Akleek, Ahmed S. Shatnawi, and Eyad Taqieddin

Abstract

Despite the security policies that organizations follow to defend against cyber crimes, phishing attacks are still among the most popular ways the criminals use to steal user's credentials. Spear phishing, fake websites, fraudulent emails, smishing, and vishing all fall under the umbrella of phishing attacks. Recent procedures followed by many organizations tend to develop anti-phishing tools that identify fraudulent emails and websites, which are embedded either implicitly within the web browsers and email applications or explicitly as an online service. In this research, we have evaluated the effectiveness of six online checking tools in detecting potentially malicious websites. Six URL scanning engines from the best well-known engines in the VirusTotal website were tested on a list of legitimate and malicious URLs, which were collected from well-known anti-phishing frameworks, including PhishTank. In order to find the most efficient anti-phishing tool, the detection accuracy, precision, recall, and F1-Score were calculated for each engine. The results showed that among website checking tools, Sophos achieved a detection accuracy of 99.23% and a precision value of 100%.

Keywords

Phishing attacks · Anti-phishing tools · Online checkers · Malicious website detection

R. Al-Share · F. Abu-Akleek · A. S. Shatnawi (✉) · E. Taqieddin
Jordan University of Science & Technology, Irbid, Jordan
e-mail: fraalshare19@cit.just.edu.jo; fkabuaqleek18g@cit.just.edu.jo;
ahmedshatnawi@just.edu.jo; eyadtaq@just.edu.jo

20.1 Introduction

One of the most common risks that have been defined by the Open Web Application Security Project (OWASP) is the broken authentication and session management vulnerability, which comes in the second-order of the OWASP top 10 vulnerabilities since 2013. By leveraging this type of vulnerability, the attacker could steal credentials of legitimate users (i.e., usernames and passwords) or even hijack their web sessions to impersonate their identities and login to the system with valid credentials [1]. The organizations that use weak authentication and session management methods are vulnerable to this type of data breach.

Over several years, phishing attacks were classified as the highest critical fraud attacks, which fall under the broken authentication security risks. The attacker usually tries to fool some users using social engineering techniques to steal their data, including login credentials or credit card information. From a personal perspective, stealing any personal information from users readily violates their privacy and may lead to impersonation attacks. Once an attacker obtains user credentials, he can impersonate their identity to launch different cybercrimes like making illegitimate financial transactions to steal money from the user's bank accounts. From a business perspective, the reputation of organizations could be severely affected as a result of phishing attacks.

For instance, when a well-known and trusted financial institution is subjected to a certain data breach involving credit card information for one million users, all at one time, customers will lose trust in such organization: hence, affecting its productivity and reputation in the market. The organizations will lose millions of dollars for multiple years in order to heal from this financial impact.

Vishing and smishing are also popular types of phishing attacks, in which an attacker uses voice calls and SMS messages, respectively, to fool the users. In smishing, the attacker tries to send professional SMS messages to the victims convincing them to click on the given malicious link in order to validate their identity. Usually, the attacker impersonates the identity of a trusted bank or organization to deceive victims. In vishing, on the other hand, the attacker aims at tricking the users through voice calls, persuading them that they are calling from a trusted organization, so they must provide their identity number, credit card information, or any other sensitive information. This information is used by the attacker, later, to steal money or compromise the user account.

Due to the highly critical risks caused by phishing attacks, many countermeasures are taken by various organizations to defend against them, including the activation of the Two Factor Authentication (2FA) property, filtering emails and websites, and the use of anti-phishing tools. Usually, the anti-phishing tools could be used either implicitly within web browsers and email applications or explicitly as an online service. In this research, we study the effectiveness of online checking tools in detecting phishing websites. The study is performed on six online detection tools and were tested against two types of website datasets to identify phishing websites. A comparison between the determined detection tools was made opposite the detection accuracy and other evaluation metrics.

The rest of this paper is organized as follows: Sect. 20.2 describes the scenario of phishing attacks. Section 20.3 provides a literature review of existing phishing detection schemes. Section 20.4 presents the online checking tools that are presented in this research. The evaluation procedure, dataset description and performance results are revealed in

Sect. 20.5, where Sect. 20.6 provides the conclusion and future work commensurate with this research.

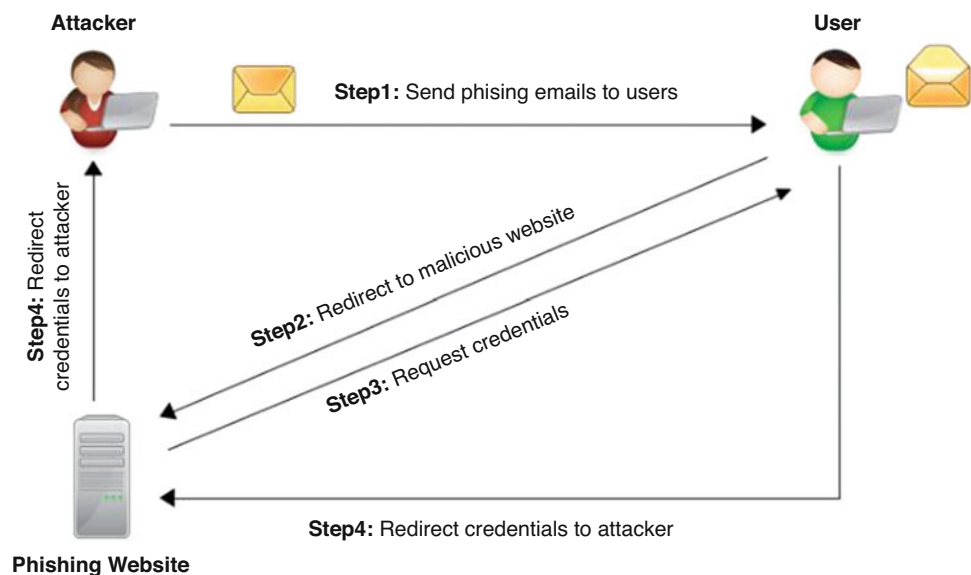
20.2 Attack Description

Generally, attackers start to launch phishing attacks by sending many fake emails to targeted or untargeted users. These phishing emails usually contain malicious URLs that lead the users to phishing websites designed specifically to steal user's credentials and access personal accounts. The most important factor that makes the operation more successful is the human factor, which is used by the attackers to collect information about users and professionally write emails that sound appealing from a user's point of view. This information could represent the user's interests, names of relatives and family members, or even the work environment the user is associated with. The whole operation of launching phishing attacks falls under the umbrella of social engineering attacks [2]. Figure 20.1 describes the attack scenario and the following stages describe specifically the operation of phishing attacks:

20.2.1 Send Phishing Emails to the Users

In the initial stage, the attackers prepare a list of phishing emails and send them to many users. Usually, at first look, the format of these emails looks interesting to the users, so they will be deceived to open the malicious URLs carried in the body of the message. At this stage, the attacker uses the information collected about the users and, thereon, applies social engineering techniques to write phishing emails.

Fig. 20.1 The scenario of phishing attacks



20.2.2 Redirect the Users to Malicious Websites

When the users click on the malicious links provided within the phishing email's content, they will be redirected to malicious websites created specifically to steal user's credentials or other secret information. These websites are carefully designed so that the users would be inclined to believe that the website is authentic.

20.2.3 Request Credentials from Users

As the legitimate website does, the malicious website requests from the users to enter their credentials, like usernames, passwords, and even credit card information. This stage is the most dangerous and undetectable action in phishing attacks, especially when the design of the website and the URL look the same as on the legitimate one. As a result, a user must be able to acquaint him/herself of the minute differences in the contents and the provided URL before releasing any secret credentials to the fake website.

20.2.4 Forward Credentials to the Attacker

Since the malicious website is created by an attacker, all the given information will be transferred to the attacking device. The attacker will be able to access the user's account and reveal any other secret information if he is successful in taking control of the username and password of the user account. Also, if he could obtain the credit card information, he will be able to transfer money from the user's bank account to his own. The actions depend on the attacker's specific aims and the activities he is associated with.

20.3 Literature Review

In this section, we present a literature review of existing detection solutions and recent survey papers which have addressed the area of phishing attack detection. \include

Authors in [3] propose a summary and classification of phishing attacks. Further, they also address certain types of phishing attacks and provide a discussion of some techniques and countermeasures to defend against them. The authors classify phishing into two categories: social engineering such as fake websites and e-mail spoofing, and technical ploy such as cross site scripting, malware phishing and session hijacking. Therein, authors have relied on two classifications in order to classify the defense solutions, the email filtering schemes and the websites phishing schemes. There are some

effective features used to discover email spoofing including body-based, header-based, URL-based and sender-based features. Pursuant with that, the authors classify email filtering into A. Network level protection, which blocks a set of known malicious domains or IP addresses as examples of it.

In order to mitigate the phishing websites authors also classify the solutions into

- A. Blacklist and whitelist where these lists include the malicious and legitimate IPS and URLs, as examples of this method Google Safe Browsing API, PhishNet: Predictive Blacklisting, and Automated Individual White-List.
- B. Heuristic Solutions which solve the problems based on previous results and experiences, examples of this method as SpoofGuard tool, Collaborative Intrusion Detection system (CIDS), PhishGuard: A Browser Plug-in, and CANTINA: A Content-Based Approach.
- C. Visual Similarities such as Visual Similarity Based Phishing Detection, and BaitAlarm.
- D. Miscellaneous Solutions that not belong to the previous categorize such as TrustBar and Dynamic Security Skin.

An empirical study was conducted in [4] to determine the main properties that are most vulnerable in webpages, to improve the selection process the authors relayed on machine learning techniques, called bagging, using the Weka program; here, the study is conducted using data of phishing and botnet malwares. Authors depend on the random tree as a classifier since it gives the best result to a high degree of accuracy; the phishing dataset include phishy behavior, suspicious behavior and legitimate behavior. While phishy behavior which aims to attack the users and steal their sensitive information to do harm, suspicious behavior is an activity which could be harmful or phishy include some malicious codes and links, and a legitimate one which is the normal case and does not contain any harmful links or codes. The authors reported on eleven features to analyze the phishing websites malware behavior some of which are IP address and port number, URL and google index, web traffic and HTTP tokens, and links pointing to pages. The dataset was analyzed, and the results were sent to an Excel sheet to make charts and draw graphs of the results. Results show that google indices are more secure to browse where just 13% of them were classified as phishy behavior, in addition to achieving an accuracy of 89% for both phishing and botnet datasets on average.

Authors in [5] propose a technique to detect phishing websites based on URL and domain name features where it can classify the websites into legitimate and fake ones. They dubbed their designed program by PhishChecker The authors set a group of rules to determine which of the websites were phished and which were legitimate; these rules were

- A. if the IP address is included within the URL then it implies a phishing webpage.
- B. if the length of the URL is greater than 54 characters then a phishing website is identified.
- C. if the URL contains dash mark (–) then a phishing website is again identified.
- D. if the URL includes @ symbol then the URL points to a phishing website.

The authors conducted an empirical study to test the proposed program. The experiment was performed on 100 different URLs; here, the results classified 68% of them as legitimate while the others as malicious with an accuracy of 96% of phishing webpage detection.

Authors in [6] propose an auto-updated whitelist to detect the phishing websites at the client side where it reports on the IP address and domain name. The normal behavior when applying the whitelists is to allow the entries involved and notify other non-included webpages whether it is deemed as legitimate or, otherwise, a phishing attack. Therefore, the idea for an auto-update is to test the webpages first by extracting the hyperlinks in the webpages and check them out, and then to decide whether to visit the website in the event that it is legitimate. This is in addition to adding it to the whitelist or, else, invoke a block action. This technique also allows for the detection of a DNS phishing attack and the Zero-hour phishing attack. In this, the system was designed using JAVA programming language, where the hyperlinks are extracted using Jsoup, where a check for the DNS domains the Google public DNS system is leveraged. The evaluation was conducted by applying an empirical study on the system by testing 1525 different webpages; the system offered an overall accuracy of 86.02%, implying a high degree of effectiveness.

Authors in [7] propose a method known as Phishidentity which relies on favicon to determine the website's identity with the help of Google image search engine which helps to use the images as a search query; here, there would be no need to using a database in addition to including the highest number of legitimate websites. Authors depend on certain features to compare the available websites with the authentic ones. The Second-Level Domain, which refers to the website name, Path in URL which refers to the location of the index.php file, the title and Snippet which refers to the text above the URL and the description under it, and the highlighted bold text (representing the text in the search results in addition to some others features) is used. These features have different weights based on their relative importance and effectiveness on the detection results then they are combined into some common frequency referred to as the final score. When this score is greater than a threshold then it is considered a phishing website. The empirical study

was performed on 10000 different websites where the results show some promising indications.

Authors in [8] propose a machine learning mechanism to detect malicious phishing emails; certain features were extracted from emails for detection purposes using Naive Bayes algorithm, followed by a set of ML algorithms as classifiers, where a final decision on the email class is examined using the decision tree algorithm. The system starts with an ML modeling which divides the input data to training and validation sets. This is used to develop the system model leveraging a Jupyter Framework Architecture. The authors built the model architecture and tested under a controlled environment. It was found that when evaluating the system model using a Random Forest algorithm a high degree of accuracy of 93.54% was reached, and when two algorithms were combined the level of accuracy attained became 96.77%.

Authors in [9] propose a deep learning framework called Conventional Neural Network Phishing Detection (CNNPD) to detect Email phishing attacks and classify them as legitimate sources or simply as phishing. In this work, this system was applied at the server side. To study and test the proposed model, the authors used different ham and phishing collected emails form PhishingCorpus and SpamAssasin public datasets. The header and body of the emails were extracted and the words were separated by spaces. Following that, deep learning is applied to the data. The deep learning first train and evaluate the input labeled dataset to build the model and get a better accuracy; then this model will be used to test the unclassified data set. CNN architecture is composed of four consecutive layers: Embedding layer, convolutional layer, pooling layer, and a fully connected layer. The email words were processed at each layer where the output of each layer is the input to the next. The output of the final layer is a classification of the email as phishing or benign. The proposed model was built in Python and the experimental test was conducted using a Google Collaboratory environment. The results yielded a 99.42% accuracy.

Authors in [10] propose a hybrid methodology to detect occurrences of phishing emails and classify them using Support Vector Machine (SVM) Probabilistic Neural Network (PNN). The proposed model architecture starts with feature extraction, feature selection to determine which features to work on. The selected features undergo a Hybrid approach which consists of both a SVM classifier and a PNN classifier to determine whether the input email is phishing or legitimate. A dataset for phishing and legitimate emails was used. The system was tested and compared with other types of classes and with static and dynamic classifier methods and the results show better performance for the proposed hybrid model with an accuracy level of 98%.

20.4 Phishing Detection Tools

Thousands of fake websites are found on the internet continuously and used to apply malicious scamming activities and steal sensitive information. The users are unaware of the real dangers that these websites can pose upon their personal lives. Many researchers have proposed several detection methods using machine learning and data mining algorithms to distinguish amongst phishing activities. The online website checkers are freely available on the internet to check for the validity of specific URLs, domains, and file contents. These usually use multiple security features that are tested against the websites to identify potentially malicious contents or behaviors.

Website checkers could be used explicitly as an online service, like the Virus Total website and Norton Safe Web, or implicitly with the web browser, itself, which includes AVAST and AVG secure browsing. In this research, we have used Virus Total (VT) to test for the effectiveness of six scanning engines on a list of legitimate and malicious URLs collected from well-known anti-phishing frameworks, including PhishTank. VT is one of the well-known websites used to check files for viruses and scan for malicious URLs using 40+ antiviruses and online URL scanners [11]. We have used the best six scanning engines from VT in this study, including Sophos, Fortinet, Avira, CLEAN MX, Kaspersky, and BitDefender.

1. Sophos: Sophos works as a real-time antivirus and web filtering engine to defend against different security threats such as Malwares, ransomwares, AI threats, and web threats [12]. Sophos uses multiple features to detect imminent threats, including Live URL filtering, documents scanning, and reputation checking of the downloaded files. For URL filtering, as such, Sophos depends on a pre-built online database that includes a set of infected websites. It proactively detects maliciously downloaded contents and reviews the reputation of downloaded files by doing deep analysis on the file's source and age. Sophos Endpoint Protection is a signature-based protection system presented by Sophos, and uses real-time threat intelligence to prevent unwanted malicious URLs, detects sudden changes, and controls flow of incoming traffic. Sophos works with different web browsers such as Google Chrome, Firefox, Edge, Internet Explorer and others.
2. Fortinet: A well-known antivirus system that protects against the newest viruses, spywares, and other malware types [13]. Its web filtering tools, including FortiGate, FortiClient, FortiCach, and FortiSandbox, all block phishing, hacked, or unwanted websites. Fortinet depends on automatic intelligent tools, periodic analysis, and continuous updates for web filtering. It uses a widely sorted and categorized list to classify websites as accepted or blocked based on threat analysis. Fortinet has an intelligent defense system, which prevents malware downloads, controls access through a policy-based control system, and blocks access to malicious websites. FortiGuard blocks over 160,000 malicious websites every minute of the day. Its deep knowledge of the threat's landscape and quick response enhances its protection efficiency [3, 4].
3. Avira: A security software that protects against different threats and cyberattacks such as web threats, identity theft, and other scamming activities [14]. Avira's threat intelligence presents file reputation, file intelligence, web reputation, and domain categorization feeds. For the file reputation feed, specific information about the hashes, sizes, and formats of malicious files is delivered every 60 seconds. More detailed information regarding related certificates, associated exploits, attack vectors, malicious procedures, malware classes, and injection impacts on the systems are fed periodically by Avira's File Intelligence Feed. For web reputation feeds, the URLs, timestamps, hashes, categories of malicious URLs are all fed and monitored. Avira blocks the known spamming, phishing, or malicious websites based on the information stored in its database. The URLs are classified into safe, malware, spam, PUA, phishing, or other types based on a real-time and site-specific query approach by using the web reputation API [5].
4. CLEAN MX: Is one of the sources that contain a large dataset of malicious verified URLs and malware sites [15].
5. Kaspersky: A cybersecurity company that uses deep threat intelligence and machine learning to present security solutions to businesses and users [16]. It protects small, medium, and large size enterprises from different cyber threats. Kaspersky Internet Security (KIS) is a product that offers protection against malwares, spams, and phishing activities. In the initial versions of this product, malware and spyware detection was done through Firewalls and Anti-Spy modules. Later, it added the sandboxing feature to its environment to run the applications on virtual machines. The URL adviser is a feature by KIS that enables the users to check websites and classify them into malicious and legitimate websites based on its database system. It color-marks these websites based on their severity level and degree of danger.
6. BitDefender: A security software that uses advanced Artificial Intelligence (AI) to detect and block the newest types of threats [17]. An Anti-Phishing protection system is used to detect online scams associated with stealing of financial credentials like credit card information and passwords. Safe online banking enables users to make electronic payments safely by securing their transactions through a dedicated browser. This solution could prevent financial frauds while doing electronic shopping. It also

uses the VPN to encrypt the internet traffic and protect banking information from sniffers. In Bitdefender, the URL Status Service identifies malicious, phishing, and fraudulent URLs and IP addresses in real-time. Bitdefender Cloud is used to scan websites without installing a scanning agent, which saves on memory and the associated processing of client's devices.

20.5 Evaluation Procedure

In order to evaluate the performance of online scanners, a list of legitimate and malicious URLs was taken from multiple data sources, including PhishTank and SophosLabs. The chosen scanning engines were tested on the collected URL datasets, and a performance evaluation was done using the confusion matrix. A detailed description of the selected datasets and evaluation metrics are listed in the following subsections

20.5.1 Dataset Description

We have applied the empirical study of URL scanning on two existing datasets collected from different sources. For malicious URLs, a dataset of 3500 websites was chosen from the sources [18, 19]. From source [18], datasets of malicious websites collected from the PhishTank data source, an open-source repository, where users submit suspicious websites continuously to the system since 2006. Additionally, we used a recent dataset that contains a list of phishing URLs collected during the period of COVID-19 pandemic by SophosLabs [19]. It has been noticed recently that the words "coronavirus" and "COVID19" were used frequently in various malicious domain names to launch phishing, malware, and spamming activities [20]. For legitimate URLs, a dataset of 3500 legitimate websites was chosen from the top 100K legitimate URLs collected in [18].

20.5.2 Evaluation Metrics

In order to evaluate the performance of online checkers, we first calculated the True positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values, which are represented in Eqs. 20.1, 20.2, 20.3, and 20.4.

Where N_p denote the total number of phishing pages, N_L denote the total number of legitimate pages, $T P$: denote the number of malicious websites correctly detected by the scanner as malicious to the total number of malicious websites, $F P$ denote the total number of legitimate websites wrongly detected by the scanner as malicious to the total number of legitimate websites, $T N$ denote the total number of legitimate websites correctly detected by the scanner as legitimate to

the total number of legitimate websites, and $F N$ denote the total number of legitimate websites correctly detected by the scanner as legitimate to the total number of legitimate websites.

$$TP = \frac{N_p \rightarrow P}{N_p} 100\% \quad (20.1)$$

$$FP = \frac{N_L \rightarrow P}{N_L} 100\% \quad (20.2)$$

$$TN = \frac{N_L \rightarrow L}{N_L} 100\% \quad (20.3)$$

$$FN = \frac{N_p \rightarrow L}{N_p} 100\% \quad (20.4)$$

From a security perspective, having large false negatives is more dangerous than having false positives, which severely degrades the defense lines for businesses. In contrast, having large false positives might not be that dangerous, but it could become pretty dangerous when users ignore phishing alerts and consider them as normal events. However, in order to keep organizations more immune to any phishing activities, the detection tools should have low false positive and negative rates.

Additionally, we have calculated the Precision, Recall, F1-score, and Accuracy (ACC) for all scanning tools represented in Eqs. 20.5, 20.6, 20.7, and 20.8. Where P denote the ratio of correctly detected phishing websites to all instances detected as phishing, R denote the ratio of correctly detected phishing websites to all instances detected as phishing or legitimate from all malicious websites, $F 1$ Score denote the harmonic mean between Precision and Recall, and ACC denote the ratio of correctly detected phishing from all inputs.

$$P = \frac{TP}{TP + FP} 100\% \quad (20.5)$$

$$R = \frac{TP}{TP + FN} 100\% \quad (20.6)$$

$$F1 = \frac{2PR}{P + R} 100\% \quad (20.7)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} 100\% \quad (20.8)$$

Table 20.1 Performance evaluation results for website checking tools

Scanning engine	Performance metric							
	TP	FP	TN	FN	Precision	Recall	F1-score	Accuracy
Sophos	98.46%	0%	100%	1.54%	100%	98.46%	99.22%	99.23%
Fortinet	59.74%	0.14%	99.86%	40.26%	99.76%	59.74%	74.73%	79.80%
Avira	77.94%	0%	100%	22.06%	100%	77.94%	87.6%	88.97%
CILEAN MX	61.29%	0.17%	99.83%	38.71%	99.72%	61.29%	75.92%	80.56%
Kaspersky	95.63%	0%	100%	4.37%	100%	95.63%	97.77%	97.81%
BitDefender	82.60%	0%	100%	17.40%	100%	82.6%	90.47%	91.30%

20.6 Testing Results and Discussion

Here, we present the results and outcomes of testing various URL scanning engines that we described earlier against the selected datasets. Table 20.1 presents the evaluation results of the six best URL scanners observed from VT website in terms of the evaluation metrics described previously. We can see from the results that Sophos outperformed other scanning engines with 99.23% accuracy and 100% precision values, which means that most of the websites were detected correctly by this scanning engine with no false positive values. Also, it achieved the highest recall value of 98.46, which indicates that the percentage of false negatives is very low compared with other scanning engines. Kaspersky comes in second place with an accuracy of 97.81%, followed by Bit-Defender (91.30%), Avira (88.97%), CLEAN MX (80.56%), and Fortinet (79.8%). All of the scanning engines achieved high precision values with more than 99% rate; Hence, almost all of them demonstrated very low false positive values. However, recall values fell in the range between 59% and 98%, depending on the false negative rates achieved by each engine.

Compared with the empirical study made in [21] for detecting phishing websites, the scanning engines that we selected in our study achieved better detection performance in terms of detection accuracy. In [21], Anti-Phishing was the best detection tool for detecting malicious and legitimate websites with 94.32% accuracy, while Sophos achieved the best accuracy value with 99.23% in our study.

20.7 Conclusion and Future Work

Phishing attacks are among the most dangerous attacks that deceive users or employees in order to steal their personal information and impersonate their identities. Not only individuals are affected by these attacks, but also the organizations involved. Any data breach in their systems can affect their reputation and productivity in the market. Many organizations tend to use anti-phishing tools that identify malicious websites and spam emails to defend against phishing. In this empirical study, we evaluated the effectiveness of online

website checkers for detecting malicious URLs. Six scanning engines from the well-known security software packages were tested on 7000 malicious and legitimate URLs. The reputations of IP addresses, DNS servers, and website contents were used as parameters for detecting the malicious URLs. The results showed that Sophos achieved the best detection accuracy of 99.23% and 100% precision value. Also, all the tested engines achieved an accuracy above 50%.

References

1. K.L. Chiew, K.S.C. Yong, C.L. Tan, A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Syst. Appl.* **106**, 1–20 (2018)
2. S. Gupta, A. Singhal, A. Kapoor. A literature survey on social engineering attacks: Phishing attack, in *2016 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 2016
3. B.B. Gupta, N.A.G. Arachchilage, K.E. Psannis, Defending against phishing attacks: Taxonomy of methods, current issues and future directions. *Telecommun. Syst.* **67**(2), 247–267 (2018)
4. A.F. Alwaghid, N.I. Sarkar, Exploring malware behavior of web-pages using machine learning technique: An empirical study. *Electronics* **9**(6), 1033 (2020)
5. A.A. Ahmed, N.A. Abdullah, Real time detection of phishing websites, in *2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEM-CON)*. IEEE, 2016
6. A.K. Jain, B.B. Gupta, A novel approach to protect against phishing attacks at client side using auto-updated white-list. *EURASIP J. Inf. Secur.* **2016**(1), 9 (2016)
7. K.L. Chiew et al., Leverage website favicon to detect phishing websites. *Secur. Commun. Netw.* **2018**, 1 (2018)
8. B. Espinoza, et al., Phishing attack detection: A solution based on the typical machine learning modeling cycle, in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2019
9. R. Alotaibi, I. Al-Turaiki, F. Alakeel, Mitigating email phishing attacks using convolutional neural networks, in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, 2020
10. A. Kumar, J.M. Chatterjee, V.G. Díaz, A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing. *Int. J. Electr. Comput. Eng.* **10**(1), 486 (2020)
11. VirusTotal (2012) [Online]. Available at <https://www.virustotal.com/gui/>. Last accessed 2021
12. Sophos (1985) [Online]. Available at <https://www.sophos.com/en-us.aspx>. Last accessed 2021
13. Fortinet (2009) [Online]. Available at <https://www.fortinet.com/>. Last accessed 2021

14. Avira (1986) [Online]. Available at <https://www.avira.com/>. Last accessed 2021
15. Clean-mx [Online]. Available at <http://support.clean-mx.de/cleanmx/viruses>
16. Kaspersky (2005) [Online]. Available at <https://me.kaspersky.com/>. Last accessed 2021
17. Bitdefender (2001) [Online]. Available at <https://www.bitdefender.com/>. Last accessed 2021
18. ebubekirbbr, 2019, phishing url detection, GitHub. <https://github.com/ebubekirbbr/phishing>. URL detection
19. SophosLabs, 2020, covid-iocs/malicious URL, GitHub. [https://github.com/sophoslabs/covid-iocs/blob/master/malicious URL](https://github.com/sophoslabs/covid-iocs/blob/master/malicious%20URL)
20. S. Gallagher, A. Brandt, *Facing Down the Myriad Threats Tied to COVID-19* [Online] (2020). Available at <https://news.sophos.com/en-us/2020/04/14/covidmalware/>. Last accessed 2021
21. H. Sharma, E. Meenakshi, S.K. Bhatia, A comparative analysis and awareness survey of phishing detection tools, in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2017

Part IV

Blockchain Technology

Dina Shehada, Maryam Amour, Suadad Muammar, and Amjad Gawanmeh

Abstract

Ensuring trust between Internet of Things (IoT) devices is crucial to ensure the quality and the functionality of the system. However, with the dynamism and distributed nature of IoT systems, finding a solution that not only provides trust among IoT systems but is also suitable to their nature of operation is considered a challenge. In recent years, Blockchain technology has attracted significant scientific interest in research areas such as IoT. A Blockchain is a distributed ledger capable of maintaining an immutable log of transactions happening in a network. Blockchain is seen as the missing link towards building a truly decentralized and secure environment for the IoT. This paper gives a taxonomy and a side by side comparison of the state of the art methods securing IoT systems with Blockchain technology. The taxonomy aims to evaluate the methods with respect to security functions, suitability to IoT, viability, main features, and limitations.

Keywords

Evaluation · Social networks · Security attacks · QoS · Overhead · Privacy · Scalability · Taxonomy · Access control · Structure

devices to provide smart services to the users. It is noticeable that IoT is used by different industries to support a wide diversity of applications such as manufacturing, logistics, food industry, healthcare, utilities, smart homes, smart cities, smart grids, smart retail, etc. [1, 2]. For example, in healthcare systems, IoT is used to assess and recognize patients afflicted with chronic illnesses [3]. IoT devices monitor and collect the physiological statuses of patients through sensors then send them to be analyzed to come up with the best diagnoses and inform physicians. Another example where IoT is highly beneficial is in supply chain management, where a product often consists of multiple parts provided by different manufacturers across countries.

IoT is made up of interconnected “smart things” that are embedded with sensors, software, and other technologies to connect and exchange data with other devices and systems over the Internet. IoT enabling technologies are grouped into three categories; the technologies that enable “things” to acquire contextual information, process contextual information, and improve security and privacy. While the first two categories can be considered as functional building blocks required for building intelligence into things, the third category is a de facto requirement to severely eliminate penetration of IoT systems [2].

21.1 Introduction

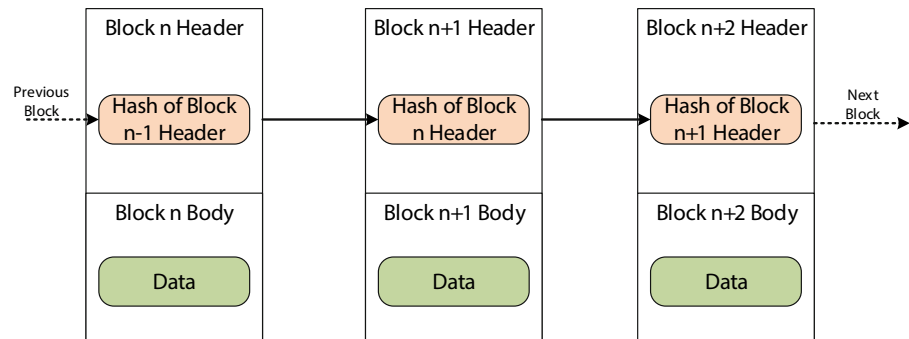
The Internet of Things (IoT) has become one of the most remarkable implemented technologies of the twenty-first century. The aim of IoT is to inter-connect all smart physical

D. Shehada (✉) · M. Amour · S. Muammar · A. Gawanmeh
 College of Engineering and IT, University of Dubai, Dubai, UAE
 e-mail: dshehada@ud.ac.ae; maamour@ud.ac.ae;
 smuammar@ud.ac.ae; agawanmeh@ud.ac.ae

21.1.1 Problem Description and Motivation

The intrinsic features of IoT cause several challenges mainly; decentralization, poor interoperability, privacy, and security vulnerabilities [1]. IoT systems are built with heterogeneous hardware and networking technologies, as a result, connecting them to the applications to extract and analyze large amounts of data is a complex task [4]. Moreover, trust management plays an important role in IoT to achieve trustworthy data collection, exchange, storage, and sharing. Therefore, there is a need to ensure trust among IoT devices. Many

Fig. 21.1 The structure of a Blockchain



researchers proposed mechanisms to ensure trust and security within different networks and applications. Some of these mechanisms include providing authentication, authorization, and trust evaluation between IoT devices. However, these trust mechanisms do not always fit the scale and diversity of the IoT, where there is no common root of trust and information extraction and processing is complex.

A decentralized architecture is crucial to enable autonomy and flexibility in the system. Therefore, to provide the best possible services in any IoT application, there is a great demand to find an efficient solution that ensures trust and privacy, suitable to the nature of IoT environments and of course feasible. Many researchers are currently interested in Blockchain technology as it provides optimization elements to overcome trust and integrity issues. Due to its promising features, the Blockchain technology attracted many researchers to integrate it into many applications including IoT applications as it is considered a promising solution that can address the challenges of IoT [1].

In this paper, we surveyed the integration of Blockchain technology within IoT and proposed a taxonomy and provided a side by side comparison of the state of the art methods to assess Blockchain based techniques in IoT from trust perspective. The comparison assesses the techniques with respect to the security functions, suitability to IoT, viability, main features, and limitations.

The rest of the paper is structured as follows. Section 21.2 discusses the background and related works in the literature. Section 21.3 explains the proposed taxonomy in detail. Section 21.4 explains how the proposed taxonomy is used to assess, analyze and compare the related works. Finally, conclusions and future work remarks are presented in Section 21.5.

21.2 Background and Related Works

21.2.1 Background

Blockchain has gained more interest in IoT applications since it brings several opportunities and presents several solutions to address existing challenges in IoT [5]. Blockchain is

a decentralized system that maintains data integrity using several consensus mechanisms [6]. It consists of several consecutively connected blocks where each block, except the first, points to its immediately previous block via an inverse reference, called the hash value of the parent block [1]. Blocks in a Blockchain containing details of transactions that have occurred within the network. The transaction information can be regarding token transfers occurring in a network, or any manner of data exchange. Each block is logically divided into two parts, namely, the header and the body. Transactions and important data are stored within the body of the block, while the header contains, among other fields, the identifier of the previous block [7]. The blocks are connected in a chain, similar to a linked list, as shown in Fig. 21.1. Transactions are verified and validated by entities in the network, then packaged into blocks and chained within the Blockchain network according to an established consensus mechanism. The fact that Blockchains use hash functions to chain blocks to ensure immutability, and encryption and digital signatures to secure information, makes Blockchain secure and tamperproof [6].

21.2.2 Related Works

In this section, we review some of the most recent related works that have addressed security issues of IoT systems by incorporating the Blockchain technology. The paper in [8] proposed the use of Blockchain to secure and ensure integrity of datasets sharing with two conceptual Blockchain approaches. A Reference integrity Metric (RIM) is used to make sure that downloaded sets are trusted according to the metrics and a central hub is responsible for dataset distribution in addition to keeping track of system members. Both the system membership information i.e. ID, address, policy, etc. and the RIM are maintained using Blockchain. The membership information is publicly available to all members making privacy a major concern.

Authors in [9] proposed a Blockchain Based Trust Management (BBTS) to maintain trust among network entities in smart cities applications such as the Internet of Vehicles

(IoV). The proposed multi-tier architecture authenticates devices and manages trust calculation and distribution using a decentralized structure that is expected to minimize delays. Road Side Units (RSUs) are devices distributed in the smart city that also have great power and fast network connections. Platoons are groups of mobile nodes with limited power and network connections. Nodes within a platoon exchange messages about their experience and trust value of a specific node. After the messages are validated by them, the updated trust value is added to the platoon Blockchain. As nodes leave one platoon into another they share the previous platoon Blockchain with the new platoon's RSU. Then, the RSU shares the new blocks with other RSUs to make sure that they all have a synchronized copy of the global Blockchain. Having nodes calculate the trust value provide a decentralized structure that ensures dynamism and flexibility of the system. However, this can cause overhead and a high computational cost on the system. Moreover, ensuring the consistency of the RSUs Blockchain is difficult due to the fact that the system is very dynamic. This also makes approximating the time needed for sharing and validating information difficult.

In [10] the proposed system is made up of different manufacturing zones. Each zone contains devices and resources such as sensors, IoT devices, in addition to authentication and trust managers (servers). Authentication managers, authenticate and authorize devices, while trust managers are responsible for calculating the final trust values of devices. Moreover, a set of miners also exist to verify trust data, create the new block and broadcast it to be added to the Blockchain network. The system has a partially decentralized structure as trust is calculated by trust managers [11, 12], moreover, the Blockchain is only shared with verified devices that were granted permission, so a private Blockchain is used.

Authors in [13] proposed a Blockchain based trust system for IoT to enhance the communication and dynamism in IoT systems. The system classifies IoT devices into three labels, lightweight, full, and coordination nodes depending on their computing capacity, storage, and connection time. Blockchain is used to manage all the unique IDs assigned to devices. Full nodes are responsible for assigning valid IDs to the devices while coordination nodes assign full nodes to lightweight nodes that newly joined the system and need to be verified and assigned a valid ID. According to the authors, it is difficult for attackers to fake IDs as the system maintains an ID management system through Blockchain. However, the authors did not explain how the global public Blockchain is maintained and synchronized to remain consistent.

Authors in [14] proposed a Blockchain based system to create end to end trust between IoT devices without relying on the root of trust. The Blockchain uses cryptographic primitives to create a credit based Blockchain with reputation. IoT devices can perform transactions using that credit, and their ability to pay back their credit adds to their reputation. Trans-

actions are validated by consumer end devices by comparing them to some set terms of use (TERMs), then are publicly recorded in the Blockchain by a powerful back end server making the system have low latency. The proposal is expected to overcome the delays that usual Blockchains have [15, 16].

Authors in [6] proposed a partially decentralized authentication and access control system for IoT, based on multi agents systems (MASs) and Blockchain. The system is supposed to provide access control and secure communication between IoT devices, fogs and cloud. The system is made up of a hierarchical Blockchain architecture made up of four layers of Blockchain managers which are local, fog, core fog, and the cloud. Managers are equipped with high resources and responsible for controlling all communications and managing the Blockchain of their layer. The proposed framework in the latter study relies on multiple private Blockchain which improves privacy but affects scalability. Also, the proposal has not been implemented in a real-world environment [17, 18].

Authors in [19] proposed a decentralized trust management system for IoV using Blockchain. The Blockchain is used to ensure the security and integrity of trust values of the different IoT devices. Blockchain blocks are managed and saved at the fogs of the system where they are the only entities that can add new blocks to the chain. The system uses a public Blockchain that is distributed to all IoT devices and is updated only upon request. Distributed data are equipped with a timestamp and are updated after a proper timeout. Though this approach may reduce the overhead in the system, it also raises a concern about setting the proper value for timeout as it can affect the validity and trust of the whole system [20].

A decentralized interoperable trust framework for health-care IoT (IoHT) is proposed in [3] to guarantee authentication and improve the trustworthiness between nodes and edges in the IoT system. Trusted zones are created after verifying the identity of the devices through cryptography and a public Blockchain. Once trusted zones are created, transactions requesting patient information are validated through a private ripple Blockchain hosted on the Health edge layer. Unlike the public Blockchains, the ripple Blockchain requires that the primary node among the group to grant access and permission to other nodes in order for them to be able to access information and write to the Blockchain [21].

An architecture for IoT combining fog computing and Blockchain based social networks is proposed in [22] to ensure the security of IoT environments. The proposal uses public Blockchain to make it easier for fogs and devices to record identities of devices and manage the relationship and trust between them. Whenever any device joins the system, it sends its information to the fogs where they are verified and recorded as a transaction. After that, the request is broadcast to miners (fogs) in the entire peer to peer network, once the

transaction is verified it is added to the public Blockchain. The system helps to establish trust in environments, by incorporating authentication and authorization mechanisms that improve the performance and security of the traditional IoT system [23, 24].

Another Blockchain based trust framework for IoT called IoT passport is proposed in [25]. The Blockchain based framework provides a trust management system among IoT devices. In addition to access control and permissions, all interactions among devices are signed by participants and recorded in the local Blockchain of the different trust domains. Later, a Blockchain node from each trust domain is chosen to synchronize and form the global public Blockchain. However, only a conceptual model is presented without any concrete implementation. Moreover, some issues such as how the global trust is formed or how Blockchain nodes are chosen are not discussed [24, 26].

The authors in [27] presented another Blockchain based trust evaluation for IoT devices, where reported histories stored in a Blockchain are used to compute the trust scores for devices. The trust score is evaluated using the proposed trust assessment where properties and capabilities set by the device manufacturer are used to find deviations and reports. Deviations define the behaviors that do not follow the capabilities and reports are the behavior feedback monitoring done by network controllers. Intelligent software

network based controllers store this information in a global Blockchain and then analyzers use it to compute the trust scores of devices and update the controllers to ensure that devices participating meet the trust level expectation. However, the proposal lacks concrete implementation and does not explain how the controllers and analyzers work together and use the information to evaluate the IoT devices [12].

In the next section, a taxonomy scheme is proposed to investigate the suitability of all the reviewed Blockchain solutions proposed in the literature.

21.3 Proposed Taxonomy Scheme

In this section, we propose a taxonomy scheme to assess the different reviewed papers and evaluate how suitable they are to IoT systems. The reviewed literature is analyzed thoroughly and classified based on the proposed taxonomy scheme. Figure 21.2, shows the proposed taxonomy and its classifications.

The taxonomy evaluates the works in the literature according to five main categories; security functions, suitability to IoT environments, the proposed solution feasibility, its main features, and limitations. Next, we explain in detail each category and how it is used to assess the literature:

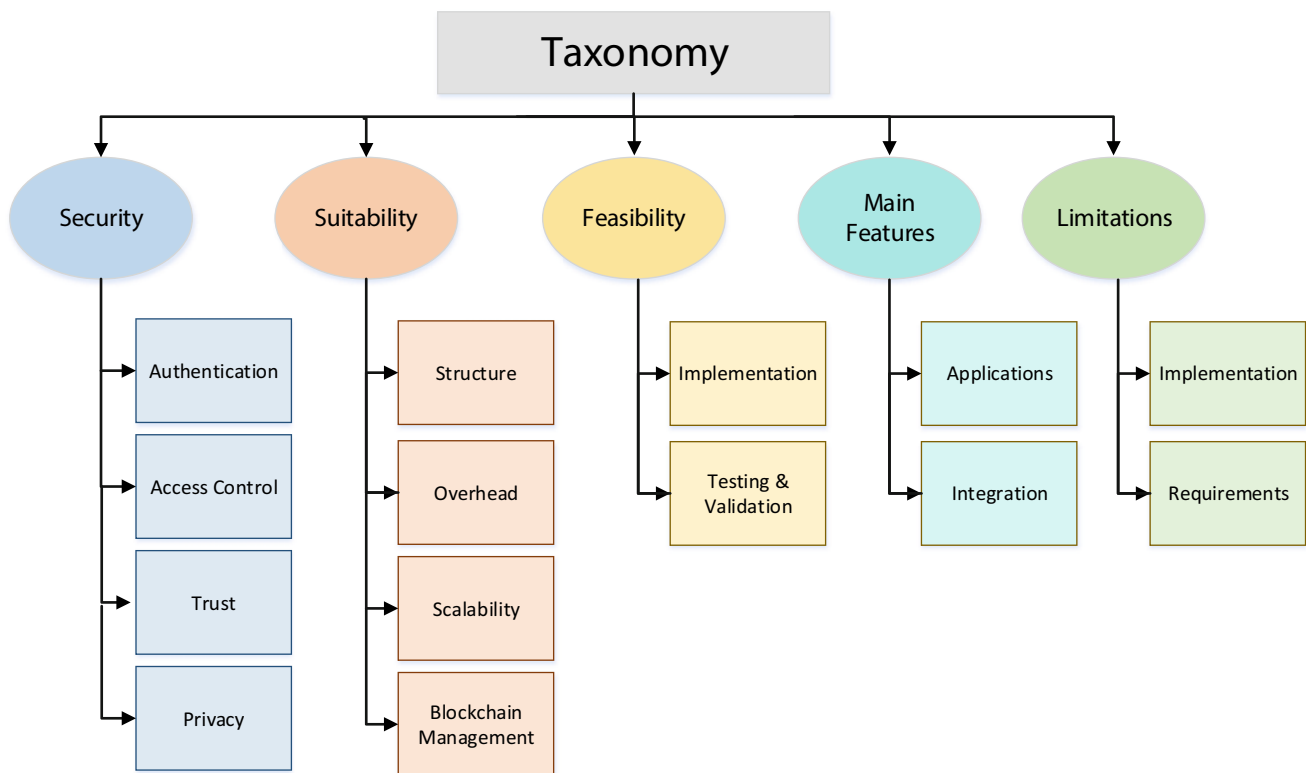


Fig. 21.2 Proposed Taxonomy for classifications and analysis of existing solutions

- The Security function: this item defines the different security requirements it provides. The main security functions provided by the works in the literature are authentication, access control, trust evaluation, and privacy. Privacy depends on the type of the Blockchain used whether it is public or private. Unlike public Blockchains, private Blockchains ensure the privacy of the end users by controlling which device can access the Blockchain. Moreover, although a public Blockchain brings numerous benefits, it can cause delays and overheads when utilizing a large amount of computational resources.
- Suitability: this item evaluates how suitable the proposed solution is to IoT system, with respect to:
 - Structure: describes whether the Blockchain is centralized, decentralized, or partially decentralized. A centralized structure affects the availability of IoT systems and suffers from the problem of a single point failure therefore not recommended for IoT, while a decentralized structure is suitable to the nature and functionality of IoT.
 - Overhead: Identifies if the solution has a high overhead that can overwhelm the system and cause delays. This is why it is important to consider and evaluate the amount of computations and communication overhead caused by the different proposed solutions.
 - Scalability: evaluating how easy it is for the system to expand. The structure of the proposed solution has a great impact on the scalability. For example, relying on the cloud makes it difficult to expand the system, while having distributed fogs responsible for registering and monitoring devices makes it easier and faster.
 - Blockchain management: defines which entity/entities have the responsibility for managing the Blockchain, its storage, and its distribution. This feature is also important as it has a direct impact on the system overhead and security and therefore on the suitability to IoT systems.
- Feasibility: this item identifies the feasibility of implementing and verifying the proposed solution.
 - Implementation: shows the feasibility of the solution in terms of implementation and deployment through practical Blockchain.
 - Testing and validation: shows the validation of the solution throughout simulation or prototype.
- Features: listing the main features that make the proposed solution unique.
 - Application: shows the type of applications proposed by the approach.
 - Integration: shows integration with other architectures such as fog and edge computing.
- Limitations: Listing any limitations that govern the proposed solution.
 - Implementation: shows limitations related to implementing the solution.
 - Requirements: shows requirements and assumptions under which the proposed solution works.

The proposed taxonomy is used to assess [3, 6, 8–10, 13, 14, 19, 22, 25, 27] from the literature. Details of the assessment are explained in the next section.

21.4 Assessment of Blockchain Based Trust Approaches in IoT

Most of the proposed works use Blockchain to save information and policies to provide access control such as [8, 13, 14]. Some provide both authentication and access control such as [3, 6, 22, 25]. While other works also save the computed trust score for entities in the Blockchain to provide trust calculation and authentication [9, 10] or access control [19, 27].

All the proposed works have a public Blockchain scheme except [6, 10] that use a private permissioned Blockchain to ensure privacy. While [3] proposed a solution that uses both public Blockchain in addition to a private ripple Blockchain. The public Blockchain is used to ensure authentication to devices, while the ripple is used to enforce authorization and access control.

The framework in [8], has a centralized structure as it relies on a single central hub to manage the Blockchain. Furthermore, the frameworks in [9, 13, 19, 22, 25, 27] have a fully decentralized structure as they use distributed fogs, nodes or devices to complete the Blockchain management and distribution. On the other hand, the rest of the reviewed frameworks [3, 6, 10, 14] are considered to have a partially decentralized structure due to their reliance on multiple but specific entities to manage the Blockchain such as agents, edge devices or servers.

When it comes to the overhead, it is important to have a lightweight trust solution to suit the nature of the IoT systems and make their expansion easier. The frameworks proposed in [6, 10, 14, 19, 22] managed to provide that by giving the system access to powerful devices or servers that can handle any computations. On the other hand, [3, 25, 27] are considered to have a relatively higher overhead.

In [3] the nature of the application requires a lot of communication between devices as any request or permission need to be managed by the primary nodes and the Health edge. In [25], heavy communications need to take place between the nodes to synchronize the Blockchain between the different trust domains. Additionally, [27], relies on controllers and analyzers to evaluate trust and generate reports. The rest are considered to have an average overhead as they either rely on broadcasting blocks to IoT devices for verification [8] or give them the responsibility for creating, sharing,

and managing the Blockchain [9] that can be overwhelming to such devices with limited power and storage. Also, in [13] the process of nodes classification, and assignment of coordination nodes to other nodes can be overwhelming on the system.

As for scalability, [3, 9, 10, 13, 14, 19, 22, 25, 27] are considered scalable solutions that are easy to expand to add more IoT devices due to the easy deployment and management of the Blockchain structure. Banerjee et al. [8], Lahbib et al. [10] have medium scalability due to the centralized structure. The framework in [6], is not considered to be scalable, as its multi layer scheme, in addition to the private Blockchain used and the issue of the big header size in their proposed framework make the system expandability complicated.

According to the proposed work in the literature the Blockchain management is done by different entities. For example, in [19, 22] the Blockchain management is conducted by the Fogs. Fogs add blocks to the Blockchain and are responsible for its distribution and storage. In [8] the Central Hub does most of the management while in [9], both nodes

and RSUs do. Full nodes manage and store the Blockchain in [13]. A powerful service provider back end server manages the Blockchain in [14]. In [6] each layer manages its Blockchain by a specific agent called Blockchain manager. In some cases, the Blockchain management is conducted by more than one entity as in the case of [3] where both the Health edge and the primary nodes carry out the task together. In [25, 27], Blockchain storage and management is done by Blockchain nodes and controllers, respectively. While in [10] miners and the trust server take care of that.

The work in [19] was implemented through Hyperledger Fabric tool. Also [3, 10] conducted a simulation using NS3 simulation and C# & JSON - LD semantic tools. While the work in [22], was validated by building a prototype using SE Linux.

Table 21.1 summarizes the assessment of the related works based on the proposed comparison scheme. Moreover, the main features, challenges, and limitations s are also highlighted in Table 21.2.

Table 21.1 Assessment of Blockchain based trust approaches in IoT

Framework	Security		Suitability to IoT				Feasibility
	Function	Blockchain type	Structure	Overhead	Scalability	Blockchain management	Implementation and testing
Banerjee et al. [8]	Access control	Public	Centralized	Average	Medium	Central Hub	No
Kandah et al. [9]	Authentication and trust score evaluation	Public	Decentralized	Average	High	Nodes & RSU	No
Lahbib et al. [10]	Authentication and trust score evaluation	Private	Partially decentralized	Low	Medium	Trust servers and miners	Simulation
Qiu et al. [13]	Access control	Public	Decentralized	Average	High	Full nodes	No
Di Pietro et al. [14]	Access control	Public	Partially decentralized	Low	High	Service provider backend server	No
Algarni et al. [6]	Authentication and Access control	Private	Partially decentralized	Low	Low	Blockchain manager at each layer	No
Cinque et al. [19]	Access control and trust score evaluation	Public	Decentralized	Low	High	Fogs	Implementation
Abou-Nassar et al. [3]	Authentication and Access control	Public and Private	Partially decentralized	High	High	Health Edge and primary nodes	Simulation
Zhu and Badr [22]	Authentication and Access control	Public	Decentralized	Low	High	Fogs	Prototype
Tang et al. [25]	Authentication and access control	Public	Decentralized	High	High	Blockchain nodes	No
Boussard et al. [27]	Access control and trust score evaluation	Public	Decentralized	High	High	Controllers	No

Table 21.2 Summary of features, challenges, and limitations of proposed Blockchain based trust approaches in IoT

Framework	Features	Challenges and limitations
Banerjee et al. [8]	<ul style="list-style-type: none"> o Secure sharing of IoT datasets o Ensures dataset integrity 	<ul style="list-style-type: none"> o Deciding time of data sharing o Cannot avoid privacy violations
Kandah et al. [9]	<ul style="list-style-type: none"> o Provides trust among IoT entities in smart cities o Single bad mouther has a minimal effect o Protection from colluding attacks if less than 33% of the vehicles are malicious 	<ul style="list-style-type: none"> o Privacy issue o Trustworthiness of connected devices need to be evaluated o Ensuring consistency among platoons by RSU o Approximating time of message sharing
Lahbib et al. [10]	<ul style="list-style-type: none"> o Blockchain based trust architecture o Collects trust evidences and gives trust scores for IoT devices o Resilient to Tamperproof, Bad Mouthing, On Off and Ballot Stuffing attacks 	Assumes that miners can all be trusted
Qiu et al. [13]	<ul style="list-style-type: none"> o Dynamic and scalable architecture for IoT o Classifies nodes based on their connectivity, computing and storage 	<ul style="list-style-type: none"> o How public Blockchain is formed by the different full nodes o Trusting full nodes
Di Pietro et al. [14]	<ul style="list-style-type: none"> o Creates endtrust between devices without relying on the root of trust o Credit based Blockchain o Low latency 	<ul style="list-style-type: none"> o Does not specify how terms are published o Privacy issues
Algarni et al. [6]	Provides secure access control based on MAS and Blockchain	<ul style="list-style-type: none"> o Delays caused due to the big header size in the proposed Blockchain structure o Difficult to implement
Cinque et al. [19]	<ul style="list-style-type: none"> o Secures the trust management of IoV o Ensures security and integrity of trust value score 	<ul style="list-style-type: none"> o Setting a suitable timeout time to update value o Does not provide protection against attacks
Abou-Nassar et al. [3]	<ul style="list-style-type: none"> o Interoperable trust framework using ripple Blockchain o Provides privacy while not limiting access to a single person 	<ul style="list-style-type: none"> o Relies on primary nodes to allow access to Blockchain information o High overhead
Zhu and Badr [22]	<ul style="list-style-type: none"> o Hybrid architecture o Fog computing enabled o Blockchain based social networks o Distributed public ledgers for recording identities 	Privacy issues
Tang et al. [25]	<ul style="list-style-type: none"> o Blockchain based trust frameworks o Cross platform collaboration between IoT devices o Interactions are recorded on the Blockchain o Interactions are utilized in an incentive plan 	<ul style="list-style-type: none"> o Performance of contemporary Blockchain infrastructure is limited o There is a need for other mechanisms to ensure synchronization of Blockchains
Boussard et al. [27]	<ul style="list-style-type: none"> o Blockchain based trust evaluation o Reported history is analyzed using simple risk assessment o Provides a trust score that is calculated using capabilities, deviations and reports 	<ul style="list-style-type: none"> o Assumes a baseline of expected network behaviour o Can be used to monitor and report IoT devices o Automated (dis)connections can be perceived as unjustified and authoritative

21.5 Conclusion

The issues related to the security and privacy of the IoT systems are immense and require careful consideration. In order to effectively counteract these security and trust issues, this research proposes a taxonomy derived from the synergy of Blockchain and trust in IoT. The taxonomy is used to assess the existing Blockchain based trust approaches in the literature. A comprehensive review of the approaches along with a detailed analysis about their security features, suitability to decentralized IoT communities, and feasibility is provided through a side by side comparison. We also highlight the main features and limitations to help the reader in effectively evaluating the different proposed approaches.

References

1. H.-N. Dai, Z. Zheng, Y. Zhang, Blockchain for internet of things: a survey. *IEEE Int. Things J.* **6**(5), 8076–8094 (2019)
2. O. Vermesan, P. Friess, et al., *Internet of Things—from Research and Innovation to Market Deployment*, vol. 29 (River Publishers, Aalborg, 2014)
3. E.M. Abou-Nassar, A.M. Iliyasu, P.M. El-Kafrawy, O.-Y. Song, A.K. Bashir, A.A. Abd El-Latif, Ditrust chain: towards blockchain-based trust models for sustainable healthcare iot systems. *IEEE Access* **8**, 111223–111238 (2020)
4. C. Sobin, A survey on architecture, protocols and challenges in IoT. *Wireless Personal Commun.* **112**(3), 1383–1429 (2020)
5. M.A. Uddin, A. Stranieri, I. Gondal, V. Balasubramanian, A survey on the adoption of blockchain in IoT: challenges and solutions. *Blockchain: Res. Appl.* **2**, 100006 (2021)
6. S. Algarni, F. Eassa, K. Almarhabi, A. Almalaise, E. Albassam, K. Alsubhi, M. Yamin, Blockchain-based secured access control in an IoT system. *Appl. Sci.* **11**(4), 1772 (2021)
7. M.S. Ali, M. Vecchio, M. Pincheira, K. Dolui, F. Antonelli, M.H. Rehmani, Applications of blockchains in the internet of things: a comprehensive survey. *IEEE Commun. Surv. Tutorials* **21**(2), 1676–1717 (2018)
8. M. Banerjee, J. Lee, K.-K.R. Choo, A blockchain future for internet of things security: a position paper. *Digital Commun. Netw.* **4**(3), 149–160 (2018). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352864817302900>
9. F. Kandah, B. Huber, A. Skjellum, A. Altarawneh, A blockchain-based trust management approach for connected autonomous vehicles in smart cities, in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)* (2019), pp. 0544–0549
10. A. Lahbib, K. Toumi, A. Laouiti, A. Laube, S. Martin, Blockchain based trust management mechanism for IoT, in *2019 IEEE Wireless Communications and Networking Conference (WCNC)* (IEEE, Piscataway, 2019), pp. 1–8
11. A. Sharma, E.S. Pilli, A.P. Mazumdar, P. Gera, Towards trustworthy internet of things: a survey on trust management applications and schemes. *Comput. Commun.* **160**, 475–493 (2020)
12. B. Shala, U. Trick, A. Lehmann, B. Ghita, S. Shiaeles, Blockchain and trust for secure, end-user-based and decentralized IoT service provision. *IEEE Access* **8**, 119961–119979 (2020)
13. H. Qiu, M. Qiu, G. Memmi, Z. Ming, M. Liu, A dynamic scalable blockchain based communication architecture for IoT, in *International Conference on Smart Blockchain* (Springer, Berlin, 2018), pp. 159–166
14. R. Di Pietro, X. Salleras, M. Signorini, E. Waisbard, A blockchain-based trust system for the internet of things, in *Proceedings of the 23rd ACM on Symposium on Access Control Models and Technologies* (2018), pp. 77–83
15. M.S. Ali, M. Vecchio, M. Pincheira, K. Dolui, F. Antonelli, M.H. Rehmani, Applications of blockchains in the internet of things: a comprehensive survey. *IEEE Commun. Surv. Tutorials* **21**(2), 1676–1717 (2018)
16. J. Huang, L. Kong, G. Chen, M.-Y. Wu, X. Liu, P. Zeng, Towards secure industrial IoT: blockchain system with credit-based consensus mechanism. *IEEE Trans. Ind. Inf.* **15**(6), 3680–3689 (2019)
17. B. Alouffi, M. Hasnain, A. Alharbi, W. Alosaimi, H. Alyami, M. Ayaz, A systematic literature review on cloud computing security: threats and mitigation strategies. *IEEE Access* **9**, 57792–57807 (2021)
18. S. Pal, A. Dorri, R. Jurdak, Blockchain for IoT access control: Recent trends and future research directions (2021). arXiv Preprint arXiv:2106.04808
19. M. Cinque, C. Esposito, S. Russo, O. Tamburis, Blockchain-empowered decentralised trust management for the internet of vehicles security. *Comput. Electri. Eng.* **86**, 106722 (2020)
20. C. Esposito, M. Ficco, B.B. Gupta, Blockchain-based authentication and authorization for smart city applications. *Inf. Proc. Manag.* **58**(2), 102468 (2021)
21. M.A. Rahman, M.S. Hossain, M.S. Islam, N.A. Alrajeh, G. Muhammad, Secure and provenance enhanced internet of health things framework: a blockchain managed federated learning approach. *IEEE Access* **8**, 205071–205087 (2020)
22. X. Zhu, Y. Badr, Fog computing security architecture for the internet of things using blockchain-based social networks, in *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (IEEE, Piscataway, 2018), pp. 1361–1366
23. H. Wang, L. Wang, Z. Zhou, X. Tao, G. Pau, F. Arena, Blockchain-based resource allocation model in fog computing. *Appl. Sci.* **9**(24), 5538 (2019)
24. P. Cui, U. Guin, A. Skjellum, D. Umphress, Blockchain in IoT: current trends, challenges, and future roadmap. *J. Hardw. Syst. Security* **3**(4), 338–364 (2019)
25. B. Tang, H. Kang, J. Fan, Q. Li, R. Sandhu, Iot passport: A blockchain-based trust framework for collaborative internet-of-things, in *Proceedings of the 24th ACM Symposium on Access Control Models and Technologies* (2019), pp. 83–92
26. C. Esposito, O. Tamburis, X. Su, C. Choi, Robust decentralised trust management for the internet of things by using game theory. *Inf. Process. Manag.* **57**(6), 102308 (2020)
27. M. Boussard, S. Papillon, P. Peloso, M. Signorini, E. Waisbard, STeward: SDN and blockchain-based trust evaluation for automated risk management on IoT devices, in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (IEEE, Piscataway, 2019), pp. 841–846

The Use of Blockchain Technology in Electronic Health Record Management: An Analysis of State of the Art and Practice

22

Henrique Couto, André Araújo, Rendrikson Soares, and Gabriel Rodrigues

Abstract

Driven by the need to offer digital solutions to the population, the healthcare sector requires computational solutions with features of security, immutability and traceability for data transactions on the Electronic Health Record (EHR). An EHR is defined as a repository of healthcare information stored and transmitted in a secure and accessible way by authorized users. To address this important area of research, this work investigated state of the art practice and studies that addressed the development and validation of computational solutions with Blockchain technology applied to the following areas of an EHR life-cycle: (i) modeling and standardization (ii) data storage techniques, (iii) standards for data interoperability, and (iv) data retrieval and visualization solutions. Based on the results found, this study presents an analysis of the main advances and opportunities identified in the use of Blockchain technology in the development of healthcare applications.

Keywords

Blockchain · Electronic health record · Health applications · Distributed databases · Review study ·

Data modeling · Data storage techniques · Data interoperability · Data security · Smart contracts

22.1 Introduction

The health sector has undergone major transformations over the past decade due to the adoption of information and communication technologies (ICT) in clinical patient care, and to the popularization of the access to Electronic Health Record (EHR) information [1]. Commonly called eHealth, this digital transformation process aims to improve healthcare practice and reduce operational costs involved in patient care [2]. The insertion of digital business models in the healthcare sector is driven by the demands of society and the global economy for functional and safe computing solutions that positively impact the population's quality of life.

In the ecosystem in which health institutions are inserted, several software systems are used to process health related data. These software systems may include an EHR, telemedicine applications, examination and diagnostic platforms, as well as mobile applications that manage and monitor an individual's clinical data. Due to the diversity of software applications present in healthcare, this sector requires solutions for the modeling and standardizing EHR data, data interoperability, and ensuring the integrity and authenticity of patient data shared in digital media [3, 4].

Advancements in computer technologies have made it easy to share data to quickly access information and provide convenience in carrying out healthcare services [5]. However, the volume of data processed by such software systems makes the authenticity of data transferred between healthcare organizations and end users an important and relevant research question within the scientific community [6]. Nowadays, accessing and sharing EHR data has become

H. Couto (✉)

Computing Institute, Federal University of Alagoas, Maceió, Brazil
e-mail: hcm@ic.ufal.br

A. Araújo

Computing Institute, Federal University of Alagoas, Maceió, Brazil

Advanced Studies in Data Science and Software Engineering, Federal University of Alagoas, Penedo, Brazil

e-mail: andre.araujo@ic.ufal.br

R. Soares · G. Rodrigues

Advanced Studies in Data Science and Software Engineering, Federal University of Alagoas, Penedo, Brazil

e-mail: rendrikson.soares@arapiraca.ufal.br;
gabriel.rodrigues@arapiraca.ufal.br

a common practice for two reasons. First, the end user, also known as e-patient, needs to access their clinical data history and grant access rights to their data to third parties if needed. Second, for administrative reasons, healthcare institutions need to exchange data with other organizations and governmental regulatory agencies.

Blockchain technology has been studied and applied in Academia and in the software industry as an alternative to guarantee the authenticity, traceability and security of online data transactions [7]. It consists of a decentralized and encrypted network that immutably certifies and stores all transactional information between the parties involved [8]. Blockchain has been used in the most diverse areas of knowledge [8], and in the healthcare sector, numerous studies have used its foundations and characteristics to provide solutions for the ecosystem of healthcare organizations [9–11]. Although this topic has been debated and applied in this sector, some Blockchain-related research issues that lack advances and investigation are reported in the state of the art, such as the generation of Blockchain transactions from EHR data, and the interoperability of clinical data between healthcare organizations [12].

In order to analyze how Blockchain technology has been applied in the healthcare sector, this paper investigates how the characteristics of decentralized storage, which promote security, immutability and traceability, have been applied in the management of an EHR lifecycle. To achieve this goal, the following steps were carried out: (i) a search strategy was defined using Blockchain and EHR related terms, (ii) the search was carried out in the main repositories covering this field, (iii) the inclusion criterias were the development and validation of computational solutions with Blockchain technology applied to data modeling and standardization, storage techniques, standards for data interoperability, and data visualization solutions. Finally, (iv) the search covered studies published between 2016 and 2021. Based on the results found, this study presents an analysis of the main advances and opportunities identified in the use of Blockchain technology aimed at the development of healthcare applications.

The remaining sections of this article are organized as follows. Section 22.2 describes the basic concepts needed to understand Blockchain technology. Section 22.3 presents the main studies identified in state of the art and practice, while Sect. 22.4 discusses the main advances found and provides an analysis of research opportunities, while final considerations and indications for future work are found in Sect. 22.5.

22.2 Background

Blockchain technology is a decentralized database that differs from others by its data storage structure. In Blockchain, data is stored in sets called blocks which are persisted in transactions, where the network nodes are responsible for authorizing the entry of these blocks into the Blockchain [13].

Blocks are arranged in a linked list, where each block has a reference to the previous block and only one new block can be added at a time. The beginning of the chain of blocks is determined by the genesis block, which stores the information that must be present in other blocks [13]. For a transaction to be carried out, the nodes that make up the network must authorize the entry of the respective blocks. In this sense, there is a consensus process between the nodes so that new blocks are added to the network.

Different types of consensus algorithms can be used, among which the Proof-of-Work (PoW), the Proof-of-Stake (PoS) and the Proof-of-Burn (PoB) stand out. In PoW, each node must solve a mathematical challenge in order to find a value for the hash code belonging to the block through the process of trial and error (or brute force) [14]. PoS is characterized by conducting a draw that determines which nodes will be responsible for authorizing a new block in the Blockchain network. In this case, users who wish to participate in the process need to deposit a certain amount of coins on the network as proof of their participation [15]. In PoB, nodes must burn some coins as proof of commitment to the network by obtaining the right to mine [16]. As the coin burning process represents virtual mining power, the more coins a user burns in favor of the system, the more mining power they have, and therefore, the better their chances of being chosen as the next validator on the block.

22.3 Blockchain-Based EHR Management Analysis

This section presents the study carried out on the use of Blockchain technology in an EHR lifecycle, and is organized as follows. In Sect. 22.3.1, the advances identified in the state of the art in the research area of this study are discussed, while in Sect. 22.3.2, the computational solutions created by the software industry and that are in use in the healthcare sector are presented.

22.3.1 An Analysis of the State of the Art

The development and maintenance process of a Health Information System (HIS) is characterized, among other things, by the need for a database project. A database design represents an important phase in the lifecycle of a software system, as it allows to identify and model EHR data requirements. In addition, it allows the design of an architecture with components that offer the best way to store and retrieve data. In this sense, the use of Blockchain technology in healthcare applications is justified by its ability to provide greater reliability in data storage. Through Blockchain, it is possible to verify the integrity and guarantee the immutability of all records stored on the network.

In order to investigate how Blockchain technology is being applied in the healthcare sector, studies that address the development and validation of computational solutions applied to the following areas of an EHR were considered in this study: (i) data modeling and standardization, (ii) data storage techniques, (iii) standards for data interoperability, and (iv) data retrieval and visualization solutions. The criteria defined above were chosen because they deal directly with an EHR lifecycle, which starts with the modeling phase, and goes on with data visualization and software applications.

Recent research involving Blockchain technology has focused on controlling access to data processed by healthcare organizations and integrating multiple Blockchain networks [17]. Gordon and Catalini [12] report that healthcare interoperability generally focuses on data exchange between commercial entities, however, solutions aimed at interoperability and data access for patients need to be developed. Pandey and Litoriya [18] propose a decentralized data storage software architecture using Blockchain technology and a computational service that allows healthcare organizations to access data without compromising the security of information transferred between stakeholders. Likewise, Zhang et al. [19] address the traffic of patient data between users and medical applications securely through Blockchain technology.

In [6], a tool is presented that aims to develop an organizational governance process to achieve the semantic interoperability of clinical models in healthcare. In [20], a cloud computing architecture to connect hospitals and provide APIs for information consultation is discussed. In the same line of research, Huang et al. [21] present a healthcare system in which records are encrypted and stored on a Blockchain network so that only patients and their healthcare providers can view or update them. The system uses a permit management system through smart contracts where patients control which institutions can access the historical record of their clinical data. Using the same approach, a hybrid architecture using Blockchain and Edge nodes is developed in [22] to

facilitate EHR access control, and a software system that uses a Cross-Blockchain platform for data sharing and privacy preservation of patient information is discussed in [23].

The use of Blockchain technology to store EHR data in cloud services requires IT resources with great processing power and high bandwidth Internet. In order to minimize the dependence on large computational resources, the solution developed in [24] uses a fog architecture to process and store data from patients who need clinical care. Following this same line of research, Nagasubramanian et al. [25] developed a framework based on a cloud architecture to ensure the integrity of the EHR. In conjunction with Blockchain technology, the framework makes use of an encryption approach called Keyless Signature Infrastructure (KSI), which is based on the hash function and provides security for the system that uses it. According to the results presented, the response time of the proposed framework with Blockchain technology is almost 50% lower than conventional techniques and the storage cost is about 20% lower for the Blockchain system compared to existing techniques. In order to minimize the use of computational resources to process Blockchain transactions, Chen et al. [26] propose a searchable Blockchain-based encryption scheme for EHR. In this solution, only one index is stored in Blockchain technology, which allows data to be searched using logical expressions.

In conjunction with Blockchain technology, healthcare standards such as OpenEHR and HL7 archetypes were used to create software applications aimed at managing EHR data. The openEHR standard addresses issues on how to standardize EHR data attributes, constraints and terminology [27], while the HL7 standard provides a set of specifications aimed at standardizing the exchange of information between software systems [28]. In this sense, the OpenEHR standard was used in [29] to create a distributed software architecture in which patients can record and consult their clinical care history from any device. The HL7 standard was used in [30] to create a Blockchain-based, globally integrated health record sharing architecture. The proposed architecture allows the patient to grant access to their clinical data and find which entities have accessed their EHR. In turn, Alamri and Margaria [31] propose a framework that makes use of HL7 to guarantee the compatibility of data exchange between different systems, Internet of Things (IoT) and EHRs, in a simple and understandable approach that saves time and costs.

To make the application of the investigated works easier to understand, Fig. 22.1 classifies the area of an EHR lifecycle for which each research solution was developed. The analysis of the advances made and the research opportunities identified are discussed in Sect. 22.4.

Research	EHR Lifecycle			
	M	S	I	V
Roehrs et. al, 2021 [17]	✓		✓	
Gordon and Catalini (2018) [12]			✓	
Pandey and Litoriya (2020) [18]		✓	✓	
Zhang et al. (2018) [19]			✓	
Carmen et al. (2016) [6]			✓	
Dubovitskaya et al. (2017) [20]		✓	✓	
Huang et al. (2021) [21]	✓	✓		
Guo et al. (2019) [22]		✓		✓
Wang and He (2021) [23]	✓		✓	
Islam et al. (2019) [24]		✓		
Nagasubramanian et al. (2020) [25]		✓		
Chen et al. (2019) [26]		✓		✓
Roehrs et al. (2019) [29]	✓			✓
Abdeen et al. (2019) [30]		✓		
Alamri and Margaria (2021) [31]			✓	
M - Modeling S - Storage I - Interoperability V - Visualization				

Fig. 22.1 Categorization of investigated works

22.3.2 An Analysis of the State of the Practice

Driven by the need to offer digital solutions to the population, the healthcare sector requires software systems with features of security, immutability and traceability for processed transactions. In this context, tools and software platforms using Blockchain technology that are directly related to EHR data management were investigated in the State of practice.

BurstIQ [32] is an ecosystem that integrates Blockchain technology and machine learning to allow data from different sources to be stored in a single data repository. Here, users have full control over their clinical data and can share or use it in exchange for access to products and services of interest to them. HealthNet [33] is a health data sharing platform focused on tracking clinical data, where it is possible to trace everything from prescription drugs to insurance claims and coordination of care between providers for patients. MediLedger [34] is a permissioned Blockchain platform developed for the pharmaceutical industry based on open standards and specifications (e.g., HL7, OpenEHR). Its goal is to enable cost savings for healthcare organizations and provide interoperability and security features.

Corel Health [35] is an iterative, interoperable, accessible, secure, and scalable health ecosystem that ensures data

security and accessibility through the use of smart contracts and Blockchain. Its goals include solving data warehousing issues, the incongruity of legacy databases, the challenges of analyzing unstructured data, high administrative costs, and lack of data security. In the same line of tools, Medicalchain [36] is a platform that uses Blockchain technology to store EHR data and allow users to share it with healthcare professionals and organizations of their choice. Furthermore, it is possible to create different health applications that use Medicalchain to perform the storage of users' health data, creating an ecosystem where EHR data can be shared between different applications. Finally, there are also initiatives such as Nebula Genomics [37], a big data company for the sequencing and health of the human genome that uses Blockchain to store data. Its goal is to build a reliable genomic data marketplace for consumers, researchers and the medical community.

22.4 Discussion

Figure 22.2 categorizes the works identified in the state of the art, observing their application of Blockchain technology in the following areas of the EHR: (i) data modeling and standardization, (ii) data storage techniques; (iii) standards for data interoperability; and (iv) data retrieval and visualization solutions. As shown in Fig. 22.2, most of the works mentioned in this study use Blockchain technology to provide data interoperability features between healthcare organizations. The large number of studies in the area of interoperability is justified by the high traffic of data exchanged between stakeholders involved in the health ecosystem, that is, private and public organizations, government agencies, health regulatory agencies and the patients themselves. All information exchanged online requires computational resources with characteristics of access control, data integrity and immutability, resources that are provided by Blockchain technology.

For data exchange to take place between healthcare industry stakeholders, storage solutions need to be developed to manage Blockchain transactions. As shown in Fig. 22.2, several studies have worked in this field of research and developed different solutions to store EHR data using Blockchain technology. However, some works only store an EHR index on the Blockchain network, while other works store all the data that characterize a patient's EHR.

Also in this field of research, it was observed that data storage with Blockchain technology commonly occurs in two ways. The first is to use platforms that are capable of creating decentralized software applications (DApps) that work on a peer-to-peer network. DApps are open source and operate autonomously and independently of central mediators [38]. Adaptations and changes to DApps only occur through

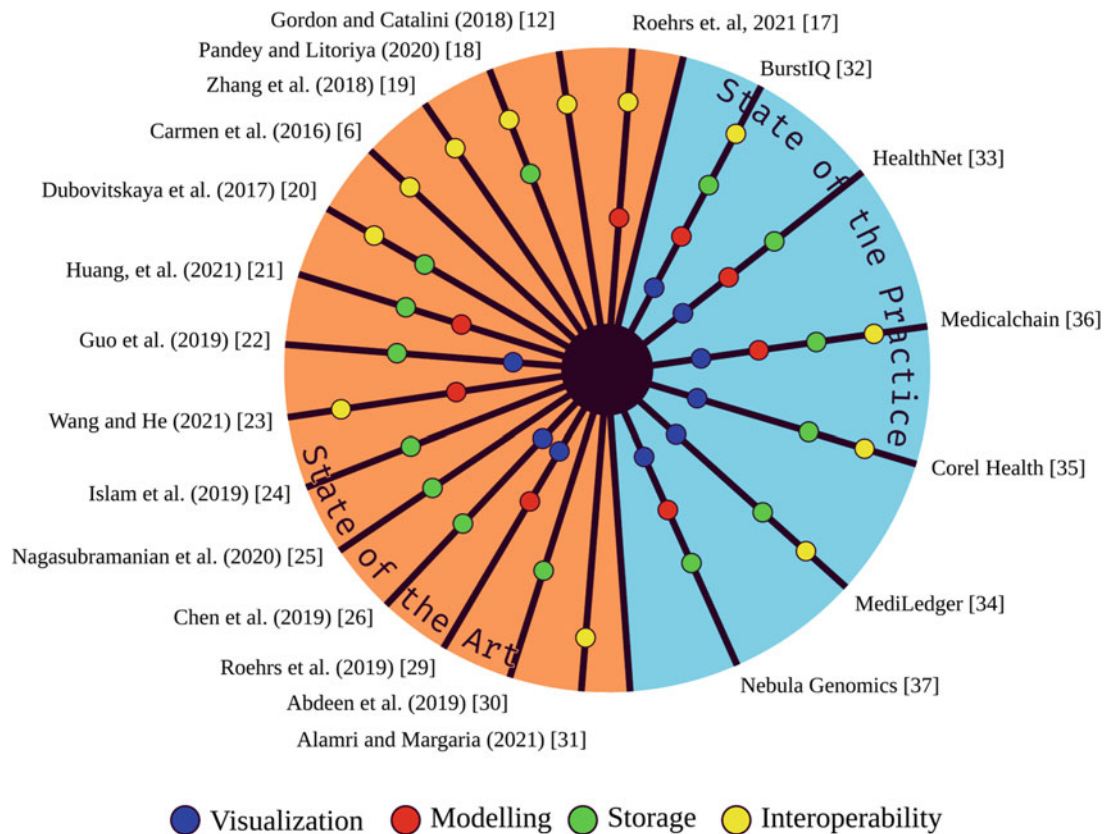


Fig. 22.2 Analysis of the investigated works by area

consensus between the parties involved. A decentralized software application interacts with Blockchain technology through smart contracts, which are self-executable protocols coded directly in the Blockchain that establish a set of rules for the interaction to occur in an autonomous and transparent way [39]. The other way is to use a DBMS that offers Blockchain technology capabilities. In this case, the application's business rules are implemented in the programming language chosen in the project, and the data persistence service is specified taking into account the data structures, software architecture and how transactions are implemented by the DBMS.

Once the analysis of the solutions developed for the areas of modeling and data visualization was concluded, the following points were identified. As seen in Fig. 22.2, few studies addressed the development of software applications with data visualization capabilities stored in Blockchain technology through graphical user interfaces. This highlights the need to develop software systems that address data visualization aspects and the construction of graphical user interfaces taking into account usability concepts for different types of applications (e.g., web, mobile). Finally, few works were identified in the area of modeling and standardizing EHR data for creating Blockchain transactions. Of the four areas of EHR management analyzed in this research, the area of

modeling and standardizing EHR data is more likely to be seen as crucial by healthcare organizations, since it directly deals with the construction of Blockchain transactions from data processed by software systems in a health organization.

The importance of this field of research is justified by the heterogeneity of data found in the healthcare sector, the complexity of managing data processed by various software systems, and the lack of uniformity in EHR data, caused by the absence of health standards (e.g. openEHR archetypes, ICD-10, Snomed) in software system development. The challenges reported above show that studies addressing the standardization of EHR data must be carried out. In particular, to support healthcare organizations in building Blockchain-based computational solutions to offer features of security and immutability for legacy EHR data.

The tools and platforms analyzed in the state of practice represent an important contribution from the software industry to the healthcare ecosystem. Business models are being transformed through Blockchain technology, which has allowed security in the storage of EHR data, access granted by patients, and the traceability of data transactions. However, although the tools and platforms presented in this study represent an important advance, it is reported in the software industry that the development of computational environments for the creation of DApps and the insertion of

Blockchain functionalities in DBMS require further research and development.

22.5 Final Considerations

This work presents a systematic review of research addressing the development and validation of computational solutions applied to the following areas of the Electronic Health Record (EHR): (i) data modeling and standardization, (ii) data storage techniques; (iii) data interoperability standards; and (iv) data retrieval and visualization solutions.

The use of Blockchain technology in an EHR lifecycle is justified by its ability to guarantee the authenticity, traceability and security of data transactions processed by software systems. In this sense, Academia and the software industry have developed computational solutions using Blockchain technology that allow healthcare organizations to improve the quality of services provided to the population and reduce the operating costs involved in patient care. In the review carried out herein, it becomes evident that many business models in the healthcare sector are being transformed, and the use of Blockchain technology has impeded fraudulent alterations in EHR data and guaranteed the traceability of service provision agreements between patients and healthcare organizations.

Analyzing the areas of an EHR lifecycle that are the object of this study, the following was found. The areas of storage and interoperability reflect most of the research analyzed, and this is justified by the need for data exchange that takes place between the stakeholders involved in the healthcare ecosystem. Meanwhile, the areas of data visualization, data modeling and EHR standardization require further investigation and the development of computational solutions.

Based on the evidence collected in this study, our indications for future work include: (i) the specification of a method to help healthcare organizations in standardizing EHR data using the openEHR standard, and (ii) the development of a computational service for generating Blockchain transactions across different networks, and database management systems with Blockchain technology features.

References

1. P. Li et al., ChainSDI: A software-defined infrastructure for regulation-compliant home-based healthcare services secured by blockchains. *IEEE Syst. J.* **14**(2), 2042–2053 (2020). <https://doi.org/10.1109/JSYST.2019.2937930>
2. K. Kaur, R. Rani, Managing data in healthcare information systems: Many models, one solution. *Computer* **48**(3), 52–59 (2015). <https://doi.org/10.1109/MC.2015.77>
3. A. Araújo, V. Times, M. Silva, A cloud service for graphical user interfaces generation and electronic health record storage, in *Information Technology – New Generations*, ed. by S. Latifi, (Springer, Cham, 2018), pp. 257–263
4. A. Araújo, V. Times, M. Silva, A tool for generating health applications using archetypes. *IEEE Softw.* **37**(1), 60–67 (2020)
5. A. Araújo, V. Times, M. Urbano, PolyEHR: A framework for polyglot persistence of the electronic health record, in *Int'l Conference Internet Computing and Internet of Things – ICOMP'16*, pp. 71–77, 07 2016
6. M. del Carmen Legaz-García, C. Martínez-Costa, M. Menárguez-Tortosa, J.T. Fernández-Breis, A semantic web based framework for the interoperability and exploitation of clinical models and ehr data. *Knowl Based Syst.* **105**, 175–189 (2016)
7. A. Clim, R.D. Zota, R. Constantinescu, Data exchanges based on blockchain in m-health applications. *Proc. Comput. Sci.* **160**, 281–288 (2019)
8. T.F. Stafford, H. Treiblmaier, Characteristics of a blockchain ecosystem for secure and shareable electronic medical records. *IEEE Trans. Eng. Manage.* **67**(4), 1340–1362 (2020)
9. I. Lima, A. Araújo, R. Souza, H. Couto, V. Times, A service-based software architecture for enabling the electronic health record storage using blockchain. *Int. J. Comput. Appl. Technol.* **65**(3), 222–234 (2021)
10. B. Shen, J. Guo, Y. Yang, Medchain: Efficient healthcare data sharing via blockchain. *Appl. Sci.* **9**(6) (2019)
11. J. Chanchaichujit, A. Tan, F. Meng, S. Eaimkhong, Blockchain technology in healthcare, in *Healthcare 4.0: Next Generation Processes with the Latest Technologies*, (Springer, Singapore, 2019), pp. 37–62
12. W. Gordon, C. Catalini, Blockchain technology for healthcare: Facilitating the transition to patient-driven interoperability. *Comput. Struct. Biotechnol. J.* **16**, 224–230 (2018)
13. S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system, *Cryptography Mailing list* at <https://metzdowd.com>, 03 2009
14. S.S. Hazari, Q.H. Mahmoud, A parallel proof of work to improve transaction speed and scalability in blockchain systems, in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 2019, pp. 0916–0921. <https://doi.org/10.1109/CCWC.2019.8666535>
15. W. Li, S. Andreina, J.-M. Bohli, G. Karame, Securing proof-of-stake blockchain protocols, in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, ed. by J. Garcia-Alfaro, G. Navarro-Arribas, H. Hartenstein, J. Herrera-Joancomart, (Springer, Cham, 2017), pp. 297–315
16. K. Karantias, A. Kiayias, D. Zindros, Proof-of-burn, in *Financial Cryptography and Data Security*, ed. by J. Bonneau, N. Heninger, (Springer, Cham, 2020), pp. 523–540
17. A. Roehrs, C.A. da Costa, R.R. Righi, A.H. Mayer, V.F. da Silva, J.R. Goldim, D.C. Schmidt, Integrating multiple blockchains to support distributed personal health records. *Health Inform. J.* (2021). <https://doi.org/10.1177/14604582211007546>
18. P. Pandey, R. Litoriya, Securing and authenticating healthcare records through blockchain technology. *Cryptologia* **44**(4), 341–356 (2020)
19. P. Zhang, J. White, D. Schmidt, G. Lenz, S. Rosenbloom, Fhir-chain: Applying blockchain to securely and scalably share clinical data. *Comput. Struct. Biotechnol. J.*, **16**, 267–278 (2018)
20. A. Dubovitskaya, Z. Xu, S. Ryu, M. Schumacher, F. Wang, Secure and trustable electronic medical records sharing using blockchain. *AMIA Annu. Symp. Proc. AMIA Symp.* **2017**, 650–659 (2017)
21. H. Huang, X. Sun, F. Xiao, P. Zhu, W. Wang, Blockchain-based ehealth system for auditable ehrs manipulation in cloud environments. *J. Parallel Distrib. Comput.* **148**, 46–57 (2021)
22. H. Guo, W. Li, M. Nejad, C. Shen, Access control for electronic health records with hybrid blockchain-edge architecture. *IEEE Int. Conf. Blockchain (Blockchain)* **2019**, 44–51 (2019). <https://doi.org/10.1109/Blockchain.2019.00015>
23. Y. Wang, M. He, CPDS: A cross-blockchain based privacy-preserving data sharing for electronic health records, in *2021 IEEE*

- 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 2021, pp. 90–99, <https://doi.org/10.1109/ICCCBDA51879.2021.9442539>
24. N. Islam, Y. Faheem, I.U. Din, M. Talha, M. Guizani, M. Khalil, A blockchain-based fog computing framework for activity recognition as an application to e-Healthcare services. *Future Gener. Comput. Syst.* **100**, 569–578., ISSN 0167-739X (2019). <https://doi.org/10.1016/j.future.2019.05.059>
 25. G. Nagasubramanian, R. K. Sakthivel, R. Patan, A. H. Gandomi, M. Sankayya, and B. Balusamy, “Securing e-health records using keyless signature infrastructure blockchain technology in the cloud,” *Neural Comput. Appl.*, vol. 32, pp. 639–647, Feb 2020.
 26. L. Chen, W.-K. Lee, C.-C. Chang, K.-K.R. Choo, N. Zhang, Blockchain based searchable encryption for electronic health record sharing. *Future Gener. Comput. Syst.* **95**, 420–429 (2019)
 27. openEHR: Open industry specifications, models and software for e-health, <https://www.openehr.org>, 2021 [Online]. Accessed 12 Oct 2021
 28. D. Bender, K. Sartipi, HL7 FHIR: An Agile and RESTful approach to healthcare information exchange, in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 2013, pp. 326–331, <https://doi.org/10.1109/CBMS.2013.6627810>
 29. A. Roehrs, C.A. da Costa, R.R. da Rosa, V.F. da Silva, J.R. Goldim, D.C. Schmidt, Analyzing the performance of a blockchain-based personal health record implementation. *J Biomed Inform.* **92**, 103140 (2019). <https://doi.org/10.1016/j.jbi.2019.103140>. Epub 2019 Mar 4
 30. A. Mohammad, R. Abdeen, T. Ali, Y. Khan, M.C.E. Yagoub, Fusing identity management, HL7 and Blockchain into a global healthcare record sharing architecture. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **10**(6) (2019). <https://doi.org/10.14569/IJACSA.2019.0100681>
 31. B. Alamri, I.T. Javed, T. Margaria, A GDPR-compliant framework for IoT-based personal health records using blockchain, in *2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 2021, pp. 1–5, <https://doi.org/10.1109/NTMS49979.2021.9432661>
 32. BurstIQ – Whitepaper. https://www.burstiq.com/wp-content/uploads/2017/09/BurstIQ-whitepaper_07Sep2017.pdf. 2017 [Online]. Accessed 12 Oct 2021
 33. HealthNet – HealthNet™: A Blockchain Enabled Population Health Platform for Enterprise Health Systems. <https://consensusnetworks.com/healthnet-whitepaper/> [Online]. Accessed 12 Oct 2021
 34. MediLedger – Mediledger 2019 Progress report. [https://assets.chronicled.com/hubfs/MediLedger%202019%20Progress%20Report%20\(1\).pdf](https://assets.chronicled.com/hubfs/MediLedger%202019%20Progress%20Report%20(1).pdf). 2019 [Online]. Accessed 12 Oct 2021
 35. Corel Health – Coral Health The Blockchain for Personalized Medicine. https://ww1.prweb.com/prfiles/2018/07/31/15707263/CoralHealth_WP.pdf [Online]. Accessed 12 Oct 2021
 36. MedicalChain – Whitepaper 2.1 – <https://medicalchain.com/Medicalchain-Whitepaper-EN.pdf>, 2018 [Online]. Accessed 12 Oct 2021
 37. Nebula Genomics – Accelerating Genomic Data Generation and Facilitating Genomic Data Access Using Decentralization, Privacy-Preserving Technologies and Equitable Compensation. <https://nebula.org/blog/wp-content/uploads/2019/05/Accelerating-Genomic-Data-Generation-and-Facilitating-Genomic-Data-Access.pdf>. 2018 [Online]. Accessed 12 Oct 2021
 38. A.M. Saghiri, Blockchain architecture, in *Advanced Applications of Blockchain Technology*, (Springer, 2019), pp. 161–176
 39. A. Rubio, G.M. Tarazona, L. Contreras, *Big Data and Blockchain Basis for Operating a New Archetype of Supply Chain* (Springer, Cham, 2018), pp. 659–669

Blockchain for Security and Privacy of Healthcare Systems: A Protocol for Systematic Literature Review

Saadia Azemour, Meryeme Ayache, Hanane El Bakkali, and Amjad Gawanmeh

Abstract

Patient's privacy, electronic health records' confidentiality, integrity and all related e-health security issues are the most critical elements of a successful digital health, as they support building trust between patients and healthcare stakeholders. Blockchain technology appears to cover a wide range of these elements. However, the use of this emergent technology in healthcare domain, still has many security and privacy challenges that need to be overcome. Recently, many research works are focusing on such issues leading to a growing literature. In the perspective to review this literature in a systematic way to deeply investigate the use of blockchain technology in healthcare for security enhancement and privacy protection, this paper proposes a protocol that could be used to conduct successfully this systematic literature review (SLR). The proposed protocol follows the PRISMA-P 2015 Guidelines. At a closer look, we indicate the use of snowballing search and automated search (on eight electronic data sources) to carry out the intended SLR, identify five pertinent research questions, and specify the related inclusion/exclusion criteria. All methods for selection process, data collection and data

analysis that will be used in the intended SLR are described in this protocol.

Keywords

Blockchain technology · Smart contract · Distributed ledger · Security · Privacy · Interoperability · Healthcare · e-health · Telemedicine · Systematic literature review protocol

23.1 Introduction

23.1.1 Rationale

Numerous systematic literature reviews deal with the use of blockchain in healthcare. For instance, [1, 2] investigate blockchain applications and platforms, and highlight challenges for possible future directions. Furthermore, [3, 4] explore the application domains of blockchain in healthcare, clarify the interaction of various healthcare stakeholders with the blockchain technology, and reveal some blockchain challenges and limitations. Moreover, [5, 6] treat briefly the security enhancement, privacy protection, scalability, interoperability, and disclose potential challenges. Many other studies explore only the security improvement based on the blockchain technology but in different domains for instance cryptocurrency, food supply chain, smart cities and Iot [7–10]. We have found one systematic mapping study that deals with the use of blockchain technology in healthcare to improve security and privacy. However, this study does not conduct security challenges related to using blockchain technology in healthcare. On the other hand, few reviews were also found in this subject, but not systematically conducted [11–13].

S. Azemour (✉) · H. E. Bakkali
Smart Systems Laboratory- Rabat IT Center Mohammed V University
Rabat, Rabat, Morocco
e-mail: saadia_azemour@um5.ac.ma;
hanan.elbakkali@ensias.um5.ac.ma

M. Ayache
STRS Laboratory, RAISS Team, INPT, Rabat, Morocco
e-mail: ayache@inpt.ac.ma

A. Gawanmeh
College of Engineering and IT, University of Dubai, Dubai, UAE
e-mail: amjad.gawanmeh@ud.ac.ae

Our final goal is to conduct a systematic literature review (SLR) to investigate the security and privacy challenges on healthcare. Therefore, and as suggested in [14–18], we started our SLR process with the planning stage that emphasize on developing a protocol for the study [19], then we will further conduct our SLR based on this protocol. Our SLR protocol follows the PRISMA-P 2015 guidelines [20,21]. Hence, we included all the 17 items that construct the PRISMA-P checklist, this will provide clarity, transparency and future reproducibility of our systematic literature review. As indicated in [22], the use of PRISMA-P 2015 facilitates the reporting of a completed systematic literature review.

23.1.2 Objectives

The main research goals of the intended SLR are: (a) to gather knowledge about the enhancement of security and patient's privacy in healthcare environment using the blockchain technology, (b) to assess the negative and positive influence of the blockchain technology in healthcare systems' security and patient's privacy, and finally (c) to spot potential research gaps and open issues that will be presented as future directions. This goal is broad enough to allow us to slice it in several more specific research questions. Using Kitchenham et al recommendations [18, 22], We have identified the five following research questions:

- What are the security and privacy issues of the current used healthcare systems?
- How the blockchain could improve security and privacy in digital health?
- What are the existing blockchain applications that propose improvements for preserving security and privacy in digital health?
- What are the limitations and the challenges of such applications (legal, human and technical aspects)?
- How can we overcome those different limitations and challenges?

The proposed protocol for our intended SLR, will allow us to further answer these research questions in the completed systematic literature review, by appealing to empirical results collected from target literature databases.

23.2 Systematic Literature Method

As indicated in Fig. 23.1 the empirical results will be collected by following several steps. The first step is the search process which utilize two types of search strategies in order to cover all relevant studies: (a) the snowballing search (backward and forward search which use 76 relevant articles as an initial set); (b) the automated search which will be

conducted on eight databases (ACM, Scopus, Springer, IEEE Xplore, Science direct, Google scholar, dblp, and web of science) using the defined search engine composed of a list of keywords related to blockchain, security and e-health.

The second step of our methodology is related to the selection process which is based on the inclusion/exclusion criteria and includes several sub-steps: (a) firstly, we will reject all studies published before 2019, then (b) we will remove all duplicates, (c) finally, we will perform a selection based on the title, abstract and keywords, and keep only relevant papers that have a direct relationship with the SLR's selected keywords.

In the third step we will perform a full screen to the studies obtained from the previous step. This screen, is also based on the inclusion/exclusion criteria, keep only relevant papers after reading the full text. To support the selection process, we will further apply quality metrics that specify the information to be obtained in order to evaluate each study and eliminate studies with low quality. The last two steps are the extraction of data and the analysis of these data.

23.2.1 Eligibility Criteria

To identify and select only relevant primary studies that focus on the area of interest, we will use the inclusion and exclusion criteria shown in Table 23.1.

A paper will be excluded when it fulfills at least one of the exclusion criteria. Any paper not excluded by the exclusion criteria and fulfills all the inclusion criteria will be included in the set of selected primary studies. The application of the inclusion/exclusion criteria will be performed by reviewing meta-data information, such as title, abstract, keywords, index terms, etc. When decision cannot be made based on the available meta-data, then, conclusion contents will be considered in order to take the final decision [18] of inclusion/exclusion.

The use of the proposed inclusion and exclusion criteria will prevent the introduction of bias into our selection process, appraise validity of the included primary studies and facilitate future reproducibility of our systematic literature review.

23.2.2 Information Sources

To ensure more complete coverage of our topic that encompasses all relevant studies, we will use two complementary search strategies snowballing search and automated search:

- We will conduct a backward and forward snowballing, following Wohlin guidelines [23]. Initial set for snowballing will consist of some papers obtained from google scholar and identified as included studies. For backward search,

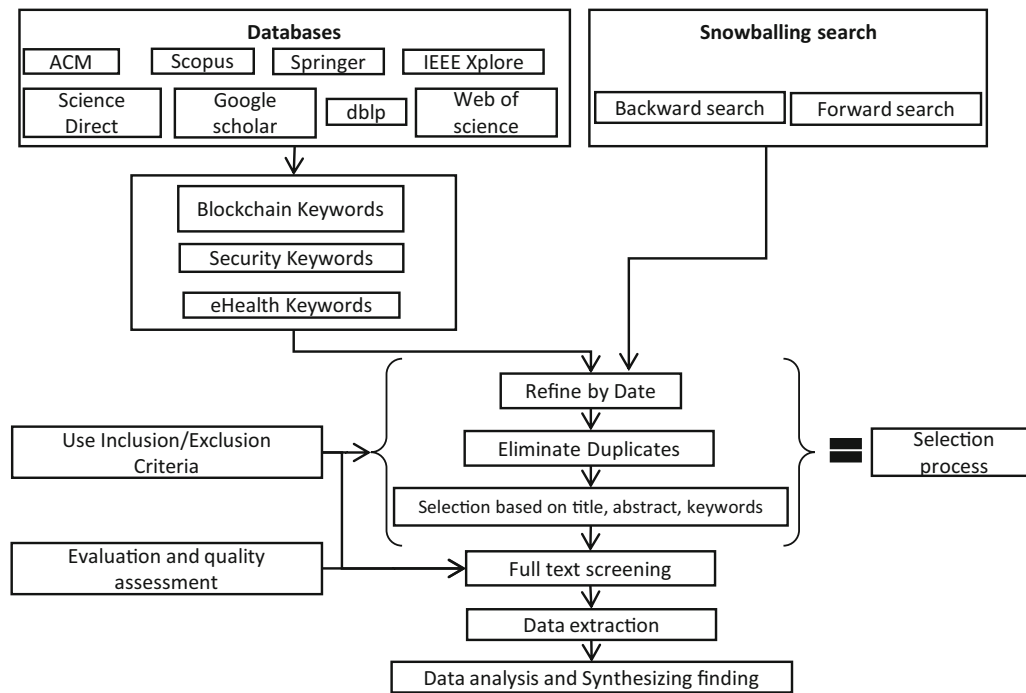


Fig. 23.1 The proposed methodology

Table 23.1 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Papers published from 2019 to 2021	Duplicated paper
Papers written in English	Papers without full text availability
Peer-reviewed papers published in an indexed conference proceeding or indexed journal	Short papers
Primary studies	Secondary studies
Papers related to the use of Blockchain technology in healthcare for security and/or privacy purpose	Papers that are technical reports or thesis
Papers that propose Blockchain frameworks or applications for enhancing security and/or privacy in healthcare	

we will use the reference section of the examined work in order to identify papers that fulfill the inclusion-exclusion criteria and exclude others. For forward search, we will use google scholar to retrieve the citations and identify new papers to include (fulfill inclusion-exclusion criteria) from those papers citing the considered work.

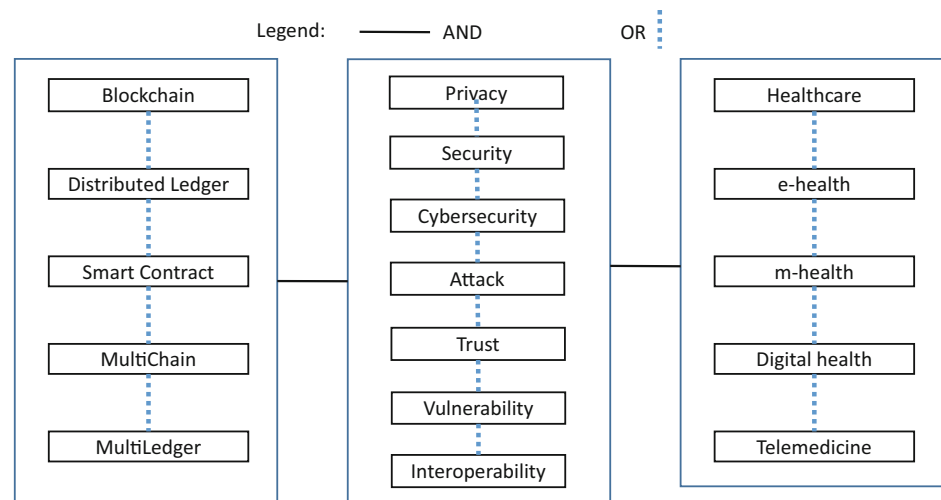
- The first step of the automated search is the selection of the Electronic Data Sources to be used [18, 22]. For us, we will use the following eight different electronic data sources: Google Scholar, IEEE Xplore, ACM Digital

Library, Scopus, SpringerLink, Science Direct, Web of science and dblp. The selected electronic data sources will cover almost every venue (Conferences, workshops and Journals) of software engineering field, which means that we can be confident that we will find almost all available evidences.

23.2.3 Search Strategy

To obtain terms for the search string we have applied two different strategies:

- firstly, We have analyzed our goal and research questions: As previously stated, our goal is to assess the use of blockchain technology in healthcare for enhancing security and protecting privacy. From this goal, we have extracted the terms: Blockchain, healthcare, security and privacy. These key terms also come from RQ1, RQ2, RQ3 respectively. In turn, RQ4 and RQ5 do not contain further key terms to add.
- In the second step, We have conducted pilot searches: we have run pilot searches on google scholar, using the previous key terms, to identify other relevant terms that are frequently used. We looked for those terms in the title, Abstract and Author’s keywords of retrieved papers. For each key term we have found a set of relevant terms:

Fig. 23.2 Search string

- Blockchain: Distributed Ledger, Smart Contract, Multi-chain, Multi-ledger.
- Security: Privacy, Cybersecurity, Attack, Vulnerability, Trust, Interoperability
- Healthcare: Healthcare, e-Health, m-Health, Digital health, Telemedicine.

We connected the extracted terms using Boolean operators, likewise we used the operator AND to connect synonyms, and the operator OR to connect the sets, as detailed below: (Blockchain OR Distributed Ledger OR Smart Contract OR MultiChain OR Multi-ledger) AND (Security OR Privacy OR Cybersecurity OR Attack OR Vulnerability OR Trust OR interoperability) AND (Healthcare OR e-health OR m-health OR Digital health OR Telemedicine). For illustration purposes, Fig. 23.2 illustrates the list of terms used and their logical connection.

Our search string was designed to find the maximum number of the known papers, However we may miss out some studies that use different terminology to describe the enhancement of security in healthcare based on blockchain.

We have transformed the constructed search string into a search script adapted to each electronic data source as shown in the following example:

- **SpringerLink:** (Blockchain OR “Distributed Ledger” OR “Smart Contract” OR MultiChain OR MultiLedger) AND (Security OR Privacy OR Cybersecurity OR Attack OR Vulnerability OR Trust OR interoperability) AND (Healthcare OR e-health OR m-health OR “Digital health” OR Telemedicine)

We will use these scripts to perform our advanced search, based on specific metadata (title, abstract, keywords) in order to find only relevant papers that will be including in the

study. We will additionally specify the date range [2019–2021] before starting the search process in each electronic data sources.

23.2.4 Study Record

23.2.4.1 Data Management

Literature search results will be uploaded to Zotero [24] tool, an internet-based software program that provide bibliography management and results organization.

We will create individual folders for each literature source on Zotero: backward-folder, forward-folder, and folder for each electronic data source (ACM-folder, Scopus-folder, etc.). Then we will create “exclusion-folder”, where we will place excluded articles, and we will specify the reason for exclusion of each article. Zotero will also help us to record the date of retrieval, abstract, keywords and notes about any other pertinent information.

23.2.4.2 Selection Process

The same study may come from different sources, in consequence, we will firstly eliminate all duplicated studies, then we will perform an initial selection of the remainder studies based only on title in order to eliminate all unrelated papers, then we will review abstract to identify relevant papers that may provide direct evidence about the research questions, and fulfill the proposed inclusion criteria. However, sometimes the information given by these meta-data are mostly insufficient to determine whether a paper should be included or excluded. So that we will additionally read the conclusion for more information before making the final decision [18].

We attend to leave the full-text review of the included studies for quality assessment and data extraction process. The review’s selection process and full-text review will be illustrated using PRISMA 2020 flow diagram.

As previously mentioned, we will record excluding papers and the reason to be excluded. The selection of all obtained studies will be conducted by One reviewer, then the selection results will be verified by two other reviewers, one of them will screen the included studies and the other one will verify the excluded studies, then a meeting will be performed to discuss the results and resolve any disagreements, this process will be confirmed by the supervisor in order to increase the reliability of our systematic literature review. Some authors could be contacted to resolve any remaining ambiguity.

23.2.5 Data Collection Process

To ensure an efficient data extraction we will carefully follow the experience-based suggested guidelines [25]. We have additionally designed a data extraction form to accurately and systematically record the information obtained from included studies. As illustrated in Table 23.2, Data extracted will include general information (authors, title, publication type, country, etc), it will also contain study characteristics (methodology, objectives, contribution, etc) and information about the purpose of our SLR (current security issues, proposed blockchain solutions, security challenges of the use of blockchain in healthcare, etc).

Table 23.2 Data extraction form

Subtopics	Data
General information	Title
	Authors
	Country
	Year of publication
	Publication type
	Publisher
	Citation count(from google scholar)
Study characteristics	Methodology
	Study objectives
	Name of approach
	Contribution
	Project suitability
	Application domain
Purpose of our SLR	Limitations
	Security issues in current healthcare systems
	Blockchain solutions
	Blockchain Features related to security improvement
	Healthcare application domain
	Blockchain security challenges
	Corresponding stakeholders
Security requirements	
Future directions	

Data value, eligible for inclusion, research question and comments will be added to the data extraction form in order to assess the data extracted. Thus we will evaluate the eligibility to defined criteria (Yes, No or Unclear). if extracted data answers to any defined research question, it will be joined to this question. Comments will be added to record the text extracted from the study and contain this extracted data.

Through the data collection process, we will read the full text of each included study, then extract relevant data and fill out the data extraction form that we have already designed, we will also add notes to each data extracted if necessary. For missing information, we will contact the included studies' authors, by sending them three email attempts as a maximum. As with the selection process, single data extraction will be employed by one reviewer, then we will carry out a set of meetings with two other reviewers in order to discuss and verify the correctness of the obtained results and resolve disagreements. finally, our data collection will be confirmed by the supervisor in order to reduce bias and errors.

23.2.6 Data Items

As previously stated, we will extract the main current security and privacy issues in healthcare system, and the blockchain security solutions that have been proposed to encounter these issues. Furthermore, we will collect blockchain security and privacy issues in healthcare and the most frequent attacks that may damage healthcare system or threat patient privacy. We will finally identify future directions. We intend to initially divide all the above-mentioned data items into five groups that present the healthcare stakeholders: patients, provider, research organization, supply chain bearer and payer. Each group will contain the corresponding current security issues, blockchain security solutions, blockchain security challenges and future directions. We will additionally categorize each study contribution as an application or framework or algorithm or any other proposal. All the blockchain security challenges will be identified based on the different security and privacy requirements.

23.2.7 Outcomes and Prioritization

23.2.7.1 Primary Outcomes

The primary outcomes will be current security issues in healthcare, proposed blockchain security and privacy solutions, and finally blockchain security and privacy challenges. We decided a priori to subdivide security issues on current healthcare systems, blockchain solutions and blockchain security/privacy challenges into different healthcare stakeholders:

- Providers that may be hospital, doctor or specialist.
- Patients.
- Supply chain bearer that may be vendors, manufactures or pharmacy.
- Research organization.
- Payers (Insurance).

Similarly, all those elements may be divided into different healthcare application domains, and may be joined to a specific security and privacy requirement, a potential scheme will be introduced to illustrate all these categorizations.

23.2.7.2 Secondary Outcomes

The secondary outcomes will be attacks related to current security issues and blockchain security challenges, contribution type, the implementation or non-implementation of the proposed solution and research limitations. These additional outcomes will be reviewed to provide complementary information for the completeness of our systematic literature review. During this planning stage we will perform a sufficient investigation to ensure that the outcomes selected above are relevant. and we anticipate including these outcomes in a summary of finding table.

23.2.8 Risk of Bias Individual Studies

We will use the Cochrane collaboration tool [26] to assess the possible risk of bias, we will additionally omit all studies that are judged to be high risk of bias. We will assess the possible risk of bias of each study by judging the extracted information as “high risk”, “moderate risk” or ‘low risk’. If there is insufficient detail reported in the study, we will judge the risk of bias as ‘unclear’ and the original study investigators will be contacted for more information. To facilitate the assessment of the overall risk of bias in included studies, we will use a checklist composed of all the predefined outcomes. We will omit all studies that are judged to be high risk of bias

23.2.9 Data Synthesis

In this stage we will report studies with low risk of bias or moderate risk of bias, we also expect to retain studies at high risk of bias only when they provide a primary outcome. To answer to the previous research questions, a narrative qualitative summary will be performed at the first time [27]. Then, we will analyze the outcomes using sub group analysis to explore the possible source of heterogeneity based on the categorization indicated in data item and outcome subsections, we will additionally use tables and schemes to summarize and explain the finding characteristics. A comparison of the results will be performed. If collected data are not

in a suitable format for analysis, we will take various steps to process these data. We intend to perform a quantitative analysis, if included studies are sufficiently homogeneous in term of design. Thus, NVIVO software will be used for completing meta-analysis [28]. If possible, we will measure number of included studies per solution type, number of blockchain solutions per healthcare domain application, number of blockchain solutions per stakeholder, number of blockchain solutions per security requirement, and number of blockchain security challenges per healthcare domain application, and number of blockchain security challenges per stakeholder. When there are missing data, we will attempt to contact the original authors of the study to obtain the relevant missing data.

23.2.10 Confidence in Cumulative Estimate and Meta-Bias

We will assess the quality of evidence for all outcomes using Grading GRADE (Grading of Recommendations Assessment Development and Evaluation), which is a system mostly used to judge the quality of evidence in healthcare literature [29]. The assessment will be performed across various domains of risk of bias, consistency, directness, precision and publication bias. we will additionally use an established approach that will assess the risk of bias across studies, and assess factors that increase the confidence in an effect based on the following quality questions:

- Are the defined research objectives really match With our research questions?
- How many research question are answered by the defined research objectives?
- Is the contribution compared to other works?
- Is the study contain future directions?

We will judge the overall quality of evidence as high, moderate, low or very low. Evidences with low and very low quality will be omitted, and we will take only evidences with high and moderate quality.

23.3 Conclusion

In this paper we proposed a protocol that we will follow to successfully conduct our intended systematic literature review that will investigate the use of blockchain technology in healthcare for security enhancement and privacy protection.

As a future work we will analyze the results of our systematic literature review conducted on the basis of this protocol. Thus, we will be able to identify the main healthcare security and privacy issues that blockchain technology

has resolved, give various categorizations of the different blockchain solutions used for this purpose, and present the most important security and privacy challenges of the use of this emerging technology in healthcare. Therefore, we believe that our systematic literature review will make a valuable contribution to identify existing research gaps for future directions.

References

1. M. Hölbl, M. Kompara, A. Kamišalić, L.N. Zlatolas, A systematic review of the use of blockchain in healthcare. *Symmetry* **10**(10), 470 (2018)
2. T.-T. Kuo, H.Z. Rojas, L. Ohno-Machado, Comparison of blockchain platforms: a systematic review and healthcare examples. *J. Am. Med. Inf. Assoc.* **26**(5), 462–478 (2019)
3. Sobia, Y., M.M. Khan, R. Talib, A.D. Butt, S. Saleem, F. Arif, A. Nadeem, Use of blockchain in healthcare: a systematic literature review. *Int. J. Adv. Comput. Sci. Appl.* **10**(5), 10 (2019)
4. L. Soltanisehat, R. Alizadeh, H. Hao, K.R. Choo, Technical, temporal, and spatial research challenges and opportunities in Blockchain-based healthcare: A systematic literature review. *IEEE Trans. Eng. Manag.* (2020). <https://doi.org/10.1109/TEM.2020.3013507>
5. M. Prokofieva, S.J. Miah, Blockchain in healthcare. *Aust. J. Inf. Syst.* **23**, 1–22 (2019)
6. A. Dubovitskaya, P. Novotny, Z. Xu, F. Wang, Applications of blockchain technology for data-sharing in oncology: 403–411. Results from a systematic literature review. *Oncology* **98**(6), 403–411 (2020)
7. A. Ekramifard, H. Amintoosi, A.H. Seno, A systematic literature review on blockchain-based solutions for iot security, in *The 7th International Conference on Contemporary Issues in Data Science* (Springer, Cham, 2019), pp. 311–321
8. Z. Yu, L. Song, L. Jiang, O. Khold Sharafi, Systematic literature review on the security challenges of blockchain in IoT-based smart cities. *Kybernetes* **51**(1), 323–347 (2022)
9. A. Hassan, M.Z. Mas'ud, W.M. Shah, S.F. Abdul-Latip, R. Ahmad, A. Ariffin, A systematic literature review on the security and privacy of the blockchain and cryptocurrency. *OIC-CERT J. Cyber Secur.* **2**(1), 1–17 (2020)
10. N. Etemadi, Y.G. Borbon, F. Strozzi, Blockchain technology for cybersecurity applications in the food supply chain: a systematic literature review, in *Proceedings of the XXIV Summer School "Francesco Turco" Industrial Systems Engineering, Bergamo, Italy* (2020), pp. 9–11
11. A. Sarkar, T. Maitra, S. Neogy, Blockchain in healthcare system: security issues, attacks and challenges, in *Blockchain Technology: Applications and Challenges* (Springer, Cham, 2021), pp. 113–133
12. M. Soni, D.K. Singh, Blockchain-based security & privacy for biomedical and healthcare information exchange systems, *Materials Today: Proceedings* (2021)
13. S. Shi, D. He, L. Li, N. Kumar, M.K. Khan, K.-K. Raymond Choo, Applications of blockchain in ensuring the security and privacy of electronic health record systems: a survey. *Comput. Secur.* **97**, 101966 (2020)
14. Y. Xiao, M. Watson, Guidance on conducting a systematic literature review. *J. Plan. Educ. Res.* **39**(1), 93–112 (2019)
15. P.V. Torres-Carrión, C.S. González-González, S. Aciar, G. Rodríguez-Morales, Methodology for systematic literature review applied to engineering and education, in *2018 IEEE Global Engineering Education Conference (EDUCON)* (IEEE, Piscataway, 2018), pp. 1364–1373
16. J.L. Barros-Justo, S. Sepúlveda, N. Martínez-Araujo, A. González-García, The use of controlled vocabularies in requirements engineering activities: a protocol for a systematic literature review (2017). arXiv:1704.00822
17. A. Kofod-Petersen, How to do a structured literature review in computer science. Ver. 0.1 (2012)
18. P. Brereton, B.A. Kitchenham, D. Budgen, M. Turner, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain. *J. Syst. Softw.* **80**(4), 571–583 (2007)
19. C. Okoli, K. Schabram, A guide to conducting a systematic literature review of information systems research. *Sprouts: Working papers on. Inf. Syst.* **10**(26) <http://sprouts.aisnet.org/10-26> (2010)
20. L. Shamseer, D. Moher, M. Clarke, D. Gherzi, A. Liberati, M. Petticrew, P. Shekelle, L.A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* **349**, g7647 (2015)
21. D. Moher, L. Shamseer, M. Clarke, D. Gherzi, A. Liberati, M. Petticrew, P. Shekelle, L.A. Stewart, Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst. Rev.* **4**(1), 1–9 (2015)
22. B. Kitchenham, *Procedures for Performing Systematic Reviews*, vol. 33 (Keele University, Keele, 2004), pp. 1–26
23. C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (2014), pp. 1–10
24. K.M. Ahmed, B. Al Dhubaib, Zotero: a bibliographic assistant to researcher. *J. Pharm. Pharmacother.* **4**, 303–306 (2011)
25. V. Garousi, M. Felderer, Experience-based guidelines for effective and efficient data extraction in systematic reviews in software engineering, in *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering* (2017), pp. 170–179
26. J.P.T. Higgins, D.G. Altman, P.C. Gøtzsche, P. Jüni, D. Moher, A.D. Oxman, J. Savović, K.F. Schulz, L. Weeks, J.A.C. Sterne, A.C. Jonathan, The cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ* **343**, d5928 (2011)
27. X. Huang, H. Zhang, X. Zhou, M.A. Babar, S. Yang, Synthesizing qualitative research in software engineering: a critical review, in *Proceedings of the 40th International Conference on Software Engineering* (2018), pp. 1207–1218
28. L.P. Wong, Data analysis in qualitative research: a brief guide to using NVivo. *Malaysian Fam. Phys.* **3**(1), 14 (2008)
29. G. Guyatt, A.D. Oxman, E.A. Akl, R. Kunz, G. Vist, J. Brozek, GRADE guidelines: introduction-GRADE evidence profiles and summary of findings tables. *J. Clin. Epidemiol.* **64**(4), 383–394 (2011)

Single Sign-On (SSO) Fingerprint Authentication Using Blockchain

Abhijeet Thakurdesai, Marian Sorin Nistor, Doina Bein, Stefan Pickl, and Wolfgang Bein

Abstract

The objective of this paper is to describe the front-end and the backend of an open-source web-application that can be integrated in any website and for which the storage of single sign-on (SSO) authentication is provided in an Ethereum network. The backend of this application is shared with a browser-based platform or Android platform. Ethereum network facilitated the implementation of peer-to-peer multi-node blockchain in distributed ledger technology. We use smart contract code for user creation and authentication. A contract is a collection of code (its functions) and data (its state) that resides at a specific address on the Ethereum blockchain. The smart contract made our proposed web app self-verifying, self-executing, and tamper resistant. The proposed software system can be used as two factor authentications in combination with passwords for servers, for payments authorizations, in banking and automotive industry.

Keywords

Android platform · Authentication system · Biometric data · Biometric trait · Distributed ledger technology · Ethereum network · Fingerprint data · Open-source

A. Thakurdesai · D. Bein (✉)
Department of Computer Science, California State University,
Fullerton, Fullerton, CA, USA
e-mail: kool7abhi@csu.fullerton.edu; dbein@fullerton.edu

M. S. Nistor · S. Pickl
Department of Computer Science, Universität der Bundeswehr
München, Neubiberg, Germany
e-mail: sorin.nistor@unibw.de; stefan.pickl@unibw.de

W. Bein
Department of Computer Science, University of Nevada, Las Vegas,
Las Vegas, NV, USA
e-mail: wolfgang.bein@unlv.edu

web-application · Peer-to-peer multimode blockchain · Secure storage

24.1 Introduction

An authentication system provides the user a full-control and distributed access to authorized services, while protecting its the privacy of its identity. Generally speaking, authentication can be described as a process in which a user offers some form of proof that he is the same user who registered the account. When developing authentication systems for current web applications, the backend contains one centralized database in which the data used for authentication is being stored; for example using SQL, MongoDB, or MEAN stack. The user authenticates by providing the requested proof of identity; it is evaluated by the authentication server (part of backend) and depending on the result of this evaluation, the authentication server may or may not grant access to the user. A proof of identity can be any piece of information that an authentication server accepts: something users have in their possession, something they know or something they are (e.g., a password or a biometric) [1]. The oldest type of proof of identity is using passwords (strings of certain length), that can be stored as plain or encrypted. A breach disclosed by Equifax [2] has exposed identity information for over 140 million individuals.

A biometric-based authentication system [3] requires first a user enrollment followed by, when needed, a user recognition. During user enrollment, a digital representation of an individual's biometric trait is generated, called biometric template, then stored in a database for future comparison. The user recognition can operate in two modes: verification and identification. In the verification mode, the system require that the user provides a biometric template and this will be

matched against a single template in the database. In the identification mode, the system require that the user provides a biometric template and this will be matched against all templates stored in the database [1]. But even if the authentication system is biometric-based system, most of the deployed systems use a centralized model. in case an attacker gains access to the web application or the biometric centralized database, s/he can extract enough information to compromise the user's digital identity [4]. If an authentication system uses a centralized architecture used for storage, then its data used for logging in and out of the system will be susceptible to attacks. And biometric data that can be used instead of text-based credentials is more valuable to attackers since it can provide access to many networks [1]. Since the user has a unique biometric trait, once extracted, it can be exploited to provide access to many networks and services.

Single sign-on (SSO) enables the users to securely authorize and authenticate with multiple web or desktop applications using only a single pair of credentials. SSO schemes rely on a third-party identity provider (IdP) to broker authentication using protocols such as SAML [5] and OpenID Connect [6]. But only about 5% of sites use any of over 50 disparate IdP [7]. Loopholes in centralized IdP-based SSO systems are exploited and result in loss of biometric data [8]. If an application does not use a SSO then authentication, authorization and web application security has to be implemented and updated by the website itself including the user management part.

SSO authentication works as follows (see Fig. 24.1):

1. A website checks if the user is already authenticated. If the user is already authenticated, then the content is delivered to the user.
2. If the user is not authenticated, then the website redirects the user to SSO login page.
3. The user enters the username and password to login.
4. The SSO identity servers verifies the credentials.
5. If the user credentials are correct, then the SSO server redirects the user to the website with appropriate HTTP code or else shows an invalid credentials error.

The SSO in generally use a database to store user credentials which is secured by a web application. The SSO also has an admin dashboard for user management. So, the admin can add or remove access to the current users. Most of the SSO implementations available in the market currently are open source. The enterprise level SSO like OKTA includes 2-factor authentication in order to stop attacks on SSO system. But still the passwords make the systems prone to the hacking. Hence using an encrypted ledger like blockchain to store the data makes an excellent SSO. The blockchain is a type of distributed ledger, comprised of interchangeable, digitally recorded data in packages called blocks. Some advantages

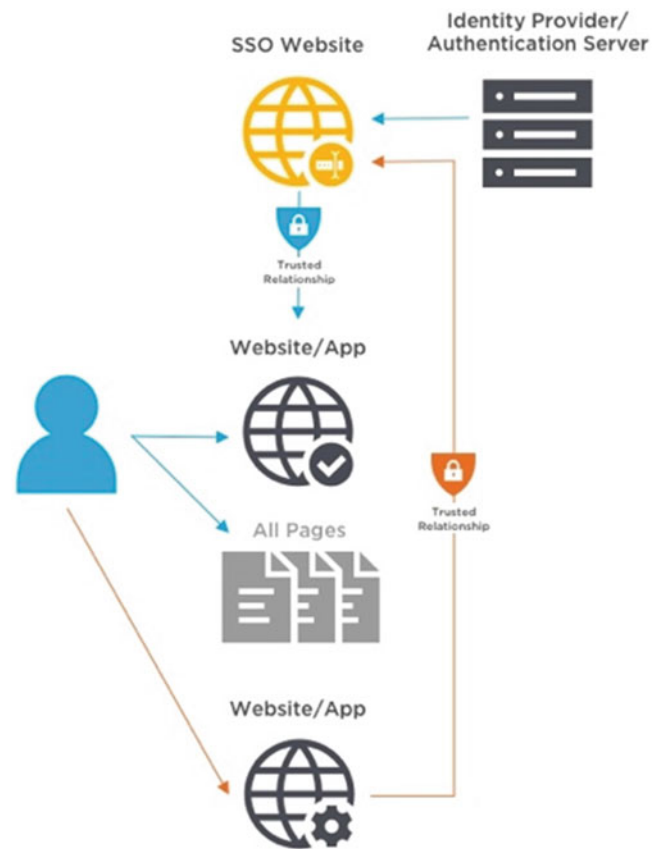


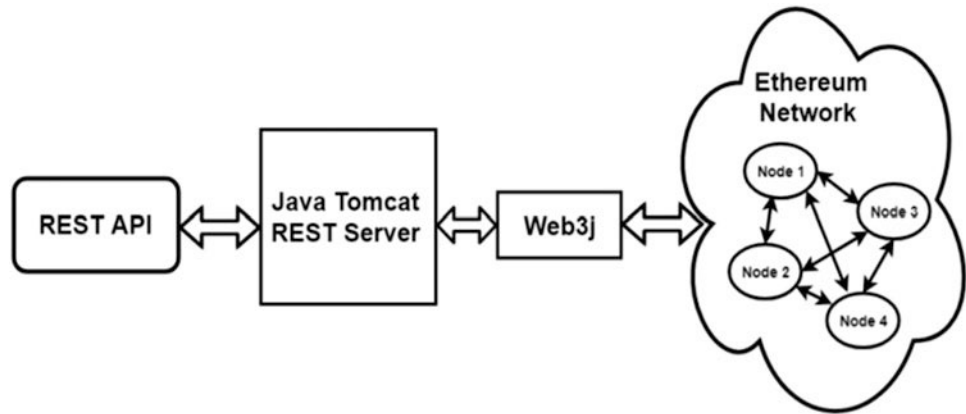
Fig. 24.1 SSO authentication flow

using blockchain for authentication include: it is impossible to tamper with the data, the central authority is eliminated, it is a decentralized ledger, and it has high concurrency to access data. In [1], the authors have proposed the Horcrux protocol as a method for secure exchange of biometric credentials within an existing standard (IEEE 2410-2017 BOPS) [9]. The proposed self-sovereign identity model uses blockchain technologies to assure exchange of verifiable credentials in lieu of 3rd-party identity providers during user authentication.

The authors of [10] propose to secure an encrypted fingerprint template by a symmetric peer-to-peer network and symmetric encryption, Advanced Encryption Standard (AES) algorithm, and then is uploaded to a symmetrically distributed storage system, the InterPlanetary File system (IPFS). The encrypted template is uploaded to the IPFS, and its returned digest is stored on the Ethereum network.

In this paper we will describe the front-end and the back-end of an open-source web-application that can be integrated in any website and for which the storage is provided in an Ethereum network (see Fig. 24.2). We use fingerprint images as a biometric data to authenticate users. Fingerprints have been used for over a decade for user verification and identification [11–13]. The fingerprint collected during the

Fig. 24.2 Back-end modules for authentication



authentication needs to have a similarity score of at least 70% with the fingerprint images stored in the Ethereum network for the user to be successfully authenticated. If no object is having similarity score greater than 70, then a null object is returned in the form of custom message ‘http 404 User not found’ (see Fig. 24.4). Using the Ethereum network in our proposed system, data are stored on nodes that keep a copy of all transactions dating back to the first one in a chain of blocks. The hash of the previous block connects each block; hence, any tampering will be noticed. This system ensures data security when a node or multiple nodes are under attack.

The major steps in developing our web app are:

1. Create a smart contract named ‘User.sol’
2. Create and set up an Ethereum Network
3. Deploy the smart contract on the Ethereum network and store the address of smart contract
4. Create a web service using Java Springboot
5. Integrate the Web3j library to communicate between nodes of the Ethereum network
6. Install Postman to test the API
7. Access the smart contract using credentials and stored smart contract address.
8. Write a REST service named `add_user` to add a user
9. Write a REST service named `verify_user` to verify a user using fingerprint data.

The entire project is stored as a Jar file.

The paper is organized as follows. In Sect. 24.2 we present the backend, followed by the frontend in Sect. 24.3. Concluding remarks and future work are given in Sect. 24.4.

24.2 Backend of OUR Proposed Web App

The backend of this application is shared with a browser-based platform or Android platform. The core logic of the backend module is designed and developed using Springboot architecture, which a scalable way of creating the web

application. By using MVC architecture, we can decompose different modules of the application using independent service components. The application development involves designing and development of frontend, Restful web services to connect the front end and back end, and the backend module in Java. The MVC design pattern keeps the monolithic backend as modular as possible. This architecture can be further extended to microservices architecture where an application is built as small services, each of them being independently deployable. These services can be written in different programming languages and can use different databases for storage.

The REST API initiates the transaction by communicating with the Java Tomcat REST server. Java Tomcat REST Server is responsible for Serving the Java web API. Web3j is intermediate component for communication between java RESTful API and Ethereum Network. Ethereum Network is the network of multiple blockchain nodes. We use smart contract code for user creation and authentication. A contract is a collection of code (its functions) and data (its state) that resides at a specific address on the Ethereum blockchain [14].

REST service `add_user` allows the admin to create a user. The admin can add users using a secure channel as follows. It first makes the request to Backend API to create a user. The backend API validates and serializes the biometric data, then it requests the Ethereum Network to store the data. An Ethereum node will mine the transaction and broadcast the resultant data to all the nodes in the Ethereum network (see Fig. 24.3).

REST service `verify_user` authenticates the user and returns the corresponding object (aka biometric data) if matched. To authenticate a user, the following steps are executed. The user trying to authenticate will initiate a transaction and click on the “Verify” button (frontend), that will generate a verify request to the backend API. The backend API will extract the features from the input biometric object and request the Ethereum nodes to verify and provide some response. Ethereum nodes will provide details of all the objects with similar biometric features. The

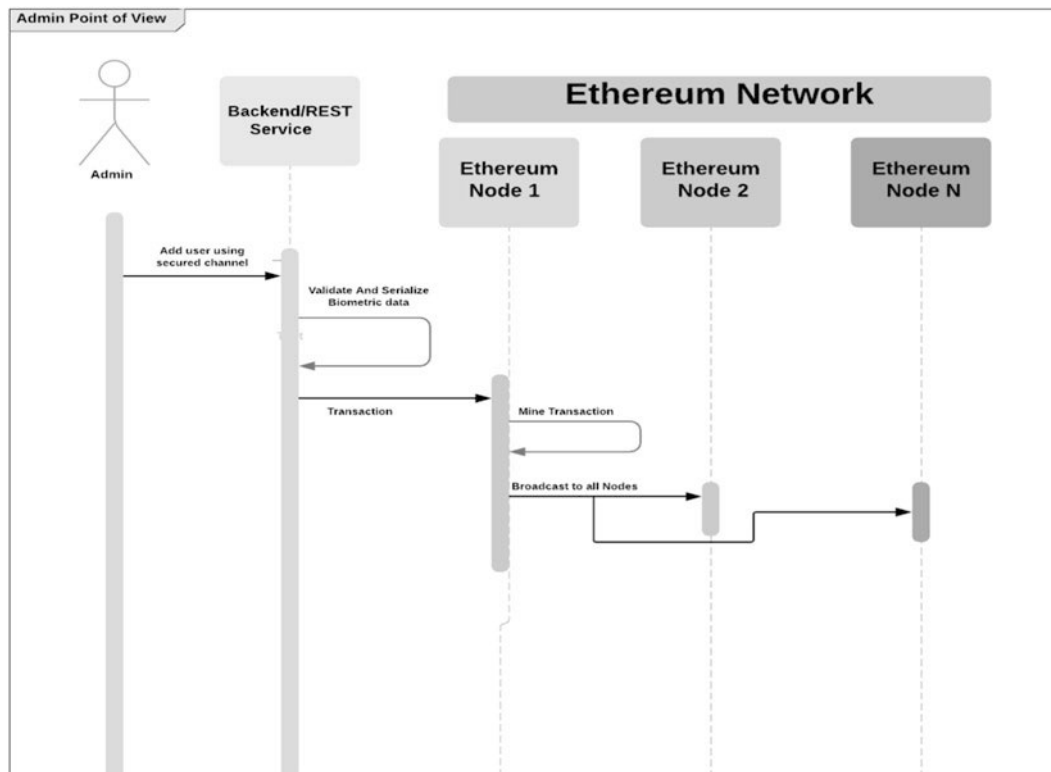


Fig. 24.3 Adding a user using an Ethereum network

backend API will then calculate the similarity score of all the returned objects and the single object having similarity score greater than 70 will be returned to the client. If multiple objects have similarity score greater than 70, then one with maximum score will be returned. If no object is having similarity score greater than 70, then null object is returned in the form of custom message 'http 404 User not found' (see Fig. 24.4).

24.3 Frontend of the Proposed Web App

We implemented a very basic frontend, that can be expanded to any type of web app that is browser-based or Android-based. Our basic frontend supports the backend functionalities, that include user creation and user authentication.

Figure 24.5 shows the user registration webform to add a user.

Once the fields are populated with necessary data, the user can be added (see Fig. 24.6).

If the admin tries to add a user who already has biometrics registered, then the system throws an error "Operation not permitted" (see Fig. 24.7). This is done in order to avoid identity theft. The authentication of a user uses an API to verify its fingerprint (see Fig. 24.8). If a user trying to

authenticate has a similarity metric of 70% or higher, the backend will return a JSON object with the user's details (see Fig. 24.9). If the details provided by the user do not manifest into the backend finding data, i.e. the user data is not present in the block chain database, the system returns a custom made error 404 error 'User does not exist' (see Fig. 24.10).

24.4 Conclusion and Future Work

In this paper we describe the frontend and the backend of an open-source web-application that can be integrated in any website and for which the storage of single sign-on (SSO) authentication is provided in an Ethereum network. The backend of this application is shared with a browser-based platform or Android platform. An Ethereum network facilitated the implementation of peer-to-peer multi-node blockchain in distributed ledger technology. We use smart contract code for user creation and authentication. The smart contract code facilitated, verified, and enforced the negotiation and performance of an agreement and transaction. The smart contract made our proposed web app self-verifying, self-executing and tamper resistant. Our Web service makes the architecture to be RESTful and flexible.

Fig. 24.4 Biometric-based user authentication

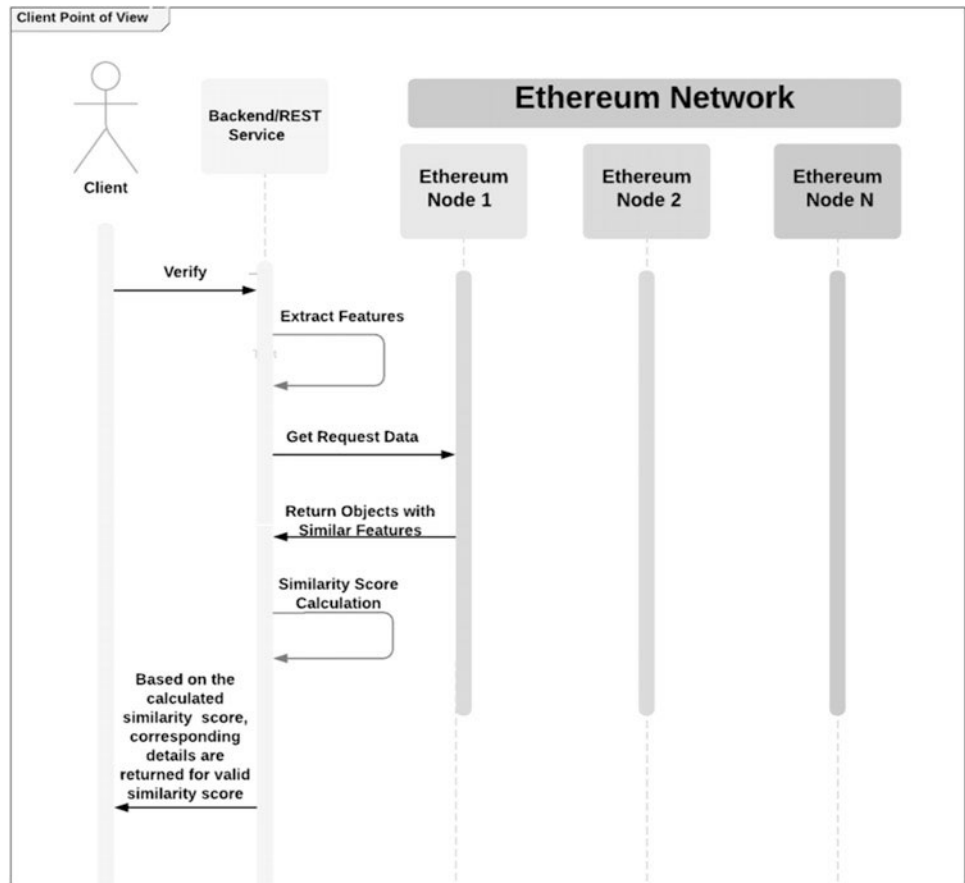


Fig. 24.5 User registration screen

[Verify API](#) [Add User API](#)

This project demonstrates use of a distributed P2P structure in distributed computing, and demonstrate how secure a biometric-based authentication system can be made that it can replace passwords. The proposed software system can be used as two factor authentications in combination with passwords for servers, for payments authorizations, in banking and automotive industry.

Two future directions are to implement microservices architecture for scaling and to implement an enterprise blockchain like IBM Hyperledger.

There are drawbacks to the proposed system that we plan to address in future work. One drawback is that, in order to keep the system running, the nodes of the system needs to be incentivized. The second drawback is that, if the private key of the seed node is lost then there is permanent loss of access.

Acknowledgement This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700.

Fig. 24.6 User registration is successful

Verify API Add User API

Name
Abhijeet

Sex
Male

DOB
06/12/1994

Select Fingerprint of the user Add User

Fig. 24.7 Denied registration of an existing user

Verify API Add User API

User Abhijeet has been created

Name
Abhijeet

Sex
Male

DOB
06/12/1994

Select Fingerprint of the user Add User

Operation not permitted. Undo

Fig. 24.8 Webform to authenticate a user



Fig. 24.9 API for the user authentication is successful

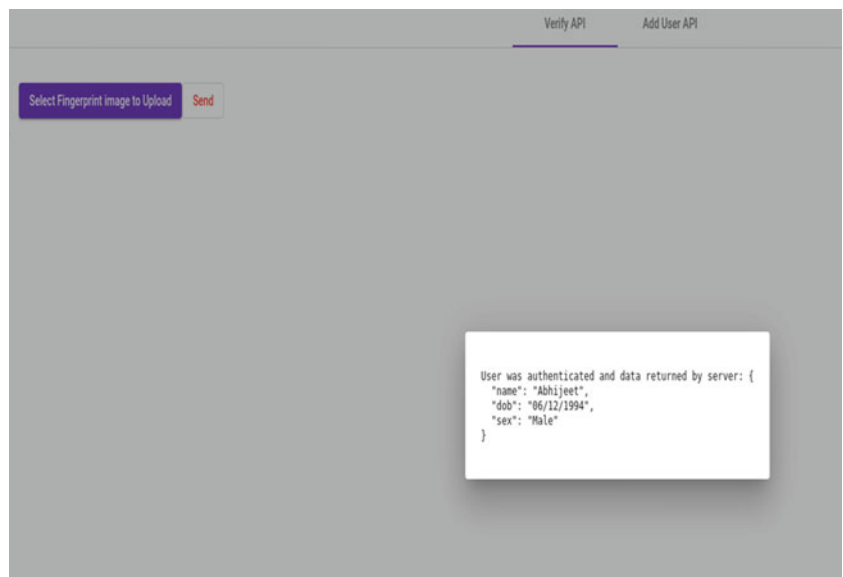
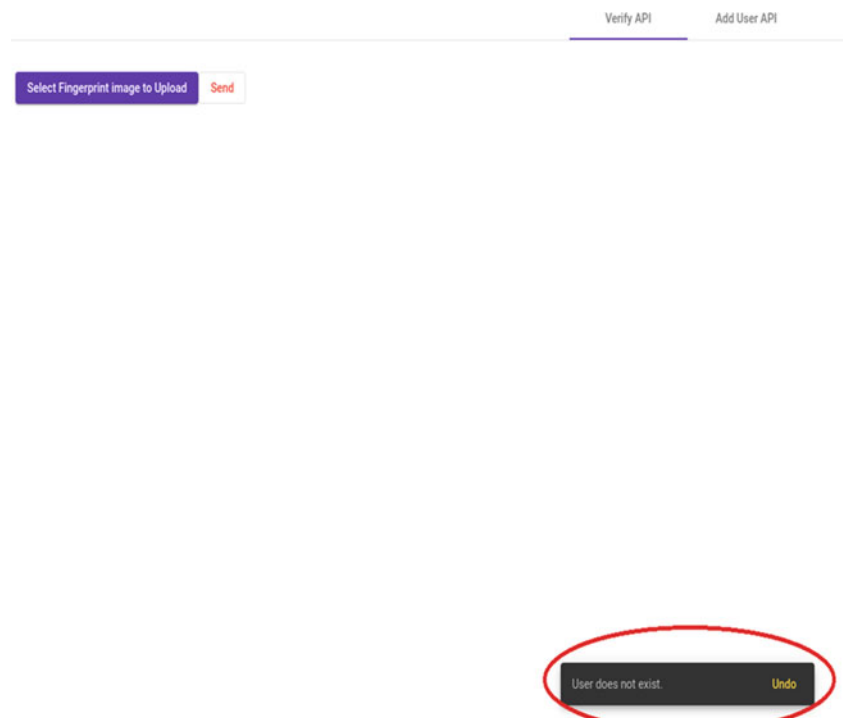


Fig. 24.10 API for user authentication does not recognize the user



References

1. A. Othman, J. Callahan, The Horcrux protocol: A method for decentralized biometric-based self-sovereign identity, in *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, 2018, pp. 1–7. <https://doi.org/10.1109/IJCNN.2018.8489316>
2. M. Hume, Identity theft cited as threat after equifax security breach. *Globe Mail Toronto A* 7 (2004)
3. A. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition. *IEEE Trans. Circ. Syst. Video Technol.* **14**(1), 4–20 (2004). <https://doi.org/10.1109/TCSVT.2003.818349>
4. A. Jain, K. Nandakumar, A. Nagar, Biometric template security. *EURASIP J. Adv. Signal Proc.*, **2008**(113), 1–17 (2008). <https://doi.org/10.1155/2008/579416>
5. J. Hughes, E. Maler, Security assertion markup language (SAML) v2. 0 technical overview. OASIS SSTC Working Draft. *sstc-saml-techoverview-2.0-draft-08*, 29–38 (2005)
6. N. Sakimura, J. Bradley, M. Jones, B. Medeiros, E. Jay. Openid connect standard 1.0-draft 18. Available online https://openid.net/specs/openid-connect-standard-1_0-18.html. Last accessed 9 Dec 2020
7. A. Vapen, N. Carlsson, A. Mahanti, N. Shahmehri, A look at the third-party identity management landscape. *IEEE Internet Comput.* **20**(2), 18–25 (2016)
8. K. Zetter, A. Greenberg, Why the opm breach is such a security and privacy debacle. *Wired*, 2015. Available online at <https://www.wired.com/2015/06/opm-breach-security-privacy-debacle/>. Last accessed 9 Dec 2020
9. 2410-2017 IEEE biometric open protocol standard (BOPS). Available at <https://standards.ieee.org/findstds/standard/2410-2017.html>. Last accessed 9 Dec 2020
10. M.A. Acquah, N. Chen, J.-S. Pan, H.-M. Yang, B. Yan. Securing fingerprint template using blockchain and distributed storage system. *Symmetry*, pp. 1–13 (2020). Available online at <https://www.mdpi.com/2073-8994/12/6/951>. Last accessed 9 Dec 2020
11. F. Hai-jian, C. Cheng-wei, Z. Chang-wei, Research on application of ethernet-based fingerprint identification system in college laboratory management, in *Fourth International Symposium on Knowledge Acquisition and Modeling*, Sanya, 1–274 (2011)
12. J. Tian, Y. Peng, Research of the Matlab application in the fingerprint identification system, in *2012 International Conference on Image Analysis and Signal Processing*, Hangzhou, 2012, pp. 1–5. <https://doi.org/10.1109/IASP.2012.6425005>
13. S. Selvarani, S. Jebapriya, R.S. Mary, Automatic identification and detection of altered fingerprints, in *2014 International Conference on Intelligent Computing Applications*, Coimbatore, 2014, pp. 239–243. <https://doi.org/10.1109/ICICA.2014.58>
14. Ethereum, Introduction to Smart Contracts. Available at <https://docs.soliditylang.org/en/v0.4.24/introduction-to-smart-contracts.html>. Last accessed 9 Dec 2020

Part V

Health Informatics

A Detection Method for Early-Stage Colorectal Cancer Using Dual-Tree Complex Wavelet Packet Transform

25

Daigo Takano and Teruya Minamoto

Abstract

Colorectal cancer is a major cause of death. As a result, cancer detection using supervised learning methods from endoscopic images is an active research area. Regarding early-stage colorectal cancer, preparing a significant number of labeled endoscopic images is impractical. We devise a technique for detecting early-stage colorectal cancer in this study. This technique consists of a 2D complex discrete wavelet packet transform and principal component analysis. As this technique does not require supervised learning, detection is feasible even in the absence of labeled data. In the endoscopic image, this technique correctly classifies early-stage colorectal cancer and normal regions with 92% accuracy. This approach outperforms the local binary pattern method.

Keywords

Wavelet analysis · Dual-tree complex wavelet transform · Complex discrete wavelet transform · Complex discrete wavelet packet transform · Directional selectivity · Principal component analysis · Image classification · Early colorectal cancer · Endoscopic image

25.1 Introduction

If detected early, the five-year survival rate of colorectal cancer is exorbitant [1]. However, its detection is difficult. Endoscopy is the most common method of detecting colorectal cancer, and there are no definite detection criteria

for its early-stage by visual inspection. Early-stage colorectal cancer is diagnosed in regions different from the surrounding areas. Thus, the detection accuracy highly depends on the physician's experience and knowledge.

Applying supervised learning for cancer detection has recently become a routine [2], but detecting early-stage colorectal cancer is difficult. Generally, supervised learning requires a massive amount of training data and corresponding correct answer labels. Figure 25.1 shows that we should prepare an image with markings in the case of cancer 25.1, and data need to be generated by a physician. As mentioned earlier, detecting early-stage colorectal cancer is difficult, so creating data for supervised learning is challenging. We developed a detection method for early-stage colorectal cancer in this study; this method does not require supervised data. A 2D complex discrete wavelet packet transform (2D-CWPT) and principal component analysis (PCA) were employed for feature extraction [3]. Further, to diagnose early-stage colorectal cancer, a thresholding process is applied to the obtained first principal component values. Early-stage colorectal cancer has a feature at the surface of the colon, which can be detected by observations from physicians. Therefore 2D-CWPT had directional selectivity can extract the difference in surface information between cancerous and normal regions.

We assumed that the input is a single endoscope image in our detection method. For the given endoscopic image, we separated it into various regions and performed detection for each, and cropped 50 images of size 100×100 for each region. To get the high-frequency components, 2D-CWPT was used on the cropped images and the statistics from each high-frequency component. Next, we perform a PCA to these components.

We described the 2D-CWPT used for feature extraction, and then proposed a method, which was followed by experimental results. Further, we applied our method to endoscopic images in the experiment. Using local binary pattern

D. Takano (✉) · T. Minamoto
Graduate School of Science and Engineering, Saga University, Saga, Japan
e-mail: takanod@ma.is.saga-u.ac.jp; minamoto@ma.is.saga-u.ac.jp



Fig. 25.1 Example of an endoscopic image of early-stage colon cancer marked by a physician

[4], the classification accuracy of each cropped image was determined and compared with that of the existing method.

This paper is organized as follows. First, we provide a brief description of 2D-CWPT in Sect. 25.2. We explain the proposed method and experiment for verification in Sects. 25.3 and 25.4. Finally, we describe the conclusion of the proposed method in Sect. 25.5.

25.2 2D Complex Discrete Wavelet Packet Transform

A method that applies wavelet packet to 2D complex discrete wavelet transform (2D-CDWT) is known as 2D-CWPT. However, as 2D-CWPT also decomposes high-frequency components, it is feasible to capture the variation of pixel values locally.

25.2.1 2D Complex Discrete Wavelet Transform

2D-CDWT is a wavelet transform that extends Kingsbury's dual-tree complex wavelet transform (DT-CWT) to include full shift-invariance [5, 6]. It consists of two filter banks, one for each filter bank. Similar to the DT-CWT, it has two filter bank structures, one each for the real and imaginary parts.

The filters applied in the row and column can be used to characterize the frequency components. Figure 25.2 shows the transform at level $j = 1$. The R and I represent the type of filter used in the column and row directions. Also C represents input image. For example, RI indicates that the real filter was used in the column direction and the imaginary

filter was used in the row direction. L, H denotes the low-pass and high-pass components. LL means the low-frequency component, and LH, HL, HH means the high-frequency components in the vertical, horizontal, and diagonal directions, respectively. Equation (1) shows the decomposition of the scaling factor c_j^{RI} . H and G denote the low-pass and high-pass filters, respectively.

$$\begin{aligned}
 c_{j+1, n_y, n_x}^{RI, LL} &= \sum_{k_y, k_x} h_{2n_y - k_y}^R h_{2n_x - k_x}^I c_{j, k_y, k_x}^{RI}, \\
 d_{j+1, n_y, n_x}^{RI, LH} &= \sum_{k_y, k_x} h_{2n_y - k_y}^R g_{2n_x - k_x}^I c_{j, k_y, k_x}^{RI}, \\
 d_{j+1, n_y, n_x}^{RI, HL} &= \sum_{k_y, k_x} g_{2n_y - k_y}^R h_{2n_x - k_x}^I c_{j, k_y, k_x}^{RI}, \\
 d_{j+1, n_y, n_x}^{RI, HH} &= \sum_{k_y, k_x} g_{2n_y - k_y}^R g_{2n_x - k_x}^I c_{j, k_y, k_x}^{RI}
 \end{aligned} \tag{1}$$

By performing addition and subtraction, we can assign directional selectivity to a frequency component. Examples of adding direction selectivity to high-frequency component LH in the vertical direction is shown as follows:

$$\begin{aligned}
 D_{j, n_y, n_x}^{R0, LH} &= \frac{d_{j, n_x, n_y}^{RR, LH} + d_{j, n_x, n_y}^{II, LH}}{2}, \\
 D_{j, n_y, n_x}^{I0, LH} &= \frac{d_{j, n_x, n_y}^{IR, LH} - d_{j, n_x, n_y}^{RI, LH}}{2}, \\
 D_{j, n_y, n_x}^{R1, LH} &= \frac{d_{j, n_x, n_y}^{RR, LH} - d_{j, n_x, n_y}^{II, LH}}{2}, \\
 D_{j, n_y, n_x}^{I1, LH} &= \frac{d_{j, n_x, n_y}^{IR, LH} + d_{j, n_x, n_y}^{RI, LH}}{2}
 \end{aligned} \tag{2}$$

25.2.2 Dual-Tree Complex Wavelet Packet Transform

2D-CWPT is a wavelet packet technique that is applied to 2D-CDWT [3, 7]. It decomposes high-frequency components during the conversion, unlike 2D-CDWT.

The frequency components obtained from 2D-CWPT are characterized by the index (n, m) . Here n represents the longitudinal frequency component ω_y and m represents the transverse frequency component ω_x . The higher the value of (n, m) , the higher the frequency component. Figure 25.3 shows (n, m) for $j = 1, 2$.

The conversion is conducted by 2D-CWPT according to the following rules:

- If $(n, m) \neq (1, 1)$, the real part of the filter is used.
- If either of the indices n or m is an even number, low-pass and high-pass filters are switched.

Fig. 25.2 Flowchart of 2D-CDWT at level $j = 1$ to 2

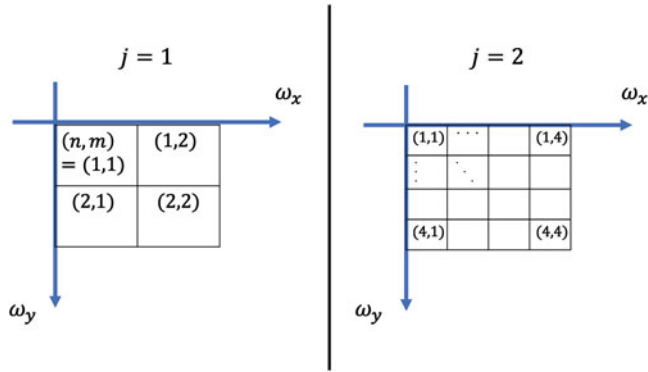
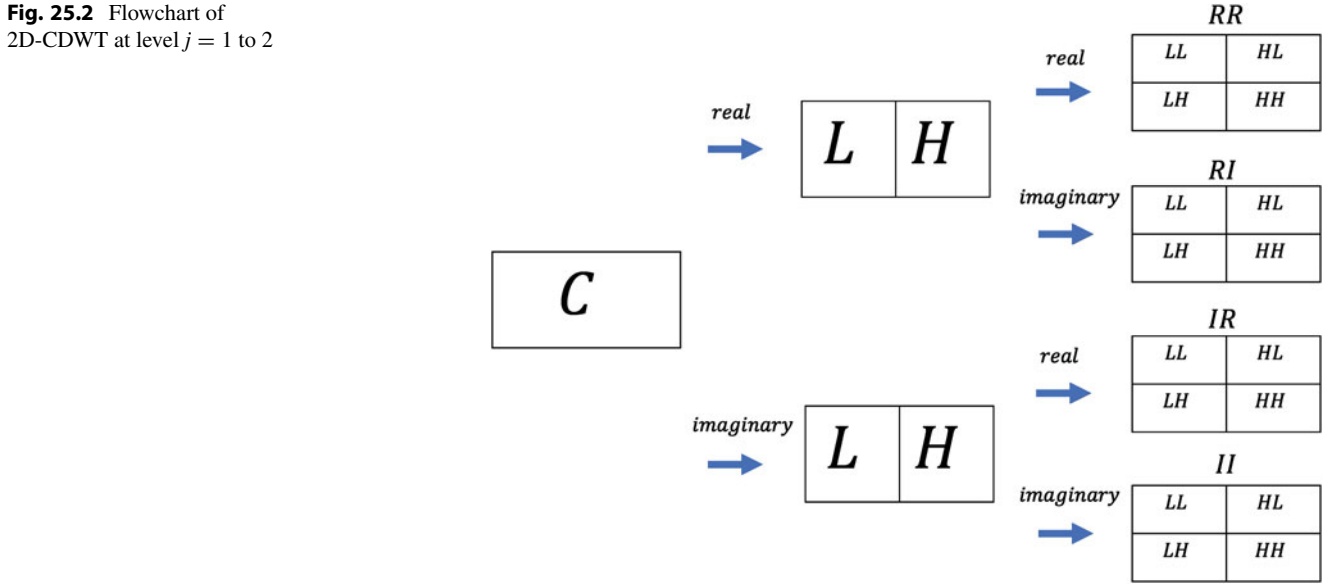


Fig. 25.3 Index of each frequency component at levels 1 and 2 [3]

- If $(n, m) = (1, 1)$, the filter used in 2D-CDWT is used.

The transformation at $(n, m) = (1, 2)$ with level $j = 2$ is shown in the following equation. The $d^{(n, m)}_j$ in the equation represents the frequency component at level j . The notation rules for R and I are the same as for 2D-CDWT.

$$d_{j+1, n_y, n_x}^{RI, (2n-1, 2m-1)} = \sum_{k_y, k_x} h_{2n_x - k_x}^R g_{2n_y - k_y}^R d_{j, k_y, k_x}^{RI, (n, m)},$$

$$d_{j+1, n_y, n_x}^{RI, (2n, 2m-1)} = \sum_{k_y, k_x} h_{2n_x - k_x}^R h_{2n_y - k_y}^R d_{j, k_y, k_x}^{RI, (n, m)},$$

$$d_{j+1, n_y, n_x}^{RI, (2n-1, 2m)} = \sum_{k_y, k_x} g_{2n_x - k_x}^R g_{2n_y - k_y}^R d_{j, k_y, k_x}^{RI, (n, m)},$$

$$d_{j+1, n_y, n_x}^{RI, (2n, 2m)} = \sum_{k_y, k_x} g_{2n_x - k_x}^R h_{2n_y - k_y}^R d_{j, k_y, k_x}^{RI, (n, m)}$$

The same procedure used in 2D-CDWT can be used to add direction selectivity to 2D-CWPT. Note, this procedure must

be performed for each index (n, m) .

$$D_{j, n_y, n_x}^{R0, (n, m)} = \frac{d_{j, n_x, n_y}^{RR, (n, m)} + d_{j, n_x, n_y}^{II, (n, m)}}{2},$$

$$D_{j, n_y, n_x}^{I0, (n, m)} = \frac{d_{j, n_x, n_y}^{IR, (n, m)} - d_{j, n_x, n_y}^{RI, (n, m)}}{2},$$

$$D_{j, n_y, n_x}^{R1, (n, m)} = \frac{d_{j, n_x, n_y}^{RR, (n, m)} - d_{j, n_x, n_y}^{II, (n, m)}}{2},$$

$$D_{j, n_y, n_x}^{I1, (n, m)} = \frac{d_{j, n_x, n_y}^{IR, (n, m)} + d_{j, n_x, n_y}^{RI, (n, m)}}{2}$$

25.3 Proposed Method

We explained the developed method for detecting early-stage colorectal cancer in this section and then described preprocessing performed on the endoscopic images. Further, preliminary experiments and their results were described.

We used coefficient p_n of the literature [5] in this study, and a range of $0 \leq n \leq 24$ to create the filter.

25.3.1 Preprocessing for Endoscopic Image

We performed preprocessing to reduce the amount of strong light in the endoscope image, and surrounding regions on the left and right sides of the image folded back to the target area. Figure 25.4 shows a portion of Fig. 25.1 with this process applied.

Using a single endoscope image, we cropped images for the proposed method. Figure 25.5 shows that we divided the



Fig. 25.4 Image after preprocessing

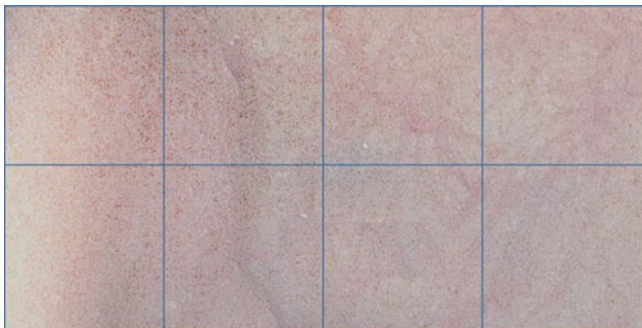


Fig. 25.5 Block division For endoscopic image

endoscope image into several regions. Further, we cropped 50 analytical images with a size of 100×100 from each region.

Each component image is converted from RGB to XYZ color space and graded to 15 grayscale as physicians perform endoscopy visually. The number of pixels, 100×100 , and the Sturges formula was used to determine the gradation number. However, Y component images are omitted from the analysis because they indicate luminance. We applied the proposed method to images created using this method to detect cancer for each region.

25.3.2 Preliminary Experiment

The ability of 2D-CWPT to capture the characteristics of early-stage colorectal cancer was confirmed, and the variance of each frequency component, calculated. From Sect. 25.2.2, 2D-CWPT generates a total of 4^{j+1} of $D_j^{(n,m)}$ in the transformation of level j . Among them, the frequency components belonging to $(n, m) = (1, 1)$ are low-frequency components. In this study, we assumed $j = 2$ and calculated the variance of the $4^3 - 4 = 60$ components, excluding the low-frequency components. Further, as we can obtain 60 variances from one endoscope image, the total number of data is $N \times 60$ when the number of images is N .

We applied PCA to the created data. Figure 25.6 shows a scatter graph of the endoscopic images, with the first and

second principal component values plotted on each axis. Normal and abnormal regions are represented by blue and red dots, respectively, and a mixture of normal and abnormal region are represented by green dots. It can be observed that the first principal component values of abnormal regions data are greater than 0.

25.3.3 Proposed Detection Method

We described our proposed method for detecting early-stage colorectal cancer in this section. We obtained an analytic image that applied the preprocessing of Sect. 25.3.1 to a single endoscopic image. Further, we applied 2D-CWPT and obtained the variance of each component with the exception of the low-frequency component. For each image, we created vectors with each variance as an element and used PCA. We calculated the probability of early-stage colorectal cancer by the percentage of the number of images with the first principal component value exceeding 0 for each region. We defined the first principal component as the degree of abnormality in this study.

25.4 Experimental Results

The outcomes of applying the proposed method are shown in this section. We tested the suggested approach on endoscopic images of colorectal cancer in its early stages, confirming that it can detect cancer. The detection accuracy of the suggested method was then compared to that of the local binary pattern [4].

25.4.1 Classification Experiment

We applied the proposed method for Fig. 25.5. Figure 25.7 shows the probabilities calculated for each region. The region surrounded by a red frame in the figure indicates early-stage colorectal cancer, and the region surrounded by a blue frame is the normal region. However, the region surrounded a green frame indicates a mixed region by early-stage colorectal cancer and the normal region. The figure shows that early-stage colorectal cancer and the normal region can be detected correctly. One area was considered to be early-stage colorectal cancer with a probability of 24%, while the other was considered to be a normal area in the region where both areas were mixed. Next, we prepared 30 endoscopic images and confirmed the average of probability of early-stage colorectal cancer by cancer region, normal region, and mixed region. Table 25.1 shows this result. The average value was higher in the cancer region than in the normal region.

Fig. 25.6 Examples of 1st to 2nd principal component values

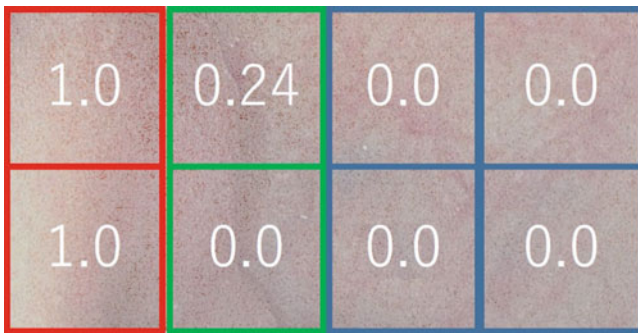
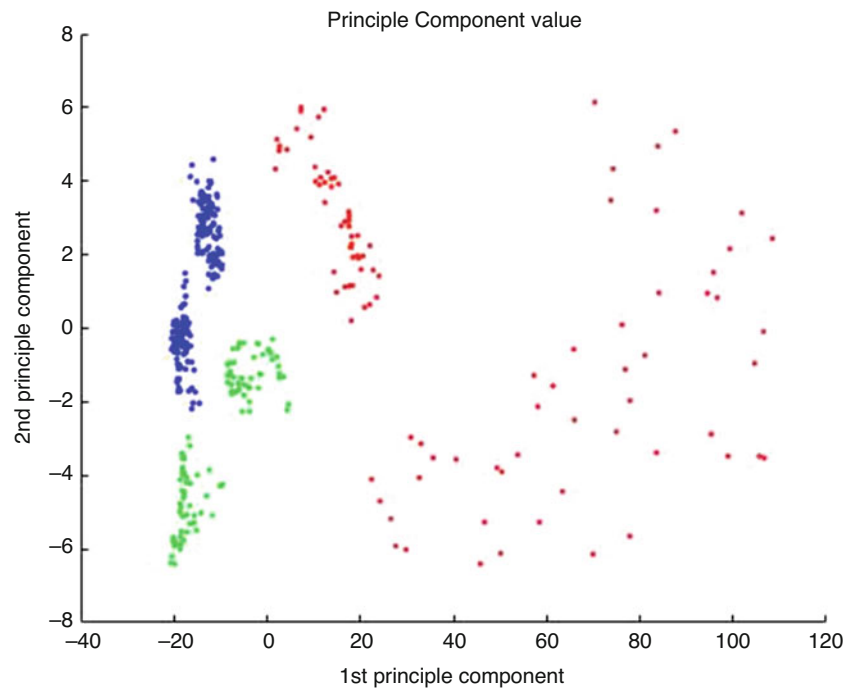


Fig. 25.7 Anomaly rate of Image in Fig. 25.5

Further, we also confirmed which frequency components affect the first principal component. The correlation coefficient between each frequency component’s variance and the value of the first principal component was calculated. Figure 25.8 shows the frequency components whose correlation coefficient exceeded 0.9. This figure corresponds to the indexing in Fig. 25.3. In the *II* component, it is obvious that there are multiple frequency components with strong correlation coefficients. In particular, the high-frequency components in the diagonal direction of the *II* components all have high correlation coefficients. Therefore, we concluded that the characteristics of early-stage colorectal cancer exist there.

Table 25.1 Probability of early-stage colorectal cancer for each region

	Normal Region	Cancer Region	Mixed Region
Average of probability	0.22	0.60	0.42

25.4.2 Comparison Experiment

The method’s accuracy was compared with that of the existing method [4]. The existing method extracts features using discrete wavelet transform and local binary pattern and then applies support vector machine for detection. Further, we evaluated the proposed method’s accuracy with that of 2D-CDWT rather than 2D-CWPT. Each method was applied after cutting 800 images from each of the two endoscope images. We confirmed the classification accuracy of each image for each method. Table 25.2 shows the accuracy of each method. We calculate accuracy(Acc), true positive(TP), and false positive(FP) for each method. Except for the TP in image 1, the suggested method outperforms the previous methods. Considering that the method [4] uses SVM for classification and has more parameters than the proposed method, it has a better detection method. We can also confirm that the proposed method outperforms the 2D-CDWT method in terms of accuracy.

Fig. 25.8 Frequency components with strong correlation

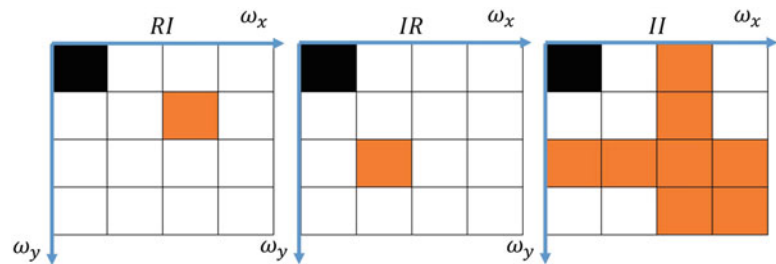


Table 25.2 Accuracy comparison between the proposed method and existing methods

	Method [4]			2D-CDWT+PCA			Proposed Method		
	Acc ^a	TP ^b	FP ^c	Acc	TP	FP	Acc	TP	FP
Image1	77.8	76.5	79.0	77.8	55.0	100	78.0	56.0	100
Image2	87.2	79.2	94.0	92.5	83.6	100	92.7	84.0	100

^aAccuracy

^bTrue positive

^cFalse positive

25.5 Conclusion

In this study, developed a method for detecting early-stage colorectal cancer using 2D-CWPT and PCA. Our method does not require labeled data because it does not involve supervised learning. However, detection is possible even when data is insufficient. Further, we have confirmed that the high-frequency components in the diagonal direction of *II* components have features of early-stage colorectal cancer. Therefore, we can explain the reason for detecting by the proposed method. We compared the proposed method's detection accuracy to that of the existing method [4] and confirmed that the proposed method's accuracy was greater. Further, the proposed method's accuracy was higher than that of a method that switched from 2D-CWPT to 2D-CDWT, confirming that 2D-CWPT was preferable for feature extraction in early-stage colorectal cancer. However, it is insufficiently accurate in regions that include both early-stage colorectal cancer and normal regions. Further, we have confirmed that the method is ineffective for images that do not contain cancer. Thus, we must find a more significant features of early-stage colorectal

cancer and develop an effective detection method for all endoscopic images.

Acknowledgments This work was partially supported by JSPS KAKENHI Grant Number 19K03623.

References

1. American Cancer Society, 5-year relative survival rates for colon cancer (2021). <https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html>. Accessed Jan 29, 2021
2. Y. Sakai, S. Takemoto, K. Hori, M. Nishimura, H. Ikematsu, T. Yano, H. Yokota, Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network, in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2018-July*, art. no. 8513274 (2018), pp. 4138–4141. <https://doi.org/10.1109/EMBC.2018.8513274>
3. T. Kato, Z. Zhang, H. Toda, T. Imamura, T. Miyake, A novel design method for directional selection based on 2-dimensional complex wavelet packet transform. *Int. J. Wavelets Multiresolution Inf. Process.* **11**(4), 1360010 (2013). <https://doi.org/10.1142/S0219691313600102>
4. S. Charfi, M.E. Ansari, Computer-aided diagnosis system for colon abnormalities detection in wireless capsule endoscopy images. *Multimedia Tools Appl.* **77**(3), 4047–4064 (2018). <https://doi.org/10.1007/s11042-017-4555-7>
5. H. Toda, Z. Zhang, Perfect translation invariance with a wide range of shapes of Hilbert transform pairs of wavelet bases. *Int. J. Wavelets Multiresolution Inf. Process.* **8**(4), 501–520 (2010). <https://doi.org/10.1142/S0219691310003602>
6. R.R. Coifman, Y. Meyer, V. Wickerhauser, Wavelet analysis and signal processing, in *Wavelets and Their Applications* (Springer, Berlin, 1992)
7. N. Kingsbury, Complex wavelets for shift invariant analysis and filtering of signals. *Appl. Comput. Harmonic Anal.* **10**(3), 234–253 (2001). <https://doi.org/10.1006/acha.2000.0343>

Visualizing 3D Human Organs for Medical Training

Joshua Chen, Paul J. Cuaresma, Jennifer S. Chen, Fangyang Shen, and Yun Tian

Abstract

Three-dimensional (3D) models have been used as essential tools in medical training. In this study, we visualize 3D models of human organs with graphics software for the purpose of training medical students. This study investigates whether 3D organ visualizations will be more recognizable to medical students than two-dimensional (2D) organ images. In our experiments, the models were shown to health science students to determine how useful they were in training and we compared the use of 3D models with 2D images. We conclude that the 3D organ models we used are more likely to be recognized by the students.

Keywords

3D · Visualization · Medical training · Blender · 3D human organs · 2D organ images · Medical education · STL files · Health Science · 3D models

J. Chen

Mililani High School, Central School District of Mililani Town, Mililani, HI, USA

P. J. Cuaresma

Guam High School, DoDEA Pacific West District, Agana, Guam
e-mail: paul.cuaresma@dodea.edu

J. S. Chen

Medical Education International, Christian Medical Dental Association, Mililani, HI, USA

F. Shen

Department of computer system tech., New York City College of Technology, New York, NY, USA
e-mail: fshen@citytech.cuny.edu

Y. Tian (✉)

Department of Computer Sciences & Electrical Engineering, Eastern Washington University, Spokane, WA, USA
e-mail: ytian@ewu.edu

26.1 Introduction

Three dimensional (3D) visualization provides a way for users to interactively examine a 3D object. Technologies such as volume renderings, are adopted to visualize CT scans as a 3D model, providing physicians with more data to determine the course of treatment for patients [1, 2]. 3D models provide surgeons with more detailed information, thus enhances confidence in their decisions when operating on the patient. For example, a patient may have a crushed pelvis from a car accident. After being able to see the patient's pelvis from all angles using a 3D model, the doctors are able to resolve the best way to operate on the patient.

Simulations based on 3D models play a critical role in the medical field. Simulation may serve as a training technique that replaces real experiences with synthesized ones [3–5]. It allows people to practice repeatedly on a computer until they have mastered the skills, which gives the students hands-on experience. In other situations, simulation can help medical students develop skills that would otherwise entail quite scarce materials or opportunities [6]. For example, simulation can be used as an alternative to using real cadavers and live patients. This is important because cadavers are scarce resources [7].

Sometimes, real patients are used to help trainees practice medical techniques. As a result, those patients may have to receive invasive operations [8]. Medical students also use simulations to practice surgical procedures in rare, unusual, or high risk situations that live patients cannot offer [9, 10].

The widely applied 3D objects and models in health science motivate us to adopt 3D visualizations techniques in fundamental training for medical students. In this work, we attempt to verify whether health science students with a basic understanding of human anatomy will show higher accuracy

and confidence when attempting to identify 3D models of human organs as opposed to two-dimensional (2D) images.

In this study, we visualize 3D models of human organs with graphics software for the purpose of training medical students. This study investigates whether 3D organ visualizations will be more recognizable to medical students than 2D organ images. In our experiments, the models were shown to health science students to determine how useful they were in training and we compared the use of 3D models with 2D images. We conclude that a majority of 3D organ models we used are more likely to be recognized by the students.

The rest of the paper is organized as follows. Section 26.2 presents our methodology and the data set in our research. In Sect. 26.3, we present and discuss our experiment results. We conclude by discussing the limitations, our conclusions and future work in Sect. 26.4

26.2 Research Methodology

26.2.1 Overview

In the project, 3D human organ models were visualized with Blender, a free and open-source 3D visualization software that can be used for modeling, rendering, and animation [11–13]. 3D models of human organs in the format of .stl files were downloaded from embodi3.com, sketchfab.com and turbosquid.com. Six 3D medical models were employed in the experiments. We created several short videos by screen-recording the rotations and translations of the 3D models when interacting with them in Blender. Then, we present to two classes of health science students the videos of the 3D models, together with some 2D images of the same set of human organs. In a survey, multiple choice questions were given to the health science students, asking what human organs they have observed in both the videos that contain the 3D models and in the images of 2D human organs. We reported the accuracy of recognizing the human organs and the confidence level of the students.

26.2.2 Experiment Setup

By default, the 3D models that we downloaded from the websites were not rendered on the screen with realistic color and lighting settings, as shown in Fig. 26.1.

We improved the 3D models with more realistic colors and lighting parameters in the visualization using Blender. A picture of the improved 3D model is shown in Fig. 26.2.

Blender software was installed and used on macOS Catalina with a version 10.15.2. The videos of visualizing the 3D models along with the survey questions were posted at the web URL: <https://create.kahoot.it/details/>

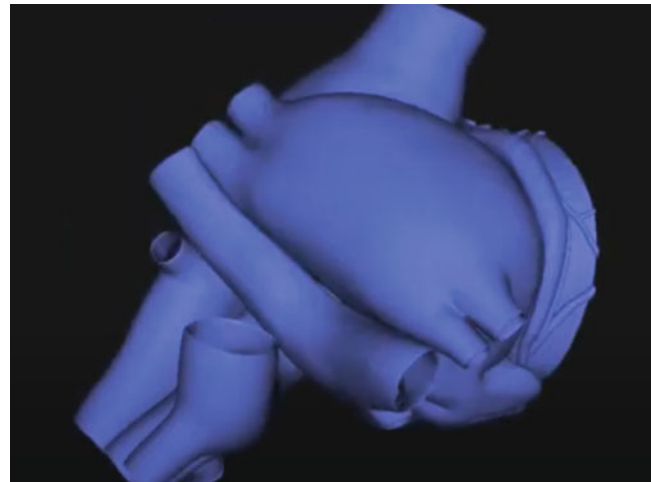


Fig. 26.1 3D model of a human heart

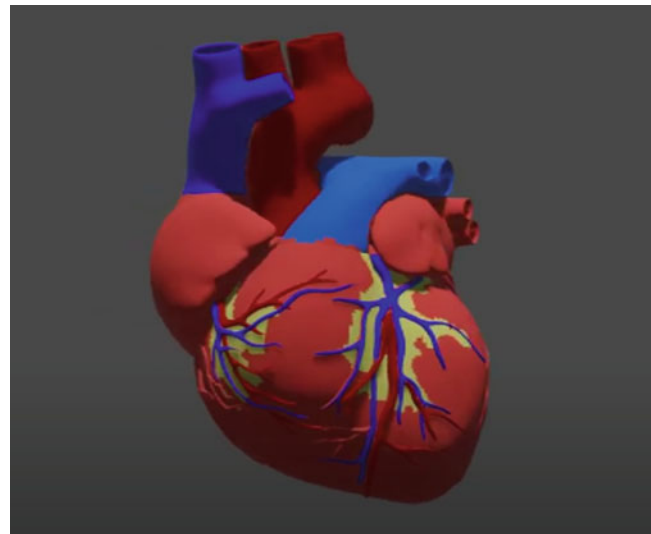


Fig. 26.2 The improved 3D model of the human heart

[group-2-a/658bdc38-06c0-40dc-9560-94a5b8964066](https://create.kahoot.it/share/group-2-a/658bdc38-06c0-40dc-9560-94a5b8964066). The corresponding 2D images of the same set of human organs along with survey questions were posted at <https://create.kahoot.it/share/group-1-a/54c345ce-a042-4d3e-b16c-3f25fa3eec88>.

26.2.3 Experiment Implementation

Two sets of surveys were given two different classes, Health Science I and Health Science II. There were 23 students in Health Science I and 6 students in Health Science II. Feedback from each class was collected and analyzed.

The specifics are described below. First, a six-second long video was recorded when a model were rotating during visualization to show all sides of the model. These videos

were uploaded to YouTube so they could be inserted into a Kahoot quiz. Second, two Kahoot quizzes/surveys were created, referred to as 3D-model survey and 2D-image survey respectively. Third, the students were informed about the survey and were given instructions before taking the surveys. Specifically, within *each* class, students were divided into two groups, group One and group Two. Group One was given the 3D-model survey, while group Two was given the 2D-image survey. This was done to minimize the effects between the students on who would be shown the 3D models first or vice versa. When the students in group One were taking the 3D-model survey, the students in group Two turned around to face away from the screen. Then, the students in group Two took the 2D-image survey, when students in group One faced away from the screen.

During the survey time, the videos and images were shown on a large screen and the students answered the questions in the predesigned survey. Each survey question were displayed for five seconds and ten seconds were given to the students to choose what they thought a 3D model or a 2D image was represented between four multiple choice answers. Then, the survey asked the students to rate their confidence level on a four-point scale within a twenty-second limit. This was repeated for every model.

Two days after the first survey were given, group One was given the 2D-image survey and group Two was given the 3D-model survey. The students had to wait two days before they took the other survey, because they needed enough time to forget what they had already seen. The students were also instructed not to share the answers to the survey questions. Seventh, feedback from each class was collected from the [kahoot.com](https://www.kahoot.com) website and analyzed.

In the experiment, the independent variables were the 3D models of human organs. The color and shape of the models can be changed to make it easier for the health science students to identify the models. The dependent variables are the accuracy of the students' answers and their confidence level. The students were a random variable because every student was different and could not be controlled. Each student had a distinct level of knowledge of the human body and a different level of confidence. The constants in this experiment were the survey that the students took and the class the students were enrolled in. Each student had the same set of models/images, amount of time, and multiple choice answers to choose from. The results were separated by different classes because every class has a different curriculum and different interests. These differences have an impact on the students' knowledge of human anatomy and would affect the results of the survey.

26.3 Experimental Results

We present the experimental results in this section, including the findings as to the adoption of the 3D models versus the 2D images.

26.3.1 Results for Our 3D Models

The recognition correctness rate are shown in Figs. 26.3 and 26.5 for two different Health Science classes. The percentages were calculated by dividing the number of students who gave the correct answer by the total number of students. The corresponding confidence level of the student participants are presented in Figs. 26.4 and 26.6.

26.3.2 Analysis and Discussion

This study attempts to verify whether 3D models will be more effective to be recognized in medical training. This was done by comparing how accurately health science students can recognize human organs represented in 2D images against 3D visualizations. As shown in Figs. 26.3 and 26.5, results indicate that the 3D models are very easily recognized because all but three of the tests of the 3D models had at least 90% of all answers correct.

Moreover, the results from the Health Science I class showed that on average, the use of 3D models achieved a marginal higher correctness rate than that of using the 2D images. The correctness rate for the 3D models compared to 2D images increased by an average of 3%, as shown in Figs. 26.3 and 26.5. For the students in the Health Science I class, as shown in Fig. 26.3, the 3D models have been recognized with a *higher* accuracy than the 2D images, except for the Lungs and Kidney.

However, as shown in Figs. 26.4 and 26.6, on average, the students who answered correctly indicated higher confidence when using the 2D image representations. The survey from the Health Science II class also showed higher accuracy rates when the 2D representations were presented. The pattern was opposite to what is expected. It was expected that the 3D models would be more accurate and have more confidence because they could be viewed from different angles. This could have occurred because the textbooks that were used in the class all show 2D representations of organs. The students

Fig. 26.3 Recognition Accuracy for the Health Science I Class

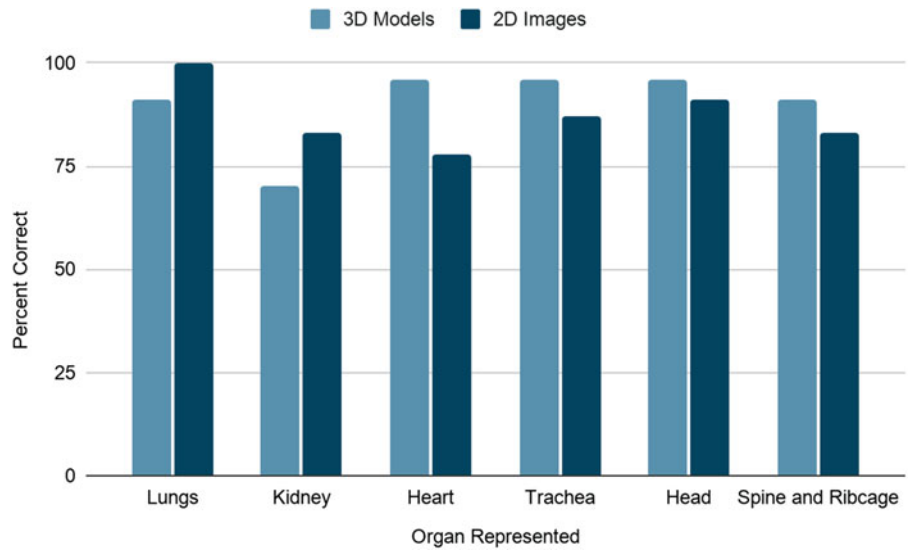


Fig. 26.4 Average Confidence Level of The Students in Health Science I

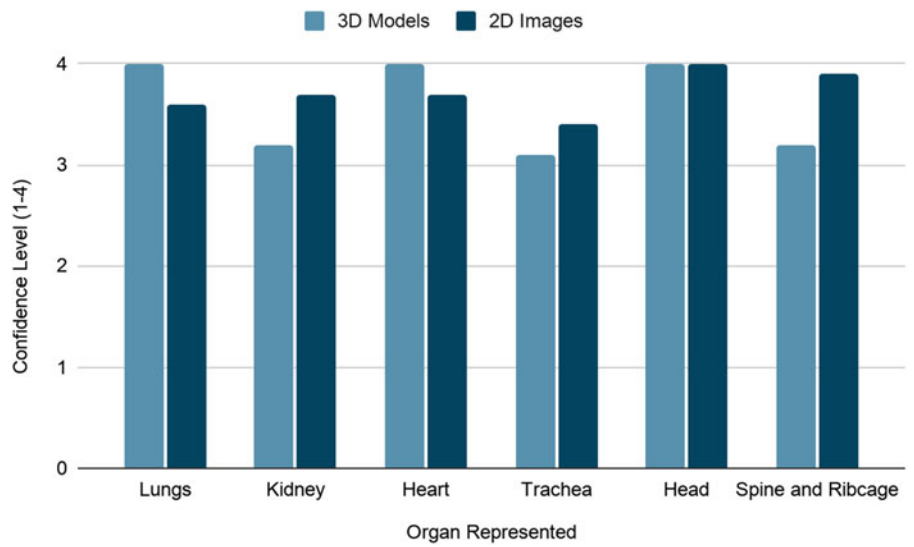


Fig. 26.5 Recognition Accuracy for the Health Science II Class

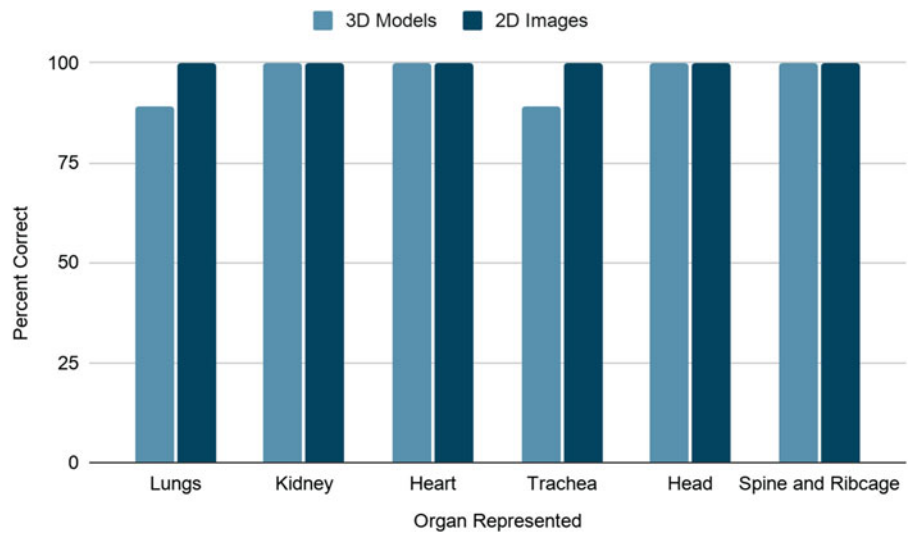
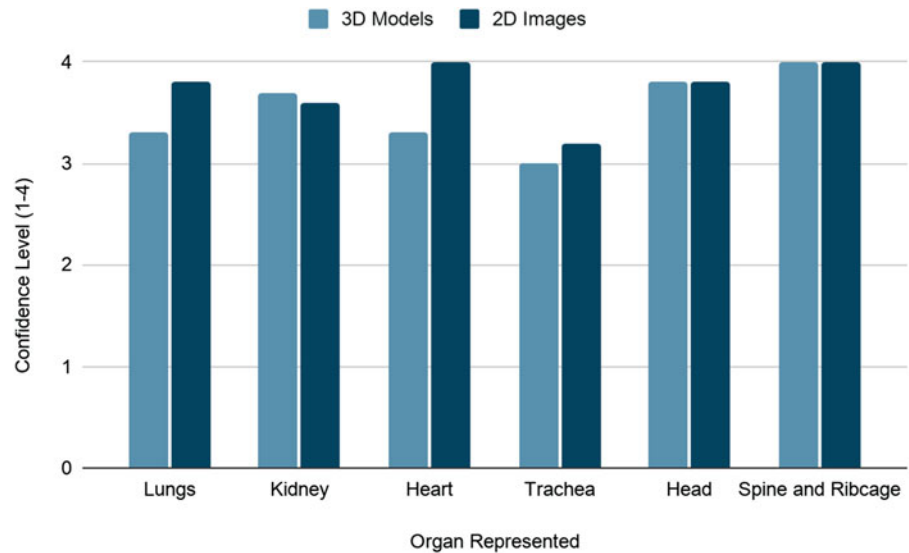


Fig. 26.6 Average Confidence Level of The Students in Health Science II



would be more used to the 2D images and were able to recognize the 2D images more easily. In addition, the 3D kidney had less accuracy than all of the other 3D organs. This could be also because the kidney was shown to the students first and the students had to get used to viewing a 3D model. The experiment could be improved by showing some 3D models that were not used in the experiment for warm-up purposes, before showing the medical models in the experiment.

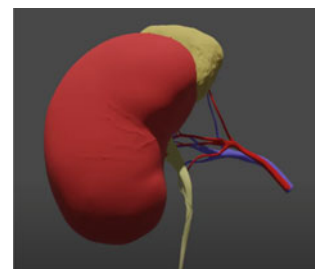
Moreover, the more advanced class was able to answer the questions with almost 100% of accuracy for both 2D images and 3D models, as observed in the Health Science II class. This could be because the advanced students had more knowledge of what organs look like and did not see a difference between the 2D and 3D representations. Because they had more experience with identifying human organs, they were able to identify the images effortlessly. This experiment could be improved by making the test harder for advanced students in order to better see the difference between 2D images and 3D models.

To summarize, on average, using the 3D representations had the same amount of or more correct answers than the 2D images. Some of the 3D visualizations are able to replace 2D images as learning tools while other 3D models still need to be improved.

lighting to make the models more realistic. For example, parts of the 3D models may be made transparent and allow the viewers to see the inside of the organ. We can also consider using other software programs to visualize the 3D models.

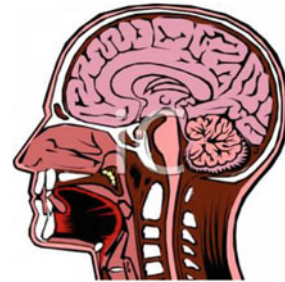
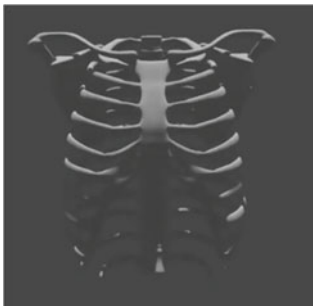
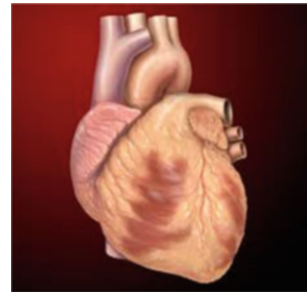
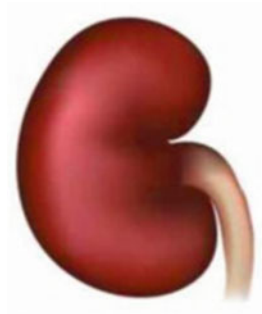
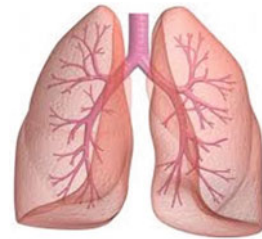
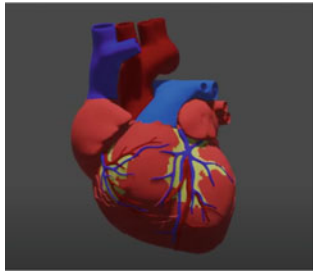
A.1 Appendices

A.1.1 Appendix A: The 3D Models Used in the Survey



26.4 Conclusions and Future Work

In this work, we visualize 3D models of human biological structures using graphics software. The results showed that the 3D models achieved the same amount of or more accurate recognition rate than that of the 2D images. In the future, the 3D models could be further improved by adding more details, specifically changing the color, shape, resolution, and



A.1.2 Appendix B: The 2D Images Used in the Survey



References

1. G. Zörnack, J. Weiss, G. Schummers, U. Eck, N. Navab, Evaluating surface visualization methods in semi-transparent volume rendering in virtual reality. *Comput. Methods Biomech. Biomed. Eng. Imaging Visual.*, 1–10 (2021)
2. Grant, Gerald T. “3D volume rendering and 3D printing (additive manufacturing), in *Emerging Imaging Technologies in Dento-Maxillofacial Region, An Issue of Dental Clinics of North America*, E-Book 62(3), p. 393. (2018).
3. O. Chernikova, N. Heitzmann, M. Stadler, D. Holzberger, T. Seidel, F. Fischer, Simulation-based learning in higher education: a meta-analysis. *Rev. Educ. Res.* **90**(4), 499–541 (2020)
4. J. Kim, C. Choi, S. De, M.A. Srinivasan, Virtual surgery simulation for medical training using multiresolution organ models. *Int. J. Med. Robot. Comput. Assist. Surg* **3**(2), 149–158 (2007)
5. H.M. Nassar, A. Tekian, Computer simulation and virtual reality in undergraduate operative and restorative dental education: A critical review. *J Dent Edu.* **84**(7), 812–829 (2020)
6. R. Umoren, S. Bucher, D.S. Hippe, B.N. Ezenwa, I.B. Fajolu, F.M. Okwako, J. Feltner, et al., eHBB: a randomised controlled trial of virtual reality or video for neonatal resuscitation refresher training in healthcare workers in resource-scarce settings. *BMJ open* **11**(8), e048506 (2021)
7. H. Brenton, J. Hernandez, F. Bello, P. Strutton, S. Purkayastha, T. Firth, A. Darzi, Using multimedia and Web3D to enhance anatomy teaching. *Comput. Educ.* **49**(1), 32–53 (2007)
8. R.J. Scalse, V.T. Obeso, S. Barry Issenberg, Simulation technology for skills training and competency assessment in medical education. *J Gen. Inter. Med.* **23**(1), 46–49 (2008)
9. P.A. Guze, Using technology to meet the challenges of medical education. *Trans. Am. Clin. Climatol. Assoc.* **126**, 260 (2015)
10. R.A. Agha, A.J. Fowler, The role and validity of surgical simulation. *Int Surg* **100**(2), 350–357 (2015)
11. E.T. Guevarra, Mendoza., *Modeling and animation using blender: Blender 2.80: The rise of eevee* (Apress, 2019)
12. M. Gárate, Voxel datacubes for 3D visualization in Blender. *Public. Astron Soc Pacif* **129**(975), 058010 (2017)
13. J.D. Durrant, BlendMol: Advanced macromolecular visualization in Blender. *Bioinformatics* **35**(13), 2323–2325 (2019)

An Information Management System for the COVID-19 Pandemic Using Blockchain

27

Marcelo Alexandre M. da Conceicao, Oswaldo S. C. Neto, Andre B. Baccarin, Luan H. S. Dantas, Joao P. S. Mendes, Vinicius P. Lippi, Gildarcio S. Gonçalves, Adilson M. Da Cunha, Luiz A. Vieira Dias, Johnny C. Marques, and Paulo M. Tasinaffo

Abstract

During the 1st Semester of 2020, 25 students from the Aeronautics Institute of Technology (*Instituto Tecnológico de Aeronáutica – ITA*) in São José dos Campos, SP, Brazil developed the academic project “Specific Technological Solutions for Special Patients with Big Data”, in Portuguese *Projeto STEPES-BD: Soluções Tecnológicas Específicas para Pacientes Especiais e Sistemas em Bancos de Dados*. They have accepted the challenge of using a technological approach to help manage and combat the Sars-CoV-2 Virus Pandemic (COVID-19). At that time, the lack of shared data between public and private agencies and the need for faster information flow were considered the main information for combating the spread of the disease to guide the development of an Information Management System for the COVID-19 Pandemic, involving the essential data from Electronic Health Records (EHRs). The combination of some emerging technologies like Big Data and Blockchain, together with the Scrum Framework (SF) and the Interdisciplinary Problem-Based Learning (IPBL) enabled those students from three different academic courses to develop a computer system prototype based on the pressing needs caused by this disease. This article describes the development of the main deliverables made by those graduate students in

just 17 academic weeks right from the beginning of the COVID-19 Pandemic crisis in the 1st Semester of 2020.

Keywords

Software Blockchain · Interdisciplinary education · Problem-based learning · Electronic health records · Scrum · Big Data · Database · Decentralized control · Digital assets · Public healthcare

27.1 Introduction

This article presents an experience of students from the Aeronautics Institute of Technology (*Instituto Tecnológico de Aeronáutica – ITA*) in the development of the academic project “Specific Technological Solutions for Special Patients with Big Data”, in Portuguese *Projeto STEPES-BD: Soluções Tecnológicas Específicas para Pacientes Especiais e Sistemas em Bancos de Dados*, applying Interdisciplinary Problem-Based Learning (IPBL) for monitoring patients with suspected COVID-19.

The STEPES-BD project aimed to develop a computer system, applying knowledge from the development of a practical and interdisciplinary project as a Proof of Concept (PoC). It involved students from 3 different courses of the Graduate Program in Electronic and Computing Engineering in the Area of Informatics (*Programa de Pós-Graduação em Engenharia Eletrônica e Computação na Área de Informática – PG/EEC-I*), using the Scrum Framework (SF) in an academic environment together with some emerging technologies like Big Data, Blockchain, among others.

This project was developed by 25 students from the following courses on the 1st Semester of 2020: CE-245 Information Technologies; CE-240 Database Systems Project; and CE-229 Software Testing. Due to the COVID-19 Pandemic,

M. A. M. da Conceicao (✉) · O. S. C. Neto · A. B. Baccarin
L. H. S. Dantas · V. P. Lippi · G. S. Gonçalves · A. M. Da Cunha
L. A. Vieira Dias · J. C. Marques · P. M. Tasinaffo
Computer Science Division, Aeronautics Institute of Technology –
ITA, São José dos Campos, Brazil
e-mail: marcelomamc@fab.mil.br

J. P. S. Mendes
State Technology Educational Center Paula Souza, São José dos
Campos, Brazil

the project was carried out remotely through meetings by video conference.

The project addressed the problem of the Sars-CoV-2 virus also known as COVID-19. To carry out efficient monitoring of infected patients, an Information Management System was developed capable of sharing data between the main actors of a Health System, aiming to provide different views on the spread of the disease and appropriate decision making.

One of the impacting factors found in the process of combating the COVID-19 Pandemic was the ineffectiveness of sharing data from the main actors of the process in the Health System: the Government and the Society [1].

This sharing involves dependencies, compilations, periodic updates, and monitoring of real data, based on real-time data collections considered essential for decision making. The impacts of this virus in the supply chain and health services caused price increases and product shortages, highlighting the need for developing a system that would provide traceability of data to make up the logistics of inputs and health products, to improve the applicability of involved resources from the public and private sectors [2].

Within this context, the STEPES-BD project was specified to provide the monitoring of infected patients, through appropriate data sharing and medical records, using Electronic Health Records (EHR) [3], creating standards and providing real-time health data analysis. This Project was conceived and developed to make data related to the hospital supply chain more transparent, providing traceability and immutability of data in a real environment, using Big Data, Blockchain, and other emerging technologies.

In addition, this prototype-level system design mainly involved: Scalability of Relational and Non-Relational Databases; Big Data; Cloud Computing; Internet of Things (IoT); as well as a real environment for applications of Software Testing methods and techniques.

27.2 Tailoring the Scrum

Due to the social distance required by the COVID-19 Pandemic, it was necessary to adjust the Totally Integrated Scrum (TIS) method to the development of the project by the students, according to the Scrum model for Distributed Agile Development [4]. In this adaptation to a TIS, the Scrum Teams (STs) were considered to be multi-functional and their members were geographically distributed, requiring frequent and periodic inter-team communications and adopting Scrum of Scrums meetings.

As it is an academic project with well-defined schedule restrictions in just 17 weeks and of variable scope, the Project management was carried out using the best practices of Scrum. For that, the time boxes, the intervals between the ceremonies, and the artifacts had to undergo adaptations, to allow students more opportunities for contacts with the agile software development management method.

To meet these adaptations, the STEPES-BD project was divided into four segments, each of which being assigned to the following STs: Patients; Physicians; Hospitals; and Suppliers. Although the STs had specific scopes in their segments, they started to work together since the Sprint 2 in an interdisciplinary effort to address activities related to integrations, involving the different needs of infrastructure and technologies of Big Data, Blockchain, among others.

The selection and distribution of the students by the STs were carried out with the participation of students and professors soughing to balance quantitative and qualitatively the members of each ST, adjusting their skills, qualifications, and previous experiences. In addition, the following Scrum Roles exercised by the students were also adapted, when two new roles were created and other two renamed within the project:

- **General Product Owner (GPO)** – The Product Owner (PO) role was renamed as General Product Owner (GPO);
- **Backup of General Product Owner (Bkp of GPO)** – This new academic role was created to replace the GPO and provide immediate answers, in case of absences of the General Product Owner (GPO);
- **Team Scrum Masters (TSMs)** – The role of Scrum Master (SM) was renamed as TS Master (TSMs); and
- **Backup of Team Scrum Master (Bkp of TSM)** – This new academic role was created to replace the TSM and to provide more immediate responses in the eventual absence of the TSMs.

Professors acted as Stakeholders and monitored the development of the computer system by reviewing the artifacts generated by the students, periodically auditing: the Product Backlog; Sprints' Backlogs; Sprints' Kanbans; Test Cases' Spreadsheets; among the minimum, necessary, and sufficient documents, by using Burndown Charts metrics.

To assess the levels of maturity reached by the STs in the development of their Segments, involving the ceremonies performed and the preparation of the artifacts, the Tuckman Maturity Model (TMM) was used. Thus, the STs based on self-management began to be periodically checked and monitored for their evolution in maturity levels, as the project was being developed, allowing behavioral realignments of its members throughout the four stages of the TMM: Forming, Storming, Norming, and Performing [5].

27.3 Background

This section describes the built architecture and highlights the main technologies and tools used in the development of the STEPES-BD project, in terms of its Proof of Concept (PoC).

27.3.1 The STEPES-BD Architecture

The conceived architecture was based on the N-Tiers concept (development in layers) [6] and allowed the division of the Project into the Layers of: Interface; Business; and Persistence, as shown in Fig. 27.1.

1. **Interface Layer:** This layer was responsible for the interaction and presentation of the system to the user and had modules addressed to the functions of Registries, Medical Assistances, Hospital Supplies, and Dashboards.

To provide data sharing between the STEPES-BD subsystems, the interoperability was implemented by using the Representational State Transfer (REST), a style of software architecture that defines a set of restrictions for the creation of Web Services in Application Programs Interface (APIs), as shown in Fig. 27.2.

2. **Business Layer:** In this layer, it was possible to achieve data interdependence for the completeness of the processes proposed by the STEPES-BD project, through flexibilities obtained from the open API routes in the subsystems, providing fast and efficient exchanges of information managed by the development teams.
3. **Persistence Layer:** In this layer, involving data transacted by domain entities, the MySQL Relational Database was used and the BigchainDB was introduced, as a database structured in Blockchain that made possible the use of this technology in data management and information recording, foreseeing controls through a ledger, for future conferences and guarantees of data traceability.

The system proposed in the STEPES-BD project made extensive use of the data managed in the Blockchain blocks and the mechanisms for searching, consulting, and locating medical records and contracts for the purchase and sale of hospital supplies have used the tools provided by the BigchainDB API.

The BigchainDB worked on the STEPES-BD project in a MongoDB database that, through its infrastructure, made it possible to store a framework that makes up the Blockchain blocks. For that, the BigchainDB has created

on MongoDB the collections of: Assets, Transactions, and Metadata.

The BigchainDB also enabled a wide use of the concept of Assets, representing any item of value [7]. In this sense, the Collection Assets of the database kept the most valuable records inside. In this location, materials transacted in the hospital supply logistics chain were stored. From then on, all items could be properly identified by serial numbers, lots, or any other forms of identification; and by medical records generated in consultations, containing relevant data such as identification of those involved, diagnoses, hospital items consumed during consultations, among others.

Subsequently, the Assets were encapsulated within structures in the JavaScript Object Notation (JSON) format to be traded via Blockchain, thus creating blockchains that guarantee data immutability.

Because it is a database structured in Blockchain, BigchainDB provided the STEPES-BD project with a decentralized and secure architecture system, ensuring its scalability even beyond a simple PoC. This makes it possible to use a wide network, with several nodes taking full advantage of the consensus algorithms used in BigchainDB, an intrinsic characteristic of a network supported by Blockchain technology.

Another relevant factor found was, because it is a database structured in Blockchain, with its core functioning in the MongoDB database, the query tools available in MongoDB's documentation have become important to facilitate the location of relevant information.

It is important to emphasize that data manipulation to insert new information or modifications directly into the MongoDB database is not allowed. This is because all data changes in the Collections used in the Blockchain must be managed by the internal algorithms at the service of BigchainDB and, if there was direct data manipulation in MongoDB, there will also be a change in the identification hash of the block that will contain this record. As a result, a blockchain violation will occur. These facts, traceable by the mechanisms of BigchainDB, are prevented through the consensus algorithms obtained in the implementation of a distributed system.

27.3.2 The Infrastructure for the STEPES-BD Project

The infrastructure used, for example by the Patient and Physician STs, for the outlined architecture, was built on a low-cost dedicated server composed of: an Intel I5 2.5 Ghz processor; 8 GiB of RAM; 500 GiB HDD; with the Ubuntu 18 Operating System; and responding on the Internet.

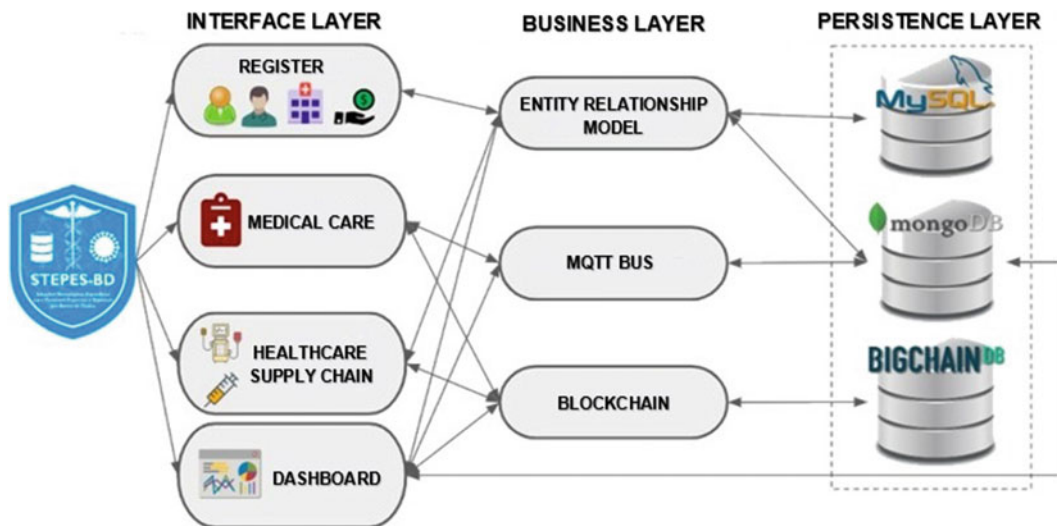


Fig. 27.1 The 3-Tier Architecture of the STEPES-BD project

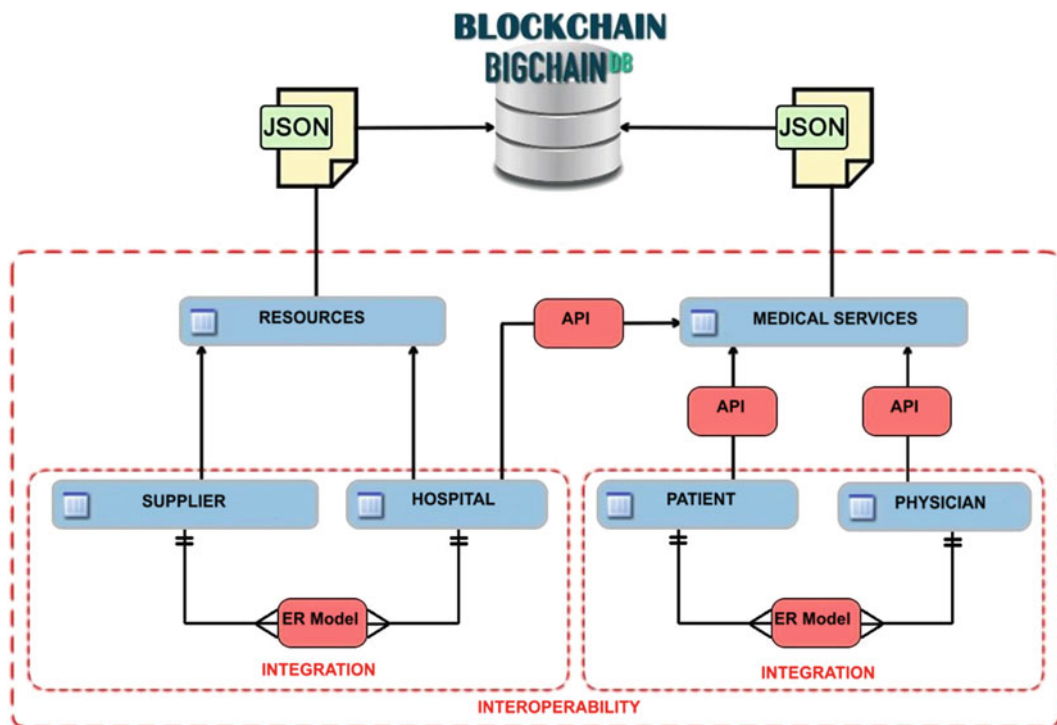


Fig. 27.2 The Data Interoperability in the STEPES-BD project

In addition, software layers were created in the Docker Container with MySQL and InfluxDB, and the APIs were developed with Python, using the micro framework Flask.

The STs of the Hospital and Supplier Segments used an infrastructure based on Platform as a Service (PaaS) with the services of Heroku and Python anywhere, to host services related to the segments. In both STs, the Google Cloud platform was used for storage in an integrated relational database.

27.3.3 The Software Application

In the front end layer, several integrated technologies were used to assist in meeting the planned requirements.

To simulate a complex application environment, where each consumer of the STEPES-BD project system (Hospitals, Doctors, Patients, Suppliers, or Public Agencies) could have the freedom to develop for their own business needs, we sought the use of data with guaranteed interoperability in the use of the RESTful API. In this sense, there was flexibility

in the use of technologies by the development teams to build the front end and back end of their applications.

Another factor that influenced the adoption of the technologies used was the diversity of skills of the team members, making interoperability the best solution so that there was not a big learning curve on homogeneous technologies for all teams and the project could be completed on fixed time.

For the back end creation to the communication between the buses of the business layer and the interface layer, the following technologies were used: Python, for the development of business rules; together with Django, for communication with the application layer. Technologies such as NodeJS, Express, and Sequelize were also used to create the back end of the Hospital follow-up and provided the integration between the Business Layer and the Presentation Layer, through technologies such as Handlebars and Bootstrap.

27.3.4 The Data Interface

For the data persistence layer, in order to guarantee data integrity, reliability, and security, Relational (SQL) and Non-Relational (NoSQL) database technologies were used.

Among them, the following stand out: MySQL, for record storage of domain entities in the Patient, Physician, Hospital, and Supplier segments; and BigchainDB, for medical care record stores and hospital supply purchases.

The BigchainDB, in addition to its technological base, contains distributed Big Data architecture inherited from the MongoDB technology (NoSQL), and it also has, as additional features, the following characteristics provided by Blockchain technology: decentralized control; immutability of data; and transferring digital assets.

The InfluxDB was used to store vital information (heart rate, blood pressure, and body temperature) of Patients monitored at a distance during the quarantine.

27.3.5 Testing the Application

In parallel with the development of the STEPES-BD project, the course CE-229 Software Testing was also taught. In it, students were introduced to the needs for software testing and the types, techniques, examples, and existing and updated tools for the elaboration of Test Cases for the Project, using the Test Driven Development (TDD).

TDD is a software development technique related to the concepts of verification and validation. This technique is based on a short cycle of repetitions, where developers initially write an automated test case and/or define desired improvements as new system functionality. Then, a code is produced to be validated in test cases, for later restructuring

and/or refactoring, using acceptable standards and transforming USs into Project functionalities.

The State Transition Test was one of the types of tests used in this Project. It was used to analyze and validate the behavior of the developed System Prototype, concerning the defined scope and the fulfillment of the specified requirements. A tool used to carry out this type of test was the State Transition Diagram, used to model certain discrete behaviors of the system, involving an initial state, states traveled, exit points, transitions defining paths to go between states and a final state of termination of the entire testing process.

Because this Project was based on Big Data, it became necessary to carry out some stress tests to predict how the system would behave in the face of a large sets.

Another type of test also used in this Project was Cloud Testing, remotely mobilizing various computing services, in terms of hardware and software. In this type of test, at least three essential services were tested for performance, security, and functionality, involving Hardware, Infrastructure, and Software: PaaS – Platform as a Service; IaaS – Infrastructure as a Service; and SaaS Software as a Service.

In this Project, the Pairwise Testing technique was also applied, coming from mathematical bases where, through orthogonal matrices, it became possible to stipulate fewer numbers of tests to be performed, ensuring better results when performing only the combinations of tests possible.

27.4 The Proof of Concept (PoC)

This section describes the functionalities of the STEPES-BD computer system presented as deliverables performed incrementally over its four Sprints

The Sprint Zero: This Sprint was dedicated to the identification of existing training and to the leveling of knowledge among students. On this Sprint zero, students took several extra-curricular courses, free of charge, via Internet. In addition to identifying the main skills of the students, initial guidelines were given for the formation of the Teams, defining the Scrum Roles to be performed, in order to allocate all members of the three Courses CE-245, CE-240, and CE-229 into the four STs of: Patients, Physicians, Hospitals, and Suppliers.

The Sprint 1: In this Sprint, the Scrum Teams developed the functionalities so that it was possible to manage the registration data of Patients, Physicians, Hospitals, and Suppliers. Before, however, it was necessary to carry out a survey on data to be persisted and to develop data models.

This Sprint considered the following deliverables: the conceptual models and the entity-relationship models for all four segments; the development of the back end and the front end of the Registration Screens for all four segments; the

creation of servers to support the computational system such as: web server; database; and the elaboration of test case for 33 artifacts.

The Sprint 2: This Sprint had three main objectives: to start the inter-team integration; to develop the pre-service process, via Web and remote Patient monitoring process; and to use the Blockchain technology to provide traceability and immutability in the transfer process of hospital supplies.

In this Sprint 2, the STs Patient and Physician started the integration of their segments, through the development of the Remote Pre-service functionality, in which Patients inform their symptoms and the System indicates if a Patient must attend a Hospital in person, to avoid unnecessary travel to Hospitals during the pandemic period. In case of need to travel to a Hospital, a Patient will be able to choose, at a distance and via Internet, the most convenient date and time for a face-to-face consultation.

At the end of the procedure in the Face-to-face attendance, the STEPES-BD system allows the generation of a Medical Report with its digital signature, using its identification keys provided by Blockchain technology.

To verify the use of the Blockchain technology in this Project, the concept of ownership over Assets was implemented. In this case, the Asset considered was a Medical Report. When creating this Asset, a Physician performs, via Blockchain, a Transaction of the CREATE type, in which the Medical Record is effectively inserted into the Blockchain.

At the end of an appointment, a physician transfers possession of the record to the Patient, characterizing the practical application of the Medical Record concept. With this, it is possible to prove in this PoC the realization of Transactions of the type TRANSFER from Physicians to Patients, thus completing the implementation of a Face-to-face Care Process in the STEPES-BD project system via Blockchain.

After Patients have attended the consultations in person, they were able to be sent to home quarantines in situations where the symptoms were of low or medium severity.

Also in this Sprint 2, the members of the STs were able to implement a Remote Monitoring Process in real time, containing vital information for this ST of Patients. In this case, remote monitoring of disease evolution in Patients became possible, collecting data on heartbeat, blood pressure, and body temperature, via Internet applications.

The Remote Monitoring functionality was implemented using the InfluxDB Time Series Database and viewed through the Grafana application. However, for data generation, no real Internet of Things (IoT) devices was used, because implementations in hardware devices were out of the scope of this Project.

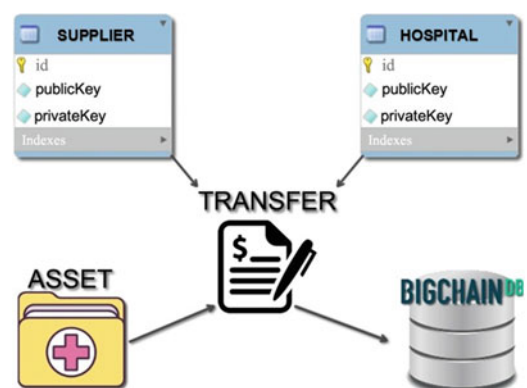
In this Sprint 2, the members of the Hospital ST and Supplier ST were also able to perform integration through the Resource View. In this case, Hospital can request from Supplier new resources such as equipments, medical supplies, and medicines, and this transfer of resources can also be registered, via Blockchain, providing guarantees of immutability and traceability.

In this case, access to the Blockchain was also accomplished through the BigChainDB tool, which abstracted all the complexity of managing the blockchain, as shown in Fig. 27.3.

In addition, the ST of the Hospital segment also implemented a register of hospital beds, so that it was possible to know, in real-time, the situation of hospital beds in the city of São José dos Campos, to support decisions of the municipal public authorities.

The Sprint 3: The main objectives of this Sprint 3 were: to carry out the integration between all the STs; to generate a large data set of fictitious data to justify the use of Big Data technologies; to provide records of face-to-face assistance; to implement and demonstrate Dashboards involving contagion maps that can be updated in real-time; to provide the creation of visualizations, assisting managers in their decision making; and to promote the exploration and implementation of EHR concepts with the traceability and immutability of Blockchain technology.

Fig. 27.3 The transfer of hospital supplies using Blockchain technology



To achieve the objectives proposed for this Sprint 3, the main deliveries made by the members of the four STs were simulated and loaded with 1M of Fictitious Patient Records and 2,000 Physician Records. With that, it was possible to generate data to update a Dashboard, synthesizing information through procedures implemented using the Spark tool.

For Proof of Concept (PoC) purposes, during the implementation of the Face-to-Face Attendance functionality, a Physician selected a Patient from his/her agenda for a Consultation, then he wrote a diagnosis in a textual field.

Based on pre-programmed options, the system made it possible for the physician to opt for the Hospitalization of the Patient in a given Hospital, and to request and register all the disposable items needed to be provided by certain Suppliers and used in a Consultation, including for example, tests of the COVID-19 virus, masks, disposable gloves, among others.

In this case, the data related to Face-to-face care involving Patients, Physicians, Hospitals, and Suppliers of materials and medicines were also consolidated through the use of Blockchain.

The Dashboard built in the Front End provided a visualization of the panorama of the COVID-19 Pandemic and made it possible to identify the critical regions of the city of São José dos Campos, with the highest incidences of those infected, as shown in Fig. 27.4.

In addition, some indicators shown in Fig. 27.5 were also implemented, to provide an overview of the COVID-19 Pandemic in the city. The front end containing the indicators

was written in React and it consumed data provided by Python APIs that was consolidated from the Patient and Physician segments, and also in NodeJS consolidated from Blockchain.

Also in this Sprint 3, the integration of functionalities between Online Pre-Attendance and Face-to-face Attendance provided the effectiveness of the initial proposal for Medical Care Records Stored in the Blockchain. This integration made it possible to track any relevant information in medical care because, through public and private keys and hash mechanisms, it became possible to identify those involved, related medicines, participating suppliers, and the attendance location.

Thus, it was ensured that all documents generated by the Prototype of the System could be certified with digital signatures, such as the Fictitious Medical Record and the Fictitious Death Certificate, as shown in Fig. 27.6.

In addition, the Electronic Death Certificate shown in Fig. 27.6 could also be implemented and issued right after the confirmation of the death of a Patient, avoiding family stresses, possibilities of fraud, or delays in communications to public authorities.

27.5 Conclusion

This article described the development of a Management Information System for the COVID-19 Pandemic using Big

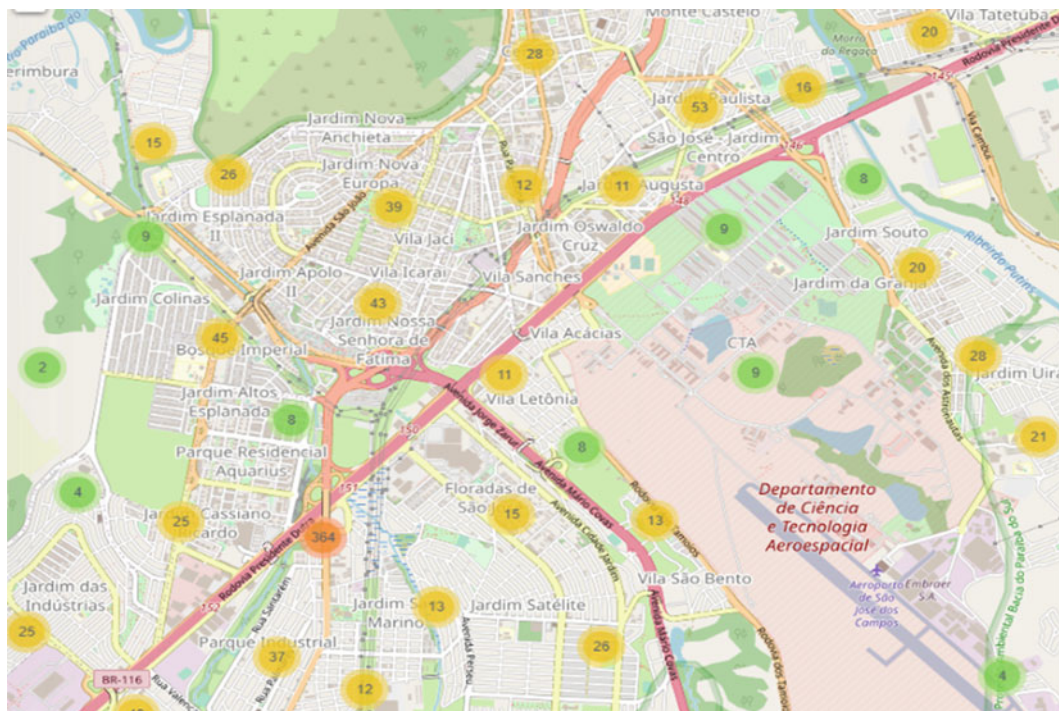


Fig. 27.4 Simulated contagion map of the city of São José dos Campos

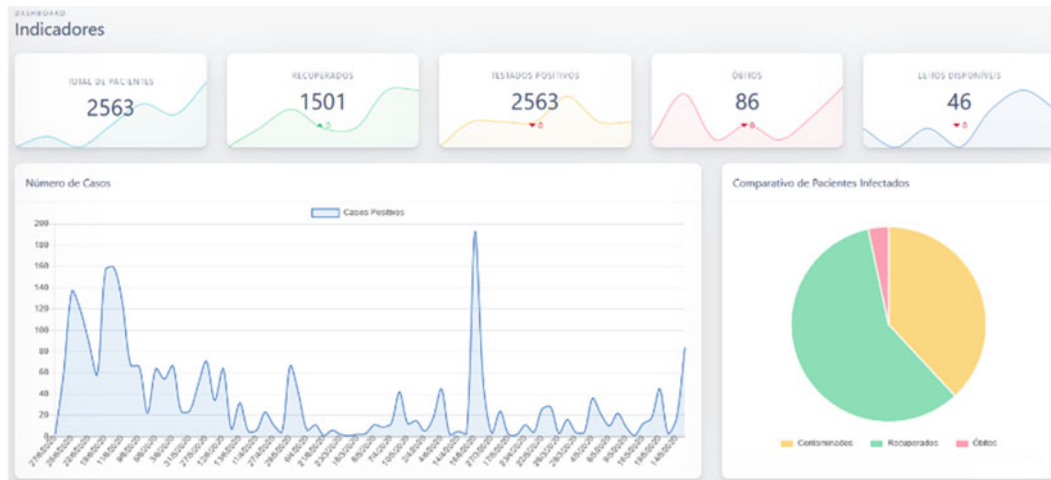


Fig. 27.5 Dashboard indicators consolidated from the physician and hospital segments and Blockchain



Fig. 27.6 An example of death certificate digitally signed.

Data, Blockchain, and other emerging technologies applied to the STEPES-BD project.

The use of these emerging concepts and technologies has provided students with the opportunity to deepen their practice in topics theoretically learned from the courses of: CE-245 Information Technologies; CE-240 Database Systems Project; and CE-229 Software Testing.

At the end of the 1st Semester of 2020, students presented the results of the STEPES-BD Project as a Proof of Concept to professors and guests from the industrial and academic sector in the city of São José dos Campos.

The developers of this Project believe that the application of several emerging technologies aimed at information

management to help combat the COVID-19 Pandemic, as a real problem that afflicts the whole world, represents an important opportunity for updates and transfers of experiences and information between academia and society.

At the end of this report involving the development of a prototype of an academic system as a Proof of Concept, the authors of this article recommend that the technologies used in the next Projects be defined, preferably, in Sprint 0, to provide better theoretical use and practical application by members of the Scrum Teams (STs).

It is also recommended that public health stakeholders should be involved in the next Projects, aiming at providing

technological transfers and generating greater benefits to society with academic projects like this one.

In the next Projects, the use of the Cloudera Cluster is also recommended, which is at an advanced stage of implementation in the Computer Science Division of ITA, to support even more scalable experiments with larger data sets. It is believed that this academic alternative is viable, instead of using commercial clusters, such as Amazon Web Servers (AWS), among others, which will make it possible to expand Proofs of Concepts and new and emerging Research and Development (P&D) involving Information Technologies (ITs) such as Big Data, Blockchain, among others.

As the base of the project is a Document-oriented database Health Level Seven (HL7) or other international standards weren't a worry in this Proof of Concept because NoSQL databases offer flexible queries and other resources and also give a good answer for scaling big types of files like CT Scans, MRI and others converting them to Base64 type.

However, BigchainDB makes the perfect union of big data distributed database and blockchain characteristics (decentralized control and immutability) not preventing cyber attacks but it is enough to guarantee a high level of security by using Interledger Protocol (ILP) and its specific conditions configurations like the minimum validators necessary for a transaction could be processed that may be set up.

For future work, the authors suggest the natural continuation of the STEPES-BD project, integrating Patient monitoring through sensors capable of measuring, in real-time, heartbeat data, blood oxygenation, among others.

It is also suggested, for future work, the creation and development of appropriate policies for accessing the STEPES-

BD Project or similar ones to provide visualizations of general data and also specific information about Patients, Physicians, Hospitals, and Suppliers to better assist managers of the Health Municipal, State, Federal, and other stakeholders.

Acknowledgments The authors of this article would like to thank the Brazilian Air Force, the Brazilian Aeronautics Institute of Technology, the Ecosystema Negocios Digitais Ltda (WiBOO/Wibx), and the Casimiro Montenegro Filho Foundation, for the support and infrastructure offered during the development of this ALFA Project (MEC-ITA).

References

1. J. Kent, Could covid-19 help refine ai, data analytics in healthcare. Accessed 24 June 2020 [Online]. Available <https://healthitanalytics.com/features/could-covid-19-help-refine-ai-data-analytics-in-healthcare>
2. P. Mirchandani, Health care supply chains: Covid-19 challenges and pressing actions. Accessed 06 May 2020 [Online]. Available <https://www.acpjournals.org/doi/10.7326/M20-1326>
3. P.K. Sinha, G. Sunder, P. Bendale, M. Mantri, A. Dande, *Electronic Health Record: Standards, Coding Systems, Frameworks, and Infrastructures* (Wiley, 2012)
4. J. Sutherland, A. Viktorov, J. Blount, N. Puntikov, Distributed scrum: Agile project management with outsourced development teams, in *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE, 2007, pp. 274a–274a
5. B.W. Tuckman, Developmental sequence in small groups. *Psychol. Bull.* **63**(6), 384 (1965)
6. T.C. Shan, W.W. Hua, Solution architecture for n-tier applications, in *2006 IEEE International Conference on Services Computing (SCC'06)*. (IEEE, 2006), pp. 349–356
7. Key concepts of BigchainDB. Accessed 13 July 2020 [Online]. Available <https://www.bigchaindb.com/developers/guide/key-concepts-of-bigchaindb/>

Jai Santosh Mandava, Abhishek Verma, Fulya Kocaman, Marian Sorin Nistor, Doina Bein, and Stefan Pickl

Abstract

The objective of this paper is to develop a Machine learning model that can classify cancer patients. Gene mutation-based treatment has a very good success ratio, but only a few cancer institutes follow it. This research uses natural language processing techniques to remove unwanted text and convert the categorical data into numerical data using response coded and one-hot encoded. Then we apply various classification algorithms to classify the training data. The proposed system has the advantage of reducing the time to analyze and classify clinical data of patients, which translates into less wait time for patients in order to get results from pathologists. The results of our experiment will demonstrate that the Stacking Classifier algorithm with One-Hot encoding and Term Frequency – Inverse Document Frequency (TF-IDF) techniques perform better than other Machine Learning methods with around 67% accuracy on the test data.

Keywords

Gene mutation · Cancer patient · Machine learning · ML · Text data classification · Natural language

J. S. Mandava · F. Kocaman · D. Bein (✉)
Department of Computer Science, California State University,
Fullerton, Fullerton, CA, USA
e-mail: mandavajsantosh@csu.fullerton.edu;
fulyakocaman@csu.fullerton.edu; dbein@fullerton.edu

A. Verma
Department of Computer Science, California State University,
Northridge, Northridge, CA, USA
e-mail: abhishek.verma@csun.edu

M. S. Nistor · S. Pickl
Department of Computer Science, Universität der Bundeswehr
München, Neubiberg, Germany
e-mail: sorin.nistor@unibw.de; stefan.pickl@unibw.de

processing · NLP · Random forest classifier · Stacking classifier · Personalized medicine

28.1 Introduction

A sizable set of people suffer from various types of cancer every year, and nearly half of them die without proper diagnosis. According to the National Cancer Institute [1], gene mutations will enhance the treatment of disease by supplying effective medicines and decreasing the number of deaths. A considerable amount of scientific evidence collected over the last decades can be used to enhance cancer detection. Data on the health of past patients with diagnostic medicine are found in DNA or gene mutation. Based on the previous results, this approach can help physicians to treat future patients. Gene mutations are carried out by manual techniques. An experienced health care professional has to manually apply knowledge to track specifics and suit the gene of the patient. Using machine learning algorithms, the clinical data analysis can be automated to improve the diagnosis and treatment accuracy and time. Artificial Intelligence can predict the gene mutation in the patient groups by implementing machine-learning algorithms. In gene mutations linked to lung cancer, family genetics could play a significant role in analyzing the likelihood of cancer risk [2].

While gene mutations are known for lung cancer diagnosis, they can be used to collect the data and compare them manually. However, the alternative is to replace it with greater precision through the use of machine learning algorithms. This research applies various natural language processing algorithms to the clinical dataset collected from lung cancer patients to perform data cleaning. Furthermore, the paper focuses on applying machine learning classification.

According to Kourou et al. [3], “Machine Learning (ML) Algorithms can detect and classify trends and

interrelationships from complex data sets when effectively predicting future results from cancer.” The preparation, testing, and prediction are three phases of ML. Algorithms do software cleaning, replication, redundancy, and data preparation, whereas the ML techniques are used in the validation process for test data using specialized expertise. To ensure the correct algorithm is chosen, the accuracy should be determined. Next, the algorithm is used for the prediction of new data.

The proposed system was implemented in the Jupyter notebook using Python language. The problem encountered in the gene mutation treatment is that any manual process takes more time to analyze the patient’s data than an automated one. For the past several years, there has been much improvement in machine learning algorithms, parallel processing, and the ability of algorithms to process data at scale. We implemented machine learning algorithms on the dataset provided by Memorial Sloan Kettering Cancer Center [4], which has around 3200 data with the clinical text of each patient. We analyze two separate files, described as follows:

- Training variants – the file that includes the definition of genetic variations used for training. Training variants Fields are IDs (including the gene ID for the row of clinical evidence used to link the mutation), gene (gene where the genetic mutation is located), variations, class (includes 1–9 classes).
- Training text – contains the clinical text of the patients with their ID.

The proposed system has the following advantages. First, it reduces the time to analyze and classify the clinical data of patients, so this translates into less waiting time for patients to get results from pathologists. Second, our model can handle a large amount of data at a time.

The paper is organized as follows. In Sect. 28.2, we present background information, followed by the proposed system model in Sect. 28.3. Simulation results are shown in Sect. 28.4, concluding remarks and future work are presented in Sect. 28.5.

28.2 Background

In the eighteenth century, cancer was treated using surgery, which was considered then as primary treatment. Post-surgery herbal therapy, castor oil, and arsenic were administered to the patient. Radiation therapy came into existence in 1895 but only cured a few cancer types [5]. In the treatment of lymphoma cancer with the subsequent

approval of the FDA, immunotherapy was introduced at the end of 1987. In 1990, gene-based treatment was approved by the FDA for immunodeficiency disorder. Later, gene-based treatment was used on cancer patients and successfully mitigated different types of cancer such as brain tumors, acute lymphocytic, and others. According to Memorial Sloan Kettering Cancer Center [4], genetic tests are conducted in various cancer types and are divided into nine different classes for treatment. Each class is a cluster of gene-related treatments in which patients get treated according to the class. Patient classification happens through three phases, and eventually, the appropriate patient class is determined. The first phase is where a collection of the patient’s records is used, such as PET scan, CT scan, and previous health reports. Then pathologist looks for the clinical evidence from the medical literature and converts it into clinical text for further analysis. In the last phase, pathologists classify the patient into nine different classes by analyzing the clinical text. The final phase usually takes one to three days to classify the patients into a particular category.

Lung cancer is a type of cancer that can be traced from family history. Although according to Memorial Sloan Kettering Cancer Center [4], many people do not get proper treatment because the selected medication does not work on them, this is where gene mutation-based treatment can be helpful. The use of personalized precision medicine improves patients’ survival and quality of life [6]. However, the patient’s clinical data needs to be analyzed manually, and that requires the time of experienced staff to make the decisions.

The authors in [7] have used the Cancer Genomic Atlas results, including patients who have lung adenocarcinoma (LUAD). In order to acquire genes for LUAD patients, the author applies a machine learning algorithm. Further information on the dataset is given by using the KNN algorithm, decision trees, SVM, and Naïve Bayes. They show that ZNF560 and DRD3 are the positive genes capable of surviving LUAD. In order to assess the patient survival rate, the top six genes are later tested by the research model. The findings have been comparable to the original.

The authors in [8] implemented deep learning algorithms to classify lung cancer into gene mutations. They considered two major lung cancer types: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). According to the report, conventional neural networks are used to study the images and classify the genes. After training the model, they researched in the real world through means of new data of the patient for testing. The results were compared to the results of the pathologist from different sources, which show there is slight misclassification of the model. Their research does not involve other types of lung cancer and focuses on only two major types.

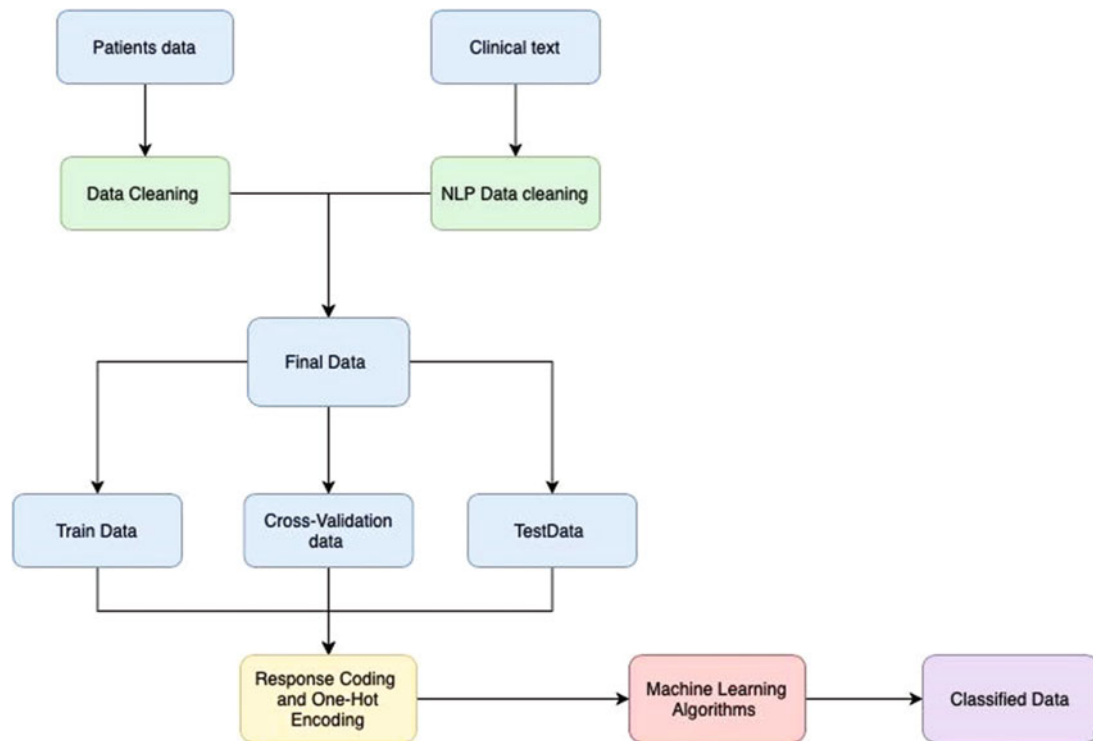


Fig. 28.1 System architecture

Table 28.1 Response encoded table

Gene	Class 0	Class 1	Class 2
ALK	4/20	5/20	11/20
CBL	14/35	10/35	11/35
KIT	4/10	3/10	3/10

28.3 Proposed System Architecture

Our proposed system architecture is shown in Fig. 28.1. We implemented our program in Jupyter Notebook using Python language. Unfortunately, the final phase of the gene mutation-based process takes more time, so we replace that with the machine learning algorithm.

We have approximately 3200 data points [4] divided into three parts: training data, cross-validation data, and test data. Before dividing the data, we need to clean the data and remove empty rows from the data. Using Natural Language Processing (NLP), we removed unwanted text in the clinical data of the patients. After splitting data, we implemented Response Coding and One-Hot encoding to convert the text data to numerical form for fitting the model in machine learning algorithms. Then the genetic mutations were classified into one of nine classes.

There are two columns of categorical data, which are genes and variations. To categorize the data, we implement one-hot encoding and response coding. One-hot encoding is

implemented using the count vectorizer in the scikit-learn library. Response coding is part of the machine learning technique used for categorizing the data. This technique is implemented to represent the data point by calculating the probability for each class by category. For example, suppose we have 233 unique genes for nine classes. We get 233 features by calculating the probability for each class. Suppose we take the three genes as three classes in the data. Table 28.1 describes how response coding is done for this example.

There are three types of genes such as ALK, CBL, and KIT. There are twenty types of patients with genes as ALK, and twenty have different classes. For example, in Table 28.1, Class 0 has 4 ALK genes. The probability is calculated by the total number of genes in the class, divided by the total number of the specific gene contained in the data.

According to [9], combining Natural Language Processing with Machine Learning algorithms helps search for patterns in text data. We implemented both Response Coding and scikit-learn's TF-IDF vectorizer methods for the text data. We then combine the features using One-Hot encoding on gene and variations with TF-IDF on the text data in each data set separately. We also combined all three features, where Response Coding was applied to gene, variation, and text data. The following classifiers with different NLP techniques were tested with different values using the CalibratedClassifierCV from the scikit-learn library. The best hyperparameters based on the smallest log-loss for each classifier were found during cross-validation. We then implemented

the following classifiers using the best hyperparameters for each one.

We use the following classifiers:

- Naïve Bayes Classifier is based on Bayes theorem for classification technique and assumes that predictors are independent. Thus, the existence of a specific feature in a class is irrelevant to any other feature in Naive Bayes classification. MultinomialNB is implemented in this research, and it can be imported from the scikit-learn library.
- K – Nearest Neighbors (KNN) can be used for statistical problems of classification as well as regression.
- Linear Support Vector Machine (SVM) is a linear model of a supervised machine learning algorithm which can be implemented in regression and classification modes. This is used primarily in the classification model. The value of each data element is the value of each coordinate in the SVM algorithm as a point in the n-dimensional field. Then we define the hyperplane, which differentiates between the two groups very well.
- Random Forest Classifier is a guided algorithm of learning. It can be used both for regression and classification. It is based on the idea of bagging. Many weak classifiers named decision trees pooled together are better than a single strong classifier. The random forest generates randomly selected data samples for decision trees, obtains predictions of each tree, and selects by voting the best solution. It also gives a very clear indication of the value of the feature.
- Stacking Classifier is a technique for combining many classification models with a meta-classifier. The individual classification models are trained based on the entire training set. The meta-classifier is then installed on the output from the individual model classifications within the ensemble-meta-features. The meta-classifier may also be educated on the predicted class labels or ensemble probabilities.

28.4 Experimental Results

We implemented various machine learning algorithms on the dataset provided in [4], which has around 3200 data with the clinical text of each patient. Our results are displayed in Fig. 28.2. Best accuracy is obtained from the random forest and stacking classifier. Table 28.2 shows each model's misclassification accuracy and log-loss on the test data we use in this project.

Table 28.2 Classification results

ML classifier	Mis-classification test accuracy	Test log-loss
Naïve Bayes with One-Hot Encoding and TF-IDF	38.25%	1.2939
Naïve Bayes with Response Coding	60.09%	1.7293
KNN with One-Hot Encoding and TF-IDF	39.91%	1.2935
KNN with Response Coding	47.74%	1.5251
Logistic Regression (Balanced)	34.64%	1.0891
Logistic Regression (unbalanced)	34.49%	1.0850
Linear SVM(Balanced)	35.84%	1.2016
Random Forest	33.58%	1.0831
Stacking Classifier	33.37%	1.0183

28.5 Conclusion and Future Work

In this paper, we describe Machine learning models that can classify cancer data of the patients. Our approach focuses on replacing the final phase of the gene mutation treatment for cancer using a machine learning algorithm. This implementation helps in improving the accuracy of the analysis and reduces the wait time for classification. The classification was performed using several popular machine learning algorithms. The test log-loss of the Stacking classifier was close to 1. The classification accuracy came out to be about 67% on test data. The problem is reasonably challenging, and more work is needed to reduce the error rate of the model so that the model accuracy can match the real-world accuracy of experienced pathologists.

Future work for the project will mainly focus on improving the accuracy of the model. In addition, we would expand the dataset beyond 3200 data points to implement deep learning algorithms that require larger amounts of training data and may produce improved classification accuracy. Furthermore, such a model can also be adapted to a triage process. For example, it might reduce the time to analyze and classify the triage protocols and the exiting clinical data of patients to improve the treatment response.

Acknowledgment This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700.

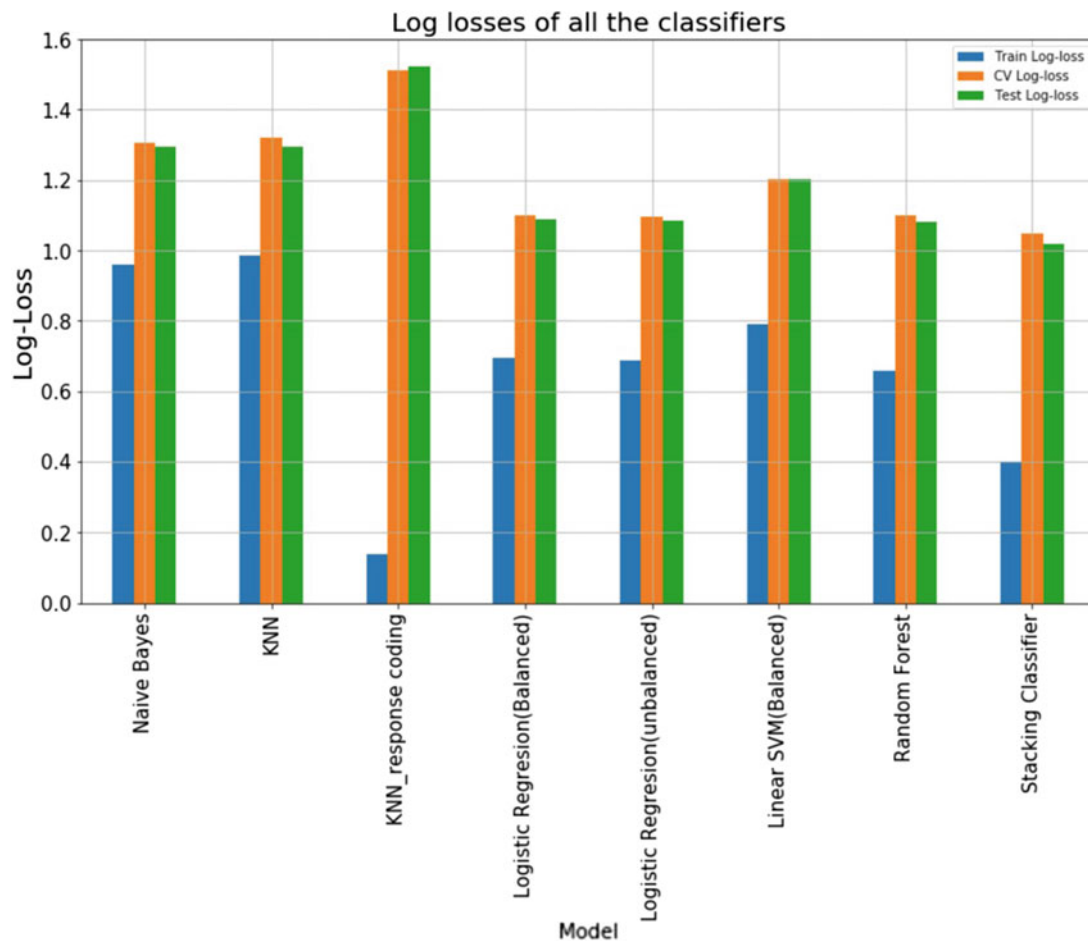


Fig. 28.2 Comparison of log-loss performance for various machine learning classification algorithms

References

1. National Institute of Cancer, *The Genetics of Cancer*. October 12, 2017. Available online at <https://www.cancer.gov/about-cancer/causes-prevention/genetics>. Last accessed 10 Dec 2020
2. Genetics Home Reference, (2019). Available online at <https://ghr.nlm.nih.gov/condition/lung-cancer#sourcesforpage>. Last accessed 10 Dec 2020
3. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J* **13**, 8–17., ISSN 2001-0370 (2015)
4. Memorial Sloan Kettering Cancer Center, *MSK-IMPACT: A Targeted Test for Mutations in Both Rare and Common Cancers*. Available online at <https://www.mskcc.org/msk-impact>. Last accessed 10 Dec 2020
5. M.H. Amer, Gene therapy for cancer: Present status and future perspective. *Mol. Cell. Ther.* **2**, 27 (2014). <https://doi.org/10.1186/2052-8426-2-27>
6. F.B. Marques, G.F. Leal, G.N. Bettoni, O.N. de Souza, *Integration of bioinformatics and clinical data to personalized precision medicine, ITNG 2021 18th International Conference on Information Technology-New Generations*, Advances in Intelligent Systems and Computing, vol 1346 (Springer, Cham). https://doi.org/10.1007/978-3-030-70416-2_23
7. H.-J. Cho, S. Lee, Y.G. Ji, D.H. Lee, Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. *PLOS ONE* **13**(11), e0207204 (2018). <https://doi.org/10.1371/journal.pone.0207204>
8. N. Coudray et al., Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**(10), 1559–1567 (2018)
9. C. Vielma, A. Verma, D. Bein, Single and multibranch CNN-bidirectional LSTM for IMDb sentiment analysis, in *ITNG 2021 17th International Conference on Information Technology-New Generations*, Advances in intelligent systems and computing, vol. 1134, (Springer, Cham, 2020). https://doi.org/10.1007/978-3-030-43020-7_53

Part VI

Machine Learning

Performance Comparison Between Deep Learning and Machine Learning Models for Gene Mutation-Based Text Classification of Cancer

Fulya Kocaman, Stefan Pickl, Doina Bein, and Marian Sorin Nistor

Abstract

Identifying genetic mutations that contribute to cancer tumors is the key to diagnosing cancer and finding specific gene mutation-based treatment. It is a very challenging problem and a time-consuming job. Currently, clinical pathologists classify cancer manually, and they need to analyze and organize every single genetic mutation in cancer tumors from clinical text. The text data analysis can be automated using Deep Learning and Machine Learning classification techniques to ease the manual work needed to extract information from clinical text. This paper aims to analyze the performance of Machine Learning and Deep Learning methods to classify cancer from gene mutation-based clinical text data. This paper uses Natural Language Processing techniques, namely, CountVectorizer and TfidfVectorizer, and Keras API's One-Hot encoding and to-categorical utility, to vectorize the categorical and text data and transform them into numerical vectors. Machine Learning classification algorithms and Deep Learning methods are then applied to the extracted features, and the most accurate combination of feature extraction and a classifier is discovered. Keras API's Embedding Layer (Word Embeddings) and Bidirectional Long-Term Short-Term Memory (Bidirectional LSTM) techniques using original and augmented text data from NLPAug library are applied as Deep Learning methods. The Keras Word Embeddings using augmented text data has performed the highest with an accuracy of 80.67%, the weighted average

precision of 0.81, recall of 0.81, F1 score of 0.81, and the log loss of 0.6391. As for the Machine Learning classification algorithms, Random Forest and Stacking classifiers are explored within this paper, and the highest accuracy of 67.02% is achieved from the Random Forest classifier with the weighted average precision of 0.70, recall of 0.67, F1 score of 0.65, and the log loss of 1.0523.

Keywords

Machine Learning · Random Forest · Deep Learning · Embedding Layer · Keras API · Text data classification · Gene mutation · NLP · NLPAug · Personalized medicine

29.1 Introduction

Cancer is one of the primary causes of death globally, and every year millions of people die of cancer-related diseases [1]. Specific alterations in protein-carrying genes can lead to abnormal cell growth and become cancer. These genetic mutations could damage the DNA of a cancer patient, and consequently, their descendants may inherit a similar type of cancer. Therefore, these gene mutations are used in targeted treatments for practical medicine. For example, targeted treatments would currently work on specific gene mutations linked to non-small cell lung cancer with EGFR, KRAS, ALK, and other gene mutations [2]. With the help of DNA sequencing and bioinformatics analysis methods, doctors and researchers can identify specific genetic mutations linked to certain types of cancer for early diagnosis and personalized treatments that might work best for their patients.

This experiment uses Count Vectorizer, and TF-IDF from the scikit-learn library and Keras API's One-Hot encoding and to-categorical utility to vectorize the categorical and text

F. Kocaman · D. Bein (✉)
Department of Computer Science, California State University,
Fullerton, Fullerton, CA, USA
e-mail: fulyakocaman@csu.fullerton.edu; dbein@fullerton.edu

S. Pickl · M. S. Nistor
Department of Computer Science, Universität der Bundeswehr
München, Neubiberg, Germany
e-mail: stefan.pickl@unibw.de; sorin.nistor@unibw.de

data and transform them into numerical vectors. We then analyze and identify genetic mutations from text-based clinical literature using Deep Learning and Machine Learning classification methods. We built Deep Neural Networks using Embedding Layer and Bidirectional LSTM using original and augmented text data in Keras API. We used text augmentation using the NLPAug library with Bidirectional Encoder Representations from Transformers (BERT) Contextual Embeddings. As for the Machine Learning classifiers, we used Random Forest and Stacking Classifiers.

The dataset provided from Memorial Sloan Kettering Cancer Center (MSKCC), one of the most extensive private cancer treatment and research institutes in New York, is used in this paper. MSKCC has provided an expert-annotated precision oncology knowledge base with thousands of mutations manually annotated by world-class researchers and oncologists for studying gene classification. The dataset contains 3321 patients' data and has the clinical text of each patient. Following describes the two files used in this paper:

- The patients' data (training/test variants) with their IDs contain Gene, Variation, and Class (1–9 classes).
- The clinical text (training/test text) of the patients with their IDs provides the clinical evidence that researchers and oncologists use to classify the genetic mutation.

The classification methods are evaluated using classification accuracy and log-loss.

Even though this problem can be considered a text classification, it is more complex and challenging than other traditional text classifications because each clinical text data entry contains very long and detailed medical terminologies. Moreover, some samples share duplicate data entry, while having different class labels [3]. Figure 29.1 also depicts how unbalanced the distribution of classes is over the entire data set.

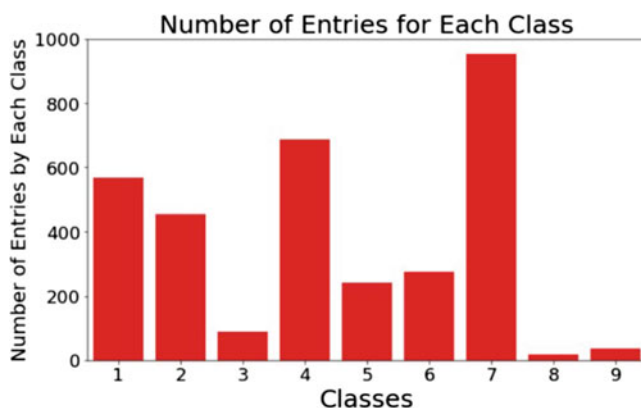


Fig. 29.1 Distribution of classes

The paper is organized as follows: Background information is presented in Sect. 29.2. Then, Sect. 29.3 describes the proposed system architecture, while Sect. 29.4 shows the results of our experiment. Finally, Sect. 29.5 concludes this paper with a conclusion and future work.

29.2 Background

For the last several years, there have been considerable improvements in genetic testing, which detects cancer from mutations in specific genes. The genetic test called DNA sequencing test can read DNA and compare the sequence of DNA in cancer cells with that in normal cells such as blood samples to identify genetic changes in cancer cells. It can also reveal inherited genetic mutations associated with an increased risk of cancer [4]. According to Memorial Sloan Kettering Cancer Center [5], many people do not get proper treatment because the selected medication does not work on them; this is where gene mutation-based treatment can be helpful. The data from DNA sequencing helps doctors make a proper prognosis, and it offers them targeted treatment options to increase the survival rate of their patients. However, there are thousands of genetic mutations of a sequenced cancer tumor, and these mutations need to be differentiated from neutral mutations [6]. The classification process is challenging and time-consuming because a clinical pathologist manually analyzes these genetic mutations from clinical text.

Like many other fields, the medicine and health industry relies on developing the latest technology to improve the accuracy of digital techniques. Correctly diagnosing a patient and prescribing the proper treatment are the essential parts of the medicine domain to cure diseases consequently [7]. Therefore, there is a growing trend of using Deep Learning and Machine Learning to assist medical personnel in predicting and detecting cancer [8].

Machine Learning classification techniques can ease the manual work needed to extract information from medical records and clinical text on unwanted, unorganized, and multi-dimensional data [7]. Applying Deep Learning and Machine Learning techniques have recently become a widespread application in cancer diagnosis and precision medicine for gene mutation-based treatments. It is mainly due to the improvements in hardware capabilities and cloud computing to store vast quantities of data and software capabilities to efficiently manage and process big data in the last several years. Therefore, an automated clinical data analysis can improve the diagnosis and treatment in terms of accuracy and time using Machine Learning algorithms. Furthermore, an automatic Machine Learning Algorithm implementation would also decrease the mistakes caused by classifying mutations manually and further support the medical domain.

Li and Yao [9] used Term Frequency-Inverse Document Frequency (TF-IDF) as a Natural Language Processing (NLP) technique for feature extraction and conversion and label encoding to convert genes and their variations into feature vectors for training to employ normalized non-linear features. They implemented Extreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM) for classification on the dataset from Memorial Sloan Kettering Cancer Center (MSKCC) through the Kaggle [6] competition. The result of their experiment showed that XGBoost outperformed SVM and produced better predictive results. Furthermore, their investigation revealed that XGBoost outperformed SVM with an accuracy of 66%.

Along with using TF-IDF to transform the clinical text into a numerical feature vector, Waykole and Thakare [10] used a one-hot encoding technique to extract features from genes and their variations from the same MSKCC dataset. They applied a basic Logistic Regression technique to the extracted features and achieved an accuracy of 64%. They also emphasized how difficult it is to classify genetic mutations based on medical literature for human experts due to decoding clinical evidence.

The authors in [11] experimented with many different Machine Learning classifiers in their study of predicting and early detecting diseases. After the feature selection, they used Logistic Regression with the selected feature, SVM with a linear kernel for their analysis to classify data to maximize the margin between the hyperplane, AdaBoost, Random Forest, and Decision Tree classifiers. They concluded that Logistic Regression performed on the heart dataset had the highest accuracy, SVM performed superior to others in their diabetes prediction, and AdaBoost classifier performed the best for breast cancer detection among the other machine learning algorithms they used. However, they also concluded that the Random Forest and Decision Tree methods did not perform better than other classifiers on three separate datasets.

Zhang et al. [3] explored Word2Vec (Skip-Gram), a predictive neural word embedding model that captures the semantic relationship of words in a text from the same MSKCC dataset. They developed an ensemble of nine primary gradient boosting models applied to the combined features, which showed the best performance in their evaluation with an accuracy of 67%.

The authors in [12] researched four different Machine Learning classifiers on the MSKCC cancer dataset to automate the process. The Logistic Regression classifier with class balancing and one-hot encoding scheme was found to be the most efficient model with an accuracy of 65.23% in their work.

Saxena & Prasad [8] experimented with five Machine Learning algorithms: Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, and Maximum Sensitivity Artificial Neural network (ANN) in their paper.

They concluded that Maximum Sensitivity ANN using Multilayer Perceptron produced the most accurate results based on less sensitivity variation in predicting and detecting cancer disease.

29.3 Proposed System Architectures

29.3.1 Deep Learning Neural Network Models

The following Deep Neural Network models are implemented in Google Colab Pro with P100 GPUs using Python language and Keras API. These models are trained only using the clinical text data to capture the semantic relations in the text. After the initial cleaning, the text transforms before getting fed into a Neural Network (NN). Keras Tokenizer was used to convert the entire text into numeric tensors by tokenizing into tokens, stripping from special string characters, then turning them into lists of integer indices, and finally turning the lists of integers into a 2D integer tensor of shape. We then split the text into training and testing sets and encoded each text data sentence into integer sequences. Moreover, we finally padded the sequences to have the same length and turned the integers sequences into tensors to fit into a NN [13]. We also chose the Panda's `get_dummies` function and Keras' `to_categorical` utility to vectorize the labels for our following Deep Learning models.

After repeatedly modifying the models, training them, and evaluating them on the validation data with different combinations of layers, hyperparameters, activation and loss functions, and batch and epoch sizes, the following Deep Learning networks gave the most optimal configuration for our problem. As the author in [13] suggested, once we developed a satisfactory model configuration, we trained the final model using both training and validation data and evaluated it on the test data.

- Embedding Layer uses dense word embeddings to learn from the data. With the help of backpropagation, the Embedding Layer learns a new word embedding at each layer. The Embedding Layer takes the input integer sequences (2D integer tensor) and turns them into embedded sequences (3D float tensor) [13].

Our Embedding Layer network learns 100-dimensional embeddings for each word, turns 2D integer tensor inputs into 3D embedded tensors in the Embedding Layer, flatten the tensor to 2D, and trains two Dense layers with 100 and 9 layers on top for classification. These Dense layers use activation functions of relu and softmax, respectively. In addition, one Dropout function is used to overcome overfitting. The summary of this network is shown in Fig. 29.2.

Fig. 29.2 Summary of the network used with Embedding Layer

Layer (type)	Output Shape	Param #
embedding_5 (Embedding)	(None, 10000, 100)	14936400
dense_8 (Dense)	(None, 10000, 100)	10100
flatten_3 (Flatten)	(None, 1000000)	0
dropout_7 (Dropout)	(None, 1000000)	0
dense_9 (Dense)	(None, 9)	9000009
Total params: 23,946,509		
Trainable params: 23,946,509		
Non-trainable params: 0		

Fig. 29.3 Summary of the network used with Bi-LSTM

Layer (type)	Output Shape	Param #
embedding_11 (Embedding)	(None, 10000, 128)	12800000
bidirectional_9 (Bidirection	(None, 128)	98816
dense_13 (Dense)	(None, 9)	1161
Total params: 12,899,977		
Trainable params: 12,899,977		
Non-trainable params: 0		

- Bidirectional Long-Term Short-Term Memory (Bidirectional LSTM) encodes the input sequence in both states; information from the past and the future states using two independent LSTMs with different parameters. It is known to be helpful in Natural Language processing and aims to improve the performance of one-directional LSTM [13]. Our Bidirectional LSTM network shown in Fig. 29.3 learns 128-dimensional embeddings for each word in the Embedding Layer and trains a Bidirectional LSTM with 64 layers and a Dense layer with 9 layers on top for classification. The Dense layer uses the activation function of softmax, and a Dropout function is implemented to mitigate overfitting.
- Text Augmentation using NLPAug library with Bidirectional Encoder Representations from Transformers (BERT) Contextual Embeddings was also explored to handle the overfitting issue with the Deep Learning models. After splitting the entire text data into training and testing sets, we set aside the testing data and augmented some text data using only the training set. During the BERT Augmentation, it substituted 100 words in each text. That data was added to the entire training data frame, and accumulated training text data was shuffled at the end of each text generation [14, 15]. As a result, we augmented around 9000 new text data using only the training data set, which gave us about 11,000 training text data. Figure 29.4 shows the distribution of the text data over each class

after the BERT augmentation. We implemented the same Neural Network structure as the Embedding Layer model presented in Fig. 29.2. After trying out many different configurations of hyperparameters and evaluating with K-fold cross-validation, that model was adequate for our augmented text data set.

29.3.2 Machine Learning Algorithms

Figure 29.5 shows the architecture of the following Machine Learning algorithms used in this paper. First, we analyzed the input data using Exploratory Data Analysis to visualize the input data (both data of training variants and clinical text). During preprocessing, the text data is cleaned, unwanted text in the patient's clinical data is removed. Then, the final data is split into three sets: training, cross-validation, and test data. Features are then extracted using various NLP and word embedding techniques, giving the model a simpler and more focused view of the clinical text. Finally, the extracted features are evaluated against multiple Machine Learning classification algorithms, and the most accurate combination of feature extraction and a classifier is discovered based on their classification accuracies. As a result, the genetic mutations are classified into one of nine classes.

The NLP techniques for feature extraction applied in the following Machine Learning algorithms are One-Hot encod-

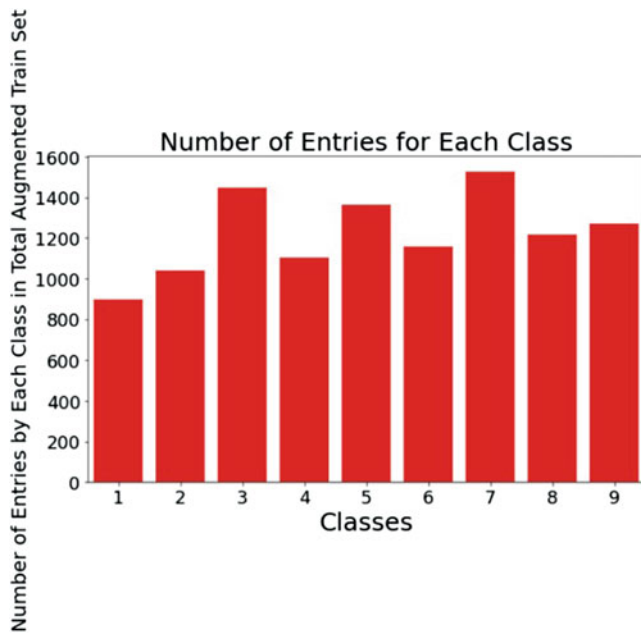


Fig. 29.4 Augmented text data distribution over each class

ing using the Count Vectorizer for gene and variation data sets and TF-IDF Vectorizer for text data. These two vectorizers are both from the sklearn library. Once these three features are vectorized on each training, test, and cross-validation data set separately, we combined the three features coming from each data set.

- Random Forest Classifier uses the bagging method and creates randomness from a standard dataset. It takes bootstrap samples from the dataset for each tree to create a forest so that the trees are trained on a somewhat different dataset. Each node of the tree gets a random subset of the features, and to add randomness to the decision tree, it can only pick from that subset rather than from the whole set [16].

In our Random Forest classifier model, we implemented the CalibratedClassifierCV using the sigmoid function from the sklearn library to calibrate probabilities in the Random Forest classifier. Multiple copies of different combinations of the number of trees and the depth of each tree are fitted in the CalibratedClassifierCV for our model. The best hyperparameters for this classifier are chosen by finding the minimum log-loss of the cross-validation data set. They are finally used to implement our Random Forest model.

- Stacking Classifier is an ensemble learning technique, where combining multiple classifiers is used as a meta-classifier. The models from each classifier are trained based on the entire training set, and the meta-classifier is then installed on the output from the individual model

classifications within the ensemble-meta-features. The meta-classifier can either be trained on the predicted class labels or probabilities from the ensemble.

Our Stacking Classifier is an ensemble of Naive Bayes, Logistic Regression, Linear Support Vector Machine (SVM), and Random Forest classifiers. We first calibrated each classifier using CalibratedClassifierCV with the best hyperparameter values that we discovered previously. As the optimization method for our model, we then implemented the Stochastic Gradient Descent - Classifier (SGD-Classifier) with Logistic regression from the sklearn library using all four calibrated classifiers. Multiple alpha values are fitted in the SGD-Classifier for our model. Moreover, the best alpha for this classifier was chosen by finding the minimum log-loss of the cross-validation data set. It was then used to implement our Stacking Classifier model.

29.4 Experimental Results

Attempts to automate classifying cancer from genetic mutations using Machine Learning algorithms with the dataset from MSKCC have been tried several times for the last five years. However, Deep Learning methods have rarely been applied to this particular dataset. Therefore, the goal of this paper was first to come up with ways to implement Deep Learning methods for the clinical text data and compare the results with more traditional Machine Learning classifiers.

After vectorizing the text data using the top 100,000 most common words and limiting maximum features to 10,000 text labels, the results from our Deep Learning models are below.

- Embedding Layer Results: Figs. 29.6 and 29.7 show the K-fold = 3 accuracy and log-loss of the embedding layer models with Panda's get_dummies function and Keras' to_categorical utility, respectively.

Overall, Panda's get_dummies function performed with 61.79% accuracy, which was 1–2% better than Keras' to_categorical utility that we implemented for this paper. Figure 29.8 displays a sample graph of an Embedding Layer network of training and validation loss on the test data, whereas their accuracies are shown in Fig. 29.9.

As seen from the large gap between training and validation accuracy on the test data in Fig. 29.9, we clearly had an overfitting issue to overcome. To develop an ideal model when tuning the hyperparameters on the validation text data, we modified the Embedding and Dense layers, experimented with different batch and epoch sizes, used other activation functions and optimizers, and applied regularization techniques like Dropout and SpatialDropout1D functions. As a

Fig. 29.5 Architecture of the proposed technique for Machine Learning algorithms

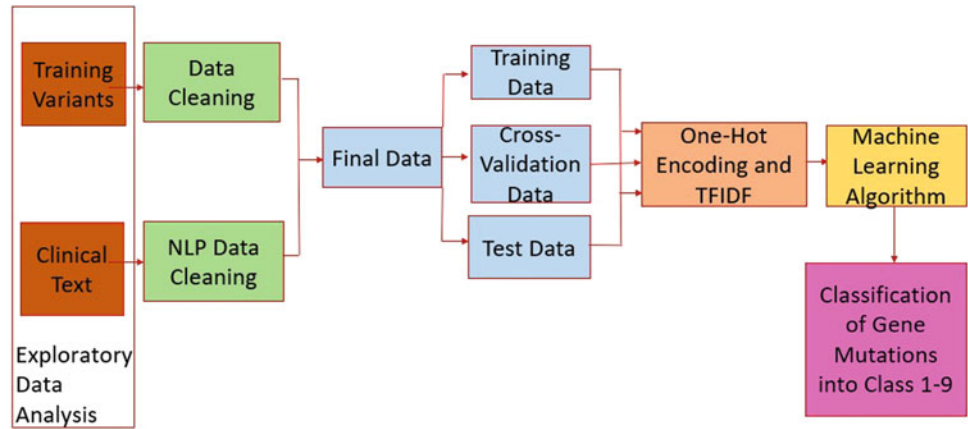


Fig. 29.6 Results from Embedding Layer model using Panda's get_dummies

```

Score per fold
-----
> Fold 1 - Loss: 1.2967923879623413 - Accuracy: 62.65822649002075%
-----
> Fold 2 - Loss: 1.1996276378631592 - Accuracy: 60.723984241485596%
-----
> Fold 3 - Loss: 1.2492704391479492 - Accuracy: 61.990952491760254%
-----
Average scores for all folds:
> Accuracy: 61.79105440775553 (+- 0.8022022612766903)
> Loss: 1.2485634883244832
    
```

Fig. 29.7 Results from Embedding Layer model using Keras' to_categorical

```

Score per fold
-----
> Fold 1 - Loss: 1.3953533172607422 - Accuracy: 58.22784900665283%
-----
> Fold 2 - Loss: 1.1765302419662476 - Accuracy: 60.36199331283569%
-----
> Fold 3 - Loss: 1.2667471170425415 - Accuracy: 62.44344115257263%
-----
Average scores for all folds:
> Accuracy: 60.344427824020386 (+- 1.7210531068666777)
> Loss: 1.2795435587565105
    
```

result, the model presented in Fig. 29.2 gave the best results from iterated K-fold validation, as shown in Fig. 29.6.

- **Bidirectional LSTM Results:** Since we could not solve the overfitting issue in the Embedding Layer model, we thought that our 3316-text data was the problem causing the models not to function well. Instead, we decided to experiment using Bidirectional LSTM because they are useful on Natural Processing problems and could work well with smaller text data [13].

Our Bidirectional LSTM model has an accuracy of 58.89% with a log-loss of 1.2205. Figures 29.10 and 29.11 exhibit the graph of log-loss and accuracy, respectively. Our Bidirectional LSTM model, like the Embedding Layer model, struggled with the overfitting problem.

- **BERT Text Augmentation Results:** We realized that we needed more text data to mitigate the overfitting issue

from the two models above. Neural Network models perform better with having a large amount of data. The new text data augmented with BERT Contextual Embeddings from NLPAug library was used in the Embedding Layer. K-fold = 3 cross-validation was applied to evaluate this model. The K-fold combines the augmented train text data with the original test text data. Then the K-fold splits the entire input into training and testing sets at each fold. It trains on the training set and evaluates the test set every time. As Fig. 29.12 shows, the average accuracy was about 80.67%, with a log-loss of 0.6391, the highest accuracy we have achieved so far. Moreover, the weighted average precision, recall, and F1 scores were all 0.81. We also tested our augmented text embedding layer model only on the original test data, which gave us about 62.65% accuracy with a log-loss of 1.2061.

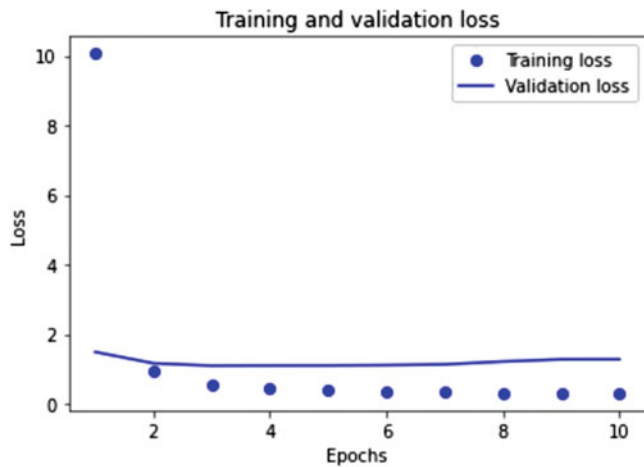


Fig. 29.8 Embedding Layer test data los-loss graph

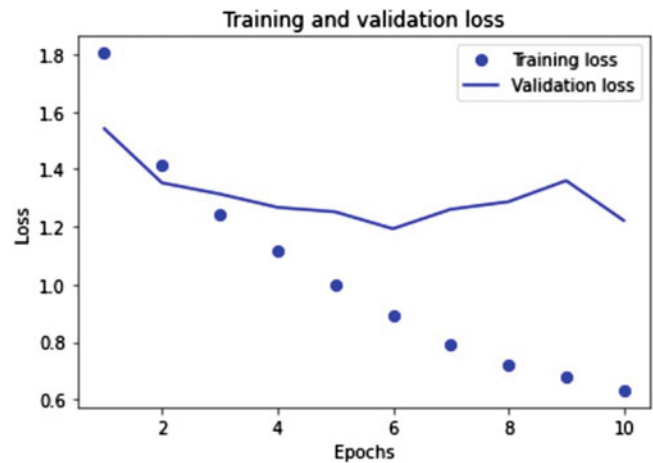


Fig. 29.10 Bidirectional LSTM test data los-loss graph

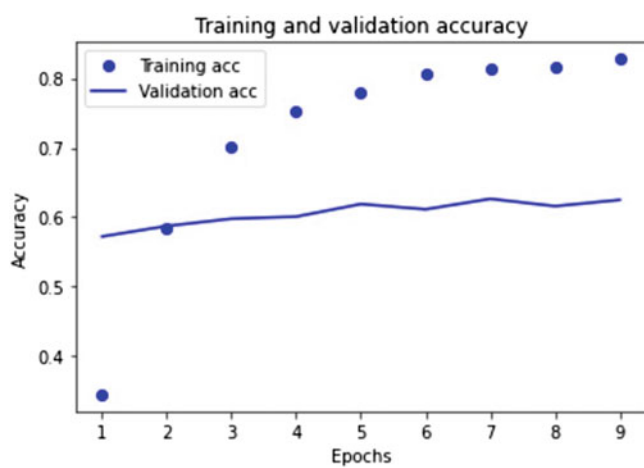


Fig. 29.9 Embedding Layer test accuracy graph

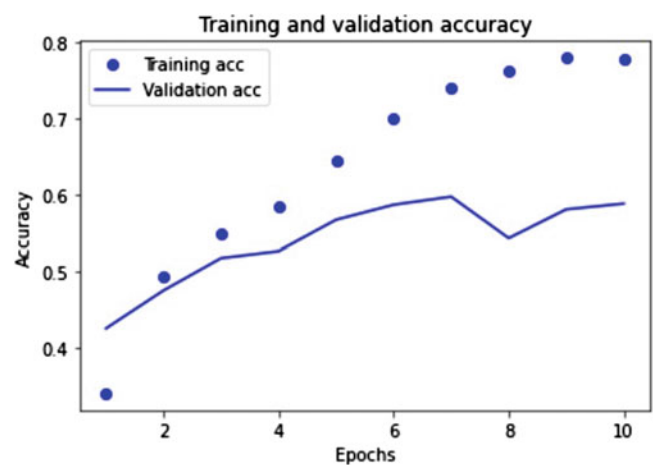


Fig. 29.11 Bidirectional LSTM test data accuracy graph

- Machine Learning Classifier Results: Our Random Forest classifier performed with an accuracy of 67.02%, the weighted average precision of 0.70, recall of 0.67, F1 score of 0.65, and the log loss of 1.0523, whereas the accuracy from the Stacking classifier was 66.27% with the weighted average precision of 0.68, recall of 0.66, F1 score of 0.66, and the log loss of 1.0344.

29.5 Conclusion and Future Work

Table 29.1 shows the comparison of the models we implemented in our work. Again, traditional Machine Learning algorithms prevailed over more complex Deep Learning methods. However, the Deep Learning algorithm with BERT augmented text data exhibited much better accuracy.

Future work involves using pre-trained word embedding techniques such as GloVe to capture semantic relationships within the text data, experimenting with other powerful Machine Learning algorithms like Extreme Gradient Boosting (XGBoost), and exploring other text augmentation techniques. Combining Machine Learning algorithms applied to the clinical text data with medical image processing techniques would further improve advancements in the diagnostic process and support all medical personnel.

Moreover, when there is more study on classifying gene mutation for cancer in personalized medicine with the help of Machine Learning algorithms, it would also open doors to other areas of medicine such as curing diseases other than cancer and even early detection of an anomaly in infectious diseases like Covid-19.

Acknowledgment This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700.

Fig. 29.12 Results from Embedding Layer augmented text model

```

Score per fold
-----
> Fold 1 - Loss: 0.6599963307380676 - Accuracy: 80.7238221168518%
-----
> Fold 2 - Loss: 0.6232505440711975 - Accuracy: 80.80081939697266%
-----
> Fold 3 - Loss: 0.6340599656105042 - Accuracy: 80.48780560493469%
-----
Average scores for all folds:
> Accuracy: 80.67081570625305 (+- 0.13317073796426218)
> Loss: 0.6391022801399231

```

Table 29.1 Classification test results

Classifier	Feature extraction	Test accuracy	Test log-loss
Embedding Layer	Keras Tokenizer/panda's get-dummies	61.79%	1.2486
Embedding Layer	Keras Tokenizer/to_categorical	60.34%	1.2795
Bidirectional LSTM	Keras Tokenizer/panda's get-dummies	58.89%	1.2205
Embedding Layer using Augmented Text	Keras Tokenizer/panda's get-dummies/K-fold=3	80.67%	0.6391
Embedding Layer using Augmented Text	Keras Tokenizer/panda's get-dummies/Testing on Original Test data	62.65%	1.2061
Random Forest	One-Hot encoding/TF-IDF Vectorizer	67.02%	1.0523
Stacking Classifier	One-Hot encoding/TF-IDF Vectorizer	66.27%	1.0344

References

- National Cancer Institute (NCI), *Cancer Statistics*, September 25, 2020. Available online at <https://www.cancer.gov/about-cancer/understanding/statistics>. Accessed 4 Nov 2021
- S. Watson, *Guide to Lung Cancer Mutations*, July 16, 2020. Available online at: <https://www.healthline.com/health/nsclc/guide-to-lung-cancer-mutations>. Accessed 4 Nov 2021
- X.S. Zhang, D. Chen, Y. Zhu, C. Che, C. Sue, S. Zhao, X. Min, F. Wang, *A Multi-View Ensemble Classification Model for Clinically Actionable Genetic Mutations*. arXiv:1806.09737v2 [cs.LG], 17 Mar 2019
- National Cancer Institute (NCI), *The Genetics of Cancer*, October 12, 2017. Available online at: <https://www.cancer.gov/about-cancer/causes-prevention/genetics>. Accessed 4 Nov 2021
- Memorial Sloan Kettering Cancer Center (MSKCC), *MSK-IMPACT: A Targeted Test for Mutations in Both Rare and Common Cancers*, 2021. Available online at <https://www.mskcc.org/msk-impact>. Accessed 4 Nov 2021
- Kaggle, *Memorial Sloan Kettering Cancer Center (MSKCC) Personalized Medicine: Redefining Cancer Treatment*, 2017. Available online at: <https://www.kaggle.com/c/msk-redefining-cancer-treatment>. Accessed 4 Nov 2021
- P. Singh, S.P. Singh, D.S. Singh, An introduction and review on machine learning applications in medicine and healthcare, in *IEEE Conference on Information and Communication Technology (CICT)* (2019)
- S. Saxena, S.N. Prasad, Machine learning based sensitivity analysis for the applications in the prediction and detection of cancer disease, in *2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. <https://doi.org/10.1109/DISCOVER47552.2019.9008083> (2019)
- G. Li, B. Yao, Classification of Genetic mutations for cancer treatment with machine learning approaches. *Int J Design Anal. Tools Integr. Circ. Syst.* **7**(1) (2018)
- R.N. Waykole, A.D. Thakare, *Intelligent Classification of Clinically Actionable Genetic Mutations Based on Clinical Evidences*, 978-1-5386-5257-2/18/IEEE. (2018)
- P.S. Kohli, S. Arora, Application of machine learning in disease prediction, in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. <https://doi.org/10.1109/CCAA.2018.8777449> (2018)
- A. Singh, S.K. Jain, A personalized cancer diagnosis using machine learning models based on big data, in *Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* IEEE Xplore Part Number:CFP20OSV-ART; ISBN: 978-1-7281-5464-0 (2020)
- F. Chollet, *Deep Learning with Python*. ISBN 9781617294433 (2018)
- R. Sangani, *Powerful Text Augmentation Using NLPAUG*, June 11, 2021. Available online at: <https://towardsdatascience.com/powerful-text-augmentation-using-nlpaug-5851099b4e97>. Accessed 4 Nov 2021
- Nithilau, *A Python Library to Augment Your Text Data*, August 25, 2021. Available online at: <https://www.analyticsvidhya.com/blog/2021/08/nlpaug-a-python-library-to-augment-your-text-data/>. Accessed 4 Nov 2021
- S. Marsland, *Machine Learning an Algorithmic Perspective* (2nd edn). Chapman & Hall/CRC Machine Learning & Pattern Recognition Series (2015)

Rahul Chauhan, Marian Sorin Nistor, Doina Bein, Stefan Pickl,
and Wolfgang Bein

Abstract

In this paper we present a Stock Backtesting Engine which would test historical data using pairs trading strategy. We implemented pairs trading strategy and ran it on historical data. We collect the S&P 500 data from the Internet and store it in a database. We then allow users to enter a set of stocks to find cointegrated pairs among them. We also provide an option to find the cointegrated pairs in all of S&P 500 stocks. Once the cointegrated stocks are selected we run the backtesting algorithm on these pairs and find from a given set of stocks, all the pairs of stocks that exhibit cointegration properties. Once such pairs are identified, this program would use pairs trading methods to calculate trades for each stock. Finally we provide the analysis of trades executed by the algorithm with average and daily data, and plot a chart of daily profit and loss with pairs trading strategy to showcase the effectiveness of the trading strategy.

Keywords

Algorithmic trading · Automated trading · Investment bank · Pair trading · Stock market · Stock backtester · Stock correlation · S&P 500 stocks · Trading strategy · Heat map

R. Chauhan · D. Bein (✉)
Department of Computer Science, California State University,
Fullerton, Fullerton, CA, USA
e-mail: rahulchauhan@csu.fullerton.edu; dbein@fullerton.edu

M. S. Nistor · S. Pickl
Department of Computer Science, Universität der Bundeswehr
München, Neubiberg, Germany
e-mail: sorin.nistor@unibw.de; stefan.pickl@unibw.de

W. Bein
Department of Computer Science, University of Nevada, Las Vegas,
Las Vegas, NV, USA
e-mail: wolfgang.bein@unlv.edu

30.1 Introduction

Every year, investment banks spend billions of dollars to invest in technology with the objective of facilitating efficient flows of funds globally and providing world class services. For example, in 2019 JPMorgan, the highest spender, has a \$11.4 billion technology budget this year, a 5.6% uptick from last year's \$10.8 billion. Bank of America's IT spend was second, at \$10 billion, followed by Wells Fargo at \$9 billion and Citigroup at roughly \$8 billion [1]. Besides investing in infrastructure, such as network security and large-scale database management, automated trading algorithms are also an increasingly mission-critical area of development.

The internet has changed traditional stock market trades to electronic trades. Automated trading is now the most important innovation in the field. To have bots trade efficiently and effectively in stock market, they need to be able to test out their strategies and hypothesis in a safe environment. A stock backtester provides a perfect sandbox environment where these strategies can be tested out with historical data. It provides an important reference to sharpen the strategies before they can be applied in real world, where the risks are much higher.

One of the most popular and simple strategy is of pairs trading [2, 3]. Pairs trading belongs to the statistical arbitrage class of trading strategies and is a simple trading strategy which involves finding a pair of stocks that exhibit similar historical price behavior, and then betting on the subsequent convergence of their prices in the event that they diverge [4, 5]. Since this trading style is very flexible and applicable to all asset classes, the methods to identify pairs trading signals is a widely researched topic. Pairs trading belongs to the statistical arbitrage class of trading strategies [6, 7].

In this paper we present a Stock Backtesting Engine which would test historical data using pairs trading strategy. This

program would find from a given set of stocks, all the pairs of stocks that exhibit cointegration properties. Once such pairs are identified, this program would use pairs trading methods to calculate trades for each stock. Finally these trades would be analyzed and presented to user in the form of a simple report as well as visualization of different aspects of effectiveness of the trading strategy.

The paper is organized as follows. In Sect. 30.2 we present background work, followed by our proposed system architecture in Sect. 30.3. Results for eight major companies for which we used pair trading are presented in Sect. 30.4. Concluding remarks and future work are given in Sect. 30.5.

30.2 Background Work

Algorithmic trading can be defined as having a machine systematically send orders to the trading exchange to be executed on behalf of human traders. Algorithmic trading uses computer codes and chart analysis to enter and exit trades according to set parameters such as price movements or volatility levels. Once the current market conditions match any predetermined criteria, trading algorithms can execute a buy or sell order on your behalf. Algorithmic trading allow automating a repetitive and non-value adding task in stock testing. This frees up humans in salespeople and traders to focus only on Selling and trading. Most importantly, having tools that can provide insight to trades to make informed decisions is a huge value added by technology. Furthermore, computers are far superior to perform such tasks which have low margins like arbitrage strategies to humans.

Professional algorithmic traders often employ two major classes of algorithmic trading strategies, namely high frequency trading (HFT) and statistical arbitrage trading. HFT is characterized by high turnover rates and high order-to-trade ratios that leverage high-frequency financial data, usually in the scale of milli- to micro-seconds. In this paper our focus is mostly on the second method which is Statistical arbitrage trading.

Statistical arbitrage involves simultaneously buying and selling assets that normally have the same price trends but whose prices diverge in the short term. This form of trading relies less on the speed of sending orders and more on the mean reverting nature of financial products, i.e., the phenomenon in which asset prices and returns eventually return back to the long-run mean of the historical time series. Pairs trading is a one of the statistical arbitrage class of algorithmic trading strategy. The rationale for it is that for similar companies or financial products, they should be priced similarly based on no-arbitrage pricing principle [8]. In other words, products with the same risk should have the same reward.

Algorithmic Stock trading is quite popular nowadays to check the accuracy of the newly created strategy or make sure

if the existing strategy works well on the given data stock backtesting engine also gains its popularity. Like the stock trading bot, the stock backtesting software comes in both paid and free software to help traders. The popular backtesting engine uses multiple trading strategies and, on its use, can have more than one option.

Tradingview is a web-based platform for stock backtesting. Tradingview also provides to check the strategies into the real-time data similar to paper trading. Tradingview provides price chart data of given stock data, the total profit, trades performance on the charts, the percentage of profit, and provide advice to buy and hold. It is easily accessible because the web-based platform, accurate financial data, provides backtesting not only on stock but also for Forex and cryptocurrencies.

The MetaTrader5, also known as MT5, with backtesting also provides expert advisors. The client must select the trading strategy before, and the corresponding advisor will get selected. The advisor program needs to run on historical data with some initial parameters. The advisor then runs a trading strategy several times with different settings on historical stock data. The process later allows the user to select a trading strategy from different combinations. MT5 Strategy backtesting engine works on most of the currency. MT5 is a program that needs to install on the user's computer and generally installed by one of the MT5 agents on the user's computer.

MS Backtesting is another popular stock backtesting engine. It allows users to do 58 different backtesting. Once the backtest is complete, MS backtesting will show the use list of every buy or sell trade and also provide a portfolio chart.

Interactive Brokers' Portfolio Builder is created by Interactive Brokers, it is meant to be used for testing different strategy on historical data. This tool starts with creating an investment strategy, where we can choose different rules. It then allows you to select initial parameters like investment amount and long/short leverage. It is a powerful tool to visualize the investment strategy.

Trend Spider backtester uses a different approach to backtesting. This backtesting engine is built in a way it can detect trend lines and Fibonacci patterns automatically. TrendSpider allows use to write code for trading strategy and run test it.

Backtesting a trading algorithm means to run the algorithm against historical data and study its performance. A backtesting can provide confidence for any trading strategy because giving a positive result on previous data can also mean the same for future data. Backtesting can provide multiple data which can help to improve any strategy without using real money investment it can give for each trade will have an associated profit or loss and total of this profit/loss over the profit of the strategy. Some of major reason to perform backtesting are:

- Filtration – At the beginning any trading strategy starts with generalization with backtesting one can remove all unnecessary conditions which are not useful in data and also can add any critical condition for any set of data.
- Modeling – backtracking allows trading in any particular time frame to test specific conditions of any trading strategy.
- Optimization – allowing to perform algo trading in large data can also give ideas for what can happen in future and if there is any way to optimize current conditions.
- Verification – As mentioned multiple times in this documentation, without checking a new strategy performed in the real time market will be a disaster. If the technique is already performed in some data can give positive results and can give developers confidence to use it for future trades.
- Net profit or loss – Net percent gained or lost for given stock.
- Averages – Percentage average gain and average loss.
- Ratios – Wins-to-loss ratio
- Money estimation – In backtesting use can set desired amount backtesting will provide the result for only that given amount.

Pairs trading is a market-neutral profits strategy that uses statistical and technical analysis. The pairs trading, as the name suggests, involves stocks in pairs. Pairs trading need two or more stocks with high correlation. Two companies with similar characteristics and work in a similar sector can have a high correlation between stocks. The strategy involves taking a long position in one stock and a short position in another stock. The following example can explain a pairs trading strategy. Assume there are two stocks X and Y which belong to similar market sectors and both financially identical in terms of revenue earned and expense structure. Both companies, X and Y, are in similar market sectors. Because they carry similar risks, so in theory, both should have similar prices. Although both companies X and Y carry a similar price due to fluctuations in supply and demand, the prices of X and Y may change, which leads to a temporary diverge in the price of X and Y. The pairs trading strategy takes advantage of this situation by selling the stock, which is relatively overpriced stock, and buying the other stock, doing similar when other stocks become overpriced and buy the underpriced stock. The pair trading strategy takes advantage of the mean-reverting behavior of the difference between the prices of the stocks X and Y. Other than market-neutral strategy, another reason why pairs trading is popular is that once a trader finds a pair, they only have to look for the relative price of a pair. This is much easier than trading with one stock with the help of a directional view.

A real-world example of the correlation between the price trends of the stock are the stock trends of Pepsi and Coca Cola

Stocks (see Fig. 30.1). Both companies are very similar and their primary product is soda pop.

Some of the other pairs can be Domino's Pizza (DPZ) and Papa John's Pizza (PZZA), Target Corporation (TGT) and WalMart (WMT).

Cointegration tests analyze non-stationary time series, processes that have variances and means that vary over time. In other words, the method allows you to estimate the long-run parameters or equilibrium in systems with unit root variables [9].

Two sets of variables are cointegrated if a linear combination of two variables is lower than the order of integration. Following is the Definition of cointegration: x_t and y_t are said to be cointegrated if there exists a parameter α such that $u_t = y_t - \alpha x_t$ is a stationary process.

The cointegration exists between two sets of variables if a set of I(1) can create linear combinations with I(0). I(1) can give us information that a single set of differences can transform the non-stationary variables to stationary. Three main methods can perform cointegration test are: Engle-Granger Two-Step Method, Johansen Test, and Maximum Eigenvalue test.

In this paper we perform cointegration by the Engle-Granger method. The Engle and Granger two-step method was developed by Robery Engle and Clive Granger in their seminal paper. One of the strengths of this procedure is the ease of testing. The two-step procedure is easy to follow and paints a good picture of the general idea behind cointegration. The idea is to first verify that the individual series are indeed integrated of order I(1) (being non-stationary). Then we regress one on the other using standard OLS and check if the residual series is integrated of order I(0) (suggesting stationarity). If the residuals are stationary, then we can extract the coefficients from the OLS model and use those to form a stationary linear combination of the two-time series.

Suppose X_t and Y_t are the natural logarithms of the price series of 2 assets, X and Y. The time series X_t and Y_t is integrated order 1, and if there is a linear combination that exists

$$Z_t = X_t - \beta Y_t - \alpha$$

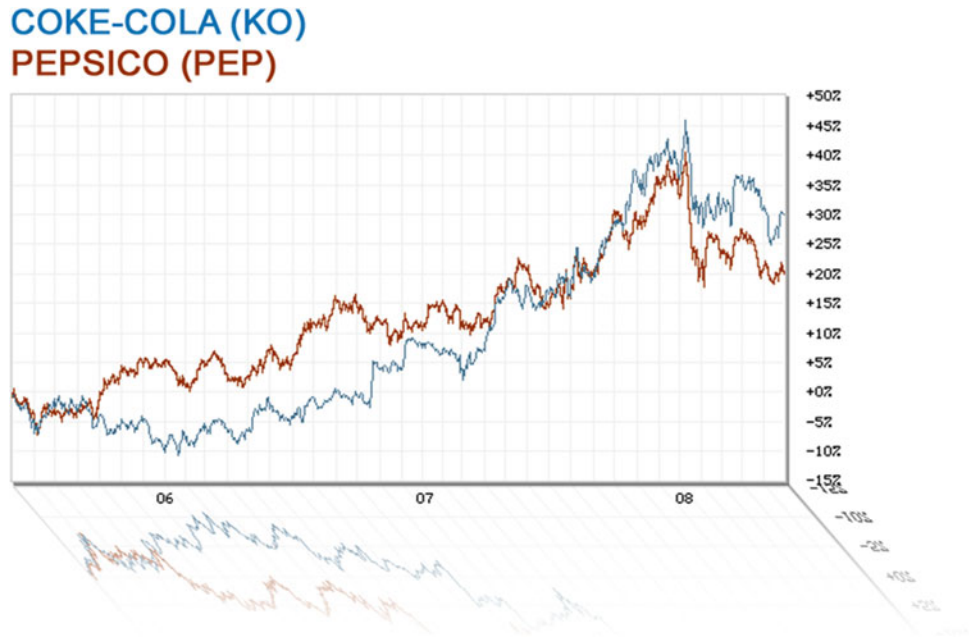
where Z_t is a stationary time series.

To check if time series X_t and Y_t are cointegrated in Engle and Granger two-step test -

1. If both X_t and Y_t are non-stationary time series, first estimate the β and α using ordinary least squares (OLS) regression.
2. Then Z_t is tested for stationarity where the Augmented-Dickey-Fuller.

If Z_t is stationary, then we can say that X_t and Y_t are cointegrated.

Fig. 30.1 Stock trends of PEP and KO. (Image is reference from The Street.com)



There are a few shortcomings of the Engle-Granger procedure. It is incapable of simultaneously testing for cointegrating relationships among multiple series. In Engle and Granger method, the choice of which series to regress on the other is somewhat arbitrary. Depending on which variable we regress on we would receive a different cointegrating vector. Finally, Engle and Granger method overlooks the underlying error-correction model influencing the spread series.

All of the above shortcomings can be addressed through the Johansen test. This procedure estimates cointegrated VARMA(p,q) in the VECM (vector error-correction model) form for m cointegrating relationships between k different series in x_t .

$$\Delta x_t = \alpha \beta' x_{t-1} + \sum_{i=1}^{p-1} \Phi_i^* \Delta x_{t-1} + a_t - \sum_{j=1}^{q-1} \Theta_j a_{t-j}$$

The notations, α and β are both $k \times m$ matrices, Δx_t represents the first difference as $\Delta x = x - x_{-1}$, Φ_i are the AR coefficients, and Θ_j are MA coefficients. The cointegrating equation is defined by $\beta' x\{t - 1\}$, where β contains the coefficients for the m cointegrating vectors.

30.3 Our Proposed System

We implemented pairs trading strategy and ran it on historical data. We collect the S&P 500 data from the Internet and store it in a database. We then allow users to enter a set of stocks to find cointegrated pairs among them. We also provide an option to find the cointegrated pairs in all of S&P 500 stocks. Once the cointegrated stocks are selected we run

the backtesting algorithm on these pairs. Finally, we provide the analysis of trades executed by the algorithm with average and daily data, and plot a chart of daily profit and loss with pairs trading strategy. The general architecture is shown in Fig. 30.2.

The Main Module acts as a broker between different modules and is also a configuration loader. It loads the configuration files in their objects and passes these configuration objects to different modules. The primary functions of the main module are:

1. Load configuration files into configuration objects.
2. Load data from the Internet to the database.
3. Find correlated pairs among the list of stocks using PairFinder Module.
4. Generate trade files and analysis using Backtest module

The database DB Module provides a repository for the entire stock data. The stock data is used in both finding the cointegrated pairs as well as calculating trades from the daily trade data for backtesting. The DB modules load the stock data and ticker from the Internet using Wikipedia and Yahoo finance. For efficiency it uses a local database for cache to store the trading data. The primary functions of DB modules are:

1. Create local database and tables.
2. Get all the S&P 500 tickers from Wikipedia.
3. Load daily stock data for these tickers from Yahoo finance into the local database tables.
4. Provide functions to other modules for calculating data from local database.

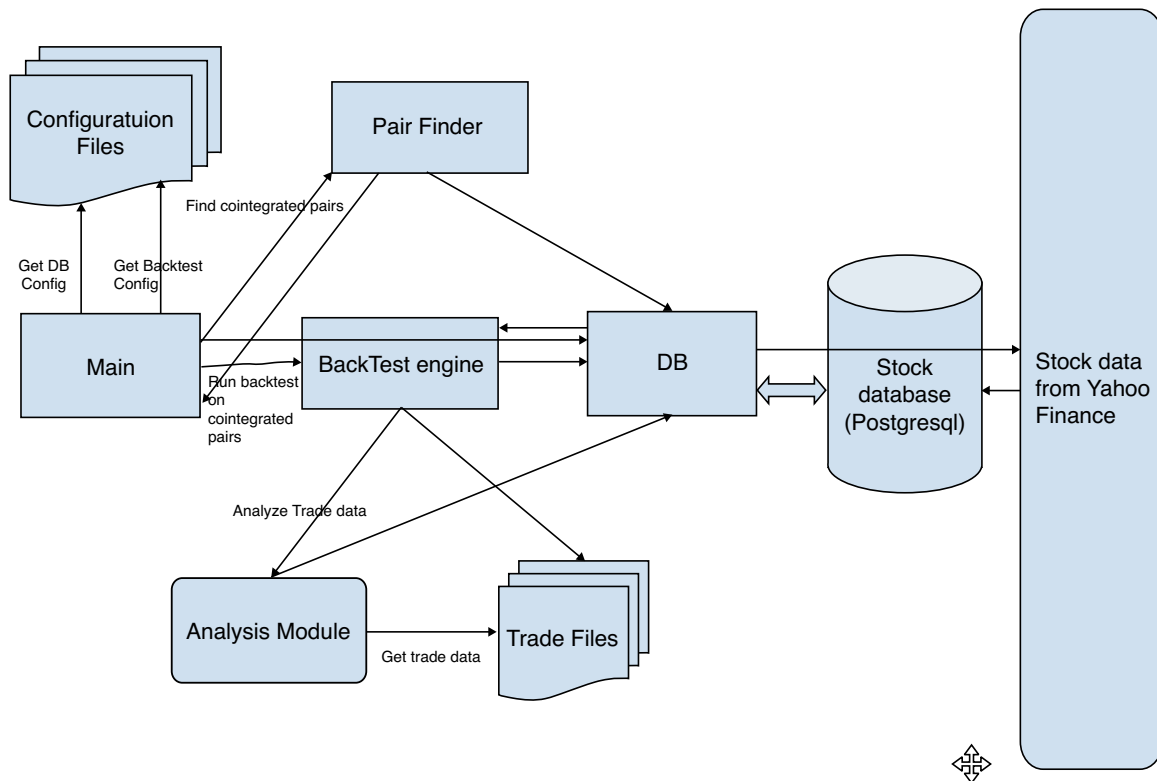


Fig. 30.2 Detailed design diagram for the system

Pair Finding Module uses stats models tools to find correlation between stock data time series. The correlation is measured as pValue. Once the pValues are calculated we measure it against a pValue threshold and decide whether two stocks are cointegrated. This module also displays a heat map of the pValues of multiple stocks, which gives a nice visualization of which stocks are correlated. The primary functions of this module are:

1. Get all the data for user given Stock tickers from the db module.
2. Calculate pValues for each stock pair.
3. Create a heat map for every sector of stocks.
4. Return correlated stock pairs.

Backtest Module takes the correlated pairs and generates trade files using the pairs trading algorithm. The central tenet of this algorithm are zValues. The current zValues for each pair are calculated using short and long look up windows and standard deviation. These zValues are then compared with the threshold values set in configuration. Based on these the backtest module produces trade signals and writes trades in Trade files. The primary functions of Backtest Module are:

1. Calculate zValues for each pair.

2. Using zValues and thresholds create trade signals and make buy/sell decisions.
3. Write these decisions into Trade files for persistent storage.
4. Call Analysis module to use Trade files and generate summary of trade data.

Analysis Module collects all the trade information in the Trade files and then generates a summary of all the trades executed. It provides an average trade summary as well an average daily trade summary. It also generates a chart for daily profit and loss with Pairs trading strategy. The primary functions of Analysis modules are:

1. Collect trade data from trading files.
2. Calculate overall and daily profit/loss and averages for trades.
3. Plot daily profit and loss on chart.

The configuration for database connection and backtesting engine is stored using YAML files for easy readability. This also allows us to modify program config without recompiling it. We use PyYaml for parsing the config files into data classes and pass the objects of these data classes to other modules.

To find pairs we use the StatsModel python library. This provides a function to calculate pValue between two given

time series. The higher the pValue the less correlated are the two time series. We compare these pValues to a configurable threshold to make sure we find two correlated pairs.

Backtesting module takes the list of pairs and for each pair calculate their zValues. The zValues are calculated by finding ratios between the two stocks of the pair and doing a rolling average and standard deviation calculation at each data point (day). It uses stock data ratios and rolling mean and standard deviation to calculate the zscore. Once we have the zValues for Stocks at every data point, we compare the value with configurable upper and lower limits of zValues and check what kind of trade we are in right now. If we are in a short trade and the logic dictates that we should continue to be in a short trade we do not create another short position. Similarly for a long trade we do not want to write in trade files multiple long trades. To make sure that it does not create extra trades and reduce the effectiveness of our trading strategy, we must maintain the state to store our current trade. Which means if we are in a short trade and get a short signal we ignore it, similarly if we are in a long trade and get a long signal, we ignore it as well. Trading signals are only used to either enter or exit from a trade. So if we are in a particular trade and get a signal for its reversal or lack of effectiveness we exit from such trade.

After the trades are calculated and trade files are generated, the analysis module just collects all the trade data and computes daily and overall average, profit and loss, winning trades etc. to display the summary of success of strategy to the user. This also generates a chart for the daily profit and loss. We use Python's matplotlib library to render the charts. We also compute some moving averages to show on the chart, which makes it easy to track the profit/loss progress on the chart.

We also plot moving averages on this chart. Moving averages are defined by average of different subsets of a given data set. These subsets are determined by the window size. In finance it is usually a price indicator that can be used for technical analysis. The primary reason for plotting a moving average on a chart is to smooth out a big fluctuation. This is precisely why our program plot Moving averages in this chart. We want to be able to see the long-term price action of our trades, and moving averages help us to ignore the daily noise and smooth out the graph to see the effectiveness of our trade. We plot two different moving averages in this chart, one with back looking window of 50 data points, while the other is for 200 data points. Each data point in this graph refers to a day. These two Moving averages provide a clear and visual identification of the effectiveness of our pairs trading strategy.

30.4 Results Based on Yahoo Finance Data

We use the Stats Model tools to find cointegrated pairs of stocks. This is done by generating pValues of the stocks. These values are displayed in a heat map for each stock of the same sector. In Fig. 30.3 we show the heatmap for correlation pValue score between following tech sector stocks {'GOOG', 'FB', 'TWTR', 'AMZN'} representing Google, Facebook, Twitter, and Amazon.

Different colors represent different level of correlation between stocks. We note that, lower the value of pValue between the stock pairs, the more correlated they are, and make good candidates for pairs trading strategy. In heatmap the higher to lower correlation is shown with a green to red spectrum. As we can see in the above example, two stock pairs, GOOG and AMZN are most correlated, while GOOG and FB are only slightly less correlated. Both pairs are good candidates for pairs trading strategy. On the other hand, TWTR stock is not correlated with any other ticker, which is shown by all red heat zones in TWTR row.

Our system also displays the price data for these stocks to provide a visual confirmation of a good correlation (see Fig. 30.4). We note that GOOG, AMZN and FB, GOOG pair really do correlate, and this confirms the correctness of heatmap generated.

A similar example is shown for the retail sector {'WMT', 'TGT', 'LOW', 'HD'} representing Walmart, Target, Lowe's and Home Depot. Figure 30.5 the heatmap for correlation pValue score between following four retail sector stocks

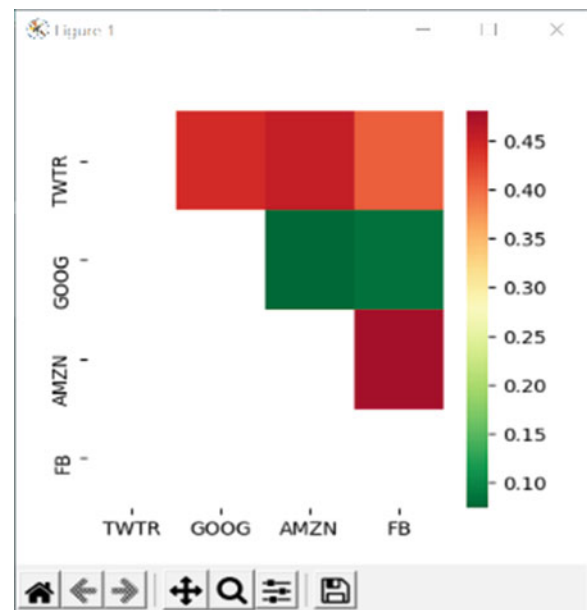
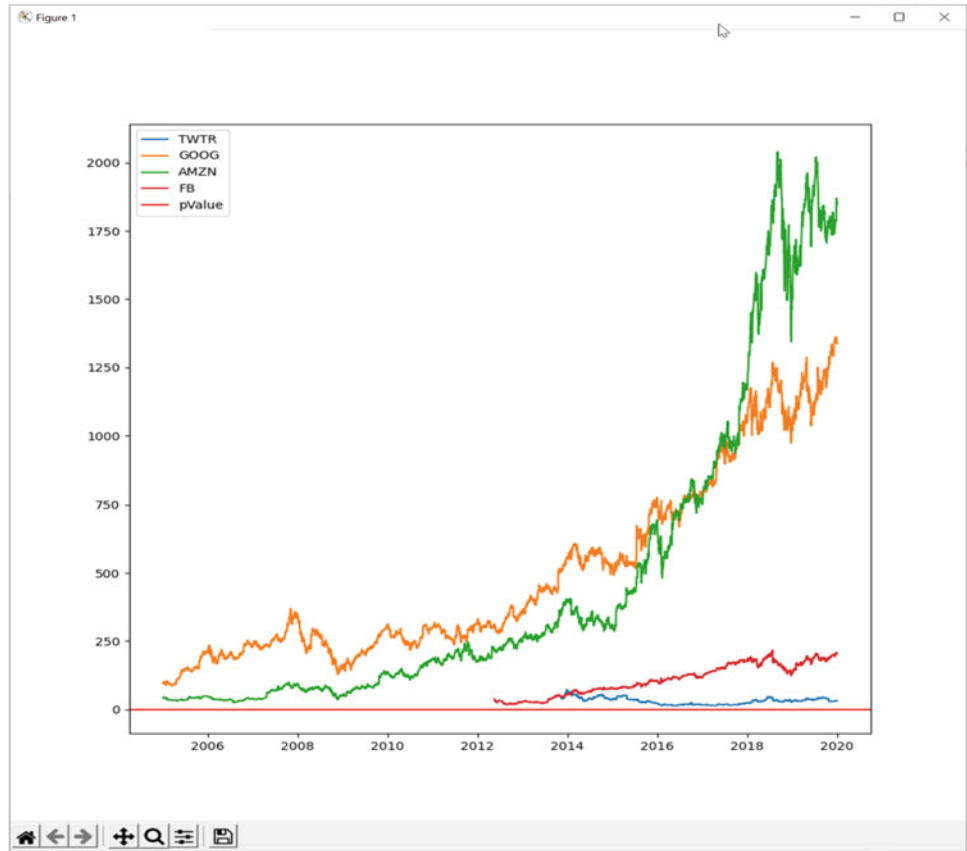


Fig. 30.3 Heatmap of correlation between four tech sector stocks

Fig. 30.4 Price data for the four tech sector stocks



{‘WMT’, ‘TGT’, ‘LOW’, ‘HD’}. Different colors represent different level of correlation between stocks.

The price data is shown in Fig. 30.6. Both price data as well as heat map, LOW and HD are the only highly correlated stocks in this set of stocks.

The system creates a trade summary from analyzing the master results and pair files. The daily profit and loss chart is also plotted (see Fig. 30.7). This is a daily profit / loss chart that shows at every data point (day) how much profit or loss our trading strategy has yielded. The daily profit and loss chart shows how effective our trading strategy is on a day-to-day basis. It is important to realize at this step is that no trading strategy could produce winning trades every datapoint or even guarantees winning trades as a higher percentage of total trades. Which is why we also generate a report that shows how many winning trades were performed by our strategy. This is only the indication of total number of winning trades and not whether the strategy was successful or not. This is because it is possible to wipe out a lot of winning trades with a single losing trade and so while the winning trades would be a higher percentage of all trades, the total initial capital might still be reduced.

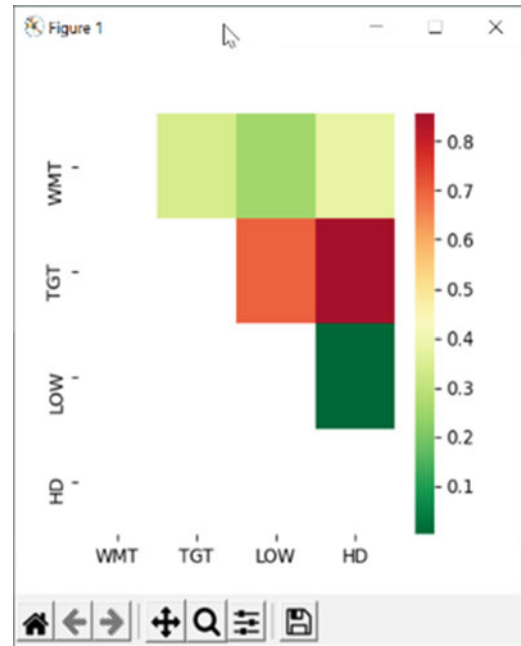
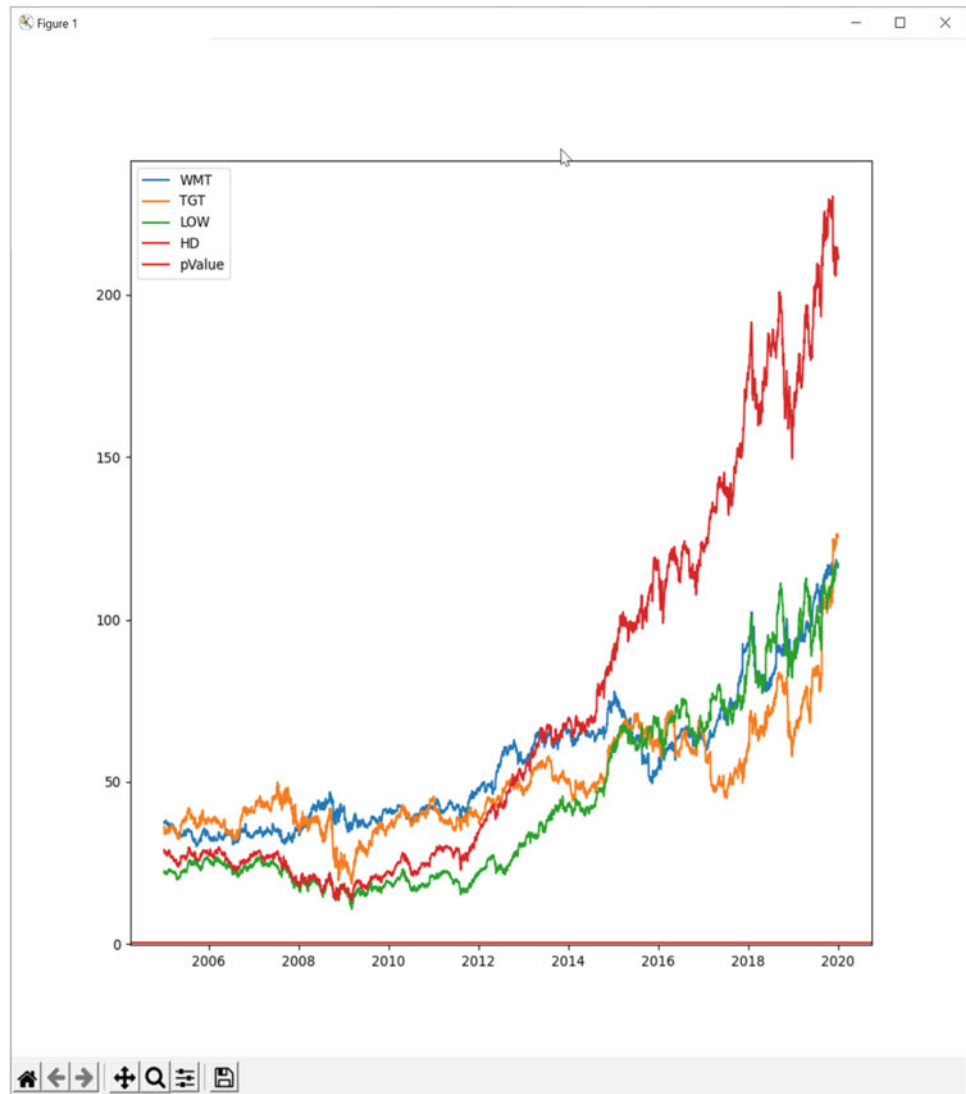


Fig. 30.5 Heatmap of correlation between four retail sector stocks

Fig. 30.6 Price data for four retail sector stocks



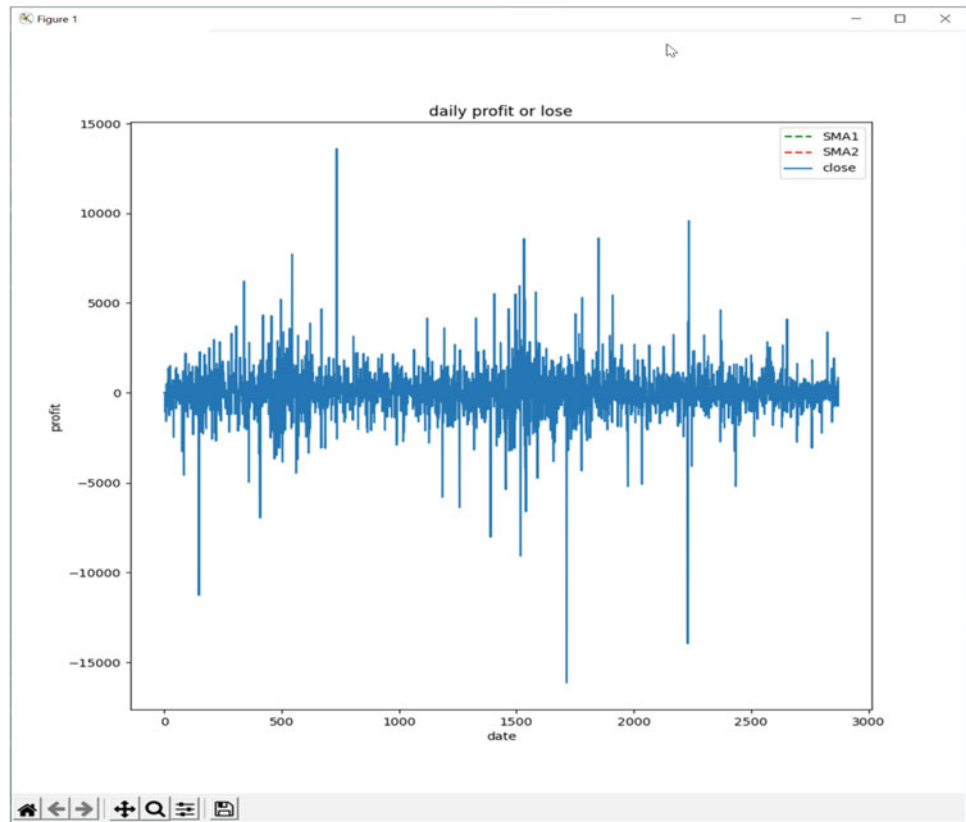
30.5 Conclusion and Future Work

This system has some limitations. Our system program does not allow for plug in strategy backtesting since our backtesting algorithm is currently hard coded. We currently do not support any machine learning or deep learning for making future predictions. This program is strictly for backtesting the pairs trading hypothesis. It takes a significant amount of time to calculate suitable correlated pairs out of a large set of Stocks. We have the option to generate the correlated pairs out of all of S&P 500 stocks, indexed by sectors in the pair generator module. However, running that takes a very long time and so its much more convenient to provide a smaller set of stocks, from which the pairs can be computed quickly. We currently write the files with trading information. This could be improved by natively writing the files to a cloud service like S3. We do not have the option to go back and look at heat

map charts again. Adding an option to persist the image files generated during the running of system is desirable.

Adding the forward trading to the existing project will become the full algorithmic trading bot. Forward trading is similar to back trading. Both forward/back trading use the same algorithm and logic. Cointegration can be improved by adding a Bayesian Kalman filter. A Bayesian Kalman filter could help improve the trading results by cutting down statistical noise and make more accurate predictions of future state of the service from historical data. We could convert the system into a web service that generates the trading data and send it back with REST APIs. This data could then be captured in a Frontend web framework like React and plotted on the browser. This would make it possible for the program to run on a more powerful backend fixing the limitation of taking a long time and we can still access it from any computer connected to it. We could train a model using different machine learning algorithms to make accurate predictions

Fig. 30.7 Daily profit and loss chart



and then reinforce it by the results of these predictions. The idea of reinforced learning would make the trading bot much more accurate and increase trust in the system.

Currently our backtesting engine only uses Pairs trading strategy. This pairs trading algorithm is also hardcoded. We could easily make the Engine separate from the actual trading algorithm so its easy to put different trading strategies to test and compare their data.

Acknowledgments This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700.

References

1. R. Shelvin, *How Much Do Banks Spend On Technology? (Hint: It Would Weigh 670 Tons In \$100 Bills)*, (2019, Apr 1). From Forbes.com <https://www.forbes.com/sites/ronshelvin/2019/04/01/how-much-do-banks-spend-on-technology-hint-chase-spends-more-than-all-credit-unions-combined/?sh=1066d949683a>
2. [https://en.wikipedia.org/wiki/Pairs_trade%23%20text=A%20simplified%20example,-Pepsi%20\(PEP\)%20and&text=If%20the%20price%20of%20Coca,to%20their%20historical%20balance%20point](https://en.wikipedia.org/wiki/Pairs_trade%23%20text=A%20simplified%20example,-Pepsi%20(PEP)%20and&text=If%20the%20price%20of%20Coca,to%20their%20historical%20balance%20point). (n.d.). From Wikipedia
3. <https://www.fidelity.com/learning-center/trading-investing/trading/pairs-trading>. (n.d.). From Fidelity
4. <https://www.investopedia.com/articles/trading/04/090804.asp>. (n.d.). From Investopedia.com
5. <https://www.investopedia.com/terms/p/pairtrade.asp>. (n.d.). From Investopedia.com
6. <https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.coint.html>. (n.d.). From Statsmodels.org
7. J.T. Peng, Research of the Matlab application in the fingerprint identification system, in *International Conference on Image Analysis and Signal Processing, Hangzhou*, (2012), pp. 1–5
8. S.H. Penman, *Accounting for Value* (Columbia Business School Publishing, 2011)
9. B. Rao, *Cointegration: For the Applied Economist*, 2nd edn. (Springer, 2007), p. 279

Rachana Chittari, Marian Sorin Nistor, Doina Bein, Stefan Pickl,
and Abhishek Verma

Abstract

Quora is an online platform that empowers people to learn from each other. On Quora, users can post questions and connect with others who contribute unique insights and quality answers. But as with any other social media or online platform, there is the potential for misuse. A key challenge in maintaining the integrity of such an online platform is to classify and flag negative content. On Quora, the challenge is to identify insincere questions. Insincere questions could be those founded upon false premises, are disparaging, inflammatory, intended to arouse discrimination in any form, or intend to make a statement rather than look for helpful answers. We propose to develop a text classification model that correctly labels questions as sincere or insincere on the Quora platform. For this purpose, we used the Quora Insincere Questions Classification dataset, which is available on Kaggle. We first trained classical machine learning models such as Logistic regression and SVMs to establish a baseline on the performance. However, to leverage the large dataset, we used neural network-based models. We trained several models including standard neural networks, and LSTM based models. The best model that we obtained is a two-layer Bidirectional LSTM network that takes word

embeddings as inputs. The classification accuracy and F1-score for this model were 96% and 0.64, respectively.

Keywords

Bi-directional LSTM model · Insincere questions · Logistic regression · Machine learning · ML · Negative context · Text classification · Quora dataset · LSTM model · SVM model

31.1 Introduction

In addition to the search engines such as Google and Bing, there are websites known as question forums that one can use to gain knowledge. A question forum is an online discussion site where people can hold discussions in the form of posted messages. They have gained a lot of popularity due to their easy to use and understand methodology. Quora, Stack Overflow, Yahoo Answers are some examples of question forum websites.

Quora is a popular online platform for users to ask simple, personal, professional questions and generally obtain well thought-out answers to [1]. While most of the questions are asked in good faith, there are several instances where people tend to ask questions that are inappropriate or inflammatory. They may be targeted at a specific group of people, intend to make a false statement, or sometimes make no sense. These questions tend to create havoc and threaten the integrity of the platform. Questions like these are termed “insincere” and they deviate from the main purpose of an online forum, which is simply to help users share knowledge.

To ensure the safety of the users and the integrity of the forum, it is thus extremely important to classify, flag and remove such insincere questions before they can cause any significant damage. In the past, Quora made use of human

R. Chittari · D. Bein (✉)
Department of Computer Science, California State University,
Fullerton, Fullerton, CA, USA
e-mail: chittari.rachana@csu.fullerton.edu; dbein@fullerton.edu

M. S. Nistor · S. Pickl
Department of Computer Science, Universität der Bundeswehr
München, Neubiberg, Germany
e-mail: sorin.nistor@unibw.de; stefan.pickl@unibw.de

A. Verma
Department of Computer Science, California State University,
Northridge, Northridge, CA, USA
e-mail: abhishek.verma@csun.edu

reviewers to label questions as sincere or insincere. However, due to the tremendous growth in the number of users and the number of questions posted daily, it becomes intractable to maintain a human based review system. Thus, it becomes important to develop an automated system to perform this classification task. One viable approach is to use a machine learning based text classification approach. We propose to build a supervised machine learning model that can classify questions as sincere or insincere.

The paper is organized as follows. In Sect. 31.2 we present background work, followed by our proposed system architecture in Sect. 31.3. Results from applying various machine learning algorithm are presented in Sect. 31.4. Concluding remarks and future work are given in Sect. 31.5.

31.2 Background Work

Detecting inappropriate and negative content online is a highly relevant problem today. This is relevant not just to Quora, but to all social media platforms and online forums. Examples of classifying negative content include analyzing movie reviews on IMDB, entries on Wikipedia, and tweets on Twitter. These are essentially problems in the domain of natural language processing.

A lot of work has been done using machine learning and deep learning models. The basic idea is that supervised machine learning algorithms can be trained on a labeled set of content to correctly identify negative examples. To this end, several companies have shared large pools of data on platforms such as Kaggle [2], and challenge machine learning researchers to develop state of the art classification models.

In the past, traditional machine learning approaches were used to tackle these classification problems. The important consideration with these approaches is how to represent a sequence of words. Once a good set of features have been extracted, they can be used with almost any classification model. Typically, representations such as bigrams, n-grams, and bag-of-words, have been combined with classifiers such as Logistic Regression and Support Vector Machines. A few examples of these are the following. Vectorization using bag-of-words was combined with Logistic Regression and Naïve Bayes classifiers for tweet classification in [3, 4]. A Support Vector Machine model was used to classify BBC documents into five categories in [5], and n-grams were combined with SVMs for emergency event detection on social media in [6].

Deep learning [7] is a broad family of machine learning models that use artificial neural networks [8, 9]. The term “deep” refers to the use of multiple layers of neurons in the network. The availability of massive data sets and high-speed computational resources such as Graphics Processing Units have enabled us to train deep neural networks to perform a wide variety of machine learning tasks.

Deep learning models have achieved near human performance in various tasks such as object recognition and speech recognition. They also provide the best results for natural language processing tasks such as machine translation, text classification and sentiment analysis.

One of the important advantages of using neural network-based models for natural language processing is that they can be used to learn word embeddings. A word embedding is a mapping from a word to a real-valued vector that is thought to encode the meaning of the word. While several methods exist to generate such embeddings, deep learning methods have provided some of the most widely used word embeddings, such as GloVe (Global Vectors for Word Representation) [10]. These word embeddings can be thought of a representational layer in a deep learning architecture.

Several types of deep learning architectures have been considered for text classification. These include convolutional neural network models, sequence models such as recurrent neural networks, and attention-based models [11] provides a comprehensive review of more than 150 deep learning models for text classification that were developed in the last few years. In this project, I have mainly focused on using recurrent neural networks and their variants for insincere question classification.

We used an open dataset available on Kaggle [2]. There are about 1.3 million examples in this dataset. Each example consists of 3 fields: q_id, question_text and target:

q_id: Unique question ID assigned to each question
 question_text: a question posted on Quora in the English language
 target: takes the values 0 or 1. The values 0 and 1 correspond to sincere and insincere questions, respectively.

Here are a couple of example questions and their target values:

Sincere (0): “How did Quebec nationalists see their province as a nation in the 1960s?”

Insincere (1): “Which babies are more sweeter to their parents? Dark skin babies or light skin babies?”

Many of the insincere questions are discriminatory in nature, inflammatory, politically polarizing, and potentially harmful to raise on the Quora platform. It is thus crucial to correctly identify and weed out these questions.

We analyzed was the distribution of the target values. We found that the Kaggle dataset is highly imbalanced. Only about 6.2 % of all the examples are labeled as insincere. This implies that we should not use classification accuracy as a metric to evaluate our model. For instance, a naïve model that always outputs 0 as its prediction would already achieve a classification accuracy of 93.8%, but importantly it would never be able to correctly identify an insincere question (with target value 1). Therefore, it is important to use metrics such

as precision and recall to correctly evaluate the performance of a model on this dataset.

We applied the following preprocessing steps:

- Convert all text into lowercase
- Remove all special characters and non-English characters that appear in the dataset
- Remove punctuation characters
- Remove stop words or common English filler words like “and”, “an”, “the”
- Split the question strings into word tokens

We utilized the Natural Language Toolkit (NLTK) [12] to perform several of these preprocessing steps. NLTK is an open-source platform for building Python programs to work with natural language data. It provides a corpus of English stop words and tokenization functions that I used to clean my dataset. Alternatively, the Torchtext library of PyTorch can also perform the same set of preprocessing steps. PyTorch is an open-source machine learning framework that can be used to quickly build deep learning models.

The next stage was to extract features from the cleaned dataset that can be used to train the machine learning models. For text classification, we must first convert the text into a vector representation. Word vectorization is the process of converting text into a numerical representation and is an important first step in any Natural Language Processing task.

We have considered the following two approaches for feature extraction:

- (a) Word count vectors. A word count vector of a question or sentence is a representation that specifies how many times each word in a given vocabulary appears in the sentence. This is essentially a bag-of-words representation, which disregards word order but keeps multiplicity. To implement the word count representation, we first built a dictionary of all the unique words that appear in my dataset. Each question was then converted into a vector of word counts. The length of this vector is equal to the number of words in the dictionary. Consider the following example for illustration. Suppose that a dictionary comprises the following words: {“hello”, “who”, “how”, “are”, “you”, “me”, “doing”, “today”}. Let us see what the word count vector for a sample question, “Hello, hello, how are you?” would like. Since the dictionary has 8 words, the sample question is represented by an 8-dimensional vector, with each element corresponding to the count of a particular word in the dictionary. The word count vector for our example would be: (2, 0, 1, 1, 1, 0, 0, 0). While word count vectors are useful in some cases, they do have a few limitations. When the vocabulary is large, the word count vectors are sparse and high dimensional. The Quora dataset, for instance,

has about 100,000 unique words. And each question in the dataset typically has only 10–20 words. Further, the word count vectors ignore word order, which is certainly important for assessing the nature of a question.

- (b) Word embeddings are mappings from words to N -dimensional real-valued vectors such that any two words that are closer in the vector space are similar in meaning [13]. Various methods can be used to generate these mappings, including deep learning techniques. Further, in deep learning models, the embeddings can be thought of a representational layer in the architecture. There are several pre-trained word embeddings that are now available. A few examples are wiki-news-300d-1M, GoogleNews-vectors-negative300, paragram_300_sl999, glove.6B.100D, and glove.840B.300D. For this project we have used the GloVe embeddings. GloVe or Global Vectors for Word Representation [10] is an unsupervised learning algorithm for obtaining vector representations for words. The glove.840B.300D embeddings output a 300-dimensional vector for each word in the vocabulary. To combine the embedding vectors for words in a question, I considered 3 methods:

1. Computing the mean of embeddings vectors of all words in a sentence
2. Concatenation of the embedding vectors of all the words in a sentence
3. Use these embeddings as inputs to the sequence models

The first approach is obviously very lossy. In the second approach, the dimensionality of the feature vector increases with the number of words in the question. We explored using the concatenated features as inputs to a few basic classifiers.

31.3 Machine Learning Models Applied

To obtain a baseline on the performance we started with Naïve Bayes Classifier, Logistic Regression, and Support Vector Machines.

A Naïve Bayes classifier is a probabilistic classifier based on Bayes theorem, which makes the strong assumption that the features are independent of each other [14]. Given an input feature vector $x = (x_1, x_2, \dots, x_N)$, we must compute the probability of every class given x . A vector x is assigned to the class C_i which has the highest probability.

Logistic regression is a binary classification algorithm, used to model the probability of a certain class given a set of input features. In logistic regression, the probability of the target given input features is computed as follows:

$$P(y = 1|x) = \sigma(w^T x + b)$$

where w is a weight vector, b is a bias term and σ is the sigmoid function given by

$$\sigma(z) = \frac{1}{(1 + e^{-z})}$$

Let us call the prediction $\hat{y} = P(y = 1|x)$. We can then compute the cross-entropy loss function for one example as follows:

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

The weight vector and bias term are then learnt such that they minimize the average cross-entropy loss over all the training examples. This optimization is typically performed using gradient descent-based methods.

For the logistic regression classifier, we used a concatenation of 300-dimensional Glove word embedding vectors as inputs. We first computed the histogram of the number of words in each question for the entire dataset. Based on this histogram, I chose the maximum number of words, N_{max} , to represent each question. Then, each question was represented by a feature vector of size $300 N_{max}$. For questions with fewer words than N_{max} , we zero padded the feature vector to ensure they are of the same length for all examples.

A Support Vector Machine (SVM) is a binary classification algorithm that tries to find the separating hyperplane/decision boundary with the maximum margin [15]. The SVM is a linear classifier, but they can also perform non-linear classification by implicitly mapping the input features into higher-dimensional feature spaces. This is known as the kernel trick. We applied SVM for question classification with a concatenation of word embedding vectors as input features.

We implemented the Naïve Bayes, Logistic Regression, and SVM classifiers using the scikit-learn, which is an open-source machine learning library in Python.

One of the crucial limitations of the methods described previously is that they do not explicitly consider the sequence of words in a question. A family of artificial neural networks known as Recurrent Neural Networks (RNNs) can handle sequential inputs and are perfectly suited for the question classification task. A Recurrent Neural Network (RNN) is a class of neural networks with self or recurrent connections between nodes. They are derived from feed-forward neural networks, which have only forward connections between nodes in different layers of a network. The hidden layer in RNNs can be thought of as memory states, which are continuously updated by the inputs. Further, RNNs can be used to handle input sequences of variable lengths. This makes RNNs ideally suited for handling applications with sequential inputs such as machine translation, time-series prediction, speech recognition, and indeed text classification. Next, let us look at how a RNN can be trained.

Backpropagation is a widely used algorithm for training neural networks [16]. In supervised learning, training a model requires us to update its weights or parameters such that they minimize some loss criterion. This requires us to compute the gradient of the loss function with respect to all the model parameters.

Backpropagation is an algorithm for computing the gradient of the loss function with respect to the weights of a feed-forward neural network for a single input-output example. In contrast to a naïve direct computation of the gradient with respect to each individual weight in the network, backpropagation uses the chain rule of calculus to compute the gradients one layer at a time. It iterates backwards from the output layer to avoid redundant computations of intermediate terms. This results in a highly efficient algorithm for computing gradients in a neural network. This efficiency makes it feasible to use gradient methods for training multilayer neural networks.

The backpropagation algorithm was originally designed for feedforward neural networks. But they can be adapted to train recurrent neural network as well. A recurrent neural network can essentially be unfolded in time to obtain a feed-forward network with tied weights. This intuition was used to develop the Backpropagation Through Time algorithm (BPTT) [17].

Given an input sequence, BPTT works by unrolling all input timesteps. Each timestep of the unrolled recurrent neural network can be considered as an additional hidden layer, which also receives as input the hidden state from the previous time step. BPTT calculates and accumulates errors across each time step to compute the gradient of the loss function with respect to the network's weights.

BPTT is a useful algorithm for training recurrent neural networks. However, it can become computationally expensive as the number of time steps in the input sequence increases. Further, when the input sequences are long, the number of derivatives required for a single weight update will be high. This can cause the gradients of the weights to become diminishingly small depending on the nature of the activation functions used in the network. This problem is known as the vanishing gradients problem. For some activation functions, we can also observe the exploding gradient problem, which would result in numerical overflow problems.

The vanishing gradient problem is a major drawback of the BPTT algorithm especially for standard recurrent neural networks that use sigmoidal activation functions. The vanishing gradients make learning extremely slow. Another consequence of the vanishing gradient problem is that it does not allow learning of long-range dependencies within input sequences. Learning dependencies among different parts of a sequence is vital for several tasks such as machine translation and text classification.

Variations of BPTT such as Truncated BPTT were developed to solve the vanishing gradient problem. But perhaps

the best solution was networks that used gating mechanisms, such as Long Short-Term Memory networks [18].

Long Short-Term Memory (LSTM) is a type of recurrent neural network architecture that uses gating mechanisms to regulate the flow of information through a neuron [19]. A basic LSTM unit is composed of a memory cell c , hidden state h , and input, output, and forget gates that allow the unit to remember values over arbitrary times by regulating the flow of information in and out of the unit.

In theory, standard RNNs could keep track of arbitrarily long-term dependencies in the input sequences. However, the problem with standard RNNs is computational in nature. While training RNNs with BPTT, we run into the issue of vanishing gradients. RNNs using LSTM can alleviate this problem because LSTMs can remember values across several time-steps if necessary. This would allow the gradient to flow unchanged during backpropagation. This makes LSTM based networks well-suited to problems involving data where there can be lags of unknown duration between important events in the input sequence.

There are other types of architectures that use gating mechanisms such as Gated Recurrent Units (GRU) [20] that are based on the same intuition.

Recurrent neural networks are ideally suited for processing sequential inputs. However, in normal RNN architectures the computation at any time step depends only on the inputs up to the current time. However, for several tasks it might also be important to consider input data in future time steps as well.

This is crucial in applications such as natural language processing. In many cases, to understand the context in which a word is used, it is important to know what comes before and after it in the sentence. Consider the following two sentences:

- (a) She said, “Teddy bears are for sale”
- (b) She said, “Teddy Roosevelt was the 26th president of the United States”

A recurrent neural network processing the two sentences would perform the exact same computations for the first three words. However, in the first sentence the word “Teddy” refers to a toy, and in the second it refers to a person. This information can be obtained only by using the words that follow “Teddy” in the two sentences. Bidirectional RNNs were developed to address exactly this issue.

A bidirectional RNN is simply an RNN with two groups of neurons in the hidden layer. One group serves to process the input sequence in the forward direction, and the other for the backward direction (see Fig. 31.1). The two groups of neurons are then connected to the same output neurons, thereby conveying relevant information from both directions of processing. Bidirectional RNNs can be trained

very similarly to unidirectional RNNs, because the forward and backward states do not directly interact with each other.

Bidirectional RNNs are especially useful when the context of an input is needed like in the Quora insincere question classification task. We decided to use a bidirectional LSTM (Fig. 31.2) for this classification task. This combines the advantages of both a bidirectional architecture and LSTM units.

At each time step, the embedding vector for a word in the question text is provided as input to the network. The embeddings can be considered as an additional representational layer in the architecture. These embeddings can be learnt while training the network. Pre-trained embeddings can also be used (i) by keeping them fixed, or (ii) by fine-tuning them during training. The embedding vectors at each time step act as input to the forward and backward hidden states. The hidden states are usually initialized using random values. After the entire sequence has been processed in both directions, the final hidden states in both forward and backward directions are fed as inputs to an output neuron. The activity of this neuron is then passed through a sigmoid function to obtain the probability of the given question being insincere.

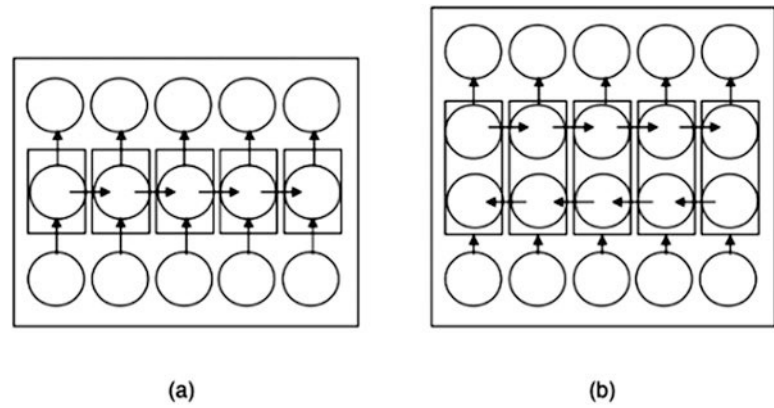
Fig. 31.2 illustrates an architecture with one hidden layer. But this could be generalized to having multiple hidden layers. RNNs with multiple hidden layers are known as stacked RNNs. In a stacked RNN architecture, the hidden states in a given layer are fed as inputs to the hidden layer above it. Increasing the number of layers or depth of the network enhances its representational power, similar to conventional feedforward networks. It is also thought that stacking allows the hidden states at each level to operate at different timescales, and thus allow the network to learn both short-range and long-range dependencies in a sequence [22].

31.4 Results of Various Machine Learning Algorithms

We used Google Colaboratory, or “Colab” or short, to develop and train the bidirectional LSTM model. Colab is a product from Google research that allows one to write and execute Python code through the browser. Colab also allows one to use Graphics Processing Units (GPUs). GPUs can be used to train deep learning models much faster than on CPUs.

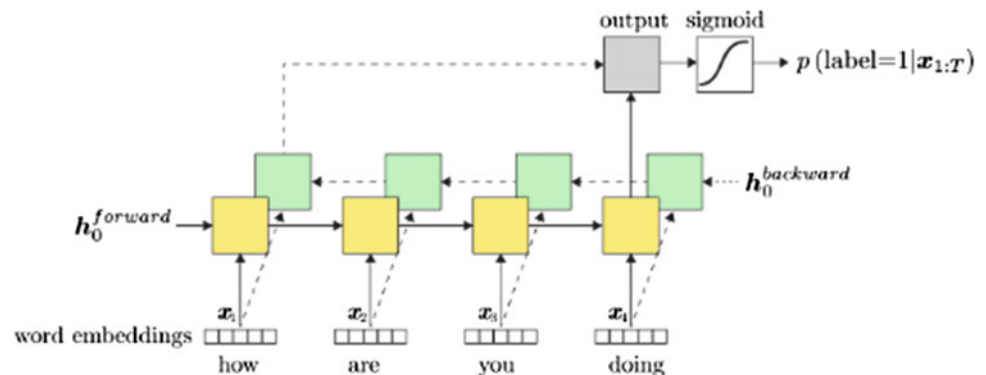
Since the Quora Insincere Questions dataset is highly imbalanced, the classification accuracy is not a good metric to evaluate a model’s performance. A confusion matrix is a better way of summarizing the results. We can use the entries of the confusion matrix to compute three metrics: precision, recall, and F1-score. For our classification task, it is very important to correctly flag all the insincere questions. Therefore, recall is probably the more important metric to

Fig. 31.1 Comparison of a unidirectional vs. bidirectional RNN [21]



Structure overview
(a) unidirectional RNN
(b) bidirectional RNN

Fig. 31.2 Bidirectional LSTM network architecture



consider. When we need a balance between precision and recall, we can use the F1-score.

We used a subset of the full data to train the classical models. The full dataset has roughly 1.3 million examples. Using word count vectors or concatenation of word embeddings results in very high dimensional features. For instance, even for a subset of 200,000 examples, the size of the input matrix of word count vectors for all examples was $200,000 \times 85,903$. The number of unique words in the dataset is 85,903. As the number of examples used increases, we can expect the number of unique words to grow as well. Such high dimensional features result in memory constraints. Thus, to obtain a feasible implementation as well as a rough lower bound on the performance, I trained the models using small subsets of data.

We used a subset of size 50,000 for Naïve Bayes, and size 20,000 for Logistic Regression and SVM. In each case, we split the data into train and test with a split ratio of 0.8, i.e., 80% of the data was used for training and the remaining 20% of the data was used for evaluating the model performance. Just for the Naïve Bayes classifier, we managed to train the model using a subset of size 200,000.

We implemented the Bidirectional LSTM using PyTorch on Colab. PyTorch is an open-source machine learning plat-

form for building and training deep neural network models. One of the most powerful features of PyTorch is Autograd, which is an automatic differentiation engine that powers neural network training. In PyTorch, we just need to implement the forward pass of the neural network model and provide the loss criterion. The Autograd engine can automatically compute the gradients of the loss function with respect to all the weights in the network. This makes the training of a neural network using backpropagation almost trivial.

The torchtext package in PyTorch consists of data processing utilities for natural language. We used torchtext to preprocess my data and created a PyTorch dataloader.

We used the GloVe.840.300D embeddings to map input words to vectors. This is treated as an embedding layer in the neural network architecture. Our implementation also has the option to fine-tune the embeddings if required. We used cross-entropy as the loss function with the Adam optimizer for training my model. Adam [23], derived from Adaptive Moment Estimation, is a gradient-based optimization algorithm that computes adaptive learning rates for each model parameter. Finally, for training my neural network model, we split the dataset into training and validation sets with a split ratio of 0.9.

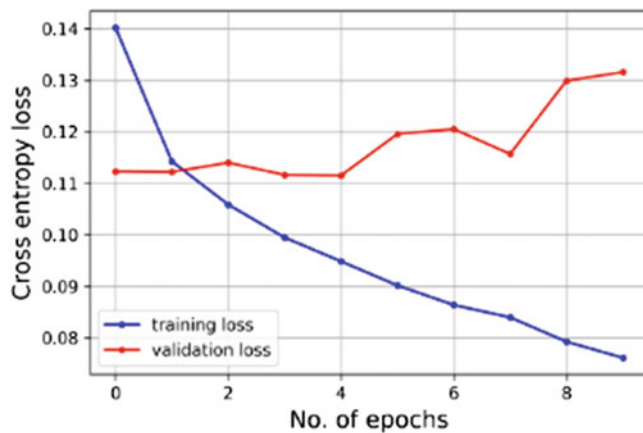


Fig. 31.3 Cross entropy loss with 64 hidden units

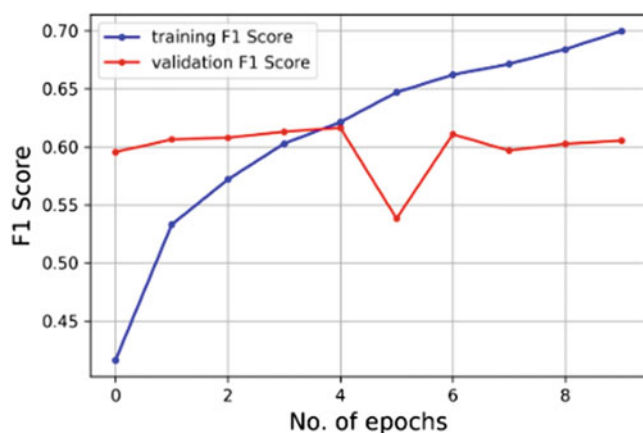


Fig. 31.4 F1-scores with 64 hidden units

We started with 500,000 examples for training the model. We used a 2 hidden layer architecture throughout. For the first pass, we used 64 units in each hidden layer, and the default learning rate of 0.1 for the Adam optimizer.

Figures 31.3 and 31.4 illustrate the cross-entropy loss and F1-scores vs. no. of training epochs. One epoch corresponds to iterating through all the mini-batches of the dataset.

In Fig. 31.3, although the mean training loss (blue) was continuously decreasing with the number of epochs, the validation loss increased after 4 or 5 epochs. This suggested that the model was in the overfitting regime. We obtained similar overfitting effects for different values of the number of hidden units. In order to address overfitting, we plan to use Dropout regularization [24].

31.5 Conclusion and Future Work

The neural network model that we used for the Quora Insincere question classification task was a Bidirectional LSTM

network with two hidden layers. We obtained better recall and F1 score than the baseline set by the traditional ML methods.

The model most likely benefits from the advantages of the bidirectional architecture and the LSTM. This could be analyzed by comparing the performance of our model with that of a unidirectional architecture.

While the performance of my model is quite good, there are several things that could be done to potentially improve the classification performance. The first of these is to perform a precision-recall analysis to find the best threshold. In classification problems such as these, there is always a trade-off between precision and recall. The precision-recall analysis would allow us to pick a threshold based on the criteria we want to satisfy.

Acknowledgment This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700.

References

1. Quora, [Online]. Available: <https://www.quora.com/>
2. Kaggle, [Online]. Available: <https://www.kaggle.com/>
3. O. Aborisade, M. Anwar, Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers, in *IEEE International Conference on Information Reuse and Integration (IRI)* (2018). <https://www.fidelity.com/learning-center/trading-investing/trading/pairs-trading> (n.d.). From Fidelity
4. S.T. Indra, L. Wikarsa, R. Turang, Using logistic regression method to classify tweets into the selected topics, in *International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (2016)
5. A.W. Haryanto, E.K. Mawardi, Influence of word normalization and chi-squared feature selection on support vector machine text classification, in *International Seminar on Application for Technology of Information and Communication* (2018)
6. L. Yanfang, J. Niu, Q. Zhao, J. Lv, S. Ma, A novel text classification method for emergency event detection on social media, in *IEEE SmartWorld, Ubiquitous intelligence computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation* (2018)
7. Deep learning, [Online]. Available: https://en.wikipedia.org/wiki/Deep_learning
8. Artificial neural network, [Online]. Available: https://en.wikipedia.org/wiki/Artificial_neural_network. S.J. Selvarani, Automatic identification and detection of altered, in *International Conference on Intelligent Computing Applications* (Coimbatore, 2014), pp. 239–243
9. Unfolded basic recurrent neural network, [Online]. Available: https://en.wikipedia.org/wiki/Recurrent_neural_network#/media/File:Recurrent_neural_network_unfold.svg
10. J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014)
11. S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning – Based text classification: A comprehensive review, in *ACM Computing Surveys (CSUR)* (2021)
12. S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (O'Reilly Media, Inc, Beijing, 2009)

13. Word embedding, [Online]. Available: https://en.wikipedia.org/wiki/Word_embedding
14. Naive Bayes Classifier, [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier
15. SVM, [Online]. Available: https://en.wikipedia.org/wiki/Support-vector_machine
16. D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors. *Nature* **323**(6088) (1986)
17. P.J. Werbos, Backpropagation through time: What it does and how to do it. *Proc IEEE* **78**(10) (1990)
18. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**(8) (1997)
19. LSTM, [Online]. Available: https://en.wikipedia.org/wiki/Long_short-term_memory#/media/File:The_LSTM_Cell.svg
20. K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation (2014)
21. Bidirectional RNNs, [Online]. Available: https://en.wikipedia.org/wiki/Bidirectional_recurrent_neural_networks
22. R. Pascanu, C. Gulcehre, K. Cho, Y. Bengio, How to construct deep recurrent neural networks (2013)
23. D.P. Kingma, J. Ba, Adam: A method for stochastic optimization (2014)
24. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting. **15**(1) (2014)

Rohit Gund, James Andro-Vasko, Doina Bein, and Wolfgang Bein

Abstract

The objective of this paper is to the use of the Probabilistic Matrix Factorization (PMF) model which scales linearly with the number of observations and performs well on the large, sparse, and imbalanced music/movie dataset. In this project, we compare various PMF-based models and apply them to the recommendation system. We design and develop a Mix Probabilistic Matrix Factorization (MixPMF) model for music recommendation. This new model will take advantage of user network mapping and artist tag information and forms the effective rating matrix and thus will be efficient in recommending music/movies to new users. Simulation results show the advantage of our model.

Keywords

CPMF · KPMF · Machine learning · ML · Probabilistic Matrix Factorization · PMF · Rating matrix · Music recommendation system · Music streaming service · Movie recommendation system · Video streaming service

R. Gund · D. Bein (✉)
Department of Computer Science, California State University,
Fullerton, Fullerton, USA
e-mail: rohitgund@csu.fullerton.edu; dbein@fullerton.edu

J. Andro-Vasko · W. Bein
Department of Computer Science, University of Nevada, Las Vegas,
Las Vegas, USA
e-mail: androvas@unlv.nevada.edu; wolfgang.bein@unlv.edu

32.1 Introduction

Recommendation System in Machine Learning (ML) context is a class of learning algorithms that offers relevant suggestions to the users. It is useful in many different applications like for music/movie recommendations. Existing approaches for recommendation systems like collaborative filtering cannot handle large datasets and cannot deal with music or movies which have very few ratings.

There are over two hundred video streaming services and more than ten music streaming services available in the USA. Out of those hundreds of services, only a handful can recommend good content to its users. This makes the recommendation system extremely essential these days.

A perfect recommendation system is one that can recommend the best possible options as per the users. With conventional recommendation systems like collaborative filtering, the main problem is they cannot linearly scale with data and often lack accuracy with sparse and imbalanced data. This proposed recommendation system is made using Probabilistic Matrix Factorization (PMF) [1, 2] and can deal with sparse and imbalanced data effectively. Existing available methods for recommendation systems include *collaborative filtering* and *content-based filtering*. In collaborative filtering, If a person ABC has the same opinion as a person XYZ on a topic, ABC is more likely to have XYZ's opinion on a different topic than that of a randomly chosen person.[3]. Content-driven filtering approaches are based on an object definition and a consumer expectations profile. These methods are best suited to situations where information about an item is known (name, location, description, etc.), but not about the user. Content-based recommendations treat recommendations as a user-specific classification problem and learn a classifier based on the functions of an item for the user's likes and dislikes. [4] These methods do not work well when data starts

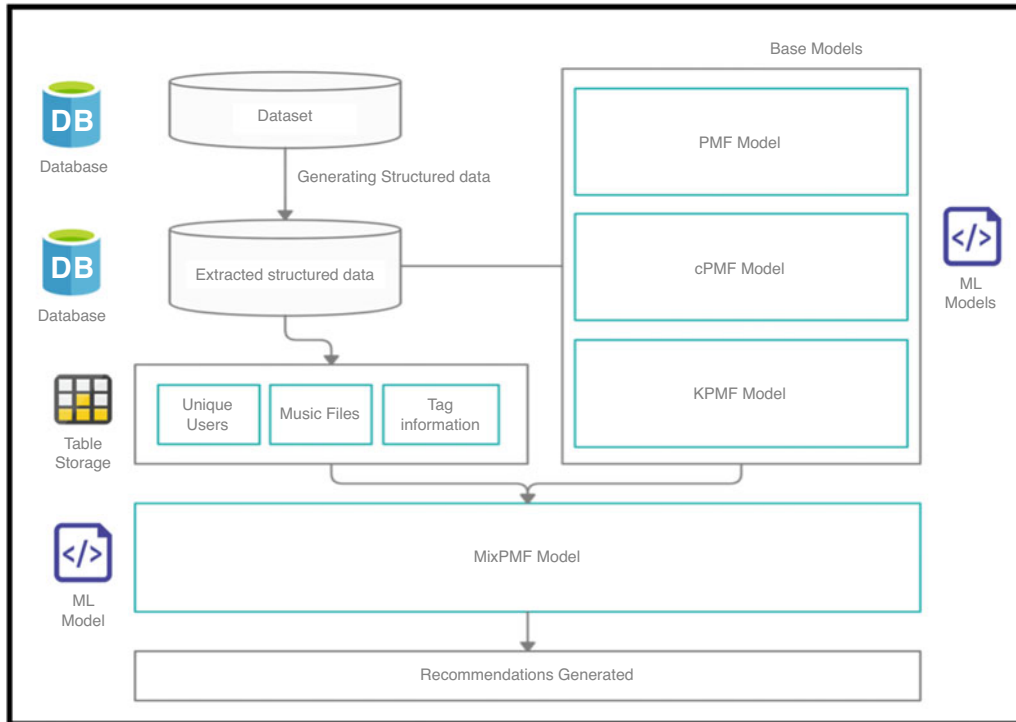


Fig. 32.1 Architecture of the MixPMF recommendation system

scaling linearly. To meet this data-intensive demand PMF was introduced by Mnih in 2008 [5].

Conventional recommendation systems built using Collaborative Filtering are effective, but when it comes to imbalanced data, it gets very challenging and difficult to produce good results. To engage the users with music/movies streaming service, it is very important to recommend the content they might like and simultaneously deal with the content which has no ratings and new users onboarding on the platform. A Mix Probabilistic Matrix Factorization (MixPMF) model can deal with this problem and can generate effective music/movie recommendations.

After studying the existing PMF models like PMF, constrained PMF and kernelized PMF and understanding their results, we can conclude that the additions of user network and artist tag information introduced in KPMF help us to improve the recommendation. Also, the constraint matrix introduced in CPMF helped to form a better rating matrix. We will combine these two approaches in our MixPMF model to recommend better music. The architecture of our proposed system is shown in Fig. 32.1 and has three main components:

1. Dataset “hetrec-2011” [6],
2. PMF, KPMF and CPMF models are the base models for understanding effectiveness of PMF in the recommendation system, and
3. MixPMF Model that can generate recommendations.

The paper is organized as follows. In Sect. 32.2 we present the background work, followed by our proposed model in Sect. 32.3, and model training and evaluation in Sect. 32.4. Concluding remarks and future work are given in Sect. 32.5.

32.2 Background Work

We studied three models:

- Probabilistic Matrix Factorization (PMF)
- Constrained Probabilistic Matrix Factorization (CPMF)
- Kernelized Probabilistic Matrix Factorization (KPMF)

The notations used in models are explained below:

N = Number of users

M = Number of items

$R \in R^{N \times M}$ = Rating Matrix

$U \in R^{N \times D}$ = User Latent Matrix

$V \in R^{M \times D}$ = Items Latent Matrix

D = Latent Dimension

W = Similarity Matrix

Y = Effect of artist ratings

$K_u \in R^{N \times M}$ = Covariance matrix for rows of R

$K_v \in R^{M \times M}$ = Covariance matrix for columns of R

$S_u \in R^{M \times N}$ = Inverse of K_u

$S_v \in R^{M \times M}$ = Inverse of K_v

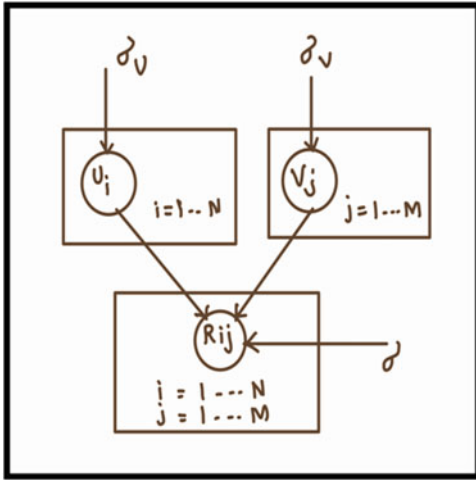


Fig. 32.2 PMF diagram

$I \in R^{N \times M}$ = Indicator matrix taking 1 if $R_{i,j}$ is an observed entry, 0 otherwise.

α = learning rate

$N(\mu, \sigma^2)$ = Normally distributed with mean μ and variance σ^2 . The notation $X \sim N(\mu, \sigma^2)$ means that X is distributed $N(\mu, \sigma^2)$

The Probabilistic Matrix Factorization (PMF) was introduced by Minh and Salakhutdinov in 2008 [5] and has shown flexibility and effectiveness in terms of large, sparse, and imbalanced data. PMF is a generative algorithm (see Fig. 32.2), and it assumes the following generative process for Rating Matrix R as follows:

1. Generate $U_i \sim N(0, \sigma_U^2 I)$ for $i \in \{1, 2, 3, \dots, N\}$
2. Generate $V_j \sim N(0, \sigma_V^2 I)$ for $j \in \{1, 2, 3, \dots, M\}$
3. For each non-missing entry of $R_{i,j}$, generate $R_{i,j} \sim N(U_i \cdot V_j^T, \sigma^2 I)$

Our ratings are logarithms of the user’s music listening time or video viewing time. So, we will use dot products of U and V directly. The log-posterior over the latent matrices U and V is given by following formula:

$$\log p(U, V | R, \sigma^2, \sigma_U^2, \sigma_V^2) = \log \log p(U, V, \sigma^2) + \log \log p(U, V, \sigma_U^2) + \log \log p(V | \sigma_V^2) + C$$

where C is constant and does not depend on U and V.

MAP estimate will be performed to learn U and V which maximizes by stochastic gradient descent. For each non-missing, update rule is as follow:

$$err_{i,j} = R_{i,j} - U_i V_j^T$$

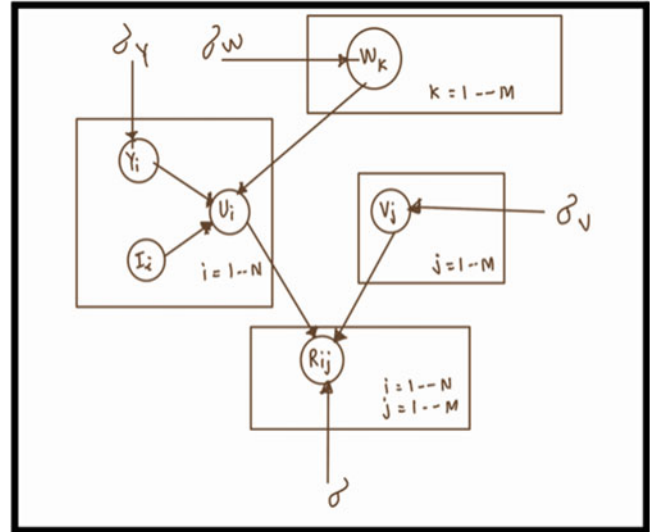


Fig. 32.3 CPMF diagram

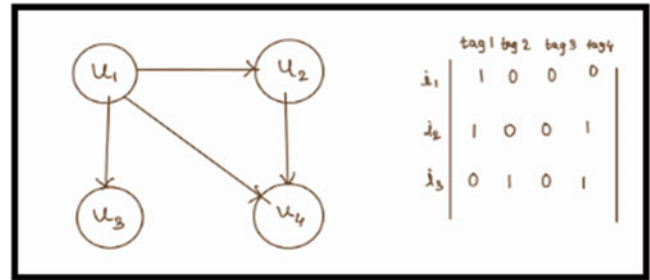


Fig. 32.4 Additional input for KPMF model

$$U_i = U_i + \alpha \left(err_{i,j} V_j - \frac{\sigma^2}{\sigma_U^2 \sum_{p=1}^N I_{i,p}} U_i \right)$$

$$V_j = V_j + \alpha \left(err_{i,j} U_i - \frac{\sigma^2}{\sigma_V^2 \sum_{p=1}^N I_{p,i}} V_j \right)$$

Mnih [5] enhanced the performance by trying to find other users who listened to the same artists as user ‘i’ and constrain the latent vector of similar users to the model and named it as CPMF (see Fig. 32.3), which contains U with latent similarity constraint matrix $W \in R^{M \times D}$ [7].

KPMF models were introduced by Zhou in 2012 [8], these models allow U and V to capture the covariances between any two rows of U and V by assuming the columns of U and V are generated from zero mean Gaussian Process (GP). In KPMF models, we will use two additional inputs: user network and artist tag information (see Fig. 32.4).

The generative process for KPMF model (see Fig. 32.5) is as follow:

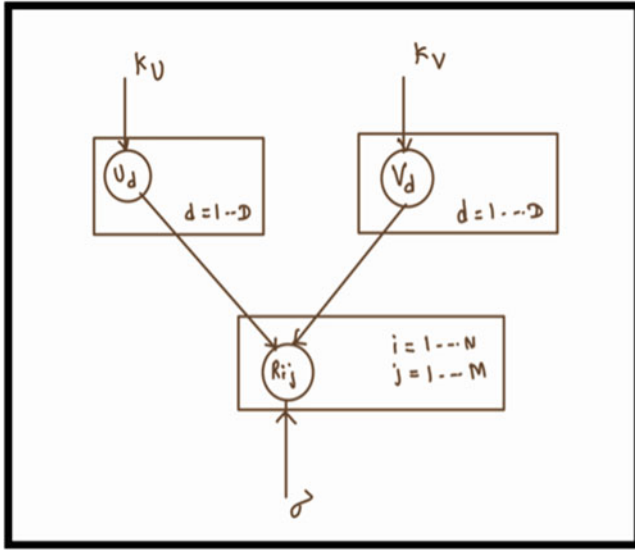


Fig. 32.5 KPMF diagram

1. Generate $U_d \sim GP(0, K_u)$ for $d \in \{1, 2, 3, \dots, D\}$
2. Generate $V_d \sim GP(0, K_v)$ for $d \in \{1, 2, 3, \dots, D\}$
3. For each non-missing entry of $R_{i,j}$, generate $R_{i,j} \sim N(U_i, V_j^T, \sigma^2)$

32.3 Our Proposed MixPMF Model

We used CPMF and KPMF models for building our own variation of the PMF model and calling it the MixPMF model. The KPMF model introduced the user network information which proved to be very effective. We used the constrained vector reference from the CPMF model and combining methods of CPMF and KPMF.

The steps of our model are as follows:

1. Generate $W_k \sim N(0, \sigma_W^2 I)$ for $k \in \{1, 2, 3, \dots, M\}$
2. Generate $Y_i \sim N(0, \sigma_Y^2 I)$ for $i \in \{1, 2, 3, \dots, N\}$
3. Generate $V_d \sim GP(0, K_v)$ for $d \in \{1, 2, 3, \dots, D\}$ [9]
4. Generate indicator matrix I such that $I_{i,j} = 1$ if $R_{i,j}$ is observed or $I_{i,j} = 0$ otherwise
5. For each non-missing entry of $R_{i,j}$, generate

$$R_{i,j} \sim N\left(\left(Y_i + \frac{\sum_{k=1}^M I_{i,k} W_k}{\sum_{k=1}^M I_{i,k}}\right) V_j^T, \sigma^2\right)$$

Variable Description:

V_d = Latent Dimension Vector
 W = Similarity Matrix
 Y = Effect of artist ratings Matrix

For implementing our MixPMF model, we used NumPy and Pandas to perform the mathematical computation for data and

Table 32.1 Stats of dataset

#	Item	Stats
1	User	1800
2	Artists	1000
3	Ratings	55000
4	Rating Density	3.02%
5	tags	85000

Sci-kit is used to implement the machine learning part. The proposed MixPMF model is shown in Fig. 32.6.

We used the following files:

- MixPMF.ipynb to compute the respective matrices
- Algobase function from Algobase.ipynb inherited in the cKPMF function in MixPMF.ipynb file
- `_compute_metrics(self, x)` will calculate the rmse values after passing the Numpy array
- `predict_pair()` will return the model rating prediction for a given user and item pair; the input to these function is the structured music dataset.

These functions are then externally accessed from experiment files and predicted results are stored into text file for easy reading and understanding.

32.4 Model Training and Evaluation

We used the hetrec-2011 [6] dataset which is sparse and imbalanced to prove the effectiveness of the MixPMF model. We executed PrepData.ipynb file to generate the data files suitable for this project on the .dat file from hetrec-2011 [6]. These data files give clear information about users, items, artists, ratings and rating density, etc.. We observed a total 55000 ratings with 1800 unique users, 1000 unique artists and rating density is 3.02% while the total tags remain at 85000. Stat table, generated using the PrepData.pynb file, and shown in Table 32.1, note that the distribution of 55000 ratings are highly distorted and only 1300 ratings are larger than 5000.

Due to this we opted for logarithmic count and Fig. 32.7 shows that rating distribution before and after log transformation. We note that in Fig. 32.7, only a few artists in the dataset are rated and almost 80% artists never received any ratings from users.

We note that only handful users rate to music (see Fig. 32.8).

The most used method for model evaluation is RMSE (root-mean-square error). This is defined as the square root of the average squared distance between the actual score and the predicted score:

Fig. 32.6 The proposed MixPMF model picture representation

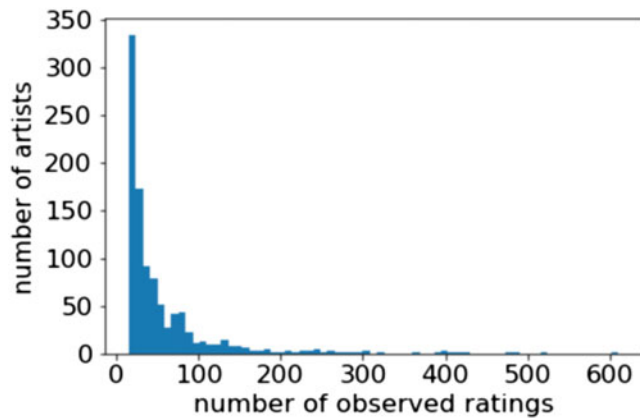
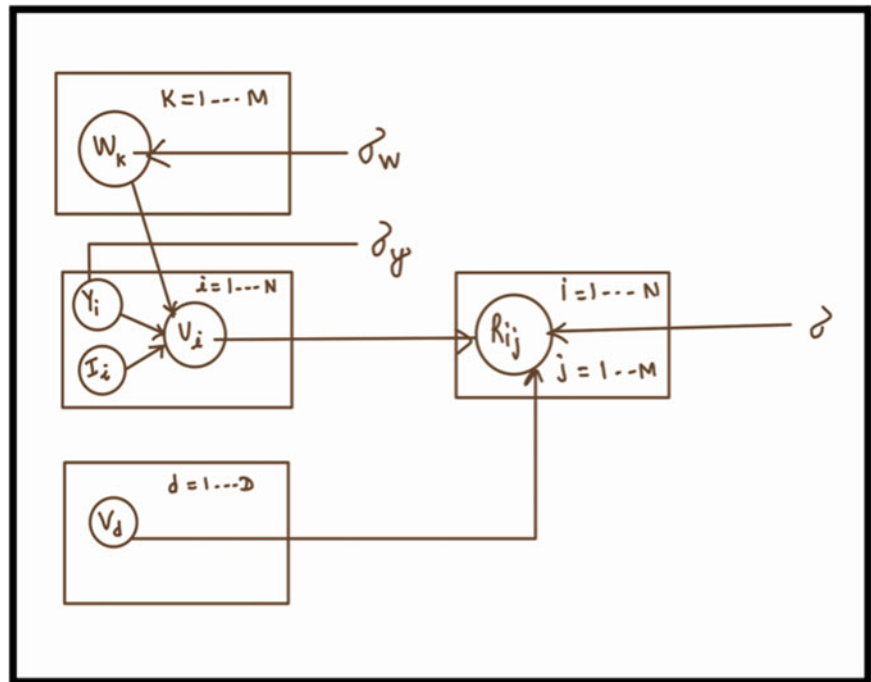


Fig. 32.7 Frequency Distribution for Artists

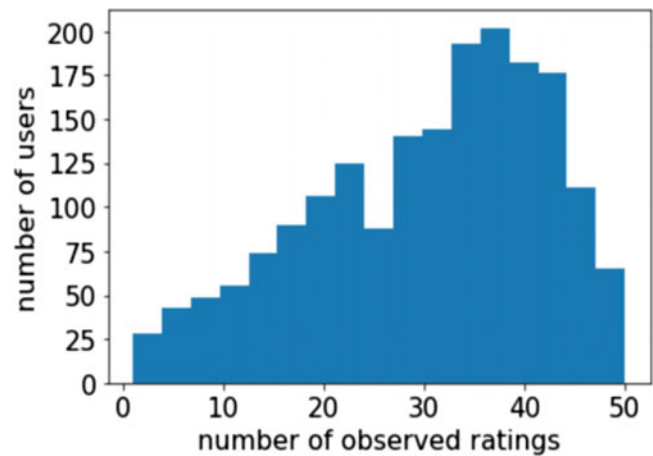


Fig. 32.8 Frequency distribution for users

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

where y_i denotes the true score for the i th data point and \hat{y}_i denotes the predicted value. This is the Euclidean distance between the vector of the true scores and the vector of the predicted scores, averaged by n , where n is the number of data points [10].

The models will inherit the PMF properties from base.ipynb file and the function AlgoBase passed as a parameter to KPMF, CPMF and MixPMF models. At the end of each iteration, the model will generate the predicted and true values using the sci-kit learn libraries inbuilt method [11]. We use an array of these values and passing

Table 32.2 RMSE value for different models

Model	On 20% dataset	On 50% dataset	On 100% dataset
PMF	2.73	2.57	2.8
CPMF	1.66	1.36	1.32
KPMF	2.01	1.98	1.99
MixPMF	0.88	0.83	0.87

it as a Numpy array to calculate RMSE values. We refer to these values in model evaluation. The final calculated RMSE is shown in Table 32.2.

As per the model evaluation method RMSE, we can conclude that our MixPMF model is efficient compared to PMF, CPMF and KPMF models. MixPMF model is working

with RMSE of 0.83 and 0.87 on 50% and 100% of the dataset respectively.

32.5 Conclusion and Future Work

We found that combining the properties of CPMF and KPMF recommendation generations is efficient and thus we used these methods to proposed our MixPMF model. The MixPMF model is evaluated and tested with well-defined test cases. RMSE of all the models compared on the same music dataset and MixPMF model stand most reliable and efficient for both partial data from dataset and entire data from dataset. Machine learning model MixPMF is capable of recommending music efficiently than PMF, CPMF and KPMF models.

As future work, developing a GUI with an automatic playlist is a future enhancement. In our project, we focused only on the back-end development. Songs generated using the MixPMF model can be analyzed for a few days and based on results they will be stored as an automatic playlist for a particular user. We also propose to investigate using a parallel processing and multithreading approach which can scale up these models to industry level applications and handles millions of files per second.

Acknowledgment This research was sponsored by the NATO Science for Peace and Security Programme under grant SPS MYP G5700.

References

1. Hui Fang, Yang Bao, Jie Zhang . Leveraging Decomposed Trust in Probabilistic Matrix Factorization for Effective Recommendation, https://people.eecs.berkeley.edu/tinghuiz/papers/sdm12_kpmf.pdf
2. PMF for Recommender Systems: <https://towardsdatascience.com/pmf-for-recommender-systems-cbaf20f102f0>
3. Collaborative Filtering Recommender Systems: <https://www.nowpublishers.com/article/Details/HCI-009>
4. Recommendation System Wiki: https://en.wikipedia.org/wiki/Recommender_system
5. Probabilistic Matrix Factorization – <https://papers.nips.cc/paper/2007/hash/d7322ed717dedf1eb4e6e52a37ea7bcd-Abstract.html>. Accessed on 20 May 2021
6. Music Dataset <https://groupLens.org/datasets/hetrec-2011/>
7. Sources and temporal variations of constrained PMF factors obtained from multiple-year receptor modeling of ambient PM2.5 data from five speciation sites in Ontario, Canada: <https://www.sciencedirect.com/science/article/abs/pii/S1352231015001806>
8. Kernelized Probabilistic Matrix Factorization: Exploiting Graphs and Side Information https://people.eecs.berkeley.edu/tinghuiz/papers/sdm12_kpmf.pdf. Accessed on 21 July 2020
9. A Visual Exploration of Gaussian Processes: <https://distill.pub/2019/visual-exploration-gaussian-processes/>
10. RMSE: <https://vijay-choubey.medium.com/how-to-evaluate-the-performance-of-a-machine-learning-model-d12ce920c365>
11. Sci-Kit RMSE: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

Aditya Dingare, Doina Bein, Wolfgang Bein, and Abhishek Verma

Abstract

Text summarization creates a brief and succinct summary of the original text. The summarized text highlights the main text's most interesting points without omitting crucial details. There is a plethora of applications on the market that include news summaries, such as Inshort and Blinklist which not only save time but also effort. The method of manually summarizing a text can be time-consuming. Fortunately, using algorithms, the mechanism can be automated. We apply three text summarization algorithms on the Amazon Product Review dataset from Kaggle: extractive text summarization using NLTK, extractive text summarization using TextRank, and abstractive text summarization using Seq-to-Seq. We present the advantages and disadvantages for these three methods.

Keywords

Machine learning · ML · Kaggle · extractive text summarization · abstractive text summarization · text summarization · TextRank · Seq-to-Seq · NLTK · document summary

A. Dingare · D. Bein (✉)
Department of Computer Science, California State University,
Fullerton, Fullerton, CA, USA
e-mail: adityadingare@csu.fullerton.edu; dbein@fullerton.edu

W. Bein
Department of Computer Science, University of Nevada, Las Vegas,
Las Vegas, USA
e-mail: wolfgang.bein@unlv.edu

A. Verma
Department of Computer Science, California State University,
Northridge, Northridge, CA, USA
e-mail: abhishek.verma@csun.edu

33.1 Introduction

There are various forms of summaries: single document, multi document, informative summary, and query focused summary [1–3, 5, 7, 8]. The type of input provided to an algorithm determines these types, so for a multi-document summary, multiple documents are used. The input for query based is focused on a specific query outcome. There are two output-based primary methods for summarizing the text: abstractive text summarization and extractive text summarization [4–6, 10, 12, 13].

In extractive text summarization, the summarized text is part of the original text as the algorithm extracts the most relevant words and sentences from the original text. For example, in Fig. 33.1 the output text is consisting of all the words from the original input text only.

Abstractive text summarization is opposite to extractive text summarization [7] as it returns the summary of the text that may consists of new word and sentence that are not part of the original text. For example, in Fig. 33.2, the output of the abstractive summarization consists of the words that are not part of the original text. Hybrid text summarization uses both abstractive and extractive text summarization techniques together [8].

While abstractive text summarization produces more substantive summarized text than extractive text summarization, it is more difficult to implement. Most of the research focuses on extractive text summarization's implementation and limitations [9–11, 14, 15, 18].

We apply three text summarization algorithms on the Amazon Product Review dataset from Kaggle [12]: extractive text summarization using NLTK, extractive text summarization using TextRank, and abstractive text summarization using Seq-to-Seq. We compare their performances over various product reviews.

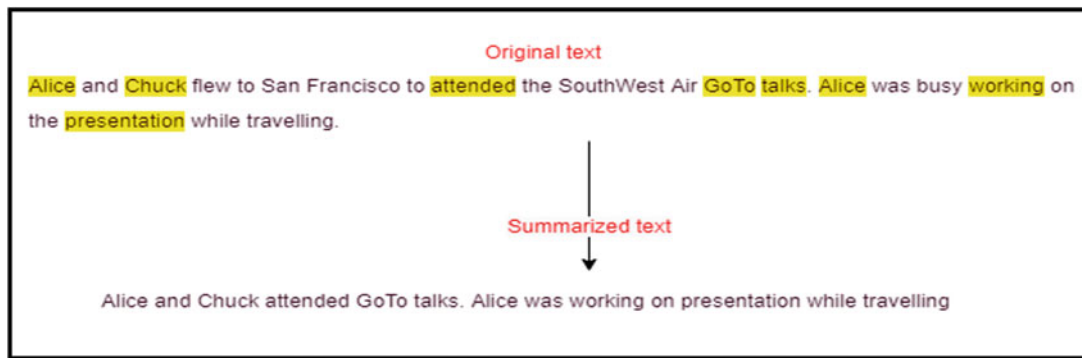


Fig. 33.1 Extractive text summarization

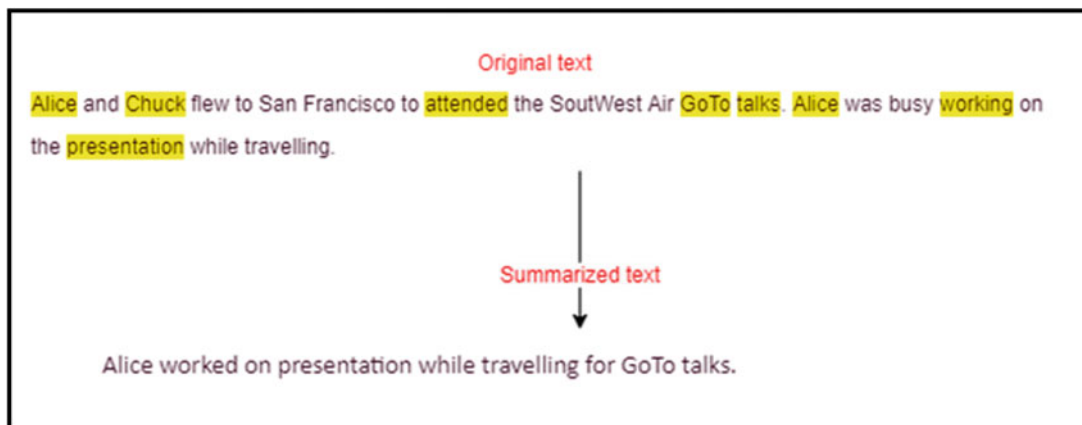


Fig. 33.2 Abstractive text summarization

The paper is organized as follows. In Sect. 33.2 we present the background and related work, followed by the description of the algorithms used in Sect. 33.3. Simulations are results are presented in Sect. 33.4. Concluding remarks and future work are given in Sect. 33.5.

33.2 Background and Related Work

Term frequency, latent frequency, and graphical extractive algorithms are the three primary types of extractive algorithms [13, 20]. The sentence that has a similar appearance to the document word has a high score in terms of frequency. The sentences are sorted first in latent variable, and the sentence with the closest representation of the latent variable is chosen [14]. A similarity matrix is constructed in the graphical process, and the TextRank algorithm is executed based on it.

The extractive text summarization can be divided into three main categories (see Fig. 33.3): term frequency, latent variable, and graphical. Term frequency and sum basic algorithm are similar: commonly occurring sentences are added together [15]. Latent variable algorithm works like as discussed above. In embedding page rank algorithm, the

embedding page vector is calculated and used during the algorithm execution [16].

Abstractive text summarization can be divided into two categories (see Fig. 33.4): semantics based, and structure based. The implementation discussed in this project is from the semantics graph-based technique.

We use Text Rank algorithm developed by Mihalcea [17]. The TextRank algorithm is like PageRank PageRank algorithm that was designed and developed by Google, but instead of web pages, it uses sentences. The similarity between two sentences is the likelihood of a web page switch. This score of similarity is stored in a square matrix [18]. The standard steps in the TextRank algorithm are to load input data and construct vectors for sentences using GloVe word embeddings. The next stage is text preprocessing, which involves cleaning the data and removing common terms such as am, an, the, for, and so on. We then construct a vector representation of sentences and a similarity matrix and next we apply TextRank algorithm.

The abstractive text summarization technique using Sequence-to-Sequence modeling (Seq2Seq model) is used to summarize the text. The standard implementation involves usage of encoder and decoder as referenced. The encoder and decoder are configured into two phases training and inference phase. The encoder reads input data and extracts the

Fig. 33.3 Algorithms to implement extractive text summarization

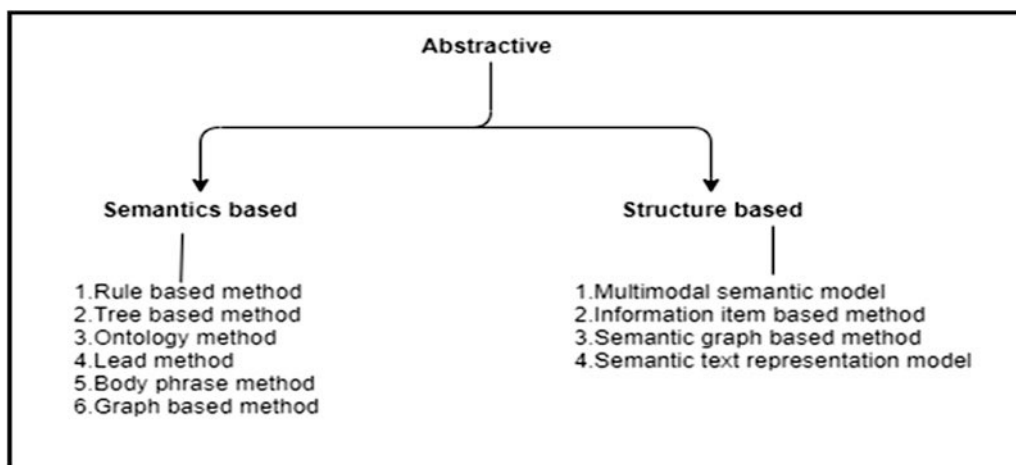
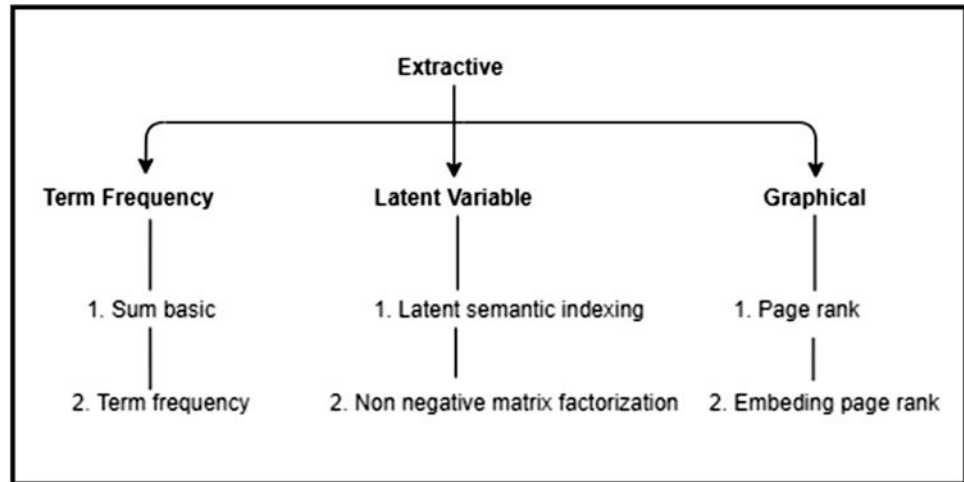


Fig. 33.4 Algorithms to implement abstractive text summarization

contextual information present in the input sequence using the Long Short-Term Memory model (LSTM). The decoder, on the other hand, uses the encoder's output as an input and is equipped to predict the next word in the series [19]. The input data used by TextRank algorithm is a single document and does not support the use of RNNs and LSTM. It is difficult for encoder to memorize the huge data size as the fixed length vector is used to store the input data. In addition, the encoder uses a unidirectional LSTM. The context cannot be captured in both directions using a unidirectional LSTM. The bidirectional LSTM, combined with global attention for the previous problem, can be used to address the LSTM issue [20, 21].

33.3 Implemented Text Summarization Algorithms

The main steps in implementation are data gathering, data cleaning, and algorithm implementation. We implement three algorithms and compare their results.

For data gathering we used Amazon Product Review dataset from Kaggle [12] that has approximately 568,455 rows and 10 columns, almost 300 MB in size. Out of 10 columns, the ProductReviewText column has detailed description of the product available on the Amazon and is mainly used by the extractive text summarization for the TextRank algorithm.

Data cleaning involves contraction mapping for handling the words with short forms like "ain't", "don't" as "do not" etc., and changing the input data into either lower or upper case, remove parenthesis, eliminate stops words (e.g. "is", "and", "are"), punctuations, and special characters like @, # etc. These two steps are common to both abstractive and extractive algorithms [22].

The abstractive text summarization uses two columns mainly ProductReviewHeader and ProductReviewText. The column ProductReviewHeader is nothing but the header line of that particular review. This column is either one or two lines of short headline of the review.

33.3.1 Extractive Text Summarization Using NLTK

We used the Natural Language Toolkit (NLTK) library for statistical language processing which include tokenization, calculating frequency of words, and calculating weighted frequency of words. Term frequency can be used to identify the keywords. Extraction of keywords in reviews enable customers in determining whether a product review is necessary and whether or not to continue reading it. Following the calculation of word frequency, a weighted frequency of each sentence in the input data column, ProductReviewText, is calculated. We calculated the frequency and weighted frequency of each word that is present in the review text. The weighted frequency can be calculated in other way by dividing the words frequency by frequency of the word that is mostly occurred [23]. The next step is to add weighted frequencies and sort the sum of weighted frequencies in descending order. The sentence with the maximum sum of weighted frequency is extracted as the summarized text. Based on the weighted frequency, the summary of the original text is returned.

33.3.2 Extractive Text Summarization Using TextRank

The TextRank algorithm depends on PageRank algorithm. The probability among two words in the sentences is calculated. For the TextRank algorithm, the input reviews data will be subdivided into text units such as keywords, key phrases and the graph model is built. In this implementation, we used an undirected weighted graph; each node represents the sentence in the review text and the edges represents the relationship between them calculated using the formula [24]:

$$WS(V_i) = (1 - d) + d * \sum_{V_j \in \text{In}(V_i)} e_{ji} / \sum_{V_j \in \text{Out}(V_i)} e_{jk} \quad WS(V_{ij})$$

Each sentence is treated as a node in the text. There is an undirected right edge between the nodes corresponding to the two sentences if two sentences are identical. The following formula can be used to check sentence similarity [24].

$$\text{Similarity}(S_i, S_j) = \frac{|\{W_k | W_k \in S_i \& W_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)}$$

where S_i and S_j are two sentences of our product review where as W_k represents word in the sentence.

The steps are:

1. Split the given text review T into complete sentences
2. Clean the data by deleting stop words, nouns, and verbs from the input data for each sentence

3. Build a candidate keyword graph $G = (V, E)$, where V is a node collection of sentences, and then draw an edge between any two points if and only if these two sentences are linked.
4. Calculate the weight repetitively using the formula mentioned below.
5. The node weights are sorted in reverse order to obtain the most relevant T terms as candidate keywords.
6. The most significant T words are extracted from #5, marked in the original text, and then combined into a keyword if adjacent phrases are created. [25]

The product reviews in column ProductReviewText is divided into small chunks of sentences only if it contains long sentences.

33.3.3 Abstractive Text Summarization Using Seq-to-Seq

The encoder and decoder are needed for extractive text summarization using the Seq-to-Seq model. In this implementation, the Long Short-Term Memory (LSTM) is used as encoders and decoders to catch the phrase dependencies in a sentence's words. To implement encode and decoder," Recurrent Neural Networks i.e. RNN" can also be used. The encoders and decoders are designed further in two stages, namely training and inference. The encoder receives the input data as input and extracts the contextual information present in the data. The timestamp is also important factor here. So, for each time stamp, each word from the product review sentence is given to the encoder which retrieves the contextual information from the product review. This is the part of the training phase. Before feeding the target sequence into the decoder, their start and end are inserted [23].

The encoder training phase is single direction (see Fig. 33.5).

The encoder receives the input word by word at each time interval and each steps of the LSTM passes the output to the next cell in the LSTM.

The decoder receives the hidden state (h_i) and cell state (c_i) from the previous steps as data. The decoder training is single direction (see Fig. 33.6) This decoder cell also takes input from the encoder and process each word.

There are total of three cells of each encoder and decoder.

To calculate the past as well as future context for each product review sequence we used a bi-directional LSTM.

The bidirectional LSTM in which the original cells as well as it's transpose are used. The single-directional LSTM has the disadvantage of being unable to predict future background information and learning all the input sequence sentences. As referenced in the Fig. 33.7, in bi-directional LSTM the output of the last cell is given back to additional LSTM.

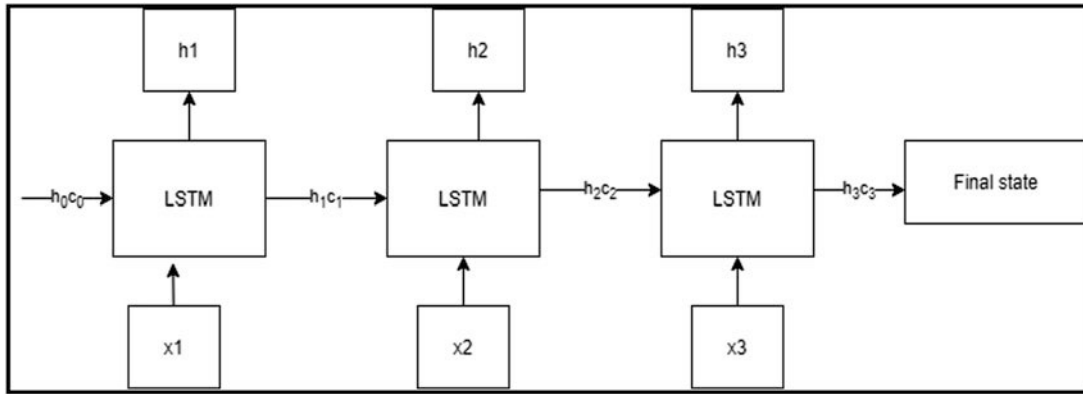


Fig. 33.5 Training encoder

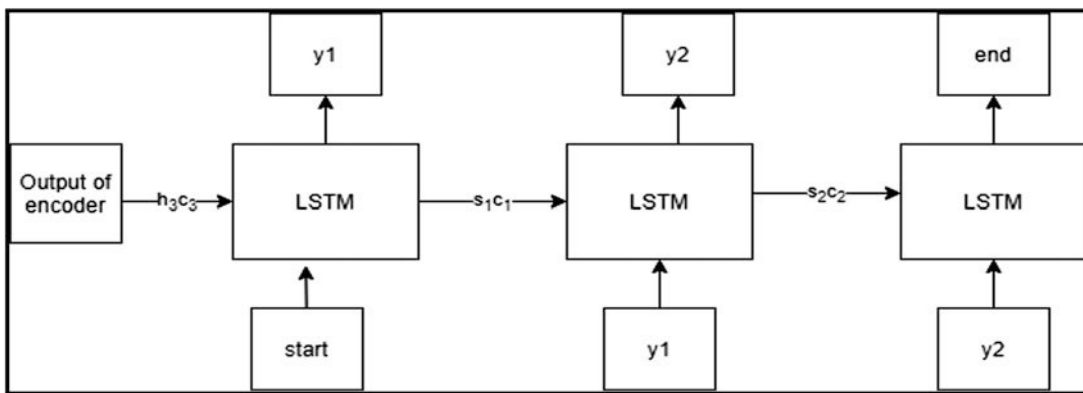


Fig. 33.6 Training decoder [7]

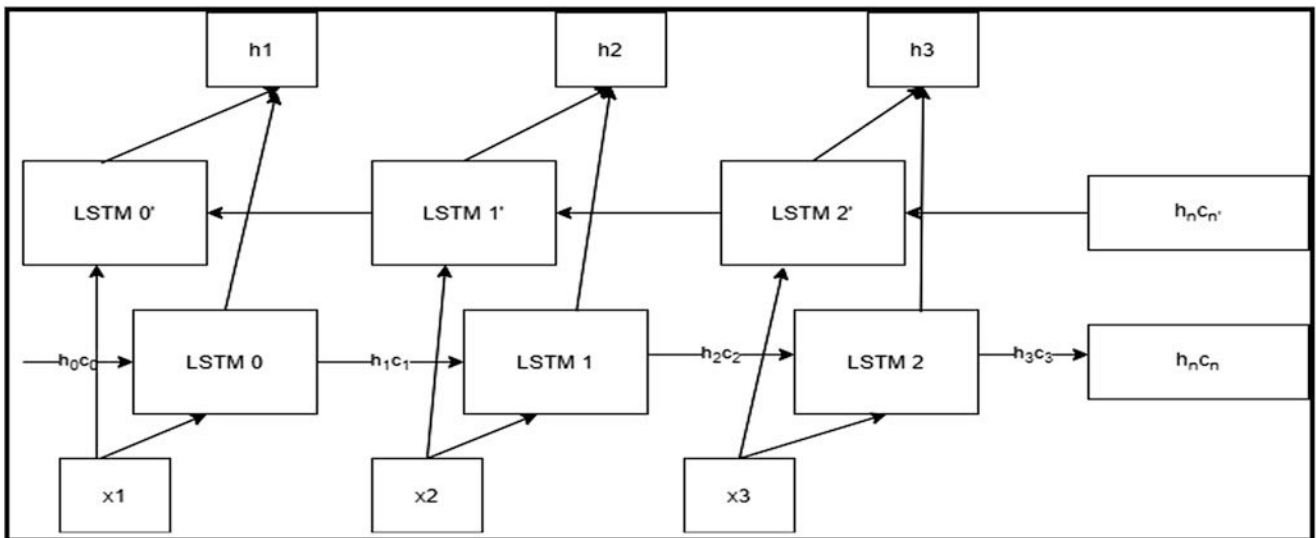


Fig. 33.7 Bi-directional LSTM

```
* It's even better than the organic, all-natural brands I have tried.
* No Coffee Shop has a better one and I like most of the other products, too (as a usually non-coffee drinker!
* The first and second cracks are distinct, and I've roasted the beans from medium to slightly dark with great results every time.
* I've been consuming various sports nutrition products for decades, so I'm familiar and have come to like the taste of the most of the products I've tried.
* I put this food on the floor for the chubby guy, and the protein-rich, no by-product food up higher where only my skinny boy can jump.
* I have switched them to a different food (due to price) a couple of times and end up going right back to natural balance.
* It requires a beverage as advertised, a glass of very cold milk, and a box of Kleenex since it will make your nose run.
* Also good for small puppies.
*German translation of the summary*
* Wir mögen den stahlgelbeschnittenen Hafer von McCann sehr, stellen aber fest, dass wir ihn nicht zu oft kochen. <br /> Das schmeckt mir viel besser als die Marken von Lebensmittelgeschäften und ist genauso praktisch. <br /> Alles, was mich zum Essen bringt Haferflocken regelmäßig ist eine gute Sache.
* Der Geschmack war erstaunlich und während ich auf das Etikett schaute und mich fragte, was diesen leckeren, neuen zuckerfreien Leckerbissen so gut schmecken lassen könnte, sank mein Herz, als ich dem Kleinen Sternchen neben dem zuckerfreien Süßstoff * bis zum Ende des Etiketts folgte und Lesen Sie "Maltitol" in kleinen Buchstaben!
* Dieses Produkt dient mir als Elektrolytquelle während und nach einem langen Lauf oder einer Radtour. <br /> Ich habe alle Geschmacksrichtungen ausprobiert, mag aber wirklich den Grapefruitgeschmack ... kein Nachgeschmack und ich mag den leichte Kohlensäure. <br /> Ich verwende andere Hammer-Produkte und mag deren gesamte Produktlinie sehr.
* Außerdem mag ich es, die Aromen jedes Mal zu mischen, da ich denke, dass die gleiche Mahlzeit Tag für Tag etwas langweilig werden könnte, also dachte ich mir, warum nicht
```

Fig. 33.8 NLTK output

LSTM0', LSTM1', and LSTM2' are all the transpose of original LSTM0, LSTM1, and LSTM2 [23].

33.4 Simulations and Results

For both the abstractive and extractive text summarization we considered reviews that have at least 250 words. The output of the extractive text summarization using NLTK applied to a sample product review and its German language translation of the summary (cross language summary) is shown in Fig. 33.8.

The output by the TextRank algorithm of the sample product review is shown in Fig. 33.9. There are approximately more than 5 reviews being summarized and their associated German translations.

Figure 33.10 shows the output of the Seq-to-Seq implementation with original text as well as summarized text.

The following limitations have been noted. The extractive text summarization implemented using TextRank algorithm does not return proper output for the duplicate words and sentences. The algorithm is modified to perform the multi document text summarization. It means the algorithm checks for the .CSV files in input directory and picks up all the file while processing. However, both the TextRank and Seq-to-Seq algorithms do not remember the input data from one source input file while processing another input file. In short, even though the multi document processing is supported, the inter dependency of input is not taken into consideration. Another drawback of our implementation is that the models we developed are unable to generate new product feedback that could be used in conjunction with summarizing the

subsequent input data. The main disadvantage of using Seq-to-Seq for abstractive text summarization is that sentences are not ranked like text summarization. This will cause us to skip over text that appears frequently in the input.

33.5 Conclusion and Future Work

In this paper we apply three text summarization algorithms on the Amazon Product Review dataset from Kaggle [12]: extractive text summarization using NLTK, extractive text summarization using TextRank, and abstractive text summarization using Seq-to-Seq. There are advantages and disadvantages to using these algorithms for product reviews summarization.

As future work, we note that the TextRank algorithm we used for extractive text summarization does not endorse Recurrent Neural Networks (RNN). The RNN is the most common algorithm for dealing with a continuous stream of data. Internal memory in the RNN aids in remembering the input and makes it ideal for deep learning problems. We could use RN to improve the current algorithm processing because the RNN remembers the import part of the input and uses it as a guide in subsequent runs. The RNN's performance summarized text looks more like text summarized by a person. However, the RNN has its own disadvantage – It fails for complex model. Its output is also poor if the input text includes duplicate words and sentences. For both the abstractive and extractive text summarization, using certain user parameters, the output of the summarized text may be further refined.

```

423 iteration
* We really like the McCann's steel cut oats but find we don't cook it up too often.<br />This tastes much better to me than the grocery store brands and is just as convenient.<br />Anything that keeps me eating oatmeal regularly is a good thing.
* The taste was amazing and while I was looking at the label wondering what could possibly make this yummy, new sugarfree treat taste so good, my heart sank when I followed the little asterisk next to sugarfree sweetener down to the very bottom of the label and read "maltitol" in tiny little letters!
* This product serves me well as a source of electrolytes during and after a long run or bike ride.<br />I have tried all of the flavors but really do like the grapefruit flavor... no after-taste and I actually like the slight carbonation.<br />I use other Hammer products and really like their whole product line.
* Furthermore, I do like mixing up the flavors each time as I think the same meal day over day might get a little boring, so I figured why not.
* I am going to try using some essential oils next and see if I can get a good chocolate/orange mix.<br /><br />All of the ingredients I mentioned are here online.
* There is no escaping the fact, however, that even the best instant oatmeal is nowhere near as good as even a store brand of oatmeal requiring stovetop preparation.
* I got this for my Mum who is not diabetic but needs to watch her sugar intake, and my father who simply chooses to limit unnecessary sugar intake - she's the one with the sweet tooth - they both LOVED these toffees, you would never guess that they're sugar-free and it's so great that you can eat them pretty much guilt free!
* That could just be me since I like my oatmeal really thick to add some milk on top of.
* On the other hand, I hate using cold milk or cream, because I like HOT coffee.<br /><br />I stumbled across this on Amazon one day and got the idea of making my own creamer.
* When you take this into account, you're actually getting more "bang for your buck" with the natural dog foods since you don't have to buy as much to last just as long as the normal dog foods... and a healthier, happier dog, to boot!
* Let it settle for a bit before opening the top though.<br /><br />This stuff tastes WAY better than the storebought creamers and it is fun to experiment and come up with your own flavors.
* In my 1300watt microwave the oatmeal cooks up in about one minute and twenty-seven seconds, so you should also watch that to get a handle on how much time and water to use.<br /> The only bad thing -- if you can consider it a bad thing -- about this offering is that you have to buy in lot so you'll end up with six ten-count boxes.
* There is another company that makes grape gummy bears that are a little bit better in my opinion, but these are well worth it for the price.
* I came home and to my surprise realized that I could save $20 each time I bought dog food if I just buy it off Amazon.<br /><br />All in all, I definitely recommend and give my stamp of approval to natural balance dog food.
* We bought it specifically for one of our dogs who has food allergies and it works great for him, no more hot spots or tummy problems.<br />I LOVE that it ships right to our door with free shipping.
* Despite the higher cost of natural dog foods, I find that he eats significantly less of the Natural Balance dog foods and still stays happy and full.
* ).<br />The little Dolche Gusto Machine is super easy to use and prepares a really good Coffee/Latte/Cappuccino/etc in less than a minute (if water is heated up).
* I found it to be a hardy meal, not too sweet, and great for folks like me (post-bariatric surgery) who need food that is palatable, easily digestible, with fiber but won't make you bloat.

```

Fig. 33.9 TextRank output

```

Review: bought several vitality canned dog food products found good quality product looks like stew processed meat smells better labrador finicky appreciates product better
Review: product arrived labeled jumbo salted peanuts peanuts actually small sized unsalted sure error vendor intended represent product jumbo
Review: confection around centuries light pillowy citrus gelatin nuts case filberts cut tiny squares liberally coated powdered sugar tiny mouthful heaven chewy flavorful highly recommend yummy treat familiar story lewis lion witch wardrobe treat seduces edmund selling brother sisters witch
Review: looking secret ingredient robitussin believe found got addition root beer extract ordered made cherry soda flavor medicinal
Review: great taffy great price wide assortment yummy taffy delivery quick taffy lover deal
Review: got wild hair taffy ordered five pound bag taffy enjoyable many flavors watermelon root beer melon peppermint grape etc complaint bit much red black licorice flavored pieces kids husband lasted two weeks would recommend brand taffy delightful treat
Review: saltwater taffy great flavors soft chewy candy individually wrapped well none candies stuck together happen expensive version fralinger would highly recommend candy served beach themed party everyone loved
Review: taffy good soft chewy flavors amazing would definitely recommend buying satisfying
Review: right mostly sprouting cats eat grass love rotate around wheatgrass rye
Review: healthy dog food good digestion also good small puppies dog eats required amount every feeding

```

Fig. 33.10 Seq-to-Seq output

Acknowledgment This research was sponsored by the NATO Science for Peace and Security Program under grant SPS MYP G5700.

References

1. <https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example/>
2. A Gentle Introduction to Text Summarization in Machine Learning. <https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/>
3. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, *Attention Is All You Need*. [online] arXiv.org (2021). Available at: <https://arxiv.org/abs/1706.03762>. Accessed 17 May 2021
4. Pranab Ghosh, *Six Unsupervised Extractive Text Summarization Techniques Side by Side* [Online]. Available: <https://pkghosh.wordpress.com/2019/06/27/six-unsupervised-extractive-text-summarization-techniques-side-by-side/>. Accessed 17 May 2021
5. J.N. Madhuri, R. Ganesh Kumar, Extractive text summarization using sentence ranking, in *2019 International Conference on Data Science and Communication (IconDSC)*, 2019, pp. 1–3. <https://doi.org/10.1109/IconDSC.2019.8817040>
6. C. Lakshmi Devasena, M. Hemalatha, Automatic text categorization and summarization using rule reduction, in *IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM –2012)*, 2012, pp. 594–598
7. <https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/>

8. S. Selvarani, S. Jebapriya, R. Smeeta Mary, Automatic identification and detection of altered fingerprints, in *2014 International Conference on Intelligent Computing Applications*, (2014), pp. 239–243
9. R. Mishra, V.K. Panchal, P. Kumar, Extractive text summarization – An effective approach to extract information from text, in *2019 International Conference on Contemporary Computing and Informatics (IC3I)*, 2019, pp. 252–255. <https://doi.org/10.1109/IC3I46837.2019.9055636>
10. S.M. Meena, M. P. Ramkumar, G.S.R. Emil Selvan, Text summarization using text frequency ranking sentence prediction, in *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 2020, pp. 1–5, <https://doi.org/10.1109/ICCCSP49186.2020.9315203>
11. X. You, Automatic summarization and keyword extraction from web page or text file, in *2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET)*, 2019, pp. 154–158. <https://doi.org/10.1109/CCET48361.2019.8989315>
12. Consumer Reviews of Amazon Products. <https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products>
13. Y. Chen, Q. Song, News text summarization method based on BART-TextRank model, in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2021, pp. 2005–2010. <https://doi.org/10.1109/IAEAC50856.2021.9390683>
14. <https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/>
15. Abstractive Text Summarization Using Transformers | by Rohan Jagtap | The Startup | Medium. <https://medium.com/swlh/abstractive-text-summarization-using-transformers-3e774cc42453>
16. D. Suleiman, A. Awajan, *Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges* (2021)
17. R. Mihalcea, P. Tarau, TextRank: Bringing order into texts, in *Proceedings of EMNLP 2004*, ed. by D. Lin, D. Wu, (Association for Computational Linguistics, Barcelona, Spain., 2004), pp. 404–411
18. N. Patel, N. Mangaokar, Abstractive vs extractive text summarization (output based approach) – A comparative study, in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 2020, pp. 1–6, doi: <https://doi.org/10.1109/INOCON50539.2020.9298416>
19. A. Rahman, F.M. Rafiq, R. Saha, R. Rafian, H. Arif, Bengali text summarization using TextRank, fuzzy C-means and aggregate scoring methods, in *2019 IEEE Region 10 Symposium (TENSymp)*, 2019, pp. 331–336. <https://doi.org/10.1109/TENSymp46218.2019.8971039>
20. Sequence to Sequence Learning with Neural Networks arXiv:1409.3215v3 [cs.CL] 14 Dec 2014
21. Get To The Point: Summarization with Pointer-Generator Networks. <https://arxiv.org/abs/1704.04368>
22. S.R. Manalu, Stop words in review summarization using TextRank, in *2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2017, pp. 846–849. <https://doi.org/10.1109/ECTICon.2017.8096371>
23. W. Jiang, Y. Zou, T. Zhao, Q. Zhang, Y. Ma, A hierarchical bidirectional LSTM sequence model for extractive text summarization in electric power systems, in *2020 13th International Symposium on Computational Intelligence and Design (ISCID)*, 2020, pp. 290–294. <https://doi.org/10.1109/ISCID51228.2020.00071>
24. T. Behere, A. Vaidya, A. Bihade, K. Shinde, P. Deshpande, S. Jahirabadkar, Text summarization and classification of conversation data between service chatbot and customer, in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 2020, pp. 833–838. <https://doi.org/10.1109/WorldS450073.2020.9210289>
25. M.R. Ramadhan, S.N. Endah, A.B.J. Mantau, Implementation of textRank algorithm in product review summarization, in *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, 2020, pp. 1–5. <https://doi.org/10.1109/ICICoS51170.2020.9299005>

Antonio Dantas, Leandro Diniz, Maurício Almeida, Ella Olsson, Peter Funk, Rickard Sohlberg, and Alexandre Ramos

Abstract

The search and identification of people lost in an emergency is a very important activity, it is carried out to assist in human lives in danger, and the unavailability of support technologies. Unmanned aerial vehicles (UAV) act directly in this activity for greater capacity in the coverage of the area in a shorter time of operation. A brief state-of-the-art survey is presented, and specific needs are raised for accuracy and practical application. In this joint work, low-processing image recognition alternatives for use in UAV will be explored to face challenges such as location in large areas, small targets, and orientation. Assessments are performed on real images using Inception, SSD, and Yolo. For tracking, MIL, KCF, and Boosting techniques are applied to empirically observe the results of test missions. Preliminary results show that the proposed method is viable for action in the search and rescue of people.

Keywords

UAV · Search · Rescue · Images · Recognition

A. Dantas · L. Diniz · M. Almeida · A. Ramos (✉)
Institute Mathematics Computing (IMC), Federal University of Itajubá (UNIFEI), Itajubá, Brazil
e-mail: antoniodantas@unifei.edu.br; d2019102840@unifei.edu.br; d2021101420@unifei.edu.br; ramos@unifei.edu.br

E. Olsson
Saab Group and IDT, Mälardalen University, Västerås, Sweden
e-mail: ella.olsson@saabgroup.com

P. Funk · R. Sohlberg
Artificial Intelligence Group, Mälardalen University, Västerås, Sweden
e-mail: peter.funk@mdh.se; rickard.sohlberg@mdh.se

34.1 Introduction

Capturing images of the earth's surface is increasingly gaining importance, particularly for the analysis of anomalies in the soil for various purposes, such as geospatial studies, fires, natural disasters, agriculture, etc. One of the most relevant scenarios is the search and identification of people who are lost or at risk, they are called search and rescue (SAR) activities, and today they are now commonly assisted by information technology [1].

Brazilian and Swedish researchers have been developing methods, techniques and tools for interpreting soil images obtained by sensors embedded in unmanned aerial vehicles (UAV), in order to help and complement the efforts of human researchers. Motivated mainly by the reduction of costs and efficiency, provided by the versatility in the process of photos captured in real time in strategic positions, in comparison with the high cost and sometimes of low efficiency, when performed in areas with few professionals or photos of satellites [2].

These systems of support for search and rescue activities, known as search and rescue unmanned aerial vehicles (SARUAV), have different obstacles to be faced in the current scenario. Adverse conditions stand out in certain regions, the area covered by the UAV camera and the change in lighting, are imposed for high-performance human detection [3]. Supporting information such as direction, speed and action can also help rescuers reach the victim more quickly and effectively [4].

In this work will present alternatives to face these challenges, for that, similar works will be presented in the recent literature, after an overview of the techniques that can be applied, then the methodology adopted in the study.

34.2 Related Works

Presented here are some recent works related to SARUAV, presenting its strengths, as well as the improvements recommended by the authors. In [4–7] the focus of the studies is on the planning of individual flight paths and resource allocation, for which the probability of targeting is performed, proportionally, avoiding redundancy in the coverage, thus correctly allocating the available resources. These methods have a support center for optimization, and people specialized in the identification of targets.

There are works in the literature using several different ways to identify people in danger. Noteworthy are the works that use artificial intelligence, in [8] implemented a system in real time, using the deep convolutional neural network of the Single Shot Detection (SSD) binary output topology, based on TensorFlow Lite, reaching a recall of 65.4 and 96.4% accuracy in images at a rate of 11 to 17 FPS, captured from a sequential search from a defined area. The HERIDAL dataset [9] was used, which has 68,750 images, 29,050 with positive examples. The author highlights the need to adjust the parameters and topology of the network, so that objects that are proportionally very small in relation to the entire captured image can be safely detected.

The work of [10] with a focus on maritime rescue, also uses convolutional networks with binary output, however the input is through data from a multispectral camera, with the filter channels blue (B), red (R), green (G), red edge (REG) and near IR (NIR), which makes the cost of the project high. The results obtained by them show that the best classification performance of 89% through the Xception and SqueezeNet topology, in images captured with about 6 FPS, was obtained by combining the G, REG and NIR channels combined with the use of sliding windows.

Finally in the paper [11] a lightweight system was proposed using the You Only Look Once version 3 (YOLOv3) algorithm, focusing on human detection in low-pixel (640×512) multispectral infrared and thermal images, that is, with use of high-cost special cameras. The results showed an accuracy of 89.26% in images at 24.6 FPS. Recent work on this topic shows that despite good results, the cost of processing, such as material and maintenance, is high. In addition, the operating scenario and path and search mode can be optimized.

34.3 Methodology

This section will present general techniques commonly used for problems in identifying anomalies in images for the most diverse purposes. The raw image, as it comes from the UAV, can be difficult to identify anomalies in the soil for both

humans and algorithms, to circumvent this image treatment techniques are employed. In this pre-processing, segmentation, resampling and cutting of the region of interest are performed, equalization and enhancement filters can also be applied, these processes and image manipulation can result in less computationally expensive processing in the classifiers [12].

Machine learning techniques are commonly used in image processing to identify subtle differences [13]. Convolutional neural networks (CNN) are a variation of the multi-layer perceptron networks (MLP) of the feed-forward type, which in turn are inspired by the biological processes of the pattern of connectivity between neurons, originating in the organization of the animals' visual cortex. It stands out because it has been successfully applied in the processing and analysis of digital images, requiring a minimum level of pre-processing when compared to other traditional image classification algorithms, since its structure allows it to correctly recognize and apply the standardized characteristic filters taken from the training images [14].

Although it is already widely used in this subject, it has recent technologies of low operational cost for use in UAV that can be applied in this context. In the Inception V2 architecture two 3×3 convolutions are used, the module's resource banks are expanded instead of deeper. This would prevent loss of image information and decrease computational time in convolution. The Single Shot MultiBox Detector (SSD) is a popular algorithm, which aims in a single shot to detect multiple objects within the image, keeping the concept of regions, unlike other approaches that need two shots, so it maintains its accuracy with less computational time. You Only Look Once (YOLO) introduced a faster way of real-time object detection, instead of using the region selection method, YOLO uses the Convolutional Neural Network which provides the bounding boxes as well as the class to these boxes, dividing the image into a constant value grid.

When classified, a tracking can be performed to monitor the detected anomaly, this ensures that a certain item is closer to something alive, as it moves. In addition, when obtaining the plotted path, it is possible to make estimates of the location at a given time. With the confirmation of the target, the processing is also lighter as it does not need to classify continuously.

Boosting is based on AdaBoost that acts in haar cascade, has stability and good detection despite the adjustment time. The Multi-Instance Learning (MIL) tracker is one of the most used algorithms, its principle is to look in a small neighborhood around the current location to generate several potential positive examples, where only one image can already be considered, it has a performance and even under partial occlusion. The Kernelized Correlation Filters (KCF)

algorithm, its principle is similar to MIL, however it applies several mathematical properties in the overlapping regions of positive samples for optimization, and therefore obtains better precision and processing speed [15].

To develop a decision support system in the search and rescue of people, it is necessary to approach the reality of the professionals involved. Thus, it was proposed to create a dataset with the most different situations. The images were obtained from a DJI Mavic drone with a simple 1/2.3 camera (CMOS), up to 12.35M pixels, and 78.8° 26 mm field of view (FOV) lenses, Distortion <1.5% and Focus from 0.5 m. Which are considerably less expensive than UAVs with multispectral cameras.

The first set featured images of a volunteer performing positions in different ways: arms open as if he were asking for help; sitting as if he were at rest; and performing ok gestures as if he were dispensing help. The second in the set of images, on the other hand, used clothes of different colors scattered in the UAV's area of operation. So that it could simulate people lying on open fields, shadows or corners.

Finally, the identification points were monitored using the UAV, so that this third set of images could determine the path taken by an individual in this environment. These image sets will undergo manipulation processing, and by the recognition techniques presented, their effectiveness will be evaluated by precision

$$P = \frac{VP}{VP + FP} \quad (1)$$

and recall

$$R = \frac{VP}{VP + FN} \quad (2)$$

considering true positives (TP), false positives (FP) and false negatives (FN).

The processing speed will be evaluated using the runtime and image flow indicators. Another proposal of the work is to analyze other forms of field of view coverage and image capture, using several low cost drones, at different heights and positions in relation to the ground. For this, a mapping of a test area will be defined and its coverage and quality of the images and identification will be evaluated, and share with other UAVs, the complete process is shown in Fig. 34.1.

This process will allow you to receive the images for analysis, where color filtering will be performed, which helps to identify specific objects and decrease processing time. From the result of the image, it is possible to cut out anomalies in the image, with a certain margin of safety, so that the detection algorithm of the specific object is applied. The location of this anomaly will be analyzed and shared with other UAVs on the team.

With classification and location of the object it will be possible to establish the direction and speed of the identified

person, from two frames in a row with a controlled interval, a certain time t will be determined, the distance will be obtained by the distance between the initial pixels of the object x_0, y_0 .

Thus, considering the height as distance from the drone to the person h in meters, the focal length f in meters, and the width of the person in Image p in pixels, it is then possible to obtain the approximate size of individual w with the formula

$$w = \frac{h * p}{f} \quad (3)$$

and also estimate the distance between points euclidean distance defined in pixels for meters d_m . The average speed v_m is then obtained by this distance by the time difference t of obtaining the two frames

$$v_m = \frac{d_m}{t_1 - t_2} \quad (4)$$

The direction will also use the initial pixels of the detected object to obtain the tangent of the angle

$$\theta = \text{atan}((y_1 - y_0), (x_1 - x_0)) \quad (5)$$

and then relate to the cardinal and collateral points. In addition to the image information, data from the Global Positioning System (GPS) can also be used to determine the individual's real position. The last step is crucial for the result of the search operation, from the data collection a specialist system will be developed to determine a suggested action. For this, a database will be developed with professionals trained from previous experiences.

34.4 Provisional Results

With the development of the study, it is expected to develop a specialist system for decision making in the identification, search and rescue of people at risk, based on images from cameras embedded in UAVs. This system will be based on the processed information of the person identified in the images, including speed, direction and position of the people. This information will be obtained when anomalies are identified in the images and classify them using machine learning as people in those images, as shown in Fig. 34.2.

It is also expected to contribute with its own dataset, so an image survey was carried out using UAV for composing the dataset. Through the Roboflow [16] each person in the image received a mark with the points regarding their location in the image, as well as position information, altitude of the UAV, situation in which they are and camera position at the time of capture, so far there are 231 images and 459 tags.

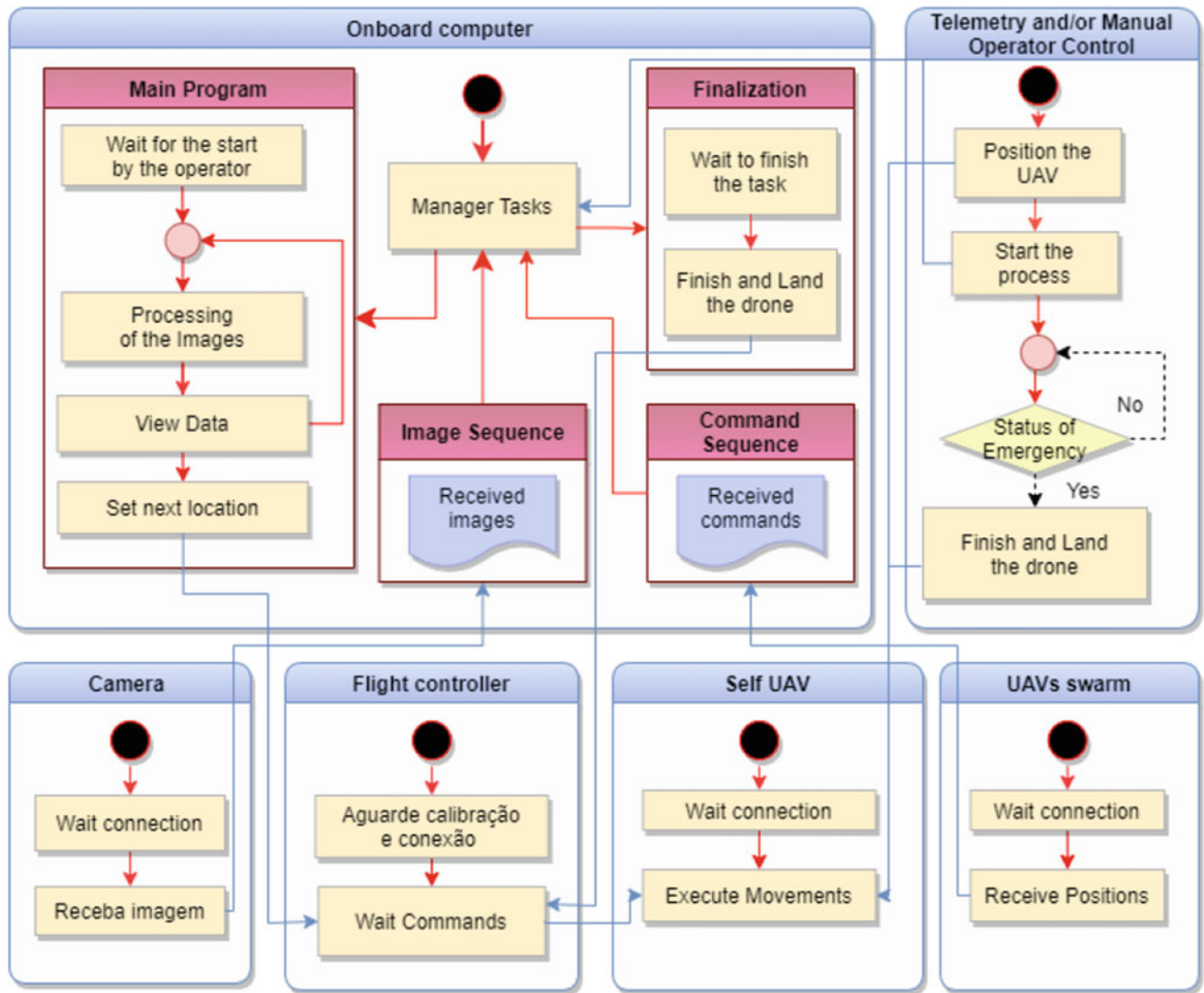
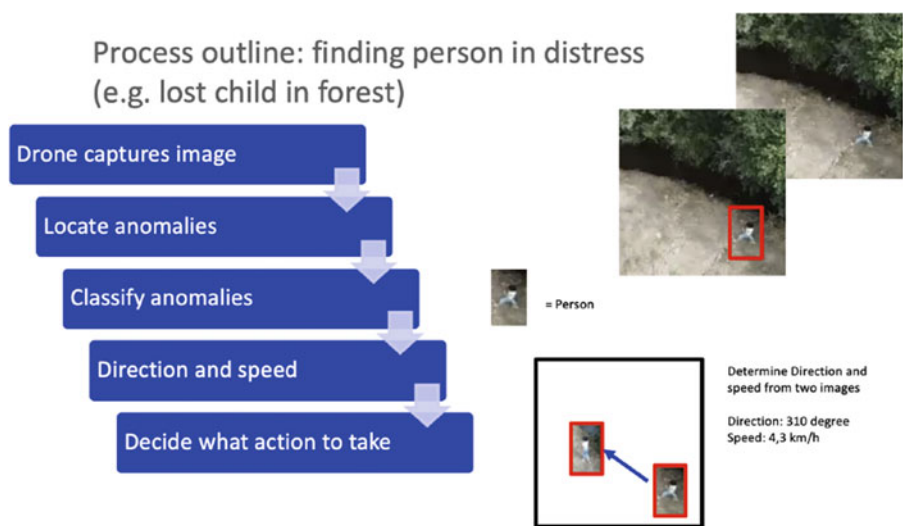


Fig. 34.1 UAV communication process flowchart

Fig. 34.2 Process outline of project (adapted from presentation by Peter Funk)



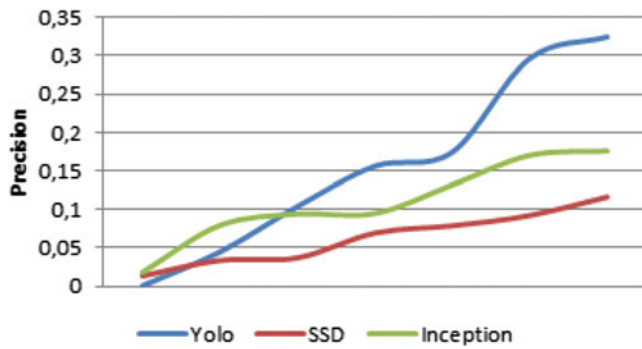


Fig. 34.3 Comparison of training of listed object detection algorithms with precision)

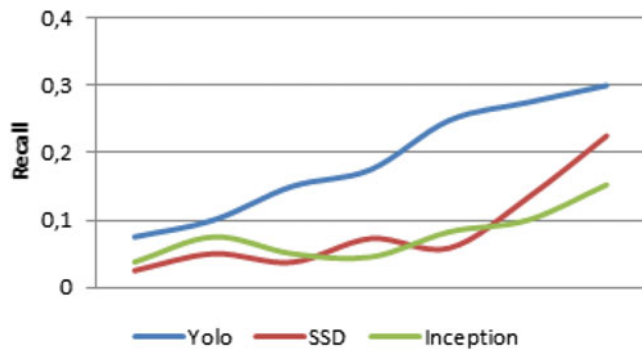


Fig. 34.4 Comparison of training of listed object detection algorithms with recall)

The development of this project was carried out in the Google Collaboratory tool. This working environment in a web browser for PyTorch codes, is based on Jupyter notebooks, does not require configuration, runs in the cloud and allows limited free use of processing units graphics (GPU), in general consists of a limited Tesla K80 GPU. Its high-level interface is based on Python and tensor computation, and makes it possible to train neural networks through acceleration via GPU [17, 18].

It is expected to apply segmentation and machine learning to obtain bounding boxes represented by the sub-regions of the original image where the object detection algorithm was applied, and thus obtain the highest probabilities associated with them using the technique with greater precision, the data from first training are shown in Figs. 34.3 and 34.4, which represent the precision and recall respectively.

SSD Mobilenet v2 performed well both in training and inferences but still identifies many false positives, while Inception v2 performed less in processing time efficiency but slightly higher accuracy, while Yolo v5 despite the good accuracy and recall, did not get good results in practical inference. It is understood the results obtained as a result of a network manager in the case of Inception and difficulty in identifying small objects in the case of Yolo.

For tracking, the 3 listed techniques were compared, and the correct performance in tracking the object determined

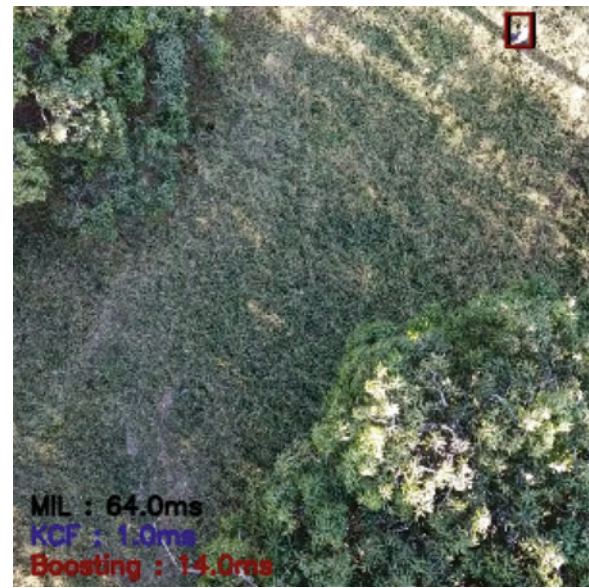


Fig. 34.5 Comparison of training of listed object detection algorithms)

in the previous detection was empirically verified. A frame obtained from the result of this test is presented in Fig. 34.5.

According to what was observed, the algorithm that managed to maintain the longest tracking time, even when there was an occurrence, was the KCF.

The Fig. 34.6 shows the provisional result of a test of the algorithm in its toolset, for a high-amplitude search and rescue video sequence.

34.5 Conclusion

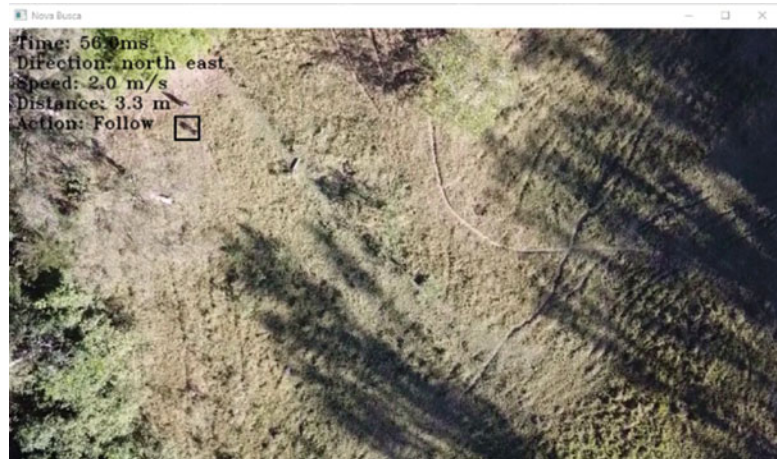
This article presented a work in progress, which already has a dataset, an initial basis for training, an architecture for experiments and, mainly, a theoretical concept for conducting a search and rescue of people using UAV images.

From the provisional results, it is possible to observe the feasibility of the procedure adopted in the study. So far, detecting people/clothes using Yolo V5, and tracking using KCF have yielded the best results for the research context.

It is expected to increase the current dataset with new public images and considerably increase the training time for long-range identification of people and objects. It is also a future objective to improve the tracking algorithm used and obtain more variables that help in the search and rescue of people in emergency situations.

Acknowledgments I am very grateful for the internship received from the Federal University of Itajubá, Brazil and for Mälardalen University, Sweden, providing a work environment to carry out the project and the Swedish Innovation Agency Vinnova (as the internship project is also a part of the Vinnova project 2018-04464 with Saab AB, Mälardalen University, Embraer SA, Instituto Tecnológico de Aeronáutica, I got the opportunity to closely work together with senior researchers and PhDs).

Fig. 34.6 Search and rescue algorithm execution with estimated results)



I would like to especially thank my supervisors Alexandre Ramos, Peter Funk and assistance from Ella Olsson (SAAB AB) for their inspiration, discussions and guidance.

The project is an important contribution to a research project at Mälardalen University funded by the Swedish Innovation Agency Vinnova. This study also was financed in part by Unifei and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Reference

1. M. Poteyeva, M. Denver, L.E. Barsky, B.E. Aguirre, Search and rescue activities in disasters, in *Handbook of Disaster Research* (Springer, New York, 2007), pp. 200–216
2. A.J. Dantas Filho, A.C. Ramos, L.D.D. Jesus, H.F.D. Castro Filho, F. Mora-Camino, A general low cost UAV solution for power line tracking, in *17th International Conference on Information Technology—New Generations (ITNG 2020)* (Springer, Cham, 2020), pp. 525–530
3. T. Niedzielski, M. Jurecka, M. Stec, M. Wieczorek, B. Miziński, The nested k-means method: a new approach for detecting lost persons in aerial images acquired by unmanned aerial vehicles. *J. Field Robot.* **34**(8), 1395–1406 (2017)
4. Z. Kashino, G. Nejat, B. Benhabib, Multi-UAV based autonomous wilderness search and rescue using target ISO-probability curves, in *2019 International Conference on Unmanned Aircraft Systems (ICUAS), Atlanta, 2019* (2019), pp. 636–643. <https://doi.org/10.1109/ICUAS.2019.8798354>
5. D. Hanna, A. Ferworn, A UAV-based algorithm to assist ground SAR teams in finding lost persons living with dementia, in *2020 IEEE/ION Position, Location and Navigation Symposium (PLANS)* (IEEE, Piscataway, 2020), pp. 27–35
6. A. Alhaqbani, H. Kurdi, K. Youcef-Toumi, Fish-inspired task allocation algorithm for multiple unmanned aerial vehicles in search and rescue missions. *Remote Sens.* **13**(1), 27 (2021)
7. R. da Rosa, M. Aurelio Wehrmeister, T. Brito, J.L. Lima, A.I.P.N. Pereira, Honeycomb map: a bioinspired topological map for indoor search and rescue unmanned aerial vehicles. *Sensors* **20**(3), 907 (2020)
8. Z. Domozi, D. Stojcsics, A. Benhamida, M. Kozlowszky, A. Molnar, Real time object detection for aerial search and rescue missions for missing persons, in *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)* (IEEE, Piscataway, 2020), pp. 000519–000524
9. Ž. Marušić, D. Božić-Štulić, S. Gotovac, T. Marušić, Region proposal approach for human detection on aerial imagery, in *2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech)* (IEEE, Piscataway, 2018), pp. 1–6
10. A.J. Gallego, A. Pertusa, P. Gil, R.B. Fisher, Detection of bodies in maritime rescue operations using unmanned aerial vehicles with multispectral cameras. *J. Field Robot.* **36**(4), 782–796 (2019)
11. I. Martinez-Alpiste, G. Golcarenenrenji, Q. Wang, J.M. Alcaraz-Calero, Real-time low-pixel infrared human detection from unmanned aerial vehicles, in *Proceedings of the 10th ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications* (2020)
12. B.S. Min, D.K. Lim, S.J. Kim, J.H. Lee, A novel method of determining parameters of CLAHE based on image entropy. *Int. J. Softw. Eng. Appl.* **7**(5), 113–120 (2013)
13. L.D. de Jesus, F. Mora-Camino, L.V. Ribeiro, H.F. de Castro Filho, A.C. Ramos, J.R.G. Braga, Greater autonomy for RPAs using solar panels and taking advantage of rising winds through the algorithm, in *16th International Conference on Information Technology—New Generations (ITNG 2019)* (Springer, Cham, 2019), pp. 615–616
14. S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 221–231 (2012)
15. T. Liu, G. Wang, Q. Yang, Real-time part-based visual tracking via adaptive correlation filters, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 4902–4912
16. B. Dwyer, J. Nelson, Roboflow (Version 1.0) [Software] (2021). <https://roboflow.com>
17. Google Collaboratory (2021). <https://colab.research.google.com>
18. A. Dantas, L. Diniz, L. Gewehr, C. Alcino, W. Martins, T. Pimenta, A. Ramos, Low-cost UAV for medical delivery. *Int. J. Dev. Res.* **11**, 21239 (2021)

Part VII

Human-Computer Interaction

An Application for Interaction Comparison Between Virtual Hands and Virtual Reality Controllers

Daniel Enriquez, Christopher Lewis, Sergiu M. Dascalu, and Frederick C. Harris, Jr.

Abstract

This paper presents an application for Virtual Reality (VR) interfaces for virtual hands which will allow us to compare interaction between virtual hands and VR controllers in Virtual Environments (VEs). Development for human-computer interaction in VEs needs improvement to accommodate the growth and need for applications inside VR. Virtual hands are growing more prevalent with many devices detecting the location and mimicry of the user's own hands inside the VE. Virtual hands can also be implemented via VR Gloves to more precisely pinpoint the movements of hands. This work also implements interaction mediums that can be used by virtual hands or VR controllers to directly manipulate and control virtual objects and virtual interfaces. Unity was used to generate the VE and to render the input mediums and interactable objects. SteamVR was used to connect the input mediums to Unity. The HTC Vive Pro Eye was used to connect the user to the VE. The two input mediums that were compared are the HTC Vive Controllers and the HI5 gloves. All of these components come together to form an immersive and consistent means to compare input mediums in different kinds of interactions.

Keywords

Virtual reality · Virtual environments · Human computer interaction · Virtual hands · Virtual interaction · HTC vive

D. Enriquez · C. Lewis · S. M. Dascalu · F. C. Harris, Jr. (✉)
Computer Science and Engineering, University of Nevada, Reno,
Reno, NV, USA
e-mail: denriquez@nevada.unr.edu; christopher_le1@nevada.unr.edu;
dascalu@cse.unr.edu; fred.harris@cse.unr.edu

35.1 Introduction

Virtual Reality (VR) has seen rapid development recently with its rise in the consumer market for entertainment and its practical application in education. VR applications have been accustomed to using tracked controllers. Tracked controllers have added drawbacks of cost, time-to-learn, and energy. These problems are beginning to be addressed with virtual hands as virtual hand tracking improves and becomes integrated onto more devices. Virtual hands have the added benefit of being able to transition between VR and augmented reality as they shift into any virtual world much easier than the controller as a physical device. Virtual hands are an emerging technology that allow for the user's hand's location and orientation to be mapped to a Virtual Environment (VE). This work implements virtual hands using the Noitom Hi5 VR Gloves [7] as a means to accurately track finger movement, hand position, and gestures. Virtual hands are typically mapped via a series of sensors, such as a visual sensor like a camera or by VR gloves that have sensors to determine where the user is moving their hands. New virtual hand technologies such as the Oculus hand tracking [2], Ultraleap [12], and VR gloves are seeing their usage in more virtual world applications. The use of virtual hands as a medium of interaction within VEs needs development as more applications in the future are likely to incorporate this more accessible method of interaction.

A key component in both the research and creation of VR interaction is analyzing how people interact with both real objects and their virtual counterparts. Two main differences exist between interacting with a physical object and a virtual object. One is the lack of physical stimuli inherent with a virtual object, for example, the weight or texture of an object. The second is the lack of direct mapping between virtual and

physical interaction. This difference is very apparent when looking at the VR controller as, due to its length, it doesn't have the same degrees of freedom that a person's hands/arms have. The controllers also don't allow the user to "grab" anything, the closest action to this is pulling the trigger on the controller. The virtual hands have this problem due to the lack of resistance a user feels, or rather does not feel, when trying to touch or grab an object. This means the user doesn't know where to position or close their hands when interacting with an object. For example, if a person is physically opening a door, they close a hand around the door handle, the door handle provides resistance, a physical stimuli, which tells a person that they have grabbed the door handle and don't need to squeeze any harder. If this same example is to be used when virtually opening a door, the user has no way of knowing when they have successfully grabbed the door handle. A VR controller uses the click of the trigger or the physical resistance of the trigger to simulate this, but with virtual hands there is no inherent indicator.

Another key component of applying virtual hands to VEs is finding a way to translate ready-made VR controller applications to use virtual hands. Many VR applications incorporate object interaction where the user needs to press a button on the VR controller to correspond to an action done onto an object in the VE. This work proposes a method to allow this shift to take place in relation to virtual menus, virtualized everyday objects, and rigidbody physics objects. Virtual menus are seeing more use in VR as they provide an easy means of selection for user interface, typically done with laser pointing using the VR controller, further necessitating the transition to virtual hands. With these methods of interactions, a direct comparison of VR interactions using virtual hand technology could be made as a contrast to the commonly used VR controllers.

The rest of this paper is structured as follows: Sect. 35.2 presents background on VE interaction mediums and the Human Computer Interaction elements (HCI) that affect VE input. It also showcases related works in the area involving Virtual Hands, VR Gloves, and virtual interaction. Section 35.3 highlights the methodology of the HCI VE elements considered in the application as well use cases of both virtual hands and VR controllers. Section 35.4 discusses the details of the application, the uses of the application, and the interaction elements that can be examined with this application. Section 35.5 summarizes the uses of the application and the methods of interaction inside VEs. Section 35.6 highlights areas of future work based on this research.

35.2 Background and Related Works

35.2.1 HCI VE Considerations

In the process of making a VR application, there are several general baselines of achieving a successful HCI interaction. Shneiderman [10] details how there are five usability measures to consider with interactions, those being: time to learn, speed of performance, rate of error, retention over time, and subjective satisfaction. This work implements the application to highlight some of these usability measures in relation to input into VEs with these usability metrics taken into consideration. The measures chosen were performance, presence, and ease-of-use, which are a slight modification of general HCI usability features proposed by Shneiderman, with a focus on VE interaction. The user should be able to perform the task required at hand in an efficient manner, achieving speed of performance and low rate of error. The user should be able to feel immersed into the VE. As a result, presence was also chosen as it highlights the user satisfaction, while simultaneously examining the user's ability to be immersed inside the VE. Finally, ease-of-use was chosen because it considers time to learn, expected outcome, and rate of error.

35.2.2 VE Interaction Design

The design of object interaction and UI interaction are the primary means of human input into a VE and are critical to the core of a virtual world. Sutcliffe et al. [11] discusses the human-computer interaction (HCI) challenges that are native to a VE experience. The authors went through multiple case studies and determine design processes and associated design trade-offs for each case study. The authors find that immersion inside of VEs is imperative. They also found that a good way to keep immersion while allowing users to read information or interact with a user interface (UI), is to integrate this UI into context-sensitive pop-ups that are translucent. They also found that audio/speech inside of these case studies should be accompanied by or replaced with pop-ups or readable interfaces.

Another consideration within interaction is whether or not an interaction medium needs remapping to occur. Actions executed in a virtual environment do not necessarily need to mirror real life in order to achieve the same goal. As a result, some interaction methods can use remapping, a method in

which an action in the VE leads to an outcome that may not always lead towards the same outcome in the real world [5]. An example of this is having a user snap their fingers to open a virtual door; in the VE the action of snapping your fingers are “remapped” to open a door, whereas in real life the user is just producing sound. A common implementation of remapping that we selected was to use a laser pointing system for the menu interaction. We also decided to implement remapping for grabbing objects. This was used to allow the user to perform the grip action, with either virtual hand or controller, to take hold of the object to interact with them.

35.2.3 Virtual Input Devices

VEs can have many methods of interaction inside the applications to immerse the user into the virtual world. These range from microphones and speakers, to eye tracking and mouth tracking. Another method of immersive interaction is by allowing the user to interact with objects using everyday hand movements. This research uses the Noitom Hi5 VR Gloves [7] to accurately map the location of fingers, hand movements, and gestures into the VE.

A significant portion of the considerations of interaction in VE is whether or not the response triggers haptic feedback. In VEs, the lack of haptic feedback is apparent because there are a limited amount of devices that can interact inside a VE [9]. We decided to create an application where the lack of haptic feedback would be apparent with the usage of virtual hands but not so apparent with the usage of controllers. The user may experience some haptic feedback with the click and the physical presence of the controller but may experience the empty hand entirely differently. We also decided that, in terms of accessibility for VR/AR devices, virtual hands should be used as they are more likely to be a common medium of interaction in VEs. Our work could also allow for the potential of comparing haptic vs. non-haptic virtual interactions to determine their level of usability

Fahmi et al. [3] preformed a user study on the experience with VR controllers, the UltraLeap’s Leap Motion Controller, and the senso glove for anatomy learning systems. In this study, they concluded that the Vive controllers [4] were rated higher in terms of user satisfaction when compared to the other two controllers. Our work aimed therefore to create an application that can allow for these differences to be highlighted and examined; also, to confirm whether or not the controllers are more satisfying to use in many diverse types of interactions.

35.2.4 VR Gloves

Bowman et al. [1], described interaction techniques used in different VEs specifically designed for *PinchGloves*. The authors pointed out limitations of VEs that are still major hurdles for any modern VR application to overcome. One such limitation being that interactions within VR/VEs should mainly be designed around physical interactions, or interactions within a menu. Data entry and the completion of certain tasks, like typing a sentence, are generally much slower to complete within VR/VEs, and should generally be avoided when designing an application.

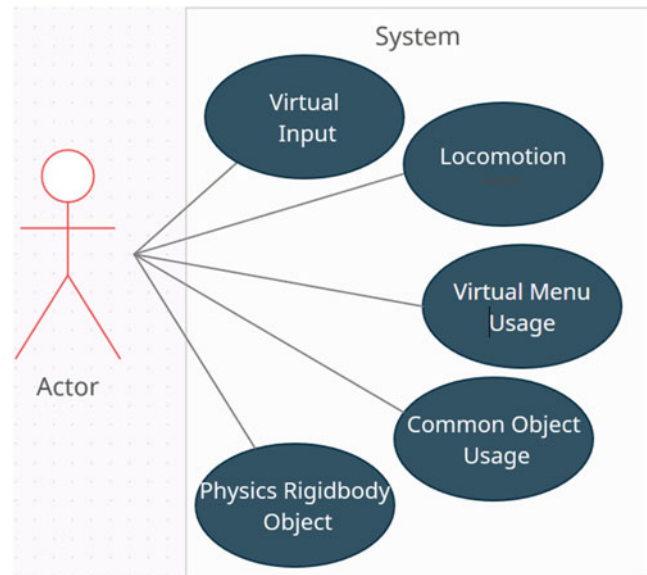
Luzanin et al. [6] discusses the use of probabilistic neural networks in the use of static gesture recognition for data gloves in VR. The authors found great accuracy within their testing of static gesture recognition, even with users who were not part of the neural network’s training. While this approach to gesture recognition seems to be very accurate, using a neural network in conjunction with a VR application with multiple input mediums is rather computationally expensive and time consuming for both the developers and the users. Thus, more gloves are being created with built-in SDKs/APIs that will do basic static gesture recognition or allow the developers using the gloves to create their own gestures. Among these, the Noitom Hi5 VR gloves [7] and the SensoryX VRFree VR gloves [8] are notable examples of gloves suited for this type of gesture recognition.

35.3 Methods

35.3.1 Virtual Input Usability Factors

There are important usability factors to consider for an interaction to perform effectively in a VE. The factors we decided to focus on are performance, presence, and ease-of-use. While there are other factors to consider when analyzing the virtual input, these were chosen as they could most likely be affected from the transition from input type. **Performance** is the ability for the user to perform a specific action, and the rate at which that action is performed. Virtual applications need to have users understand and perform any given task effectively. **Presence** is the ability to feel immersed in a virtual environment. Presence was chosen because of the need to evaluate how realistic it is to perform an action with your hand vs. the drawback of not having that touch stimuli. **Ease-of-use** was chosen to evaluate if the input actions on the virtual hand or the controller are easier to understand and use for users. It would also be useful to evaluate the intuitiveness of input design using ease-of-use as metric.

Fig. 35.1 Use case diagram showcasing the interactions shared between virtual hands and VR controllers



35.3.2 Detailed Use Cases

The use cases illustrated in Fig. 35.1 describe interactions between users and the system that are shared between virtual hands and VR controllers. This is typically illustrated with theoretical actors and things they can do within the system. These use cases were defined in relation to how a user would interact with a VE given these input devices. A more descriptive account of these actions is as follows:

1. **Virtual Input** The user will input their actions into the VE using either virtual hands, implemented via VR gloves, or VR controllers. The location and orientation of both are mapped from the physical world into the VE relative to their position of the user.
2. **Locomotion** The user will move in the Virtual Environment using some motion or action on the user's hands. Depending on the action of the user, the player will move corresponding to the input and the direction the user is facing. Using the touchpad or joystick on the controller, the user will move in the virtual environment according to the movement on the controller.
3. **Virtual Menu Usage** The user will use buttons on a virtual menu in VR. The buttons will have noticeable effects to notify the user that they have been successfully pressed. The user can point to the menu via a laser pointing system attached to their virtual hand/VR controller. The virtual menu will have UI buttons, sliders, drop-down menus, and more.
4. **Common Object Usage** The user will use and interact with virtualized common objects that people experience on a day-to-day basis in the real world. The user should

be able to intuitively interact with these common objects, like they would normally.

5. **Physics RigidBody Object** The user will be able to either throw, catch, pick up, or hit an object that moves corresponding to the physics of the VE. The user will need to grip the VR controller or virtual hand to be able to grab the object. The object's location in the virtual environment will determine its collision with the virtual object.

35.4 Implementation

The application was divided by creating three stages that encompass the core aspects of methods of virtual interaction inside VR. Each stage focuses on key aspects differentiating VR controllers and VR gloves and their usage in a virtual environment. The three stages are **UI interaction**, **ubiquitous object interaction**, and **object coordination and interaction**. By dividing the application, there could be a variety of simple, but key, interactions that could be useful for determining user input while allowing for each stage of interaction to be independent of another. The order of execution can be seen in Fig. 35.2, which showcases the steps the user will take in this application.

The following sub-sections describe in detail the processes in which a user will perform the application. Each stage allows for the analysis of usability features within the applications between virtual hands and VR controllers. Each stage also describes the purpose of the interaction method that will aid to represent the HCI elements of objects in virtual environments.

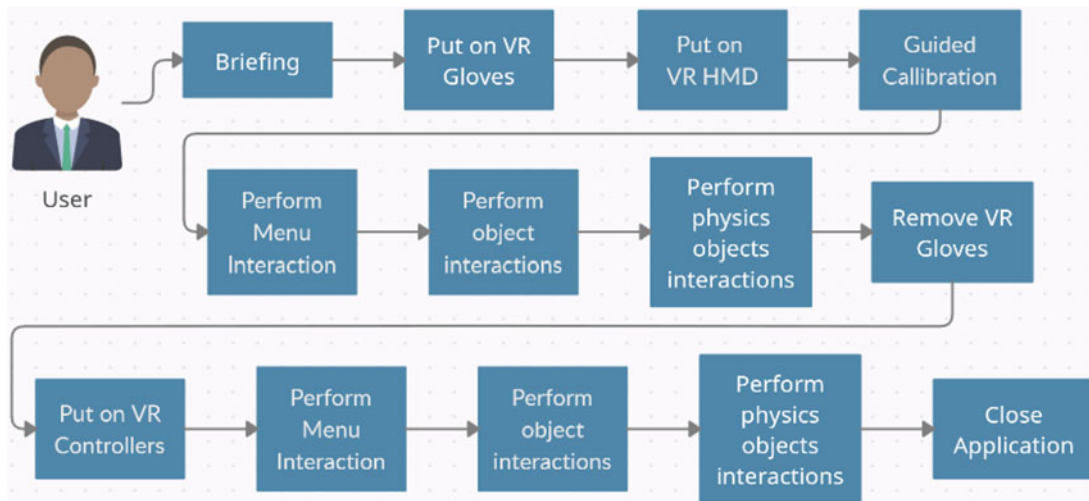


Fig. 35.2 Order of execution of the user's activities in the VE

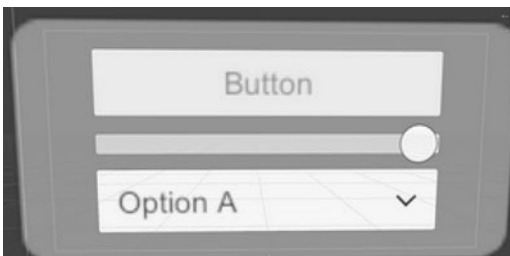


Fig. 35.3 Stage 1 menu interaction UI sample, containing a button, slider, and a drop down menu

35.4.1 Stage 1: Menu Interaction

The menu interaction of the application allowed for a pointing system using a laser system for both VR gloves and VR controllers. The VR gloves pointing system was implemented as if the user would be holding a laser pointer. To select or click on the application, the user would be required to make a fist with their opposite hand. A left mouse button click would occur as the user would make a fist. The click would remain until the user unclenches their fist. This action would be similar to clicking and letting go of a click on a mouse. From this, it was decided that this method of interaction was sufficient for interacting with the virtual menu interface, since it is rare to find VR interfaces that contain options for right click, middle mouse click, or scrolling. With the VR controllers, a laser pointing system was put in place, which allows the user to point using their left controller and allows the user to left click on the menu by pressing the controller's trigger button, after pointing at an option in the UI.

As seen in Fig. 35.3 the menu system's architecture incorporated a basic canvas UI object with three components attached to it that indicate basic user actions. The first com-

ponent is a standard button, which turns a darker color when hovered over, and an even darker color when pressed. The second component is a slider, where the user can move a small circle along a line that would similarly change color, depending on what was currently happening to it. The last component is a simple drop down menu that, when clicked, allows the user to choose between three options.

This stage was chosen to help differentiate the intuitive use of a pointer system to interact with virtual menus in VEs. This will allow for the comparison of ease-of-use when it comes to determining pointing systems. Some users may find the VR controller easier to point as it could be considered a wand while others may prefer the virtual hand as they can point with their hand as if they were to be holding a laser pointer instead of an object. This will also allow us to evaluate performance using the application. With this, we can determine if the user is able to perform the desired actions using both methods of input (Fig. 35.4).

35.4.2 Stage 2: Virtualized Everyday Objects

For the second stage, a common physical object to interact with was integrated into the virtual world that would allow the user to use both input methods. The common physical object that was implemented was a door to represent a mechanical operation. This is because a door is something that everyone is accustomed to interacting with on a daily basis in the real world, however implementation in a virtual world is an area which has not been used frequently in VR applications. An example of the usage of the door's implementation can be seen in Fig. 35.5, in which the lone button on the wall is inviting the user to press it to open the door.



Fig. 35.4 Stage 2 A user reaching for the doorknob



Fig. 35.6 Stage 3 A user preparing to catch a ball

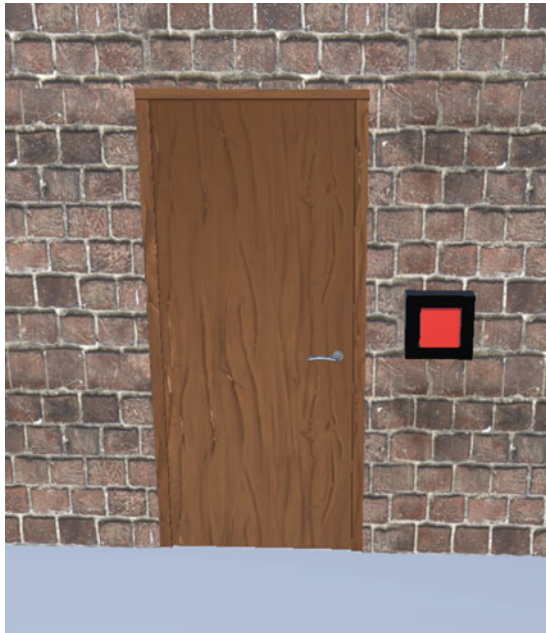


Fig. 35.5 Stage 2 virtualized everyday objects interaction with a door and an ADA button

There are three different parts to stage 2: the first, where the door is opened by turning a simple door knob; the second, where the door is opened by pressing a disability access button; and the third, where the door is opened by pulling a lever. By using a door knob, if the user is using VR gloves, they would approach their hand to the virtual door knob, as seen in Fig. 35.4 and then, by making a fist, they could turn the knob which then opens the door. The door would work in a similar way, except the player would grab the lever and pull it down. For pressing a disability access button, the user would need to push the button with their hand in the virtual world. The implementation of VR controllers is nearly

identical except that, instead of grabbing, the user would press down on the trigger.

This stage was chosen to understand how users translate real, common objects into virtual objects and to analyze if they experience these interactions differently in the virtual world between the controllers and their virtual hands. Analyzing the level of presence is a key part of this step since the virtual hands input method will have no haptic feedback, while the user may experience such feedback with the controllers. This input may lead towards a change in the user's perception of their presence in VR. Another consideration is that this stage allows the examination of the ease-of-use of real objects that are virtualized into the VE. This standardization will determine if the direct translation of real objects to virtual objects is something that the users are comfortable with using or if there should be a level of remapping that occurs to allow for interaction. With this, it could be useful to determine if having a level of remapping occur could be more useful to the user for ease-of-use, because without remapping the user may not feel the expected haptic feedback.

35.4.3 Stage 3: Rigidbody Physics Object Interactions

For stage 3, the implementation of objects and the interactions with them was done by having the user interact with a series of objects with different properties. The final stage's implementation was split into three parts. A simple ball was implemented that the user would interact with in a variety of different ways, depending on the part. The first part would have the user catch a simple ball, as seen in Fig. 35.6, the second would have the user tasked with throwing a ball, and the third would require the user to hit a ball that was bouncing in front of them.

If the user is using VR gloves, they would grab the ball to catch or throw by clenching their fist slightly. The ball would then be attached to their fist. For hitting the ball, the collision for the VR glove will be enabled that will move the ball depending on how the user hits the object. Similarly, when the user is using VR controllers, a user will hover over the ball and then click on the trigger to pick up the ball in order to catch or throw it. In the case of hitting the ball, since the controller will be displayed inside the virtual world, the ball would need to collide with the display of the controller in the virtual world.

This stage was chosen to analyze how users interact with physics objects in the virtual world. People typically have an expectation of how to catch a ball or dynamic objects in the real world, but this stage should help quantify if those expectations help the performance of that action in the virtual world. Quantifying performance with physics objects is critical in high fidelity applications of VR such as sports training. Therefore, it is important to understand if the level of remapping that occurs is relevant to the outside experiences one may reproduce in a VE. Currently, a level of remapping is required as the user expects haptic feedback when touching an object that never occurs. This effect could be mediated with haptic gloves in high fidelity scenarios.

35.5 Conclusion

Methods of interactions in VE need to be accommodated in order to account for new input medium for VEs. Virtual hands are a new input medium that may be more commonly used in the future. This work made an application to showcase the mediums of changing between virtual hands and VR controllers on interaction. This work applies HCI principles in VEs to allow for the examination of controllers and virtual hands as mediums of interaction and input in VR.

35.6 Future Work

First, a user study using this application is of the utmost importance. The user study would be used to validate, using quantitative measures, how effective this application is at comparing the two input mediums.

Second, creating locomotion techniques for the VR gloves and comparing them to common locomotion techniques

found for the controllers is likely a significant way to expand this application. Locomotion is integral to many games and applications. This would help to further compare the gloves and controllers using new metrics.

Third, almost every existing VR application can bring motion sickness to the users of that application. This application is no different. Therefore, it is important to test not just the usefulness of these input mediums, but also their effect on the persons using them.

Acknowledgments This material is based upon work supported by the National Science Foundation under grant numbers 2019609 and IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. D.A. Bowman, C.A. Wingrave, J.M. Campbell, V.Q. Ly, C.J. Rhoton, Novel uses of pinch gloves™ for virtual environment interaction techniques. *Virtual Reality* 6(3), 122–129 (2002)
2. Facebook Technology LLC, Oculus controllers and hand tracking. <https://support.oculus.com/articles/headsets-and-accessories/controllers-and-hand-tracking/index-controllers-hand-tracking/>. Accessed 29 Oct 2021
3. F. Fahmi, K. Tanjung, F. Nainggolan, B. Siregar, N. Mubarakah, M. Zarlis, Comparison study of user experience between virtual reality controllers, leap motion controllers, and senso glove for anatomy learning systems in a virtual reality environment, in *IOP Conference Series: Materials Science and Engineering* (2021)
4. HTC Corporation, Vive 1 next-level vr headsets and apps (2021). <https://www.vive.com/us/>. Accessed 28 Oct 2021
5. S.M. Lavalle, Virtual reality, p. 283 (2021). <http://lavalle.pl/vr/>. Accessed 28 Oct 2021
6. O. Luzanin, M. Plancak, Hand gesture recognition using low-budget data glove and cluster-trained probabilistic neural network. *Assem. Autom.* 34(1), 94–105 (2014)
7. Noitom, Noitom Hi5 VR gloves (2021). <https://hi5vrglove.com/>. Accessed 28 Oct 2021
8. SensoryX. VRFree gloves. <https://www.sensoryx.com/>. Accessed 28 Oct 2021
9. W.R. Sherman, A.B. Craig, *Understanding Virtual Reality, Interface, Application, and Design*. (Morgan Kaufmann, Burlington, 2003)
10. B. Shneiderman, C. Plaisant, *Designing User Interface, 6th Edition, Strategies for Effective Human-Computer Interaction* (Pearson/Addison Wesley, London, 2015)
11. A.G. Sutcliffe, C. Poullis, A. Gregoriades, I. Katsouri, A. Tzanavari, K. Herakleous, Reflecting on the design process for virtual reality applications. *Int. J. Hum.-Comput. Interact.* 35(2), 168–179 (2018)
12. UltraLeap, Digital worlds that feel humanultraleap (2021). <https://www.ultraleap.com/>. Accessed 29 Oct 2021

Andrew Muñoz, Chase Carthen, Vinh Le, Scotty D. Strachan, Sergiu M. Dascalu, and Frederick C. Harris, Jr.

Abstract

Light Detection and Ranging (LiDAR) sensors have been employed in many different ways over time and continue to be utilized today. These sensors produce point clouds which are large and complex data sets that are a collection of position points across a 3D space. As LiDAR point cloud data can be highly complex, it can often be difficult to conduct analysis and visualization of the data sets. A web tool was developed to analyze and visualize this type of data, ensuing in an interactive and readable representation of the data. The data obtained for this tool is from LiDAR sensors located on street lights directly adjacent to the University of Nevada, Reno to analyze traffic information. In order to ensure the effectiveness of the tool, a user study was conducted to test the functionality and assess possible improvements.

Keywords

IDAR · Big data · Point cloud · Containerization · Data analysis · Data visualization · User interface · Web framework · MQTT · Traffic detection

highly effective tool that can be used to create solutions for a diverse range of problems. The term LiDAR stands for Light Detection and Ranging, meaning that it uses a light source, specifically lasers, to detect both the distance and position of various objects within a real-world environment [1]. The positional data gathered from LiDAR sensors are referred to as point clouds. A point cloud data set is a group of points that are a 3D representation of a specific object or location. These complex data sets often can be difficult to analyze in their raw format and therefore need to be cleaned and processed to procure useful information.

Another field that has seen many advances is data analysis, which has grown exponentially in recent years due to the rise of big data. Big data refers to the process of extracting and analyzing useful information from large and often complicated data sets. Data analysis is a useful component of big data as it simplifies the data in an attempt to discover valuable and insightful information. A large part of data analysis is that of data visualization, which aids in communicating information found in the data in a readable and comprehensible manner. The most common types of visualizations are graphs and charts, which are extremely useful in finding and determining patterns and relationships within a data set [2].

LDAT, aka the LiDAR Data Analysis and Visualization Tool, is a web application that is used to access point cloud data from street based LiDAR sensors and extract insightful information from the point clouds gathered. The LDAT tool is built upon the Angular and Flask web frameworks to create a robust web application that is able to process, clean, compress, and output the near real-time point cloud data stream. The output is generated through a 3D render of the mesh scene that corresponds with a particular LiDAR sensor. Both the data visualization graphs and charts, as well as the 3D render, are presented within a single web page as a data dashboard to make the data easily accessible and readable. LDAT was created as a proof of concept and therefore was designed with scalability, portability, and efficiency in mind.

36.1 Introduction

Sensor technology has been around since the mid twentieth century and continues to advance in many fronts. One such piece of sensor technology called LiDAR has become a

A. Muñoz · C. Carthen · V. Le (✉) · S. D. Strachan · S. M. Dascalu · F. C. Harris, Jr.
 Computer Science and Engineering, University of Nevada, Reno,
 Reno, NV, USA
 e-mail: amunoz24@nevada.unr.edu; ccarthen@unr.edu; vle@unr.edu;
strachan@unr.edu; dascalus@cse.unr.edu; fred.harris@cse.unr.edu

The rest of this paper is organized as follows: Sect. 36.2 details the project background and provides a brief overview on some related work. Section 36.3 describes the system design, data workflow, and implementation of both the hardware and software components. Section 36.4 presents the final prototype implementation of the LDAT application. Section 36.5 details the user study conducted and the results obtained from the study. Section 36.6 lists the concluding thoughts on the application and contains study results, and outlines possible directions of future work.

36.2 Background and Related Work

Data analytics and visualization are both very important in many modern data solutions. Big data has grown exponentially in recent years as data continues to increase in amount and complexity. One such complex data type is point cloud data that is captured through LiDAR sensors. The LDAT application gathers complex point cloud data directly from LiDAR sensors to analyze and visualize street traffic information. At the time of development, there were two active LiDAR sensors being utilized at the intersection of North Virginia Street and 15th Street which is directly adjacent to the University of Nevada, Reno. The LiDAR sensors stream over raw point cloud data (.pcd) files [3], which are then cleaned and processed to be used for analysis and visualization. Before diving into the design and development of the application, next there is a discussion on some related works.

There are a few works in recent years that relate to the processing, cleaning, and visualization of point cloud data from LiDAR sensors. The first is a study [4] that examines three low cost and scalable sensor technologies that could be mounted onto streetlights to create a street traffic detection system. The three sensor types used in this study included passive infrared motion detectors (PIR), thermographic cameras, and LiDAR sensors. Each of the sensors were fixed and secured to a pedestrian overpass located six meters directly above the road. The researchers conducted tests to assess each sensors ability to detect vehicles of varying speeds and determine the count of vehicles.

Another related project is the 3DSYSTEK viewer [5], which was designed by a group of researchers to be a web-based application that makes viewing of detailed 3D point cloud models easy and efficient. It was made to address mobility and portability problems that are typically associated with remote field applications. The researchers used Terrestrial Laser Scanner (TLS) technologies, otherwise known as Terrestrial LiDAR, due to its ability to swiftly gather 3D point cloud data that included coordinate, color, and orientation information. In order to create and develop the 3DSYSTEK viewer, the research team used the open source tool WebGL

in correlation with the Three.js library which allowed for the rendering of the 3D point clouds within the web application.

In a separate study [6], a 3D annotation tool was designed to visualize point cloud scenes. The main purpose of the tool was to be able to annotate objects such as vehicles and pedestrians within a 3D point cloud render. This was accomplished using bounding boxes to annotate the objects and make them easily distinguishable in the scene. The LiDAR sensor used for this tool captured point clouds by rotating around a test vehicle to capture data from every direction. In order to visualize and annotate the scene, the researchers developed a GUI where they could load the point cloud and related image to visualize the data and add in annotations via bounding boxes.

36.3 Software Design

LDAT was built using both the Angular [7] and Flask [8] web frameworks. Each of these components were designed to work as independent components to process, clean, and visualize the point cloud data sets. In this section, an overview of the system architecture is presented along with a breakdown of the data workflow and application interface design.

36.3.1 System Architecture

The system architecture was designed primarily with performance in mind while maintaining a smooth transition to create an optimal user experience. In order to fulfill this condition, the system would need to be designed with multiple servers to properly handle, manage, and process the large amounts of data that would be coming from the LiDAR sensors. Figure 36.1 presents the system level diagram for this application that provides a broad overview of the system and indicates the different components in the system along with the relationships between them.

LDAT is comprised of four main components that make up the system architecture. The first of these components is the LiDAR sensors. The hardware used was the Velodyne Ultra Puck sensor [9] due to its 200 m range and 360° view of the surrounding area. At the sensor level, the data is gathered in its raw format and pushed to the network through the second main component, the MQTT Broker [10]. The MQTT Broker is a network publish/subscribe protocol that allows the system to subscribe and receive messages directly from the LiDAR sensors. This network protocol not only receives the messages directly from the sensors, but also securely transports the data to other servers that are subscribed to the topic that is linked directly to the sensor.

Fig. 36.1 System level diagram showcasing the main components of LDAT

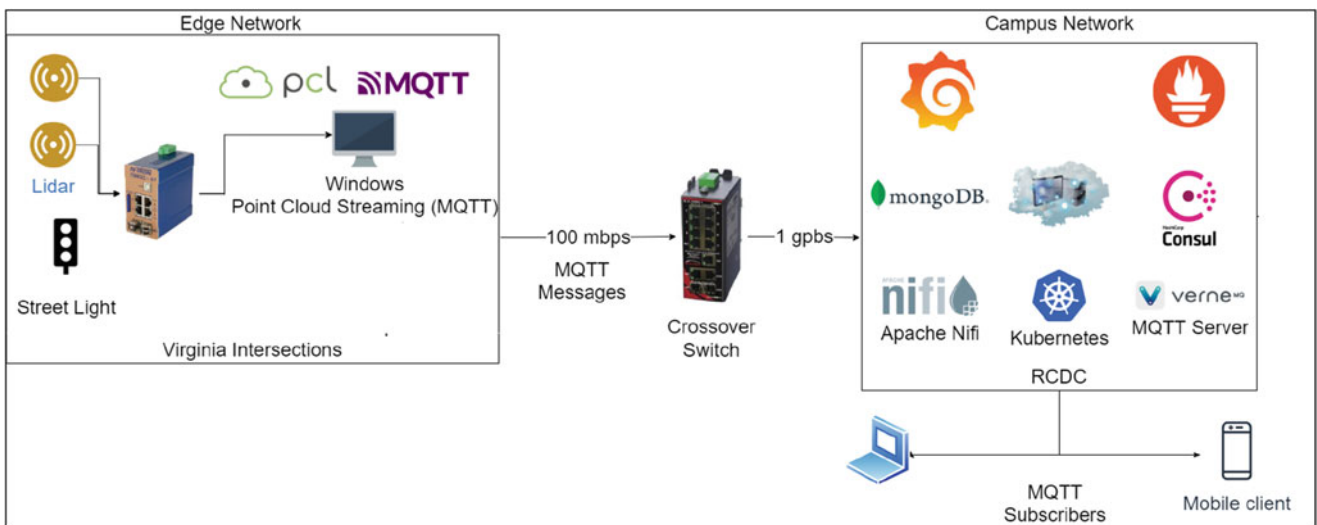
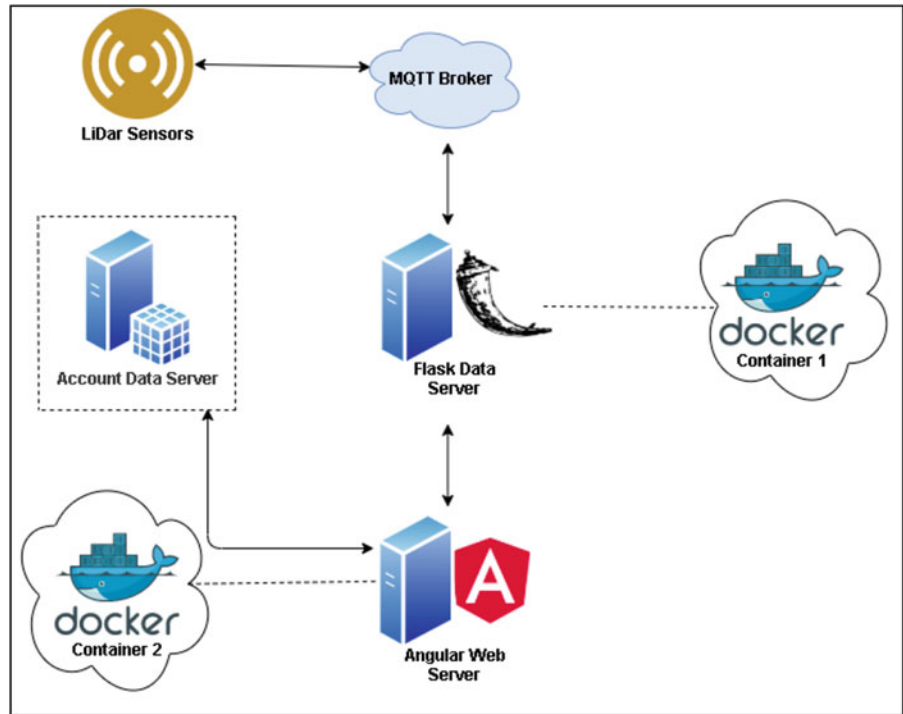


Fig. 36.2 High level diagram of the data workflow

The third component in the system architecture is the Flask Data Server which connects directly to the MQTT Broker. This data server subscribes to the topics on the MQTT Broker and the broker then pushes or publishes the data onto the server. Once on the server, the Flask app parses through the data and preps it for visualization on the fourth component, otherwise referred to as the Angular Web Interface. The Angular Web Interface acts as the main web page that is accessible to users and it is where data visualization and analysis occur. An important note is that both the Flask Data Server and Angular Web Interface are placed in separate

Docker containers [11] to increase development efficiency and portability. Additionally, there is an Account Data Server which serves as a database server for user accounts such as log in information. This server is covered in more detail in Sect. 36.6.

36.3.2 Data Workflow

The point cloud data files are gathered by the LiDAR sensors which are located on an edge network. The high level diagram [12] of the data workflow is shown in Fig. 36.2.

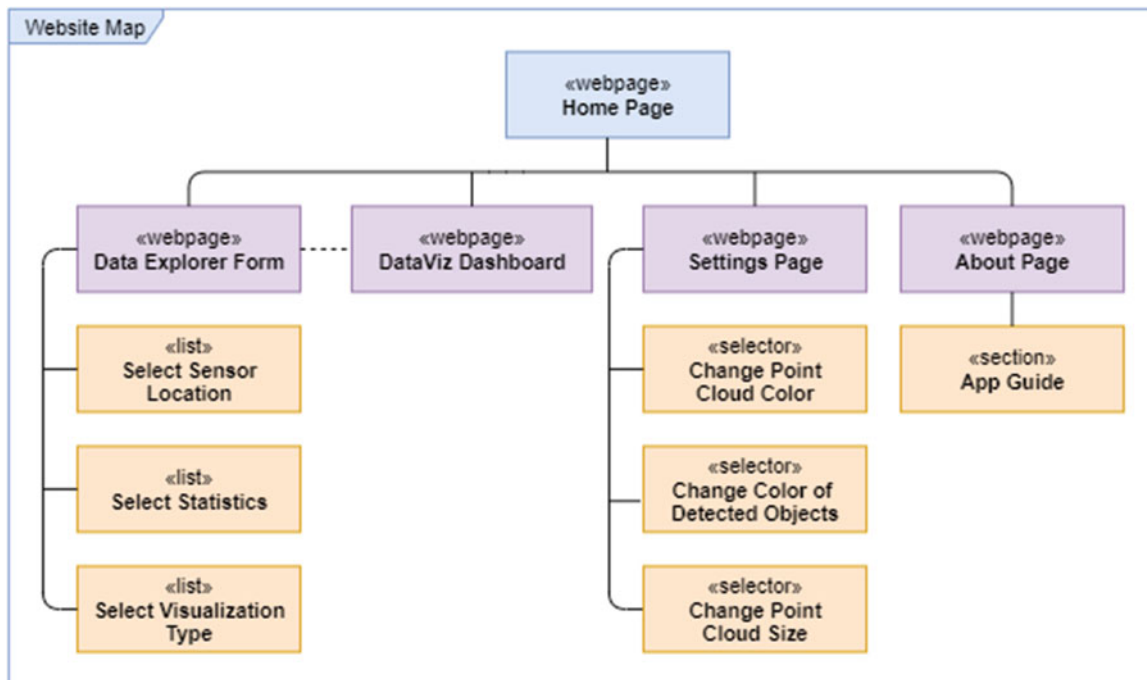


Fig. 36.3 Website map for the LDAT application

The two sensors illustrated in this figure represent the two different topics and data streams that were collected during the implementation of LDAT. Once a point cloud file is obtained by either sensor, the data is collected by the MQTT publisher [10] at the edge which then flows into the crossover switch. It is from the crossover switch that the data is published to the campus network and into Kubernetes and the MQTT Broker server. After the data stream has been published onto the MQTT Broker, the LDAT application is able to subscribe to the broker and pull in data from the two LiDAR sensor data streams.

36.3.3 Application Design

For the application design of the LDAT software we have created diagrams to exemplify the front end of the application. Shown in Fig. 36.3 is a website map of the LDAT application. This exemplifies the main UI layout of the application and some of the different features that can be found on each page of the website. The user is welcomed to the application on the Home Page and from there is able to go to the Data Explorer Form to create a data visualization dashboard. Once the form is submitted, the user can navigate to the Data Visualization Dashboard, which currently contains three different visualizations that provide the user with differing insights into the point cloud data. The user can then go to the Settings Page from the dashboard to make some adjustments to the visualizations, including changing the color and size of the points.

Lastly, the user can go to the About Page which provides a brief description of the application and a short guide the user on how to use LDAT. A more detailed breakdown along with additional screenshots of the UI are provided in Sect. 36.4.

36.4 Prototype

LDAT was developed with simplicity in mind: a web application with a focal point on the visualization and analysis of LiDAR data. The application itself is comprised of five separate web pages that are geared towards providing the end user with a clear and straightforward experience. These pages include the application home, explorer form, visualization dashboard, settings, and about pages. Also included in the application to make navigating the website quick and easy is a navigation bar which includes tabs for each page. The general navigation path begins on the home page by guiding the user in the process of accessing the data.

36.4.1 User Interface

Currently, the application runs on a local host server and is accessible through a web browser. Upon starting LDAT, the user is greeted by the Home Page which can be seen in Fig. 36.4. To proceed the user must click the “Get Started” button at the center which will take users to the Data Explorer form. This form contains three fields which cover the selection of

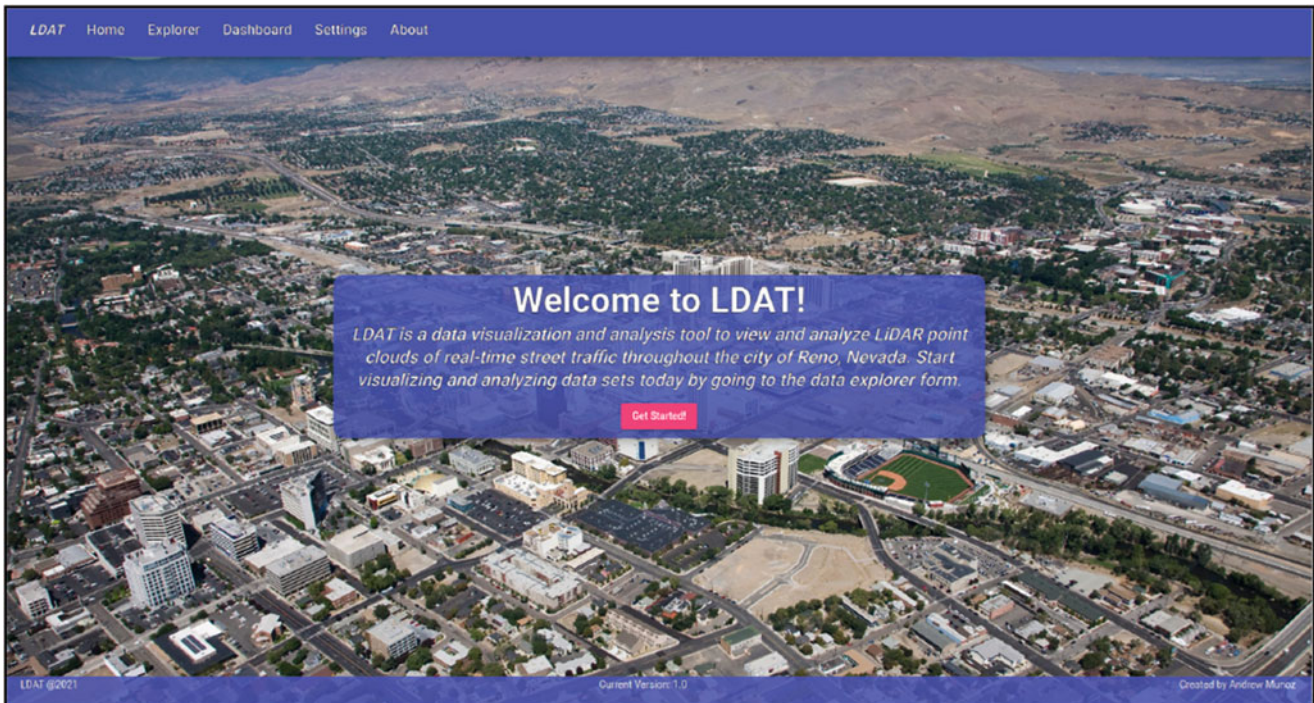


Fig. 36.4 LDAT home page

the sensor location, analysis variables, and visualization type. The sensor location is a list of the active sensors that can be selected and visualized while the analysis variables are the different types of data statistics that users can include for analysis. Lastly, the visualization type includes a list of different types of available visualization charts and graphs that the user can select to view. In this current iteration, LDAT is limited to live data charts that are updated based on live point cloud data feed.

After submitting the form, the user is routed to the Data Visualization Dashboard. The current version of the dashboard contains three different types of visualization charts and graphs. These include a bar chart, single line chart, and near real-time 3D mesh render. The charts are laid out so that the user first sees the live data coming through in the bar and line charts and can then scroll to see the 3D mesh render of the selected sensor location. In addition to the dashboard, there is a Settings Page, which allows for configurations to be made to the 3D visualization such as changing the color of the environment, changing the color of detected objects, and changing the size of the points.

36.4.2 Data Visualizations

The three main data visualizations displayed on the data visualization dashboard are all live data charts that each provide the users with a different key insight into the data.

The first visualization is a bar chart, exemplified in Fig. 36.5, that highlights the current number of objects that are detected in the scene at any given time. In the context of the data, the objects detected in the scene are specifically either vehicles or pedestrians. The line graph, presented in Fig. 36.6, displays the count per second of total objects detected in the scene. In the context of the data, the objects detected in the scene are specifically either vehicles or pedestrians. Additionally, the objects must be moving in order to be detected by the sensor.

The third and final visualization is that of the near real time 3D point cloud mesh shown in Fig. 36.7. It should be noted that the near real time 3D point cloud mesh can be configured in the Settings page through options to change the color of objects and size of the point cloud as displayed in Fig. 36.8. These changes are visualized in Fig. 36.9 with the colors changed and point cloud size changed to 0.1. The render of the 3D point cloud mesh exhibits what is happening at the street level in near real time and allows the user a quick insight into the ongoing street traffic. The scene is also able to be manipulated through mouse and keyboard input. The controls of the scene are as follows: WASD keys move the scene camera, left mouse click rotates the scene, right click pans the scene, and the scroll wheel permits zooming in/out.

36.5 User Study and Results

Software in general should be designed with the end user in mind, and this is especially important for data visualization

Fig. 36.5 Bar chart that displays the current count of each object being detected in the point cloud. The objects that can be detected are vehicles and pedestrians

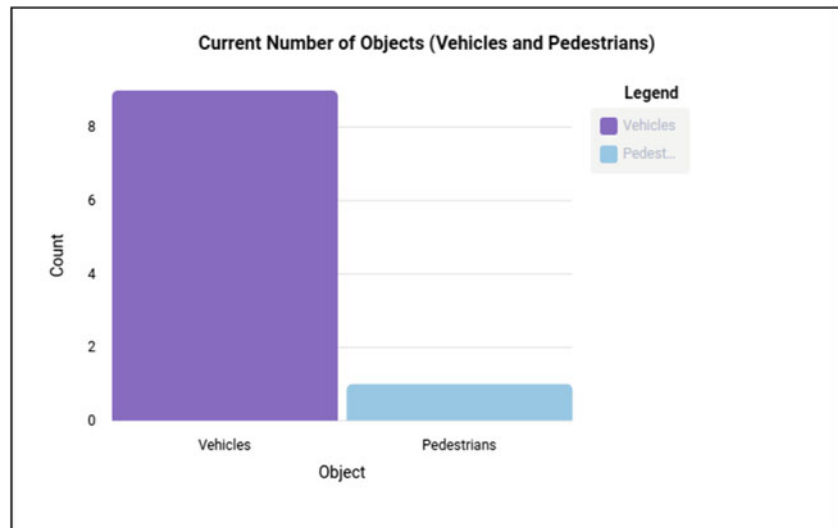


Fig. 36.6 Line graph that presents the total number of objects detected within the point cloud

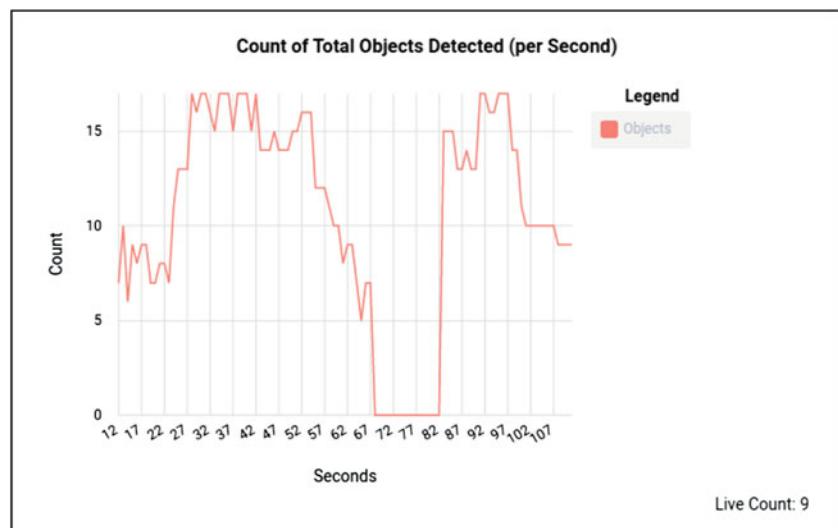
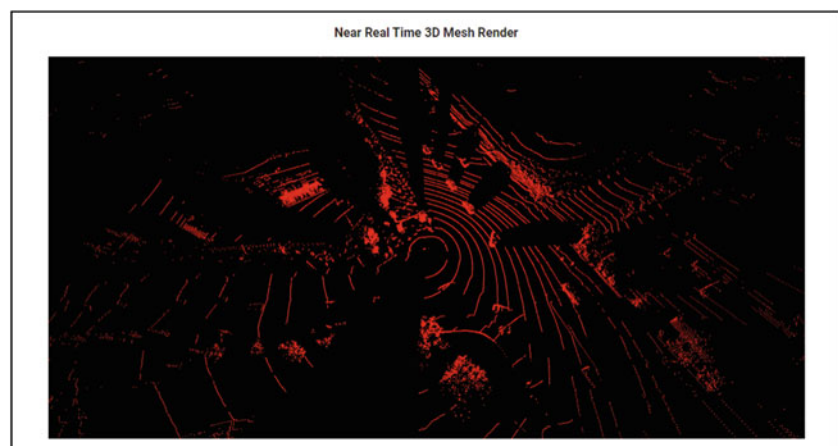


Fig. 36.7 Screenshot of near real time 3D point cloud mesh render which displays the view from the sensor located on the Southeast corner



interfaces such as LDAT. Software interfaces are built to be used by people, therefore, extensive testing is essential to ensure for optimal user experience. In order to properly test the LDAT application, a user study was conducted to

evaluate the usability of the UI. Before the user study could be conducted with human participants, a certification was needed to get approval from the Institutional Review Board

Fig. 36.8 UI layout of the configuration options on the Settings page

Appearance

Change Point Cloud Color: Red (Default) Green Blue

HEX Value: #0000ff

**Note: If the Red, Green, or Blue button is selected the color picker will be updated to match the selection.*

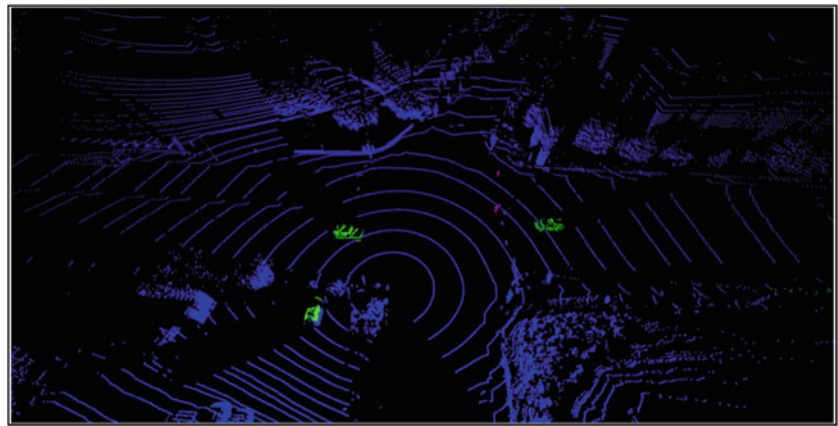
Change Color of Detected Objects:

Car:
 HEX Value: #3efe1f

Pedestrian:
 HEX Value: #ff09ce

Change Point Cloud Size: 0 1

Fig. 36.9 Screenshot of 3D point cloud mesh render scene with object colors changed and a point cloud size of 0.1



(IRB) [13]. Once the study was approved then the testing of the application could proceed.

36.5.1 User Study

There were approximately 20 participants who took part in the study, each of which came from STEM fields and were not necessarily affiliated with software development in the hopes of obtaining varied results. The study itself was conducted in the William Pennington Engineering Building (WPEB) room 436 located on the University of Nevada, Reno campus. A telecommunication option via Zoom was also provided for virtual participation due to consideration made in light of the ongoing COVID-19 pandemic.

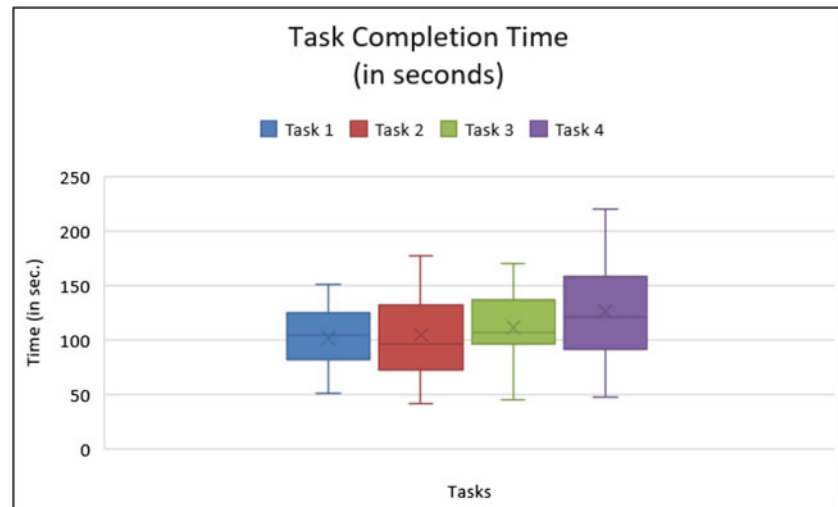
There were three main components to the study including: a pre-questionnaire, a provided set of tasks, and a post-questionnaire. The pre-questionnaire asked participants some demographic information such as their age, gender, and educational status as well as their familiarity with some of the topics covered in the study including LiDAR, data visualiza-

tion, and websites. This was followed by a set of tasks that participants had to complete to test different aspects of the interface which were as follows:

- Fill Out Form (Task 1)—Participants were asked to navigate to Data Explorer Form from Homepage to fill out the form and submit it to view visualizations.
- Change Sensor Location (Task 2)—Participants were asked to change the location of the sensor and navigate around the 3D point cloud mesh render.
- Change the Point Size (Task 3)—Participants were asked to navigate to the Settings page from the Dashboard and change the point cloud size.
- Change Color (Task 4)—Participants were asked to navigate from the Dashboard to the Settings page and change the color values for the point cloud environment and detected objects.

After the completion of the set of tasks outlined above, the participants were then asked to answer a post-questionnaire. The post-questionnaire asked the participants aspects of their

Fig. 36.10 Box plot for task completion time showing the mean average completion time in seconds and the min and max values for each task



general satisfaction with the application and whether or not they felt the application was useful in its current state.

36.5.2 Results

User performance was measured during the completion of the four main tasks which included tasks completion time, number of left mouse clicks, number of right mouse clicks, and number of mouse wheel scrolls. The two measurements that had the most significant results were the task completion time and number of left mouse clicks. Figure 36.10 presents the box plot for the task completion time of each task. At first glance, it is clear that Task 1 had the least amount of variance and therefore was fairly consistent among all the participants (it had no extreme outliers). Task 2 had slightly more variance with a larger upper bound while Task 3 had outliers mostly on the lower bound. Task 4 had the largest range between its minimum and maximum values when compared to the other tasks.

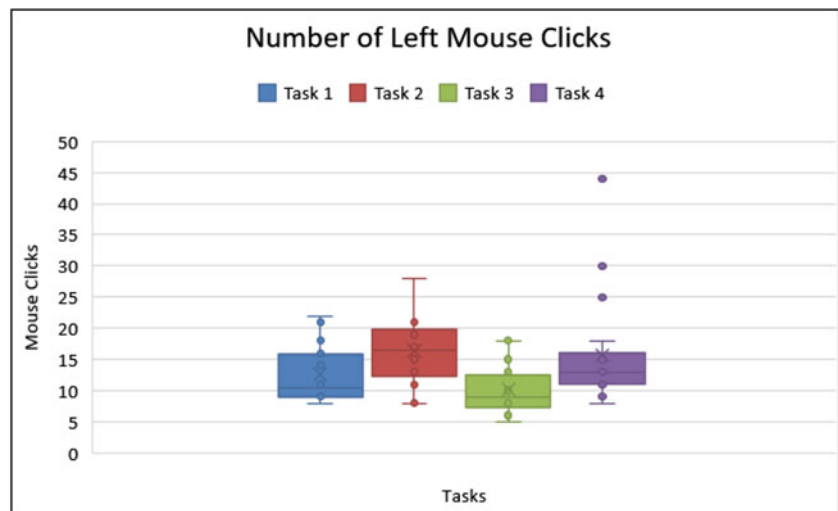
The box plot for the number of left mouse clicks is depicted in Fig. 36.11. Task 1 had only a few outliers in the upper bound while the lower bound was very small and consistent throughout the study. Task 2 had the most variance out of the four tasks with the largest range between the min and max. Lastly, Tasks 3 and 4 had a similar distribution of results. The number of right mouse clicks was the lowest of the four measurements suggesting that it was not an intuitive control mechanism for navigating the 3D mesh. Lastly, the number of mouse wheel scrolls had the most variance but was difficult to measure as it was potentially affected based on various input devices used from virtual and in-person participants.

36.6 Conclusion and Future Work

The overall goal of the LDAT application was to make the LiDAR data easily accessible for researchers and scientists while also providing the data in a format that was both readable and usable to assess traffic patterns at any given time of day. Traffic patterns were assessed through the use of object detection, which currently is able to detect moving object including both vehicles and pedestrians within a scene. The final application included a dashboard which contained three different types of visualizations: bar chart, line graph, and 3D LiDAR point cloud mesh scene render. Each of these were continuously updated in order to keep up with the near real time data stream. Configuration functionality was also added to the application to provide an element of control, customization, and ensure the user was able to adapt the visualizations based on their own preferences or need.

As for future work, there are a several different additions that have been discussed. A useful addition to this visualization would be a statistical analysis or classifier able to detect and display the speed and trajectory for any and all moving objects within a scene. A major development and very useful addition to work would be the fusing of these two streams from two sensors in order to create one large 3D mesh. Another addition could be the inclusion of a user account database which would convert this visualization application into a fully function user portal, capable of logins, controlled access, and security. Finally, a scalable database to store past point cloud captures could be implemented. This would afford users with the ability to go back and look at past data streams and conduct comparative analysis of the current data.

Fig. 36.11 Box plot for number of left mouse clicks per task



Acknowledgments This material is based in part upon work supported by the NSF OAC CC* award #1827186, NSF OAC CC* award #2019164, and the Nevada Center for Applied Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Nevada Center for Applied Research.

References

1. R. Harrap, M. Lato, An overview of LIDAR: collection to application. *NGI Publ.* **2**, 1–9 (2010)
2. C.N. Knaflic, *Storytelling with Data: A Data Visualization Guide for Business Professionals*, 1st edn. (Wiley, Hoboken, 2015)
3. Point Cloud Library, The PCD (Point Cloud Data) file format. https://pointclouds.org/documentation/tutorials/pcd_file_format.html. Accessed April 6 2021
4. K. Mohring, T. Myers, I. Atkinson, A controlled trial of a commodity sensors for a streetlight-mounted traffic detection system, in *Proceedings of the Australasian Computer Science Week Multiconference, ACSW'18* (Association for Computing Machinery, New York, 2018)
5. E. Maravelakis, A. Konstantaras, K. Kabassi, I. Chrysakis, C. Georgis, A. Axaridou, 3DSYSTEK web-based point cloud viewer, in *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications* (2014), pp. 262–266
6. S. Kulkarni, M. Chandrashekaraiyah, S. Raghunandan, 3D annotation tool using LiDAR, in *2019 Global Conference for Advancement in Technology (GCAT)* (2019), pp. 1–4
7. Angular, What is angular? (2021). <https://angular.io/guide/what-is-angular>. Accessed May 27, 2021
8. Flask, Flask: user's guide (2021). <https://flask.palletsprojects.com/en/2.0.x/>. Accessed May 27, 2021
9. Velodyne Lidar, Ultra puck (2021). <https://velodynelidar.com/products/ultra-puck/>. Accessed July 16, 2021
10. MQTT, MQTT: the standard for IoT messaging (2021). <https://mqtt.org/>. Accessed June 3, 2021
11. Docker, What is a container? (2021). <https://www.docker.com/resources/what-container>. Accessed June 4, 2021
12. Data infrastructure for advanced traffic sensor edge networks, in *PEARC'21: Practice and Experience in Advanced Research Computing*. (Association for Computing Machinery, New York, 2021)
13. Research Dataware, LLC., IRBNet: innovative solutions for compliance and research management (2021). <https://irbnet.org/release/index.html>. Accessed 26 July 2021

Autumn Cuellar, Yifan Zhang, Sergiu M. Dascalu,
and Frederick C. Harris, Jr.

Abstract

Social media is a popular pastime in our current society. There are numerous and diverse social media applications available to use. The study presented in this paper aimed to determine which application is easiest to use and most preferred by users. The apps considered were four of the most popular existing social media applications: Facebook, Twitter, Instagram, and TikTok. The participants in the study were timed while publishing a picture, a text, and a video using each application, and were asked to comment and provide their level of linking on each post they made. Post-questionnaire answers reveal that the majority of participants found Facebook the easiest and more preferable application to use. Experiment results also show that publishing videos on Facebook is quicker than on the other three media apps. On the other hand, publishing pictures and liking/commenting take about the same time on all four apps considered in our study.

Keywords

Human computer interaction (HCI) · Social media · Apps · User interface evaluation · Facebook · Twitter · Instagram · TikTok · Time-based evaluation · Heuristic evaluation

37.1 Introduction

Throughout the years the way people communicate has changed. People used to write handwritten letters or use the telephone to communicate before the creation of the Internet. The Internet allowed for the exchange of messages to become much quicker, using an electronic medium and format. This has been further supported by the creation of a vast number of social media platforms [1]. There are many different types of such platforms, catering to the many diverse interests of the users. Some applications and associated platforms include Facebook [2], Twitter [3], Instagram [4], and TikTok [5]. All these apps have one thing in common: they offer modern means of communication that greatly support interaction among users. People are able to communicate with anyone around the world thanks to these applications. Among the numerous existing apps, Facebook is the most popular, as shown in Fig. 37.1. However, considering human computer interaction (HCI) principles we still don't know exactly which application is the most preferred.

Many social media related topics have been explored and researched throughout the years [1], [6], and [7], mainly in the area of information systems. These topics vary but, in the end, undertaking them has had the purpose of improving social media. The topics included looking at the behavioral aspects of social media [8], providing reviews and recommendations, using social media for organization purposes, and using it as a marketing tool [9]. Other topics focused on online blogs and communities, risks related to them, positive and negative effects, the relationship between usage and value creation, the use of social media to share information during disasters, traditional versus social media, utilization in a political context and public administration, and looking at the existing social media models [10]. This

A. Cuellar · Y. Zhang · S. M. Dascalu · F. C. Harris, Jr. (✉)
Department of Computer Science and Engineering, University of
Nevada, Reno, Reno, NV, USA
e-mail: acuellar24@nevada.unr.edu; yfzhang@nevada.unr.edu;
fred.harris@cse.unr.edu

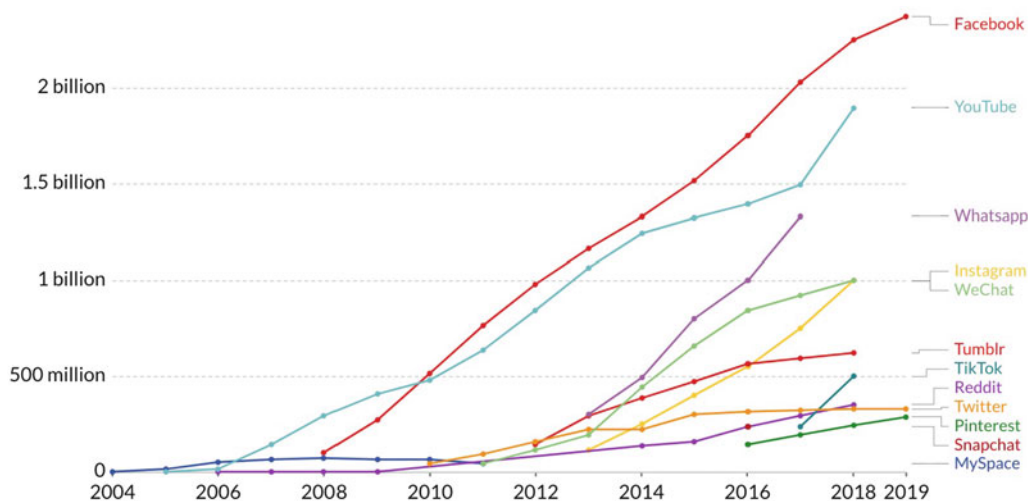


Fig. 37.1 Popular social media sites between 2004–2019 [13]

means social media has proved to be significant and useful in our society. However, there has not been a lot of research when it comes to the user interface of social media apps [11]. Yet, it is well known that the interface can make or break a user's experience. Evaluating a user interface can be done through timing tasks, but it can also be done heuristically. The heuristic evaluation consists of evaluators commenting on the interface in question. Through some experiments, Jakob Nielsen and Rolf Molich found that this evaluation is easy to plan for, inexpensive, fairly quick, and people can be more motivated to do it. The downside is since there are only comments from participants, they may not always help on how to fix potential problems [12].

We decided to use timing and heuristic evaluation in a user study focusing on social media apps. The apps that have been used were Facebook, Twitter, Instagram, and TikTok. At the same time, with the popularity of social networks, different social apps reflect different social characteristics. Take content as an example: Facebook and Twitter focus on the delivery of text messages, Instagram is used for image sharing, and TikTok focuses on the sharing of short videos. For instance, some famous people like to express their opinions using Twitter and Instagram is famous for celebrities sharing their pictures. If we take the social type as an example, Facebook and Instagram are mainly used to socialize with friends that they already know in real life, while Twitter and Tiktok are more inclined to share content between strangers. The authors of this paper wanted to see which one of these four major social media apps is the easiest to use and which is the most preferred by users. Therefore, we decided to conduct a user study involving Facebook, Twitter, Instagram, and TikTok. The main goal was to analyze the differences between these social applications and identify the deeper reasons for their popularity.

The remaining of this paper is structured as follows: The user study methodology is described in Sect. 37.2, results and discussion are provided in Sect. 37.3 and conclusions are presented in Sect. 37.4.

37.2 Methodology

There are several components of this study that we designed, as described in detail in this section.

37.2.1 Participants

Before the start of the experiment, we had some expectations for the participants. First, to avoid the impact of unfamiliarity with the tools on the experimental results, we hoped that the participants in this study had some experience using the Internet and smartphones. Second, we hoped that the participants can be distributed among different age groups. We believe that there are significant differences in the preferences and interests in social media among different age groups. In the end, we recruited a total of 10 participants—please note this study was conducted during a peak of the pandemic period, hence it was harder to recruit participants. These participants ranged in age from 18 to 45, with the majority in their 20's. Eight participants were female and two were male. The majority of the participants were recruited using Facebook.

37.2.2 Apparatus

The software environments and applications in this study were Facebook, Twitter, Instagram, and TikTok. Their versions of the software were the latest available when the user

Fig. 37.2 The picture [14] published by participants in Task 1



study took place. The functions of these four social media applications are somewhat similar. For example, they all support the publishing of text, image, and video information. They also have certain differences. Therefore, in this user study, these four apps were used as our research subjects. The study was conducted through Zoom mostly on desktop and laptop computers. This was to protect the participants from the risk of COVID 19. Smartphones were used only when the participants could not complete the user study tasks on a computer web-browser.

37.2.3 Procedure

There were several steps that each participant went through during the experiment. We first explained what the experiment is about and what are the tasks the participants will be asked to execute. The participants were informed that their participation is voluntary and they can terminate it anytime. The participants signed a consent form if they still wanted to continue after the introductory explanation. The experiment started with the participants answering an entry (pre-usage) questionnaire without instructions.

Next, we instructed the participants to publish, as a first task, a given picture and text. This was done with all four apps and time was recorded for how long the publishing process took for each app. The experiment continued onto the second task, which consisted of publishing a given video and text. This process was the same as in the previous task. When this task was finished, participants were asked, in the third task, to view, like, and comment on a publishing of their choosing that included a picture or a video. The three tasks were all done one app at a time so that the participants did not have to constantly change tabs. The experiment ended after the participants finished the exit (post-usage) questionnaire given to them.

37.2.4 Tasks

In this user study, we asked the participants to complete three tasks. The first task (Task 1) was to publish a post with a given picture and text on Facebook, Twitter, Instagram, and TikTok. The given picture and the text, which was “I am doing an HCI experiment. Social media evaluation. HCI is cool!”, are shown in Fig. 37.2. The second task (Task 2) was to publish a given video and related text content on each of these four applications. We picked a TikTok video to be published. The video, picture, and text were given to the participants through email. The final task (Task 3) was to browse and interact at will with the content published by other users on the four apps included in the study.

We provided the same text, pictures, and videos to the participants of the user study, and did not introduce the user interface of the four apps in advance. In the process of publishing pictures and videos, we divided users into experienced and inexperienced categories, and recorded the time of their publishing process. Also, we studied the user-friendliness of the different applications for publishing new content. Furthermore, we designed a questionnaire to assess the participants’ experience of interacting with other users on the four apps.

We believe that these three tasks can simulate the daily behavior of most users on social apps, so the results of this user study could be meaningful and representative.

37.2.5 Design

We designed this user study to address the research questions at hand. The independent variable was the type of social media app, with factors (or levels) the four social media apps studied. The dependent variable was the time taken to complete the posts. This metric and the participants’

written preferences helped identify which of the applications was the most preferred. We used an analysis of variance (ANOVA) [15] to determine the level of preference for each of the four applications. This user study had also some random and confounding variables. The random variable was the participants' experience with the applications. Some participants knew more about the applications than others. We wanted this because we intended to see the preferred application among all users, not just the experienced ones. The confounding variables were the internet speed and the type of technology the participants used. These two factors can cause a delay in publishing times when they should be small. This is because internet speed can slow down publishing a post to an application or a small screen can increase the typing time. We tried to minimize the influence of these confounding variables as much as possible. The user study was evaluated using the within subjects method. This required all participants to complete all the tasks on all apps.

37.3 Results and Discussion

The collected data represents the time, in seconds, it took to each participant to complete each task on each of the four social media apps. Tables 37.1, 37.2, and 37.3 show how much time took each participant to perform on each task. Figures 37.3, 37.4, and 37.5 are the graphical representations of the results shown in the tables mentioned above. Participant 1 was unable to complete any of the three tasks on TikTok. This problem resulted because we did not have all suitable materials at the time to complete the tasks. To keep this participant's data in the calculation, we decided to replace its times with the mean of all the participants' TikTok time.

The time it took for different participants to complete the same task varied greatly, most likely due to the following reasons: (1) There are different ways to try to complete the tasks. For example, Instagram does not support uploading pictures and videos on the web. The first participant used the developer mode of the Chrome browser to simulate the mobile phone on the web to complete the task, while other participants used their mobile phones to complete the task. (2) The internet speed of different participants had certain differences. For example, the Wi-Fi of the fourth participant encountered some minor problems when participating in this user study, which affected the time to complete the task. (3) Differences in the user interface are likely causes as well. TikTok's web interface is not friendly for uploading, and participants tended to fail to find the uploaded page in time. TikTok's website was updated when participants 8–10 were performing their user study. They were able to do every task on their web browsers while the other participants had to comment using their phones.

Table 37.1 Publishing image and text times (in seconds) for each participant (Task 1)

Participant ID	Facebook	Instagram	Twitter	TikTok
1	57	373	53	70
2	37	109	38	104
3	41	30	12	80
4	50	47	30	51
5	35	79	28	63
6	71	27	23	77
7	54	18	12	46
8	126	128	49	72
9	27	63	21	53
10	35	73	21	80

Table 37.2 Publishing video and text times (in seconds) for each participant (Task 2)

Participant ID	Facebook	Instagram	Twitter	TikTok
1	78	146	75	35
2	26	19	11	21
3	20	52	48	28
4	13	156	68	88
5	18	54	17	24
6	29	47	22	22
7	45	21	19	25
8	23	138	27	34
9	29	61	27	43
10	20	19	15	35

Table 37.3 Liking and commenting times (in seconds) for each participant (Task 3)

Participant ID	Facebook	Instagram	Twitter	TikTok
1	33	107	35	50
2	30	125	139	113
3	33	11	48	85
4	115	50	19	10
5	118	45	44	67
6	68	51	15	37
7	29	5	8	19
8	22	8	25	41
9	53	45	82	47
10	44	20	19	30

37.3.1 Data Analysis

From the collected data, there are certain differences in the completion of a given task on different software by the same participant. To analyze this difference, we used a one-way ANOVA test. Table 37.4 shows the one-way ANOVA analysis results of Task 1.

According to the results of the one-way ANOVA (shown in Table 37.4), the f-value was 2.54 and the p-value was 0.07. With a p-value greater than 0.05, the null hypothesis could not be rejected. Thus, we concluded that there was no significant

Fig. 37.3 The results of Task 1

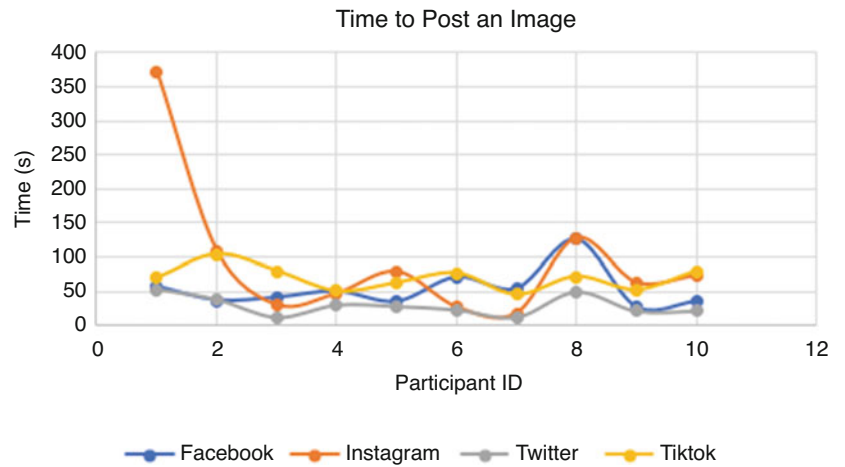


Fig. 37.4 The results of Task 2

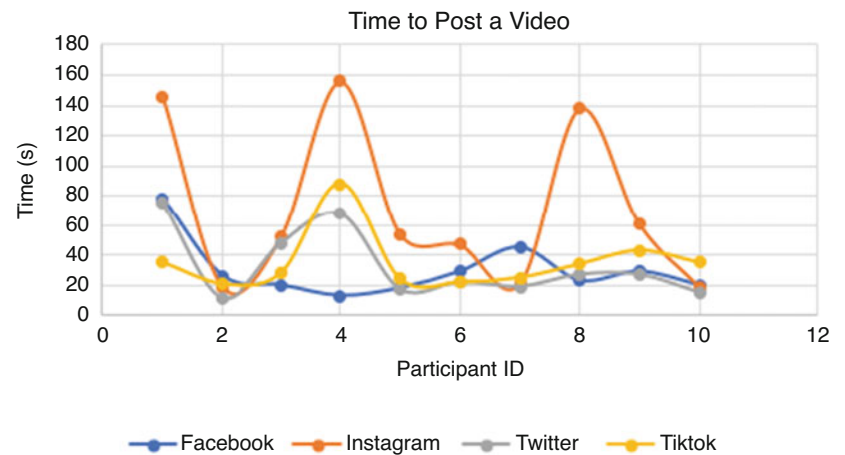


Fig. 37.5 The results of Task 3

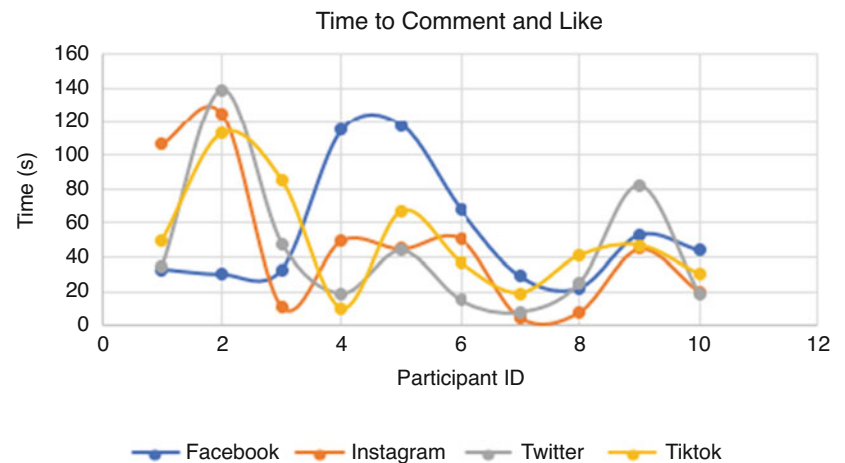


Table 37.4 Task 1 one-way ANOVA results

	SS	df	MS	F	p-value
Condition	23094.09	3	7698.03	2.54	0.07
Error	109253.80	36	3034.82		
Total	132347.89				

difference in the time taken by the participants to complete the task of publishing posts with pictures on the four apps.

By inspecting the data, we observed that it usually took 15–35 s for a participant to complete Task 1. On some apps, the participants may have spent more time due to their unfamiliarity with the user interface. But, in general, the user

Table 37.5 Task 2 one-way ANOVA results

	SS	df	MS	F	<i>p</i> -value
Condition	11294.26	3	3764.75	3.58	0.02
Error	37908.76	36	1053.02		
Total	49203.03				

Table 37.6 Task 3 one-way ANOVA results

	SS	df	MS	F	<i>p</i> -value
Condition	671.19	3	223.73	0.16	0.92
Error	49257.88	36	1368.27		
Total	49929.08				

interface of each of the four apps may have caused the participants to take more time to complete Task 1. At the same time, according to the pre-questionnaire, most participants indicated that they have more experience with Facebook than with other apps. However, from the ANOVA results, Facebook was not significantly different from the other apps. Hence, it can be considered that the support for publishing posts with a picture was very good on all four apps.

Task 2 was similar to Task 1, in that the same text was published, but the pictures were replaced by videos. Table 37.5 shows the one-way ANOVA analysis results of Task 2.

According to the results of the one-way ANOVA (shown in Table 37.5), the *f*-value was 3.58, and the *p*-value was 0.02. With a *p*-value smaller than 0.05, the null hypothesis could be rejected. This means that the time for participants to complete Task 2 was significantly different on the four apps.

By analyzing the collected data, the average time for participants to complete Task 2 on Facebook, Instagram, Twitter, and TikTok was 30s, 71s, 33s, and 36s, respectively. It can be concluded that the user interfaces of Facebook, Twitter, and Tiktok are roughly equivalent in user friendliness, while the Instagram interface is more difficult for users to publish videos. It is worth mentioning that Instagram is the only one of the four apps that does not support publishing pictures and videos on the web, so the participants had to use their mobile phones to complete this task. It can also be concluded that Facebook is slightly better at publishing videos than the other apps. Nevertheless, more research needs to be done to further support this conclusion.

Task 3 asked the participants to like and comment on posts of their interest. Table 37.6 shows the one-way ANOVA results of this task.

According to the results of one-way ANOVA (shown in Table 37.6), the *f*-value was 0.16, and the *p*-value was 0.92. With a *p*-value greater than 0.05, the null hypothesis could not be rejected. Judging from the results of the one-way ANOVA, the four apps allowed participants to easily like and comment on other people's posts. Overall, all four apps seemed to be doing a good job in allowing participants to find content that interests them.

37.3.2 Questionnaire Analysis

Besides the one-way ANOVA results of the data, we found interesting the answers to Questions 2 and 4 of the exit questionnaire. Question 2 asked which of the tested social media apps the participant prefers. Half of the responses (5 out of 10) indicated Facebook, as seen in Fig. 37.6. This is not surprising because most of the participants indicated they use Facebook regularly. The results may have turned out differently if participants were recruited using all four apps, not just Facebook.

Question 4 asked which of the tested social media apps the participant found easier to use. Figure 37.7 shows that the majority of the participants (8 out of 10) indicated Facebook. There were many reasons for this decision. Participant #2 said it was because "It's what I'm used to." Other participants, numbers #4 and #10, noted that on Facebook it is "easier to find how to post" and that its interface is "user friendly and self explanatory," respectively. This makes sense, since Facebook's publishing function was immediately available (right in front of the user) once logged in. Interestingly, Twitter has the same layout as Facebook, but only 1 of the 10 participants considered it the easiest to use. More participants, with different social media experience, are needed to fully confirm that Facebook is the most user-friendly app among those tested.

37.4 Conclusions and Future Work

Social media participation is an important personal activity in people's daily lives. Obviously, given its significance, the users need related apps that are easy to use. This user study showed that Facebook was considered the easiest to use and was the most popular among participants. The reason for this maybe because of the small participant size of this study and our recruiting of participants mainly using Facebook. Note also that based on data analysis all three other apps also fared pretty well. Thus, further research is needed to confirm that Facebook is the easiest to use, which may have also led to it being the most preferred. This extended work could be done by extending the user study described in this paper, with more participants recruited in more diverse ways. Furthermore, the study can have more applications added, such as Reddit or other apps popular in specific countries. Having more research results could be beneficial to many developers and users. The developers could use aspects from the most popular applications to create more user-friendly social media apps. In turn, the users, the ultimate beneficiary of social media apps, could utilize the most popular apps to more effectively publish their content of interest and also to get more followers.

Fig. 37.6 The results of Question 2 of the exit questionnaire (“Which of the 4 apps do you prefer?”)

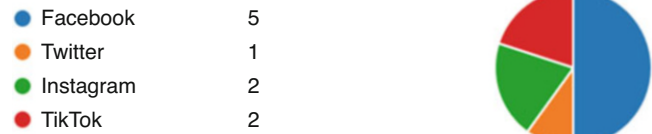


Fig. 37.7 The results of Question 4 of the exit questionnaire (“Which of the 4 apps was the easiest to use?”)



Acknowledgments This material is based in part upon work supported by the National Science Foundation under grant number IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. S. Greenwood, A. Perrin, M. Duggan, Social media update 2016. *Pew Res. Center* **11**(2), 1–18 (2016)
2. R. Caers, T. De Feyter, M. De Couck, T. Stough, C. Vigna, C. Du Bois, Facebook: a literature review. *New Media Soc.* **15**(6), 982–1002 (2013)
3. D. Murthy, *Twitter* (Polity Press, Cambridge, 2018)
4. L. Manovich, Instagram and contemporary image. Nova Iorque: CUNY (2017)
5. J. Herrman, How tiktok is rewriting the world. *The New York Times*, vol. 10 (2019)
6. A. Mayfield, *What Is Social Media*. iCrossing.com: iCrossing (2008). <https://tavaana.org/sites/default/files/what-is-social-media-uk.pdf>
7. M. Duggan, N.B. Ellison, C. Lampe, A. Lenhart, M. Madden, Social media update 2014. *Pew Res. Center* **19**, 1–2 (2015)
8. K. Casler, L. Bickel, E. Hackett, Separate but equal? A comparison of participants and data gathered via amazon’s mturk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav.* **29**(6), 2156–2160 (2013)
9. R. Felix, P.A. Rauschnabel, C. Hinsch, Elements of strategic social media marketing: a holistic framework. *J. Bus. Res.* **70**, 118–126 (2017)
10. K.K. Kapoor, K. Tamilmani, N.P. Rana, P. Patil, Y.K. Dwivedi, S. Nerur, Advances in social media research: past, present and future. *Inf. Syst. Front.* **20**(3), 531–558 (2018)
11. A. Krishnan, A. Ganesh, S. Gayathri, S. Koushik, M.V. Nair, G. Narayanan, Development of social media user interface portal for maintaining students portfolio, in *Inventive Communication and Computational Technologies* (Springer, Berlin, 2021), pp. 757–764
12. J. Nielsen, R. Molich, Heuristic evaluation of user interfaces, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1990), pp. 249–256
13. D. Chaffey, Global social media statistics research summary 2021 (2021). <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>
14. S. Hollingsworth, Top 9 benefits of social media for your business (2019). <https://www.searchenginejournal.com/social-media-business-benefits/286139/#close>
15. E.R. Girden, *ANOVA: Repeated Measures*, no. 84 (Sage, London, 1992)

Software Interfaces for New Vehicle Operating Cost Models Used in Economic Analysis of Transportation Investments: A User Study

38

Arjun V. Gopinath, Hudson Lynam, Rami Chkaiban, Elie Hajj, and Sergiu M. Dascalu

Abstract

Estimating vehicle operating costs (VOCs) allows individuals and organizations to make informed decisions about vehicle usage. As a wide variety of cars and roadway conditions exist, a relatively large amount of input must be provided to any VOC model. Developed as part of a civil engineering research project funded by the U.S. Department of Transportation, five VOC models were run initially in Microsoft Excel. While this early solution was practical and operational, to improve usability, including efficiency of data input, flexibility of running the models, and presentation of results, an alternative solution, a web-based application, was also designed and implemented. The VOC models that can be run on both Excel and the web-based application are: fuel economy, oil consumption, tire wear, mile-age-related vehicle depreciation, and repair and maintenance. This paper briefly introduces the VOC models, describes the two software interfaces created for running them, and presents the results of a user study conducted to evaluate and compare the two interfaces. The study involved 17 participants and focused on usability characteristics and the quality of the user experience. The independent variable was “user interface,” with two test conditions: Excel interface and web-based interface. The participants answered an entry questionnaire, performed tasks using both interfaces, and completed an exit questionnaire. Several dependent variables

were measured and analyzed, including task completion time, number of incorrect data entries, and number of clarification questions asked to the user study facilitator. The results obtained showed that the web-based solution consistently outperformed the Excel-based solution, although the latter received some positive feedback as well.

Keywords

Vehicle operating costs · Transportation models · Economic analysis · Web application · User interfaces · User study

38.1 Introduction

Vehicle operating costs (VOCs) are important performance indicators for transportation networks. Decreasing VOCs is a common goal for such networks, and network modifications are often judged as successes or failures based on VOC-related criteria. For example, the Federal Highway Administration (FHWA) applies the Highway Economic Requirements System (HERS) model to estimate highway conditions and performance under alternative investments in highway capacity expansion and preservation [1]. Having accurate, updated estimation models for VOCs to reflect changing technologies and research is crucial to the success of such projects. Additionally, VOCs can be used by individual drivers to plan personal vehicle usage.

The VOCs models described in this paper were developed by a team from the University of Nevada Reno (UNR) Civil and Environmental Engineering Department [2]. They originally implemented their models in Excel, but faced serious security concerns regarding its usage. The goal of the models was to be used and tested by experts and laypersons, but the models could not be freely distributed due to concerns of the

A. V. Gopinath · H. Lynam · S. M. Dascalu (✉)
Department of Computer Science and Engineering, University of Nevada, Reno, USA
e-mail: avettathgopinath@nevada.unr.edu; hlynam@nevada.unr.edu; dascalus@cse.unr.edu

R. Chkaiban · E. Hajj
Department of Civil and Environmental Engineering, University of Nevada, Reno, USA
e-mail: elieh@unr.edu

underlying data being stolen. To deal with this issue, the Civil Engineering team approached the UNR Computer Science and Engineering Department, which developed a website that allowed the use of the models as a black-box system.

Beyond security concerns, the Excel version faced inherent limitations, related in particular to scalability, distribution, and feature availability. The website was designed to overcome these limits. Websites are also known to enjoy many advantages [3] with regards to software development and deployment. The website was developed with several current web development technologies, including Ajax and JavaScript. The software development team also enjoyed frequent collaboration with the civil engineering research group, which provided vital expert feedback and quality control. The website went through testing with a selected group of users and then recently released to the public via UNR's web services [4].

A user study comparing the Excel interface to the website interface was conducted to test whether the web interface achieved its development goals. This paper reports on the results of that study, which show a decisive response in favor of the web interface. It also provides a background on VOCs and the specific models developed, and presents the structure of the two interfaces being tested.

The place of this study is in a larger context: porting applications from proprietary software such as Excel to an online environment. A primary contribution is the insight and experience gained from that process. Additionally, our study provides valuable insight into the experience of end-users with respect to informing the design and implementation of the software. Given the results of the user study, this work described in this paper can be considered successful, and the lessons learned may assist in similar development projects.

This paper, in its remaining parts, is organized as follows. Section 38.2 provides a background on VOC model development, applications, and challenges. Section 38.3 briefly explains the design and flow of the Excel and website interfaces. Section 38.4 details the user study. Section 38.5 presents the analyses performed and the results of the study. Section 38.6 contains a discussion of the results. Finally, Sect. 38.7 presents our concluding remarks and outlines possible directions of future work.

38.2 Vehicle Cost Operation Models

38.2.1 Background on VOC Research

As mentioned, VOCs are indicative of a transportation network's performance. VOCs include user costs for fuel consumption, tire wear, oil consumption, vehicle maintenance

and repair, and mileage-related depreciation. A recently completed study, which was funded by FHWA, developed improved VOCs estimation models for different vehicle types, traffic conditions, and highway design scenarios as defined by roadway properties such as lane numbers, speed limit, curvature, and grade [2].

While some advanced estimation models are currently available to examine VOCs associated with driving cycles, the current models have a limited number of driving cycles across a range of functional classes [5]. Therefore, a comprehensive database of driving cycles was established for different vehicle types, road properties (e.g., different grade levels), and traffic conditions (e.g., different level of service). This database contained more than two thousand driving cycles for an array of vehicles, including six small light duty vehicles (SLDVs), five large light duty vehicles (LLDVs), three two-axle trucks, three buses, one single-unit truck, and one combination truck with two different gross vehicle weights (GVWs). Vehicle fuel sources included gasoline, diesel, gasoline-ethanol blend of up to 85% ethanol (E85), liquefied natural gas (LNG), and hybrid-electric (HE) – resulting in 30 combinations of vehicle type and fuel source.

38.2.2 Vehicle Cost Operation Models

The approach taken for evaluating VOC was to develop full vehicle models that represent the vehicle fleet over a wide range of driving cycles. These physics-based vehicle models have the ability to include the physical characteristics of the vehicle, the driving cycles, and the physical features of the road.

The strength of this approach is that it does not require extensive retesting of the current vehicle fleet, where the data become obsolete after time. Rather, a representative fleet of simulated vehicles is developed and run through the desired driving cycles to estimate fleet averages. The effects of actual highway conditions such as grades, curvatures, and road roughness are modeled and evaluated for each vehicle. As time progresses and new technologies are developed and inserted into the fleet, the models can be updated to include the respective changes.

Due to the numerous parameters influencing the VOCs calculation, the fuel economy and tire wear models were each dissected into three separate models. The rest of the VOCs models (i.e., oil consumption, vehicle maintenance and repair, and mileage-related depreciation) depended on either fuel economy or tire wear. More details on the factors considered in the fuel economy equations that were developed as part of our study can be found in Hajj et al. 2018 [2].

38.2.3 VOC Models Application and Challenges

Since VOCs models can be used in various applications, the ability to share them with other users through an easy-to-use software becomes advantageous. Thus, our research group decided to code the models into an application to ease its accessibility and use.

Several software packages were explored, however MATLAB was the chosen software to start with. This decision was made because MATLAB was used to process and analyze the driving cycles data, and for establishing the VOCs models structure. The research group expected originally to have a smooth implementation of the various VOCs models. However, importing all the different VOCs models structures, combined with the corresponding constraints and parameters, led to complications. This was experienced because the intent of the MATLAB application was to have a console input/output with no user interface. Additionally, this type of applications require the users to install MATLAB for the application operation.

Considering all the challenges associated with MATLAB usage, our research group decided to use Microsoft Excel for the application development. This solution was adapted for several reasons: first, most of computer users around the world are familiar with Microsoft Excel; second, similar to MATLAB the application does not require creating (through code) a user interface, which makes the development easier; and third, most of the users around the world have Microsoft Excel on their devices which helps the application to be more reachable and operable by a larger group of people. Nevertheless, while implementing the VOCs models into Microsoft Excel several issues were faced, in particular the lack of predefined functions (e.g., integration), which raised concerns about future developments and updating/changing the models.

Hence, to overcome the issues of practicality, accessibility, sustainability, flexibility, and coding limitations, the research group decided to shift the development of the VOC application as a web platform. Even though this decision induced a larger workload as a user interface need to be created, all previously encountered problems were solved. In addition, the web solution comes with several other benefits, such as easy access by users worldwide, audience overview, feedback collection, and other.

38.3 Excel and Web Interfaces

38.3.1 VOC Models

The application is split up into five modules: Fuel Economy, Oil Consumption, Tire Wear, Mileage-Related Vehicle De-

preciation, and Repair and Maintenance. The Fuel Economy module must be completed to access the Tire Wear or Oil Consumption modules. This is because variables defined in Fuel Economy are required for these modules. In turn, the Mileage-Related Vehicle Depreciation and Repair and Maintenance modules require the Oil Consumption module and the Tire Wear module respectively to be completed. Figure 38.1 displays the relationships among modules using a data flow diagram unavoidable.

38.3.2 Excel Interface

The Excel interface consists of several Excel files, with a sheet within each file representing the model inputs (e.g., as shown in Fig. 38.2). Additional sheets include instructions on usage and protected sheets that contain data used to create the outputs. After entering the inputs, the outputs are displayed dynamically on that same sheet. While there are features of the modules not implemented in the Excel files that are fully implemented in the website version (such as accounting for downslope grades) only features common to both interfaces were tested in the user study presented in this paper.

38.3.3 Website Interface

Currently, the website is in development for additional advanced features (such as batch processing) but the five main modules are fully implemented and available on the public site. The site map of the public website consists of 10 fully developed web pages. The relationships between pages mirrors the requirements between modules in the VOC Excel Model. A screenshot of the web interface is shown in Fig. 38.3.

38.4 User Study

38.4.1 Participants

The user study [6] was conducted with 17 participants from social and educational circles at UNR. The only requirements to participate were a basic working knowledge of Excel and familiarity with using websites. Among the participants there were 11 men and 6 women, and their ages were between 22 and 35 years.

38.4.2 Technology

The study was conducted on a Dell Inspiron 15 5000 Series Laptop with an i7 – 8th Generation 64-bit processor. The soft-

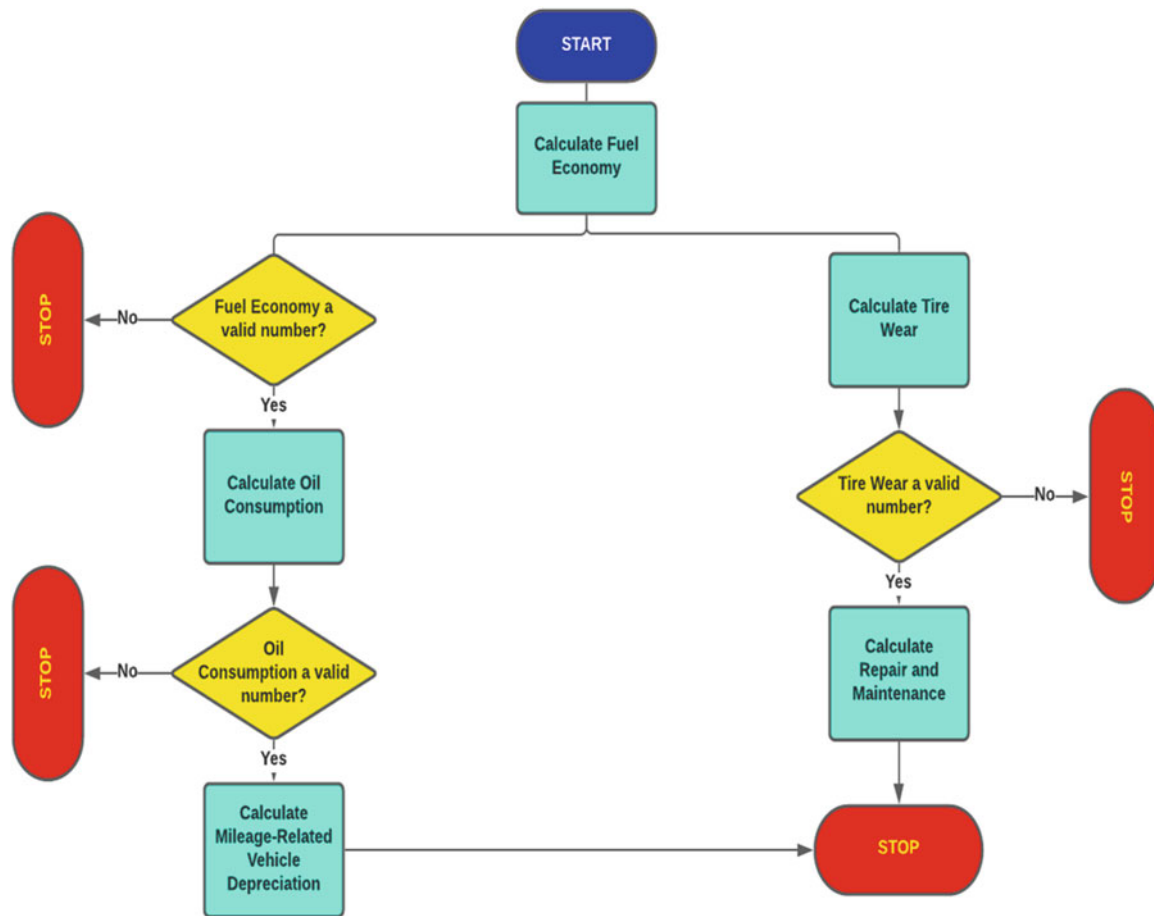


Fig. 38.1 VOC model data flow diagram

were used was Windows 10 Home, Google Chrome version 86 for the website interface, and Microsoft Excel 2016 for the Excel interface. Due to COVID-19 concerns, the study was conducted over Zoom, using screen sharing technology that allowed participants to interact remotely with the interfaces. An additional benefit of using Zoom was that it kept the Excel models on a single machine, as sending it to participants would not be possible due to security concerns.

38.4.3 Procedure

The procedure for the user study was as follows. First, each participant was welcomed to the study, briefly explained the goals and the structure of the experiment, and asked to sign a consent form. Second, the participant filled out an Entry Questionnaire. There were six questions in this questionnaire: (1) “What is your age?”; (2) “What is your gender?”; (3) “What internet browser you use most often?”; (4) “How comfortable are you with navigating websites?”; (5) “How comfortable are you using MS Excel?”; and (6) “How frequently do you drive a vehicle?”. Third, the VOC models

were explained, and any preliminary questions the participant had were answered. Fourth, each participant completed three tasks (detailed in Sect. 38.4.4), reporting the results of each task verbally before moving on to the next task. Finally, each participant filled out an Exit Questionnaire, which asked the participants to indicate, for each of the two interfaces, their agreement or disagreement on a scale 1 to 5 (“1” being strongly disagree and “5” strongly agree) with the following statements: (1) “I felt lost while attempting to complete the tasks”; (2) “I felt the interface was intuitive”; (3) “I feel confident I can perform additional tasks with this interface”; (4) “I had trouble completing the tasks”; and (5) “I had a positive experience using this interface”. There were also two open-ended questions that asked the participants to make comments and offer recommendations.

38.4.4 Tasks

The tasks completed involved three of the VOC models, specifically Task #1 Fuel Economy; Task #2 Oil Consumption; and Task #3: Tire Wear models. A task description and

Input Parameter	Value
Vehicle type	City Bus (LNG)
Grade	A ($0 \leq \text{ABS}(\text{Grade}) \leq 0.4$)
Grade (ABS(%))	0
Upslope or Downslope	U
Pavement Roughness Range (IRI)	Poor (D) (IRI > 320 inch/mile)
Average speed (mph)	60
Fuel cost per gallon (\$)	\$3.00
Degrees of curvature (degrees)	4.25
Fuel economy (mpg)	5.32
Fuel economy decrease due to curvature (mpg)	0.38
Pavement Condition Adjustment Factor (PCAF) for fuel economy	1.21E+00
Total fuel economy (mpg)	4.01
Fuel consumption (gallons per 1,000 miles)	249.10
Fuel consumption (U.S. dollars per 1,000 miles)	\$747.30

Fig. 38.2 Fuel consumption interface – MS Excel model

specific parameters were given to the participants for each task. For example, for Oil Consumption these were: Task #2: Calculate Oil Consumption. At the end, please verbally report the oil consumption (quarts per 1000 miles) of a vehicle with the following parameters. Parameters: Vehicle type: Class 5 Truck (Diesel); Grade Level: B; Grade (%): 2.3%; Pavement Roughness: Fair (B); Average Speed (mph): 45.

38.4.5 Design

Several measurements were made as each participant went through the tasks. First was the time taken to complete the task, second was the verbally reported result of the task, third was the number of clarification questions asked during the tasks, and fourth was the number of incorrect entries made into data fields during the tasks.

In terms of experiment design [6], the independent variable was the user interface. Its levels, or conditions, were the Excel interface and the Website interface. The dependent variables were the above four measured items. Several variables, including the hardware, software versions, and

experiment supervisor were kept the same throughout the study as control variables. The study used a within-subjects type of evaluation [6], with half of the participants starting with one interface and the other half starting with the other interface, as a measure of counterbalancing potential learning and fatigue.

38.5 Results

38.5.1 Results Obtained from the Entry Questionnaire

The data gleaned from the entry questionnaire can be summarized as follows. A majority of the participants were male, and the age range tended to be in the twenties with one outlier. Most participants use Google Chrome most frequently and were reasonably comfortable navigating websites (Fig. 38.4) and using Excel (Fig. 38.5), with some more comfort shown for the former. One interesting information came from answers to Question #6, which indicated a diverse range

Vehicle Model

Vehicle Type Engine Type

Road Condition

① Grade (%) Upslope Downslope Grade Level

① Horizontal Curvature: Curvature Level A (R ≥ 1,638 ft)? Yes No

Radius of Curvature (ft) Degrees of Curvature (°) Curvature Level

Pavement Surface Condition Rating	IRI Range (inch/mile)	IRI Range (m/km)
<input checked="" type="radio"/> Good	≤ 95	≤ 1.50
<input type="radio"/> Fair (A)	96-119	1.51-1.88
<input type="radio"/> Fair (B)	120-170	1.89-2.69

Fig. 38.3 Fuel consumption interface – website application

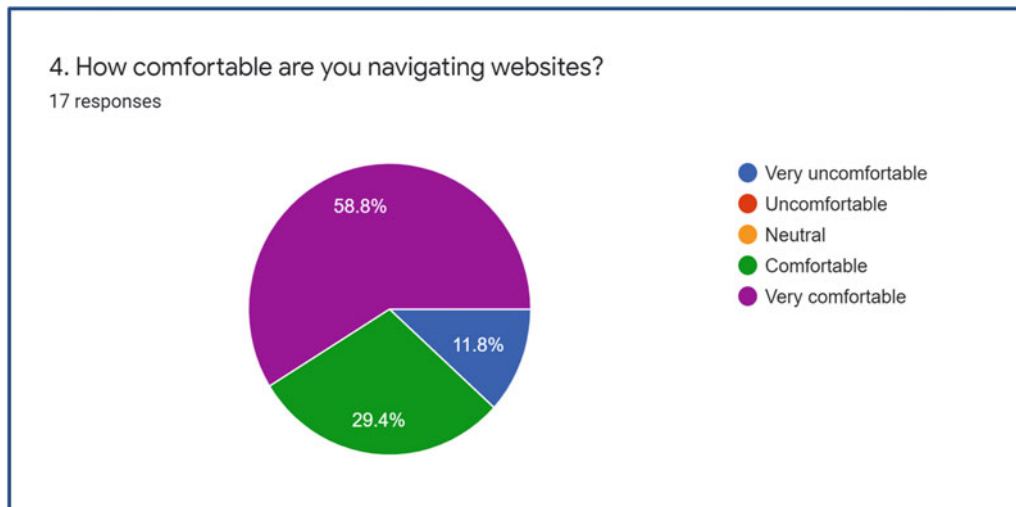


Fig. 38.4 Results of Entry Q#4: Participants' comfort with navigating websites

of how frequently the participants drive vehicles, spreading from 12% very infrequently to 23% very frequently.

38.5.2 Results Obtained from the Tasks Performed by the Participants

First, to assess whether there was evidence of a statistical difference between the two interfaces, an ANOVA test was

conducted on the dependent variable of time taken for completing each task. The null hypothesis for this test is that there is no statistical difference between the two data sets. The alternative hypothesis is that there is. Table 38.1 shows the collation of the task data and Fig. 38.6 presents the related graph. The F statistic calculated from this data was 5.82, with a target critical value of $F_c = 4.17$. As such, the null hypothesis is rejected: there is a statistical difference between the two data sets. Furthermore, given the mean times of both

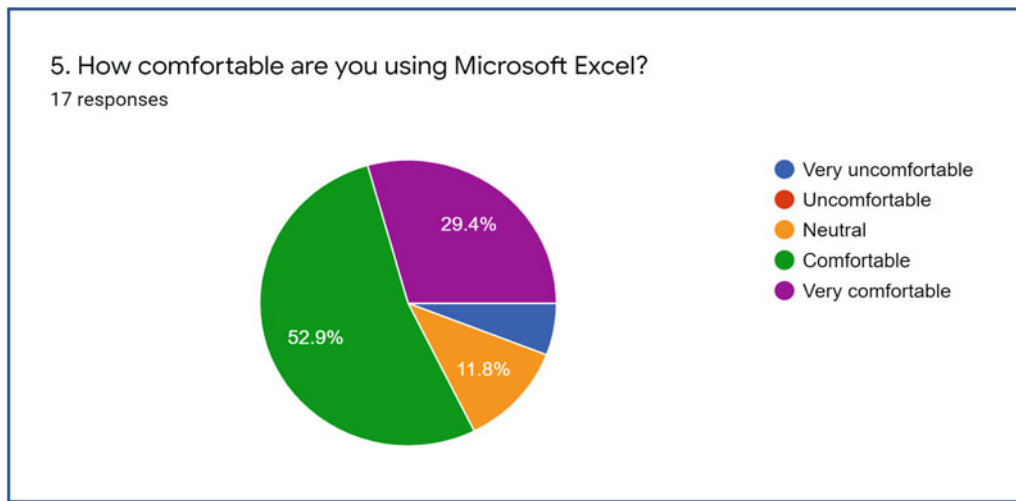


Fig. 38.5 Results of Entry Q#5: Participants’ comfort with using MS Excel

Table 38.1 Time to complete task results

Time to complete task (in scientific decimal minutes)		
Participant	Excel	Website
1	7.08	5.92
2	10.00	7.13
3	10.87	6.80
5	9.68	5.75
6	8.82	6.03
7	23.15	9.57
8	9.42	5.97
9	9.73	7.45
10	11.53	11.18
11	8.40	8.95
12	10.42	10.23
13	10.72	10.52
14	10.78	8.20
15	6.08	5.68
16	6.50	7.67
17	10.37	5.35
Mean	10.22	7.65
SD	3.68	1.86

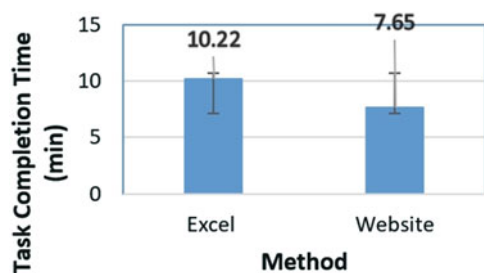


Fig. 38.6 Results for dependent variable “task completion time”

groups, it is clear that the statistical significance is in favor of the website.

Second, an ANOVA test was conducted on the dependent variable number of incorrect data entries. The null hypothesis for this test is that there is no statistical difference between the two data sets. The alternative hypothesis is that there is. The F statistic calculated from this data was 6.16, with a target critical value of $F_c = 4.17$. As such, the null hypothesis is rejected: there is a statistical difference between the two data sets. Further, given the mean times of both groups, it is again clear that the statistical significance veers in favor of the website.

Third, an ANOVA test was conducted on the dependent variable number of incorrect task results reported (that is, how many times the participants erred on obtaining the correct result value when running the models). The null hypothesis for this test is that there is no statistical difference between the two data sets. The F statistic calculated from this data was 79.71, with a target critical value of $F_c = 4.17$. As such, the null hypothesis is rejected: there is a statistical difference between the two data sets. Especially given the large F value obtained, it is again clear that the statistical significance leans massively in favor of the website.

Finally, a fourth ANOVA test was conducted on the dependent variable number of questions asked during the tasks (to the user study facilitator). The null hypothesis for this test is that there is no statistical difference between the two data sets. The alternative hypothesis is that there is. Due to space limitations the data collected are not shown here, but Fig. 38.7 presents the related graph with the results obtained. The F statistic from this data was 3.21, with a target critical value of $F_c = 4.17$. As such, the null hypothesis cannot be rejected: indeed, based on data gathered and the related analysis performed, there is no statistical difference between the two data sets. Notably, this is the only of the four variables for which statistical significance was not obtained.

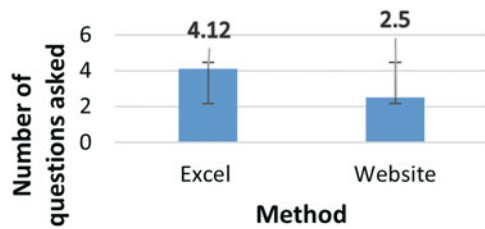


Fig. 38.7 Results for the dependent variable “number of questions asked”

Table 38.2 Answers to exit question 1: “I felt lost while attempting to complete the tasks”

Question 1		
Answer	Excel	Website
Strongly disagree [1]	2	5
Disagree [2]	5	10
Neutral [3]	2	1
Agree [4]	6	0
Strongly agree [5]	1	0
Weighted Average	2.94	1.75

38.5.3 Results Obtained from the Exit Questionnaire

The questions of the post-usage questionnaire were tailored to get a sense of whether the design goals of the website were met. With this in mind, a Chi Square test was performed on each question, to see whether there was a relationship between the answers given and the interface used. For each of these tests, the null hypothesis was that there was no relationship between the question’s answers and the type of interface. The alternative hypothesis was that there was a relationship. We note here that at the end of his/her participation one of the participants had technical difficulties answering the questions of the exit questionnaire. Thus, while 17 participants answered the entry questionnaire and performed the user study tasks, only 16 answered the exit questionnaire.

For Question 1 of the exit questionnaire, “I felt lost while attempting to complete the tasks,” the answers are tallied in Table 38.2. The Chi Square test resulted in a value of $\chi^2 = 10.28$, with a critical value of $\chi^2_{(0.05, 4)} = 9.49$. This results in a rejection of the null hypothesis. Seeing the agreement level with this statement (shown in the weighted average of the range, where Strongly Disagree is 1 and Strongly Agree is 5) we can conclude that the website was less confusing to the participants.

For Question 2 of the exit questionnaire, “I felt the interface was intuitive,” the Chi Square test resulted in a value of $\chi^2 = 10.12$, with a critical value of $\chi^2_{(0.05, 4)} = 9.49$. This results in a rejection of the null hypothesis. Again, seeing the amount of agreement with the statement on the website

Table 38.3 Answers to exit question 5: “I had a positive experience using this interface”

Question 5		
Answer	Excel	Website
Strongly disagree [1]	2	0
Disagree [2]	5	0
Neutral [3]	3	0
Agree [4]	5	9
Strongly agree [5]	1	7
Weighted Average	2.88	4.44

interface, it appears that the participants generally found the website more intuitive.

For Question 3 of the exit questionnaire, “I feel confident I could perform additional tasks with this interface,” the Chi Square test resulted in a value of $\chi^2 = 11.48$, with a critical value of $\chi^2_{(0.05, 4)} = 9.49$. This again results in a rejection of the null hypothesis. It is interesting to note that every single participant either strongly agreed or agreed with the statement about the website interface. This indicates that the learning curve for using of the website is not steep; after performing the tasks, all participants felt confident they could use the web application effectively.

For Question 4 of the Exit Questionnaire, “I had trouble completing the tasks,” the Chi Square test resulted in a value of $\chi^2 = 10.28$, with a critical value of $\chi^2_{(0.05, 4)} = 9.49$. This results in a rejection of the null hypothesis. Again, while not as dramatic as Question 3’s results, the fact that no users agreed with this statement on the website interface indicates that even for the initial tasks the participants had little trouble navigating the interface. Notably, the results for the Excel interface are not too bad either, being rather on the positive side of the feedback.

For Question 5 of the Exit Questionnaire, “I had a positive experience using this interface,” the answers are shown in Table 38.3. The Chi Square test resulted in a value of $\chi^2 = 15.48$, with a critical value of $\chi^2_{(0.05, 4)} = 9.49$. This for the fifth time in a row results in a rejection of the null hypothesis. These were by far the most dramatic results, indicating a great deal of positivity towards the website interface. As a significant design goal for the website was to encourage users to engage with and test the VOC models, these results indicate that this goal has been met.

38.6 Discussion

Going through the results, it is easy to see that the website interface generally exceeded the Excel interface in terms of efficiency and user experience. It was somewhat surprising that on every single question in the post-usage questionnaire the results indicated statistical significance. Further, the fact

that on some questions all responses veered to the positive side in favor of the website cemented the participants' apparent preference for the website interface.

In today's world, using websites is ubiquitous, particularly among the younger demographics. Given the relative youth of the participants in the study, perhaps this was a factor in their preference for the website application. However, even in evaluating the more objective dependent variable of time taken to complete all the tasks, the website's interface won out, achieving statistical significance.

Interestingly, the results for the exit questionnaire are even more conclusive than those for the dependent variables. There was a statistical difference for every single question, and the distribution of the responses for each question was again far more positive for the website interface than for the Excel interface. Notably, on exit questions #3 and #5 there were no negative or even neutral responses; all participants either agreed or strongly agreed with the statements. One possible explanation for these fairly strong results is that nowadays there may be an overall mastery in website use that does not extend to Excel. Another consideration is that when people use websites, it might be for leisure, whereas when people use Excel, it tends to be for work. In any case, the users' perception of the website interface versus the Excel interface clearly matches the results obtained by analyzing the dependent variables.

38.7 Conclusions and Future Work

In conclusion, the results were heavily stacked in favor of the website interface. In every metric pertaining to dependent variables as well as in all answers to the exit questionnaire the users seem to have had a more efficient and enjoyable

experience completing the tasks with the website. The users reported more accurate values, finished in less time, and claimed to have had a positive experience. This is highly encouraging for the website interface, and while security issues would also prevent the widespread use of the Excel models, it is evident the website provides benefits beyond that.

Future work has been planned for the website, including new features and functions such as batch processing for groups of inputs and advanced data visualization. Future user studies regarding these new features will also likely be conducted.

References

1. FHWA, *FHWA Rolls Out New HERS-ST Software – July 2002 – FHWA-RD-02-012 – Focus: Federal Highway Administration*, (2002). U.S. Dept of Transportation/Federal Highway Adm. Accessed 2/3/2021 at www.fhwa.dot.gov/publications/focus/02jul/hersst.cfm
2. E.Y. Hajj, M. Sime, R. Chkaiban, G. Bailey, H. Xu, P.E. Sebaaly, *Enhanced Prediction of Vehicle Fuel Economy and Other Operating Costs, Phase II: Modeling the Relationship Between Pavement Roughness, Speed, Roadway Characteristics 7 Vehicle Operating Costs*, Technical Report (Federal Highway Administration, Washington, DC, 2018)
3. B. Lim, H.J. Wen, Web services: An analysis of the technology, its benefits, and implementation difficulties. *Inf. Syst. Manag.* **20**(2), 49–57 (2003)
4. Vehicle Operation Costs Calculator (2021). <https://voc.engr.unr.edu/>
5. T.R. Carlson, T.C. Austin, D. McClement, S.H. Yoon, Development of Generic Link-level Driving Cycles. Sierra Research Inc., draft report prepared for the US Environmental Protection Agency (2009)
6. I. Scott MacKenzie, *Human-computer interaction: An empirical research perspective*, 1st edn. (Morgan Kaufmann Publishers Inc., San Francisco, 2013)

Vinh Le, Connor Scully Allison, Mitchell Martinez, Sergiu M. Dascalu, Frederick C. Harris, Jr., Scotty D. Strachan, and Eric Fritzinger

Abstract

When an environmental research project grows, technical concerns over system scalability, data exposure, and third-party application support are overlooked. This paper presents a system, the Microservice-based Envirosensing Support Applications (MESA), that provides a scalable environment and data infrastructure solutions for the NSF-funded Solar Energy-Water-Environment Nexus project. MESA can be broken into 4 major parts: a suite of microservices exposed over an API, an overarching service discovery, a series of tables replicated from an existing monolith, and the applications that MESA lends its support. In order to evaluate the capability of MESA, the features of this system were compared against three other existing microservice-based research systems. MESA features were more robust than two of the other systems, but was found lacking when compared to the last, as it does not lend support to advanced techniques like HPC or Machine Learning.

Keywords

Microservice · Distributed system · Data management · Containerization · Data analysis · Data visualization · User interface · Web framework · Web-based systems

39.1 Introduction

When it comes to environmental research projects, a common approach to data storage is spinning up a monolithic system consisting of one or more databases, interwoven code, and a small suite of sensors streaming in data at specific intervals. However, as data begins to accumulate and more sensors are deployed as the project begins to grow, problems emerge from this ad hoc approach.

Querying speeds, management, access, and analytics all become affected by the increase in data volume. Usually in modern software engineering practices, this implies that it is time for a system reconstruction. Although appropriate, these overhauls exact a heavy toll in time and resources, commodities not readily available for most small-to-midsize environmental research projects.

To address these problems, this paper presents the Microservice-based Envirosensing Support Applications (MESA), a distributed support system build using a Microservice Architecture style that is tailored for the use in environmental research projects. As part of this research, MESA was implemented to support the NSF Track 1 Solar Energy-Water-Environmental Nexus project and its data hub, the Nevada Research Data Center (NRDC) [12].

We evaluate the comparative merits of MESA against similar systems with a feature comparison. Using common features expected from software like this as a benchmark, MESA is shown to possess more functionality than two of the other systems. However, MESA is slightly deficient when compared to the last system due to the lack of infrastructure for Machine Learning and High Performance Computing. We address these deficiencies by indicating that these are key areas of future work further on.

V. Le (✉) · C. Scully-Allison · M. Martinez · S. M. Dascalu
F. C. Harris, Jr. · S. D. Strachan · E. Fritzinger
Computer Science and Engineering, University of Nevada, Reno,
Reno, NV, USA
e-mail: vl@unr.edu; cscullyallison@email.arizona.edu;
mitchell.martinez@nevada.unr.edu; dascalus@cse.unr.edu;
fred.harris@cse.unr.edu; strachan@unr.edu

The remainder of this paper is structured as follows: Sect. 39.2 presents the background of the NRDC and the three systems being compared against MESA, Sect. 39.3 provides details of the software specifications of MESA, Sect. 39.4 describes the various implementations of MESA's microservices, Sect. 39.5 provides a discussion on the feature comparison and other considerations, and Sect. 39.6 wraps up the paper with the conclusion, and planned future work.

39.2 Background and Related Works

39.2.1 Nevada Research Data Center

The Nevada Research Data Center (NRDC) is the central data hub of the Solar Energy-Water-Environmental Nexus project, where environmental sensor data from various research teams is collected and stored [12]. Unfortunately, the NRDC is the descendant of an older monolithic system and inherited its predecessor's rigidly interconnected approach [3]. Due to its interwoven nature, the NRDC system has great trouble even maintaining itself. As an example, the NRDC could miss several hundred data entries, which is especially disconcerting for the scientists who are expected to conduct research on the data. Furthermore, the NRDC has virtually no means to actively monitor its services' health which makes it hard for to tell if a functionality was even online. Fortunately, this paper is not the first time this problem was recognized, and there was prior work by the authors of this paper on proposing a more distributed reformation of the NRDC [8, 10].

39.2.2 Microservice Architecture

The term "Microservice" was traced back to a Microsoft Service Edge Conference presentation in 2005 by Dr. Peter Rodgers. Dr. Rodgers referred to the concept of granular web services that remained independent of each other as "Micro-Web-Services" [13]. These granular web services could then be mapped to a specific functionality and the inter-switching of them created a new option for modularity within the system. The eventual orchestration of these microservices would eventually lead to a functioning system architecture. However, the architecture has a failing in the form of it being incredibly difficult to implement. In order to deploy a microservice-based implementation, it would require the actual construction of such a system by using the concepts as a guideline. Very few software and packages exist out there that would streamline the process of creating a microservice-based system.

39.2.3 DIMMER Smart City Platform

Although not environmental in nature, a similarly developed research-oriented microservice-based system was created to support a smart city project in Europe. This system, dubbed DIMMER, collected sensor data, interfaced with a suite of applications, and used a service discovery as part of their DevOps [7]. DIMMER also featured an high performance computing (HPC) resources as part of their platform. However, because DIMMER both actively collects sensor data while also providing support to various applications, this can create significant network overhead for the researchers utilizing the system.

39.2.4 Generic Service Infrastructure for PEIS

Sharing many similarities with MESA, a microservice-based environmental research system, the Public Environmental Information System (PEIS), was established as part of a nationwide environmental research project to provide infrastructure and application support to various client applications, such as sensor networks, web applications, and mobile devices [2]. As part of PEIS' design, the system also provides support to advanced research tools such as HPC and Machine Learning. On top of this, PEIS also features modern tool integration, in the form of containerization, continuous integration, and a service discovery. The approach adopted by PEIS during the implementation of PEIS involved the complete refactoring and rebuilding of a sensor collection system. During an ongoing environmental research project, such decisions could prove to be too expensive in time, money, and data lost.

39.2.5 OceanTEA: Exploring Ocean-Derived Climate Data

On a similar research scale, a support system, dubbed OceanTEA, was designed to aid researchers with processing data from a ocean monitoring system designed by the University of Kiel in Germany [6]. This system uses microservices to structurize the data presented to researchers based on specified criteria. OceanTEA then presents this information through an intuitive and responsive web interface. However, the design of OceanTEA gives off the impression that the system is tailored only towards to the structuring and presenting of data to researchers, rather than providing the groundwork for other future tool development.

39.3 Software Specification

39.3.1 High Level Design

MESA as a system has four major components, as shown in Fig. 39.1. At the center of MESA is the Service Discovery, which serves as a both registry and monitor for all of the microservices. The Service Discovery does not actively entangle itself with any services, aside from scheduled tests, and provides to the user necessary metadata regarding location and health. The next and most crucial portion of MESA are the microservices. The microservices run independent of one another and execute specific programs and tasks for the MESA system. These are often shuffled into servers associated with their specific functionality. Their functionalities are then made available on those servers via reverse-proxy to an application through their respective HTTP APIs. Moving on to the next component, the tables of the database are not only utilized as a general data abstraction between client applications, but are also called and used in certain microservices to perform complex calculations or data management operations. Finally, the last component is the multiple applications that interface with MESA. These applications are not limited only to web applications, but also include mobile phone apps and can even be separate systems.

39.3.2 Technology Utilized

The MESA system was developed using several common web technologies compatible with the database manager used by the NRDC. The programming languages utilized include C# and Python, and the tools utilized were WCF and Flask. Windows Communication Foundation (WCF) is a toolset developed in the .NET Framework that specializes in implementing and deploying service-oriented architectures (SOA) [9]. Flask is a python-based micro-framework that supports the development of web services [1]. Database management for the NRDC is handled through Microsoft SQL Server (MSSQL). For its service discovery, MESA uses *Consul* [5]. Containerization is handled through Docker and Continuous Integration is managed with Jenkins.

39.4 Use Cases

39.4.1 SEPHAS Lysimeter Visualization

The SEPHAS Lysimeter Visualization, as shown in Fig. 39.2, is a web application developed to better visualize the environmental data gathered over the span of several years by the SEPHAS facility in Las Vegas [4]. The microservice

support from MESA played a non-vital but significant role in the visualization of the lysimeter data. By using powerful frontend visualization libraries such as D3.js, the data file could be loaded and visualized.

However, this would cause almost unbearable lag times between actions issued by the user on the visualization. Although the visualization can operate without the need for a microservice, the usage of a microservice in this case was able to cut down virtually all of the lag time between the user actions and the visualization library. The Data Visualization microservice was utilized to handle all of the data loading and transformation operation, so the web client only needed to query the microservice for all of its needs. Once the client contacts the microservice, the service will then return limited amounts of data to only preserve the shape of the visualization. However, once the user explored further into the visualization, the microservice would then alter the range and the amount of data presented to match what the user viewed. This allowed the client to levy all of its intensive actions onto the server and provide an accurate, responsive, and swift visualization.

39.4.2 NRDC Quality Assurance Application

The NRDC Quality Assurance (QA) Application, or QA App for short, is an application developed by the Cyber-infrastructure component of the Nexus project to handle metadata [11]. As part of the project, Nexus technicians often trek out to research sites for maintenance, installation, and configuration of sensor tower equipment. Technicians would have to manually write down entries on a notebook and then transcribe those notes into a database sometime after. The QA App was developed for the express reason of alleviating the troubles faced by Nexus technicians. The application narrows down metadata on research sites specific to the user and allows them to alter entries or add new ones right at the tower. Since there is limited internet access at these towers, this application stores the changes locally and syncs them to the database when appropriate internet connection is made available.

The microservices play a vital role as server backend for the QA application. The QA application upon initial activation calls upon each of the eight microservices to store a local copy of the relevant data entries within the metadata database. It is here where the microservices converts data from the NRDC and presents it to the QA application (Fig. 39.3). When changes are made in the application and the sync button is pressed, the application then uses the eight microservices alongside the Conflict Management microservice to verify and submit the changes to the database. Should the Conflict Management microservice return a merge issue, the response from the microservice is parsed and then used

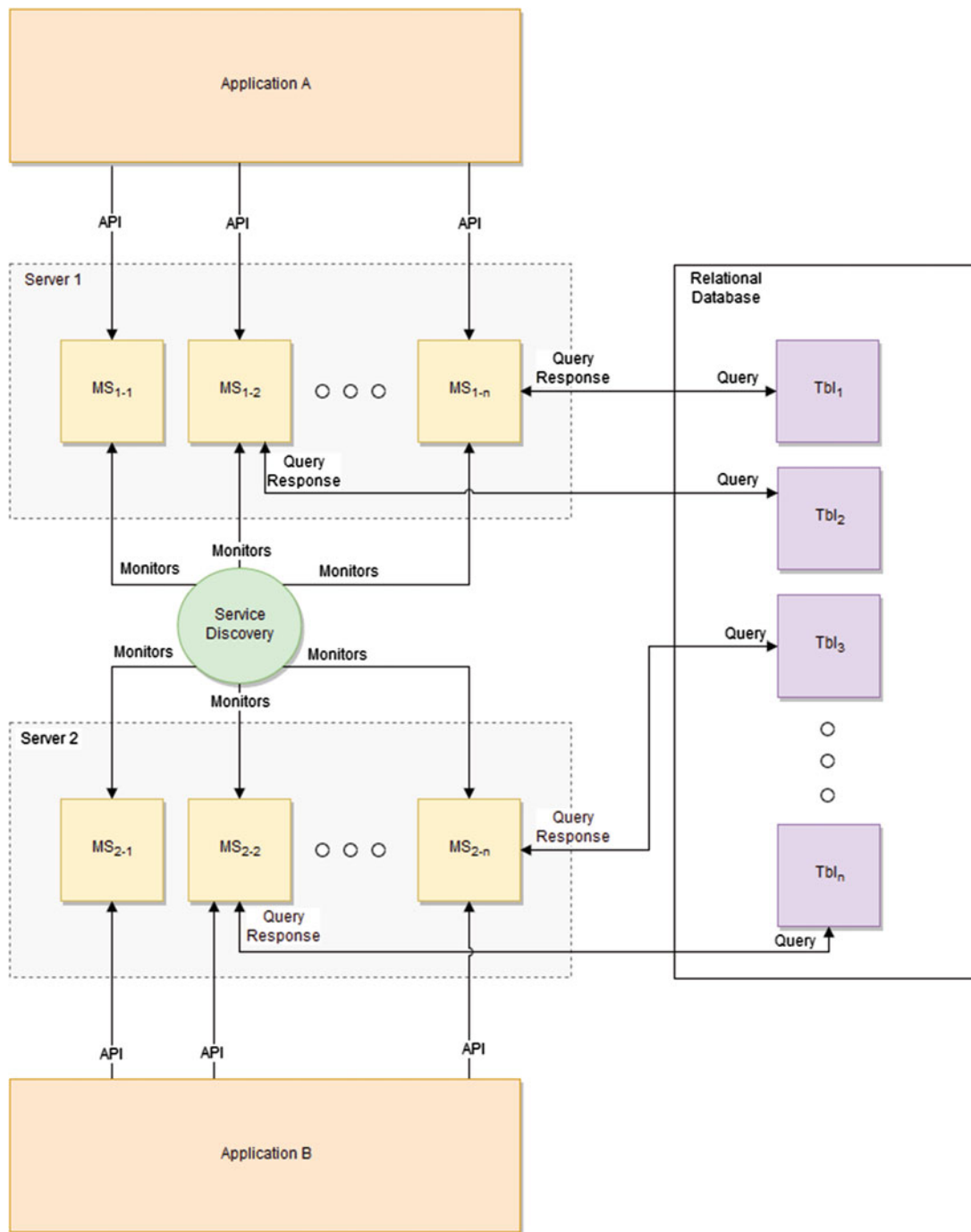


Fig. 39.1 A high-level view of the MESA system

to generate a merging interface. Additionally, the Imagery microservice is called when an entry features an image and handles the storage and retrieval of that image into the database. The Imagery microservice is also called when the a entry is viewed by the user, where it retrieves a preview image instead of the original.

39.4.3 Conflict Management

The Conflict Management microservice was created to resolve conflicts that result from multiple users enacting changes on the NRDC database with an application. Conflict Management was developed largely for the metadata

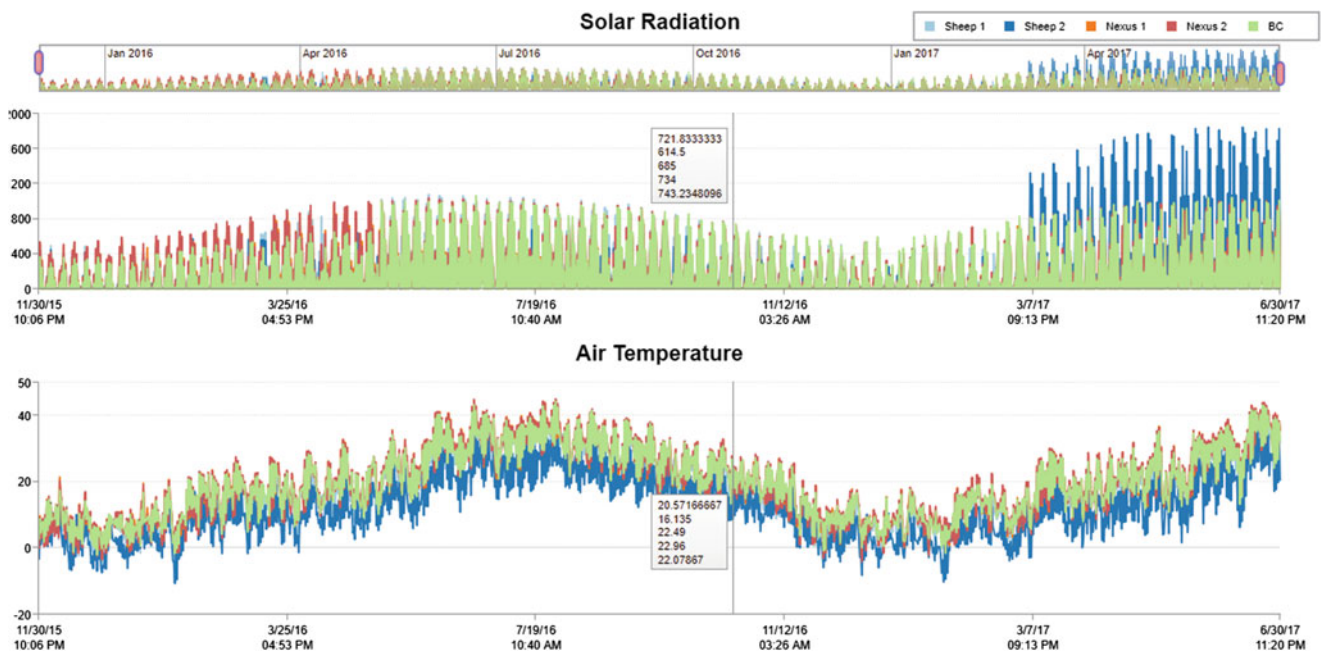
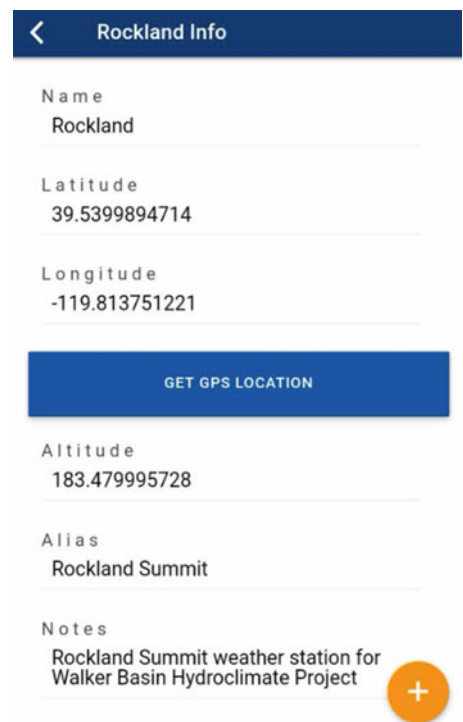


Fig. 39.2 Sample data visualization on the Lysimeter data display

Fig. 39.3 The NRDC QA application navigating through a site entry

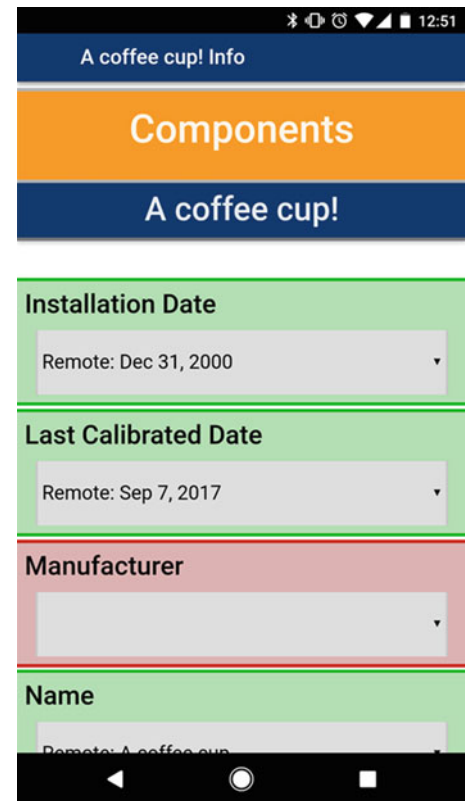


application uploading multiple entries at one time. The main process operates in a similar manner as most version control software. When the submission of a data entry whose modification date is earlier than what is listed inside the database, a conflict is flagged. Much like version control software, the user is given the option to continue with their flagged copy or merge their version with the current canon (Fig. 39.4). Once a selection is chosen, the microservices

locates the database table in which the data resides and overwrites the entry with the selection made. This is then repeated for each of the multiple entries being uploaded by the application during that one transaction.

Calling the Conflict Management microservice requires sending a POST request consisting of a list of entries to submit to the database. Also inside the JSON, metadata is given to locate the entry’s associated database table. The

Fig. 39.4 The conflict management functionality on the NRDC QA application



microservice then goes through each of the submissions and compares the modification dates. Should a conflicting modification date be found, the microservice appends that entry to a flagged list. Meanwhile, the passing submissions are added to their respective database tables via the appropriate microservices. At this point, the microservice will return a response detailing specific information about the conflict, and a copy of both what was sent and what currently exists within the database. The response returned provides the necessary information for a front end to create a conflict resolution interface. Once a finalized choice has been made, a POST request to another URI within the microservice allows for the overwriting or updating of what currently exists within that database entry.

39.4.4 Near Real-Time Autonomous Quality Control

Occasionally, sensor readings received by the NRDC from the remote research sites show signs of erroneous data. This can be missing values, values outside possible bounds, or even repeats of past values. To address these problems, the Near Real-time Autonomous Quality Control (NRAQC) System was developed for the NRDC [14]. NRAQC tests incoming data points logged autonomously at a research site to see if they meet the criteria of an invalid measurement.

The system, with aid of user specified configurations, flags all invalid measurements with metadata that specifies the nature of the invalidity. The service provided by NRAQC is necessary for the production and distribution of a quality data product. A sample of NRAQC's data visualization is shown in Fig. 39.5.

Much like the QA Application, the microservices play the vital role of server backend for the NRAQC system. NRAQC utilizes an intuitive web interface as the main client, but splits its main features into microservices that support it. These features includes the handling of differing data sources, enable autonomous flagging of measurements, interfacing with the client, enabling a data visualization, and formatting the results based on the user specifications. While the microservices deal with the computationally and memory intensive portions of NRAQC, they do not govern the entire system itself. The NRAQC client presents a number of features to the user and when the user selects a task, the NRAQC client then communicates with the microservices via HTTP.

39.5 Discussion

In the environmental field, microservice architecture is often used to drive software with narrow goals. Multiple microservices are usually developed and used to create web-based support for a singular application, such as the ones described in OceanTEA. For research outside the earth sciences, the

NRDC QC Dashboard

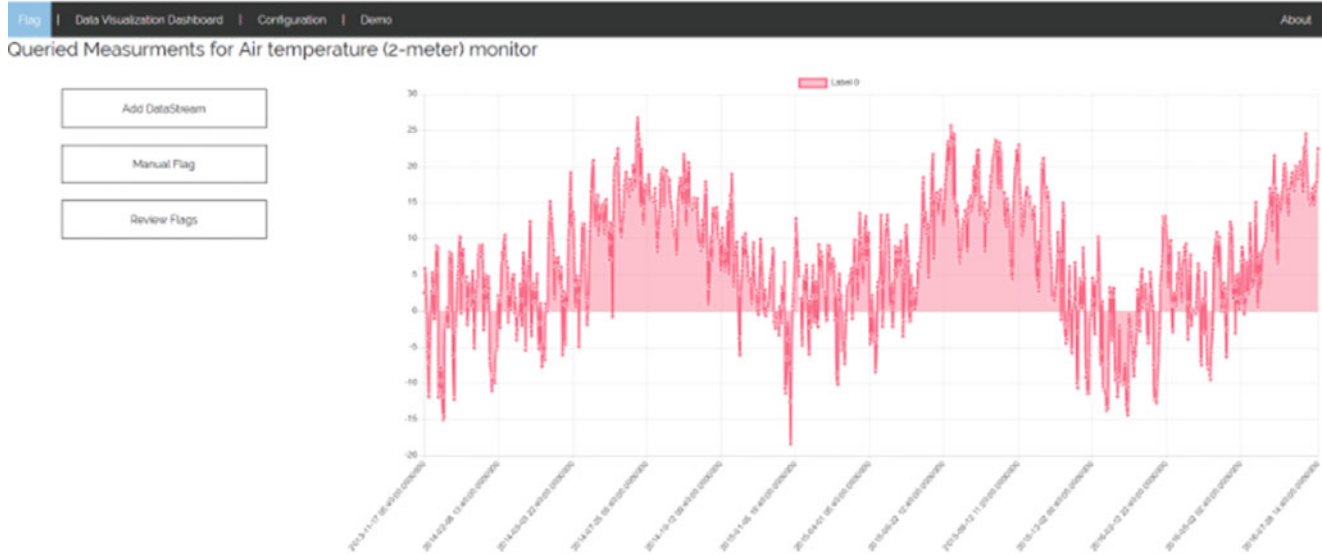


Fig. 39.5 The main visualization component of the NRAQC system

microservice architecture is often used as platform for providing an abstract data layer between an application and the data source of the project, as described in the DIMMER Smart City Platform. Interestingly enough, both research inside and outside the Earth sciences share a common trend of refactoring systems into microservice architectures when the problem involves an existing monolith. This approach usually involves the decomposition of a monolith into requirements that are mapped to microservices in the hopes of enabling scalability. This conversion can especially be seen in the PEIS system mentioned above.

While both of the described trends showcase two valid and intended use cases of the Microservice Architecture, the complete refactoring of a system from monolithic to microservice-based brings about serious concerns. In environmental research projects, especially projects ranging from a single university to an entire state, monolithic system designs are common due to unexpected growth in a project or from a sheer lack in pooled technical knowledge. Many of these projects simply do not have the resources or people to simply halt progress and perform an entire system overhaul. Additionally, many environmental projects often autonomously collect data from sensor networks and shutting down these systems, even briefly, can cause detrimental effects on the research produced by the overall project.

It is through these concerns that brings to light the novelty of MESA as a microservice-based system. MESA is designed as a scalable application development platform to support critical systems, especially monoliths, without having the need to tear down the existing system. MESA connects to a regularly updated replication of the main NRDC database

that houses copies of the incoming data, so it does not interfere with the monolith whom autonomously gathers data at set intervals. To clarify, this approach does not eliminate the option of full system migration, and provides the means to lessen the burden of the demands placed on developers until a solution is decided upon and implemented. Should a full migration be decided, MESA can be utilized to shoulder the burdens of inactive systems and can eventually become decommissioned once the migration is complete.

A feature comparison, shown in Table 39.1, was created to measure MESA's capabilities against the previously described microservice systems. MESA brings forward a unique contribution to Environmental Research in that it does not require a complete system refactoring in order to be used. Additionally, the MESA system carries with it many of the modern software features and industry practices that are present in microservice development. It is through these features that MESA is able to offer more functionality than two similar systems: OceanTEA and DIMMER. OceanTEA, although a very effective system in environmental data gathering, does not offer platform support, a service discovery, or the continuous integration features that MESA does. Similarly so with DIMMER, it does not provide as much features as MESA, lacking in areas like supporting multiple databases, containerization, and continuous integration features. However, MESA still is outperformed by microservice systems used by the larger environmental projects, like PEIS who is able to support advance features such as HPC or Machine Learning. Overall, MESA's abundance of features makes it a solid alternative to a development platform for small to medium scale environmental projects.

Table 39.1 Feature-based comparison table

Feature description	MESA	DIMMER	OceanTEA	PEIS
Requires refactoring entire system		x	x	x
Support multiple applications	x	x		x
Oriented toward environmental research	x		x	x
Service discovery features	x	x		x
Supports multiple databases	x	x		x
Uses containerization	x		x	x
Uses continuous integration	x			x
Capable of high performance computing solutions		x		x
Machine learning capabilities				x

39.6 Conclusions and Future Work

39.6.1 Conclusion

The system described in this paper, the Microservice-based Envirosensing Support Architecture, focuses on creating an alternative approach to cyberinfrastructure for environmental research projects without enforcing a costly system refactoring. The central idea of this system is to create an applications platform centralized around a regularly replicated data source without having to tear down a monolith. The option to completely refactor a system and break apart an active monolith is expensive and requires a massive amount of technical knowledge and time to execute. This application platform was created by using a series of microservices that break up business requirements into independent web services. The microservices answer to a central service discovery and are generally mapped to a feature within a client application.

MESA is a relevant and beneficial system to environmental research projects due to its ability to provide platform support to a field that is often limited by the technical aspects of software development. While the idea of switching to a distributed architecture, like microservices, can be attractive to growing environmental projects, the reality of the matter comes down to whether the project has the time to halt progress while development is made and if there are adequate resources available to achieve this result. So oftentimes, environmental scientists are forced to choose between two extremes: a limited older system or an expensive new system. MESA brings to the Earth sciences a third choice that can bridge the gap between the previous two while incorporating industry practices, such as containerization and continuous integration.

39.6.2 Future Work

Work is currently underway to make the readings gathered by the sensor towers to be more readily available and accessible as datasets for machine learning. This is a larger focus for MESA, while HPC services are currently being provided by a collaboration with the state of Nevada and Switch, a global leader in data center technologies.

To prevent the interception of data and verify that the client has clearance to interact with data, most modern RESTful-based software practices token-based authentication. Unfortunately, MESA does not utilize this industry-standard practice as of yet. Currently, the services perform this actions without verification and are highly susceptible to being intercepted. This is due to MESA being a prototype to show a proof of concept and security additions are considered secondary features. In the future, a major enhancement to the MESA system would be the application of modern security practices.

Although MESA uses containerization technology in the form of Docker, MESA only has Docker containers operating on approximately a third of the active microservices. Docker has commonly been used for Linux environments and while it does have Windows versions, it requires the deft hand of a system administrator or a DevOps engineer. During the development of MESA's Windows-based microservices, this task was deemed secondary as to allow more focus on supporting environmental research applications. However, recent talks and advancements within the project have advised steering toward a migration onto a Kubernetes environment to host the Docker containers. This would allow MESA to achieve total containerization of microservices in future iterations of the project.

Acknowledgments This material is based in part upon work supported by the National Science Foundation under grant number IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

1. R. Armin, Flask microframework. <http://flask.pocoo.org/>. Accessed Jan 12, 2022
2. E. Braun, T. Schlachter, C. Döpmeier, K.-U. Stucky, W. Suess, A generic microservice architecture for environmental data management, in *Environmental Software Systems. Computer Science for Environmental Protection: 12th IFIP WG 5.11 International Symposium, ISESS 2017, Zadar, Croatia, May 10–12, 2017, Proceedings 12* (Springer, Berlin, 2017), pp. 383–394
3. S. Dascalu, F.C. Harris, Jr., M. McMahon, Jr., E. Fritzing, S. Strachan, R. Kelley, An overview of the Nevada climate change portal, in *7th International Congress on Environmental Modelling and Software* (2014)

4. Desert Research Institute, Scaling environmental processes in heterogeneous arid soils (SEPHAS). <https://www.dri.edu/sephas>. Accessed Jan 12, 2022
5. Hashicorp, Consul. <https://www.consul.io/>. Accessed Jan 12, 2022
6. A. Johanson, S. Flögel, C. Dullo, W. Hasselbring, Oceantea: exploring oceanderived climate data using microservices, in *International Workshop on Climate Informatics (CI 2016)* (2016), pp. 24–29
7. A. Krylovskiy, M. Jahn, E. Patti, Designing a smart city internet of things platform with microservice architecture, in *2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud)* (IEEE, Piscataway, 2015), pp. 25–30
8. V.D. Le, M.M. Neff, R.V. Stewart, R. Kelley, E. Fritzinger, S.M. Dascalu, F.C. Harris, Microservice-based architecture for the NRDC, in *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)* (IEEE, Piscataway, 2015), pp. 1659–1664
9. Microsoft, Windows communication foundation. <https://docs.microsoft.com/en-us/dotnet/framework/wcf/>. Accessed Jan 12, 2022
10. R. Motwani, M. Motwani, F.C. Harris, Jr., S. Dascalu, Towards a scalable and interoperable global environmental sensor network using service oriented architecture, in *2010 Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)* (IEEE, Piscataway, 2010), pp. 151–156
11. H. Munoz, C. Scully-Allison, V. Le, F.C. Harris, Jr., S. Dascalu, A mobile quality assurance application for the NRDC, in *Proceedings of the ISCA 26th International Conference on Software Engineering and Data Engineering (SEDE 2017)* (2017), pp. 61–66
12. Nevada EPSCoR, Solar energy water environment nexus in Nevada. <https://solarnexus.epscorspo.nevada.edu/>. Accessed Jan 12, 2022
13. P. Rogers, Service-oriented development on netkernel- patterns, processes & products to reduce system complexity. <http://www.cloudcomputingexpo.com/node/80883>. Accessed Jan 12, 2022
14. C. Scully-Allison, V. Le, F.C. Harris, Jr., S. Dascalu, Near real-time autonomous quality control for streaming environmental sensor data, in *22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems* (2018)

Part VIII

Networks

Semantic Interoperability in the Internet of Things: A Systematic Literature Review

Pedro Lopes de Souza, Wanderley Lopes de Souza,
and Ricardo Rodrigues Ciferri

Abstract

The main challenges in Internet of Things (IoT) refer to the varied capabilities of things, the huge amount of data they produce, the heterogeneity of this data, and the diverse offered services. For each application domain and for each vendor, there is usually a specific and proprietary IoT platform, with no de facto standards currently being found or expected in the near future. Therefore, ensuring the semantic interoperability of things between different types of IoT platforms and applications is one of the major problems in this area. This paper aims to identify current conceptual and practical findings for the semantic interoperability problem in IoT through a Systematic Literature Review (SLR). By searching digital libraries and using the snowballing technique, 314 manuscripts were selected for data extraction from this SLR, which allowed us to come up with the current main issues and solutions related to this matter. The results obtained with this SLR and reported in this paper can help researchers and practitioners to find better solutions for this problem.

Keywords

Systematic literature review · Survey · Internet of things · Ubiquitous and pervasive computing · Semantic interoperability · IoT platforms and applications ·

Seamless communication · Ontology · Semantic web · Web of things

40.1 Introduction

The Internet of Things (IoT) terminology was first proposed by Kevin Ashton in 1999 [1]. The IoT consists of networks of everyday objects (e.g., household items, vehicles, buildings), called “things”, equipped with embedded technology, sensors and actuators, capable of collecting and transmitting data via wireless connections to the Internet. Therefore, it is a concept that refers to the digital interconnection of everyday objects with the Internet. This makes it possible to remotely control things, which can be accessed as service providers.

IoT enables the development of applications in a wide variety of domains, among which has been highlighted Home, Transport, Logistics, Automation and Industrial Manufacturing, Process and Business Management, Agriculture, and Health. The IoT’s concept as a ubiquitous network consisting of objects/resources can be seen as an advance of Ubiquitous Computing [2]. In this sense, with billions of sensors and actuators deployed and combined across multiple domain-specific platforms, Mark Weiser’s vision of a hyperconnected world can be reached very soon.

One of the big challenges in IoT is dealing with the huge amount of data produced by things, in addition to the heterogeneity of data, the diverse capabilities of things and the services offered. For each domain and each vendor, there is usually a specific and proprietary IoT platform, with no de facto standards currently being found or expected in the near future. Therefore, ensuring the semantic interoperability of things between different types of platforms is one of the problems that needs intense investigation in this area.

P. Lopes de Souza (✉)
Graduate Program in Computer Science (PPGCC), Federal University of São Carlos (UFSCar), São Carlos, SP, Brazil
e-mail: plsouza@estudante.ufscar.br

W. Lopes de Souza · R. R. Ciferri
Department of Computing (DC), Federal University of São Carlos (UFSCar), São Carlos, SP, Brazil
e-mail: desouza@ufscar.com; RRC@ufscar.com

Semantic interoperability refers to the ability of different systems to understand the meaning and context of information exchanged with each other. This encompasses the meaning of the data and the relationship between the data. Ontologies, semantic technologies, and knowledge management systems make it easy to achieve semantic interoperability.

In IoT, semantic interoperability requires a common description of data and representation structures that characterize things, their capabilities, and the data they produce. This involves developing a vocabulary to describe the exchange of data and ensure that it is unambiguously understood by things and can also be interpreted by machines.

This paper describes a Systematic Literature Review (SLR), which was conducted to explore, understand, and summarize the current main issues and solutions for the semantic interoperability in IoT. Additionally, this SLR sought to identify where and how semantic interoperability is being applied in different IoT domains, thus helping researchers and practitioners to find better solutions for this problem.

This paper is further structured as follows: Sect. 40.2 details the methodological procedures used to search and select manuscripts; Sect. 40.3 reports and discusses the main SLR results; and Sect. 40.4 presents our concluding remarks, and the current and future work of the main author of this paper.

40.2 Method

SLR is a secondary study intended to provide a thorough assessment of a research topic, using a reliable, rigorous, and auditable review methodology. According to [3], an SLR should be carried out in three main phases: planning, which involves identifying the need for a review, specifying the research questions, and developing a review protocol; conducting, which involves the identification and selection of studies, and the extraction and synthesis of data; and reporting, which involves the analysis and publication of results.

There are some tools for supporting SLRs in Software Engineering, and three of the most employed were chosen to be analyzed in [4]. Parsifal [5] was one of them. This online tool allows for geographically distributed researchers to work together within a shared workspace for performing SLRs.

We conduct our SLR according to the guidelines prescribed in [3], and employing Parsifal. In the planning phase, we define the Research Questions (RQ), and elaborate the Research Protocol (RP) containing data sources, search strings, and inclusion and exclusion criteria. In the conducting phase, we first searched digital libraries, then selected manuscripts by reading abstract, full reading, and by employing snowballing technique, and finally we extracted and synthesized data from the selected manuscripts. In

the reporting phase, we analyze and classify the selected manuscripts following predefined strategies, in order to share the results obtained with this SLR.

40.2.1 Planning the Review

In order to search for the main issues and their current solutions for the IoT semantic interoperability problem, our SLR sought to answer the following RQs:

- RQ₁** – What are the main IoT application domains?
- RQ₂** – What are the main IoT architectures and platforms?
- RQ₃** – How the IoT architectures and platforms address semantic interoperability in the IoT application domains?

The RP we developed is based on these RQs and applied in the following digital libraries: ACM Digital Library, IEEE Digital Library, Scopus, and Web of Science Bibliography.

String terms were defined according to the RQs, including synonyms and terms that frequently appear, and the RP search strings were built combining these terms with AND and OR logical operators. For RQ₁ and RQ₂, the following RP search string was built:

- (i) (“Internet of Things” OR “IoT” OR “Cloud of Things” OR “Internet of Everything” OR “Web of Things”) AND (“Survey” OR “Review” OR “Overview” OR “Systematic Literature Review” OR “SLR”).

For RQ₃, the following RP search string was built:

- (ii) (“Internet of Things” OR “IoT” OR “Cloud of Things” OR “Internet of Everything” OR “Web of Things”) AND (“Semantic” OR “Morphological” OR “Ontology” OR “Taxonomy”) AND (“Interoperability” OR “Interoperate” OR “Interoperable” OR “Integrate” OR “Integration” OR “Integrating”).

The RP inclusion and exclusion criteria were defined aiming at a qualitative selection of the related works. The following RP inclusion criteria were defined:

- (a) Title, keywords or abstract refers to a study to classify other manuscripts or is a SLR on IoT platforms, architectures, middleware and/or semantic interoperability (RP search strings (i) and (ii));
- (b) Primary and secondary studies published between 2011 and 2021 (RP search strings (i) and (ii));
- (c) Secondary studies that address IoT solutions, protocols, frameworks and/or case studies (RP search strings (i) and (ii));

- (d) Primary studies that present semantic interoperability challenges and/or solutions for IoT (RP search string (ii)).

The following RP exclusion criteria were defined:

- (a) Secondary study lacks details on methods and protocol used for its literature review (RP search strings (i) and (ii));
- (b) Primary study approaches IoT interoperability but not in the semantic level (RP search string (ii));
- (c) Repeated or duplicated manuscripts, where the most complete and comprehensive study has been considered (RP search strings (i) and (ii));
- (d) Secondary study focuses on a specific IoT domain, architecture, platform or characteristic (RP search string (ii)).

40.2.2 Conducting the Review

The IoT paradigm has been researched and analyzed from several points of view, resulting in a plethora of studies in the literature over the last twenty years. One of these is presented in [6], where a bibliometric overview of IoT from 2000 to 2019 is reported, seeking to reveal the origin of IoT, assess its main research topics, and discuss the future IoT challenges.

Our first strategy to answer RQ₁ and RQ₂ was to seek as many primary studies as possible using an unbiased search string containing terms related to these RQs. As IoT domains, architectures, platforms, and applications are broad terms, the number of returned studies was overwhelming. Considering only the year 2020, 1815 primary studies and 157 secondary studies were retrieved from the RP data sources, this already excluding duplicated studies. To avoid massive data analysis and extraction, and as the number of secondary studies was significant, we adopted a second strategy that was to seek as many secondary studies as possible to answer these RQs. For RQ₃, which has specific terms, we adopted the first strategy. Once obtained these studies, they were assessed for their actual relevance by employing the RP inclusion and exclusion criteria.

Our SLR started in January 2021 looking for manuscripts in RP data sources using RP search strings. In the first step of the selection process, the titles and abstracts of the returned manuscripts were read for pre-selection, guided by the RP inclusion and exclusion criteria, and by performing the snowballing technique [7] on the secondary studies related to RQ₃. In the second selection process step, a complete reading of the pre-selected manuscripts was carried out for data extraction, in the same way as in the previous step. Tables 40.1 and 40.2 summarizes the obtained results with this selection process for the search string (i) and (ii) respectively.

Table 40.1 Results of the selection process for RQ₁ and RQ₂

Data sources	Returned	Pre-Selected	Selected
ACM Digital Library	166	7	3
IEEE Digital Library	879	41	32
Scopus	1347	34	16
Web of Science	1038	33	17
Total manuscripts	3430	115	78

Table 40.2 Results of the selection process for RQ₃

Data sources	Returned	Pre-selected	Selected
ACM Digital Library	111	55	37
IEEE Digital Library	243	128	79
Scopus	464	58	43
Web of Science	573	94	65
Snowballing	123	40	12
Total manuscripts	1514	375	236

As shown in Table 40.1, for RQ₁ and RQ₂ 3430 manuscripts were returned from RP data sources using the RP search string (i). After the first selection process step, 115 manuscripts were pre-selected, and after the second selection process step 78 manuscripts were selected for data extraction.

As shown in Table 40.2, for RQ₃ 1514 manuscripts were returned from RP data sources using the RP search string (ii). After the first selection process step, 375 manuscripts were pre-selected, being 335 from data sources and 40 from snowballing. After the second selection process step, 236 manuscripts were selected for data extraction, being 224 from data sources and 12 from snowballing.

40.2.3 Reporting the Review

Some essential aspects of the state-of-the-art regarding the semantic interoperability problem in IoT were identified by extracting data from the selected manuscripts. In the reporting phase of this SLR, these manuscripts were categorized, classified, and analyzed for later dissemination of the obtained results to potential stakeholders. Figure 40.1 summarizes this categorization by publication date and the RQs type.

According to the distribution of the selected manuscripts shown in Fig. 40.1, IoT research can be divided into two phases in the last decade. The first one occurred from 2011 until 2015, where investment plans on IoT development were promoted by many governments, such as the “Internet of Things – An action plan for Europe” released by the European Union in 2009, and the “Twelfth Five-Year Development Plan Report on the IoT” announced by the Chinese government in 2010, both reported in [6]. As a result, the number of publications per year gradually increased during this first phase. Furthermore, two secondary

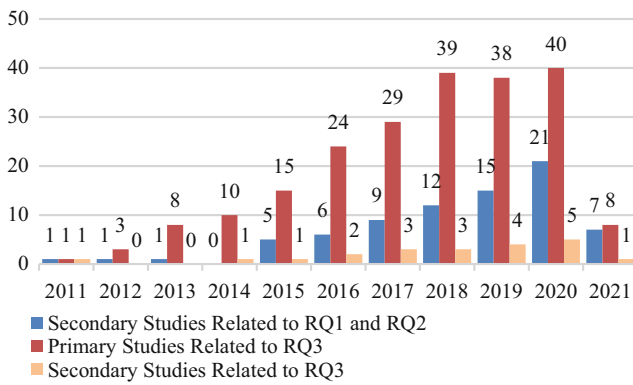


Fig. 40.1 Categorization of the selected manuscripts

studies that became the most influential in IoT-related fields were published: the most-cited study [8], among the selected manuscripts, showed that IoT embodies multiple paradigms and technologies by detailing the main challenges faced in the IoT development, and by providing relevant references for IoT research; and the second most-cited study [9] presented a Cloud centric vision for IoT implementation by highlighting IoT technologies innovative and practices. Both studies not only introduced IoT as an important technology for improving the quality of life and of environment, but also foresaw most challenges faced and still unsolved for IoT development.

The second phase started in 2016, where the number of publications per year increased rapidly, and 84.71% of the selected manuscripts belong to this phase. With respect to RQ3, which deals with IoT semantic interoperability, 82.79% of the selected primary studies and 85.71% of the selected secondary studies belong to this phase. This behavior may suggest that academic research on IoT made an influential breakthrough.

40.3 Results

The selected manuscripts were classified according to the RQs and, for space reasons, the complete list of these manuscripts, with their references and classifications, was put in a supplementary material that can be accessed at <https://bit.ly/3D7Tiyr>. The main SLR results for each RQ are reported and discussed in sequel.

40.3.1 Research Question 1

IoT has had a strong impact on several aspects of everyday life, both from a private and business point of view, as it enables the development of a wide variety of applications that improve people's quality of life in many domains.

IoT systems have been incorporated in these domains to improve communication between heterogeneous machines and between machines and humans, and to improve the performance and processing of produced and shared data. For this purpose, these systems employ intelligent and autonomous devices to sense and interact with the environment, and in some of these domains, such as Transport and Health, machine learning and deep learning techniques have been used [10].

Despite this diversity of domains, the development of IoT applications shares some common challenges and requirements: emphasis on automation and collaboration; and emphasis on data collection, monitoring and exchange. For instance, since the data collected from IoT sensors in a network are time-series, IoT systems for developing applications that deals with this data type must be adaptable to sudden changes in device configurations, to unavailability of these devices, and to network security vulnerabilities [11, 12].

The main IoT application domains of IoT systems are summarized in Table 40.3, where the first column list the main domains extracted from the selected manuscripts for RQ1 and RQ2, and the second column describes the IoT application roles in each of these domains according to [8, 9, 13, 14 and 15].

When analyzing the 215 primary studies from the selected manuscripts for RQ3 that deals with the IoT semantic interoperability, 43 (20.00%) studies are related to Health, 23 (10.69%) studies are related to Transport, 25 (11.62%) studies are related to Environment, 38 (17.67%) studies are related to Smart Home and Smart Cities, 28 (13.02%) studies are related to Agriculture, 26 (12.09%) studies are related to Industry, 3 (1.39%) studies are related to Emerging domains, and 29 (13.48%) studies do not specify a domain

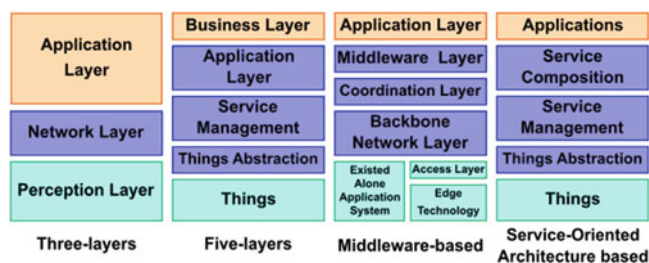
40.3.2 Research Question 2

Architectural models facilitate the understanding of systems and their behavior, and a flexible layered architecture makes it easier to deal with the interconnection problem via Internet of the large number of IoT's heterogeneous things. Furthermore, these things operate with different communication protocols, and use different data formats and interfaces, thus increasing the complexity of the architecture and not providing a clear overview for IoT application developers. As a result, an increasing number of architectures have been proposed and have not yet converged on a reference model [15, 16].

Among the selected manuscripts for the first IoT phase, some of them proposed a general purpose 3-layer architecture [8, 15, 17]. For the second IoT phase, other architectures with

Table 40.3 IoT main application domains

Domains	IoT application roles
Health	Collection, processing, analysis and monitoring of real-time medical data; Monitoring patients and medical devices (things); Aid in medical diagnosis.
Environment/Smart Grid	Distribution and energy consumption management; Transmission control; Enhanced metering architecture; Integration of renewable resources; Energy storage.
Transport/Logistic	Tracking & route mapping; Surveillance applications; Traffic control & prediction; Anomaly detection (e.g., accidents).
Smart Home/Building	Monitoring daily activities; Future events prediction; Sensitive information gatherer; Privacy & access control; Air quality; Artificial illumination control.
Smart Cities	Parking/Waste/Cleaning/Public illumination management; Emergency control.
Agriculture/Farming	Crop/Animal monitoring and protection; Weather forecasting; Improving agricultural productivity; Estimating environmental conditions.
Industry Automation	Manage and improve the global economy and performance of industries; Digital learning; Machine-to-Machine communication enhancements; Power consumption optimization.
Emerging Domains	Internet of Nano-Things (IoNT); Internet of Underwater-Things (IoUT); Internet of Flying-Thing (IoFT).

**Fig. 40.2** IoT architectures

more abstraction layers were proposed, encouraging the development of IoT platforms [16, 18, 19]. An IoT platform can be seen as a middleware layer or a set of sub-layers, placed between applications and things, that provides solutions to frequent problems, such as heterogeneity, interoperability, security, and reliability [15]. One of the most important aspects of IoT platform development, is to base it on a well-defined framework or architecture, and an extensive review on this matter is presented in [16]. Figure 40.2 provides overviews of some IoT multi-layered architectures.

IoT platforms allow for connecting different types of things, and for performing different types of operations on them, such as accessing, managing, and processing their data. With the knowledge obtained from these operations, IoT platforms can make decisions and take actions related to the connected things, playing an important role in supporting the design and implementation of IoT systems and/or applications. As a result, several IoT platforms have been developed, most for general purposes and financed by open international projects or by private industry. As for IoT architectures, the current scenario for IoT platforms is broad, heterogeneous, and characterized by a lack of standardization [16].

For answering RQ₂, we merged the results of the most recent surveys selected by our SLR and analyzed the most

cited platforms. The result of this analysis is summarized in Table 40.4, where the first column list the platforms, the second column their architectural approaches, the third column some main applications, and the last column the supported interoperability level. This table was based on [19, 20, 21 and 22], where an extensive and deep analysis of more than 40 IoT platforms can be found.

Furthermore, we compared with the 215 primary studies from the selected manuscripts for RQ₃, seeking to find the preferred IoT platforms when semantic interoperability is the main objective. Surprisingly, among these studies very few used well-known IoT platforms: 7 (3.25%) used FIWARE; 4 (1.86%) used OpenIoT; 1 (0.46%) used Xively; 1 (0.46%) used X-GSN; 97 (45.11%) developed their own platform; and the remaining 105 (48.83%) do not make use of any IoT platform. However, it is important to highlight that most papers that developed their own platform implemented specific features and used third-party IoT-cloud services platforms for storage and data processing, such as Google Cloud IoT, Amazon Web Services IoT platform, and Microsoft Azure IoT.

40.3.3 Research Question 3

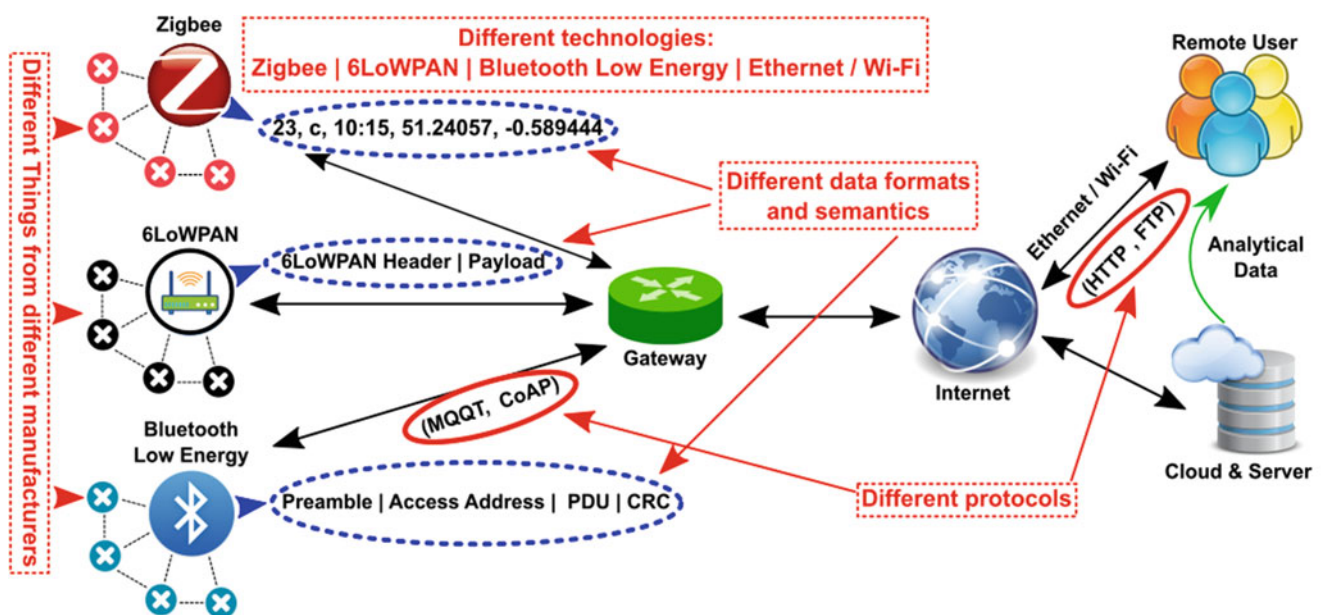
IoT is heterogeneous in terms of both hardware and software at different levels, encompassing different data formats and semantics, different devices from different manufacturers, different wired and wireless networking technologies, and different communication protocols, as shown in Fig. 40.3.

Usually, the things produce large amounts of data, which is first sent to a gateway, where it can be pre-processed, for example by mobile devices, and then forwarded to the cloud for further processing [34], as shown in Fig. 40.3.

Table 40.4 10 IoT most cited platforms

Platforms	Type	Main applications	Interoperability
LinkSmart (Hydra) [23]	Seb, Coa	Pervasive computing environments, Web services.	SeI, SyI
SENSEI [24]	Seb	e/m Health, Smart environments, Smart Home, Smart Cities.	SeI
HERMES [25]	Evb	Enabling large-scale for Smart environments, Web services.	SeI, SyI
OpenIoT [26]	Seb, Clb	e/m Health, Smart environments, Smart Home, Smart Cities.	SeI
FIWARE [27]	Evb, Coa	e/m Health, Smart environments, Smart Home, Smart Cities.	SeI, SyI
PRISMA [28]	Evb, Reo	Pervasive computing environments, Web services.	SyI
CHOReOS [29]	Seb, Cmb	Enabling large-scale for Smart environments, Web services.	SeI, NeI
Xively [30]	Seb, Clb	Home appliances connectivity and management.	NeI
C-MOSDEN [31]	Seb, Cmb	Resource constrained mobile devices.	SeI, SyI, NeI
X-GSN [32]	Seb, Cnb	Deployment and interconnection of physical and virtual sensor networks.	SeI, NeI

Abbreviations: *Seb* Service-based, *Coa* Context-aware, *Evb* Event-based, *Clb* Cloud-based, *Cmb* Component-based, *Cnb* Container-based, *Reo* Resource-oriented, *SeI* Semantic Interoperable, *SyI* Syntactic Interoperable, *NeI* Network Interoperable

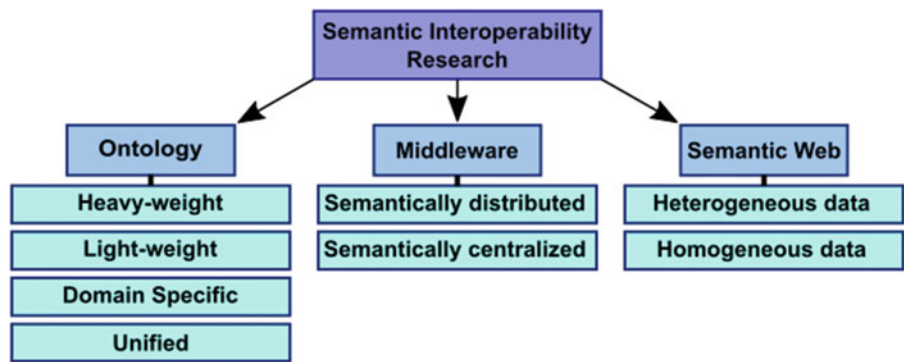
**Fig. 40.3** IoT heterogeneity levels. (Adapted from [33])

The IoT heterogeneity, in addition to making it difficult for users to access the offered services, can lead to their incorrect and/or inefficient use. One solution, for achieving a certain homogeneity in the presence of these different levels of heterogeneity, is interoperability. According to [35], IoT interoperability can be applied on five different levels:

- device*, which refers to the ability for communicating among heterogenous devices using different hardware and software technologies and different communication protocols, and for integrating new devices into any IoT platform.
- network*, which refers to the ability for enabling seamless message exchange between IoT systems through differ-

- ent wired and wireless networks, and for handling issues such as addressing, routing, security, QoS, and mobility.
- syntactic*, which refers to the ability for interoperating data type, structure, syntax, and format used in any exchange of information or service between IoT systems.
- semantic*, which according to W3C [36] refers to the ability for “enabling different agents, services, and applications to exchange information, data and knowledge in a meaningful way, on and off the Web”; and
- platform*, which refers to the ability for communicating meaningful information among several organizations, despite of their geographic locations, by integrating data from multiple domain-specific IoT platforms (e.g., Home, Health).

Fig. 40.4 Semantic models for IoT semantic interoperability



Semantic interoperability aims at providing a standard data representation for the meaningful data exchange between different applications and services. Many semantic models and approaches were proposed in the second phase of the last decade for semantic interoperability in IoT, most of them based on ontology, middleware, and semantic web, as illustrated in Fig. 40.4.

Ontology allows for storing all meta-information and knowledge related to the things, and for sharing its common understanding by different IoT applications. It allows IoT systems to describe the concepts involved in these applications without the need of a globally shared theory. Nowadays, ontologies have become very common in the World Wide Web, and their development is made mainly by domain experts.

Since IoT is a huge scale network of things in which a large number of events are generated spontaneously, it is almost impossible to establish a standard platform. In addition, the presence of heterogeneous hardware and software technologies in IoT poses new challenges for developing applications. In this context, a middleware can provide some standard services to application developers by combining these technologies. While IoT middleware helps in developing new IoT applications and services, it also poses several challenges. A huge amount of IoT middleware can be found in the literature for providing interoperability in various IoT domains, but only a few of them are using semantic models to provide semantic interoperability.

One of the major IoT issues is enabling a single thing to be used as a multimodal node for multiple IoT applications. This requires an infrastructure for connecting things to the Internet, and publish their data so it can be read by machines via the Web of Things (WoT). The semantic web allows for extracting the meaning of raw data generated by the things, and for sharing and reusing enriched information to provide a better semantic interoperability for them. Building the Semantic Web of Things (SWoT) to share universal WoT knowledge it is a huge task.

For finding the most employed semantic model to achieve IoT semantic interoperability, we analyzed the 215 primary studies of the selected manuscripts for RQ₃: 47 (21.86%)

used ontologies; 43 (20.00%) used middleware; 39 (18.31%) used semantic web; 29 (13.48%) used ontologies with middleware; 22 (10.23%) used middleware with semantic web; 16 (7.44%) used ontologies with semantic web; and the remaining 19 (8.83%) make use of others semantic models and approaches.

40.4 Conclusion

This paper presented an SLR addressing the semantic interoperability problem in IoT, which surveys the current main issues and solutions for this problem. This SLR first identified the main IoT domains, architectures, and platforms using only selected secondary studies, and then identified the main semantic models and approaches used for IoT semantic interoperability using selected primary and secondary studies.

According to the selected manuscripts for RQ₁ and RQ₂, the IoT management, security and architecture alternates as critical issues, becoming the IoT challenging main stream research topics. Furthermore, all topics incorporate IoT interoperability as part of the issue and solution. As for IoT semantic interoperability, the temporal distribution of the secondary studies of selected manuscripts for RQ₃ shows that this subject belongs to IoT management and architecture, and has been prioritized in recent years.

According to the secondary studies of selected manuscripts for RQ₃, provisioning interoperability in IoT is very challenging due to its heterogeneous nature and the lack of any standard architecture. As a consequence, extending well-known semantics models is extensively recommended. The majority of this studies suggest that a generic ontology supporting different IoT applications could provide an efficient IoT world. However, most of the primary studies of selected manuscripts for RQ₃ provided specific, isolated solutions, and only a few studies sought a broader solution.

This SLR was performed during the Problem Investigation phase of the PhD's project "An Ontology-Based Approach to Facilitating the Interoperability of IoT Applications in the Health Domain", which is being developed by the first author

of this paper. Today, this project is in the Solution Design phase, and the other phases should be developed in the next 02 years.

Acknowledgements This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We also thank the Brazilian National Council of Technological and Scientific Development (CNPq) and the São Paulo Research Foundation (FAPESP) for sponsoring our research in the context of the Brazilian National Institute of Science and Technology in Medicine Assisted by Scientific Computing (INCT-MACC).

References

1. K. Ashton, That 'Internet of Things' thing. RFID J. (2009)
2. M. Weiser, The computer for the 21st century. *Sci. Am.* **265**(3), 94–105 (1991)
3. B. Kitchenham, Guidelines for performing systematic literature reviews in software engineering, version 2.3. in *EBSE Technical Report EBSE-2007-01*, Keele University, University of Durham, 57 pages, 2007
4. D. Stefanovic, S. Havzi, D. Nikolic, D. Dakic, T. Lolic, Analysis of the tools to support systematic literature review in software engineering, in *The 9th International Conference on Engineering and Technology (ICET 2021)*, vol. 1163, pp. 012013, 09 pages, 2021
5. Parsifal. *Perform Systematic Literature Reviews*. Available: <https://parsif.al>. Accessed 4 Nov 2021
6. J. Wang, M.K. Lim, C. Wang, M.-L. Tseng, The evolution of the Internet of Things (IoT) over the past 20 years. *Comput. Indus. Eng.* **155**, 107174., 17 pages (2021)
7. C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE'14)*, paper No 38, 10 pages, 2014
8. L. Atzori, A. Iera, G. Morabito, The Internet of Things: A survey. *Comput. Netw.* **54**(15), 2787–2805 (2010)
9. J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **29**(7), 1645–1660 (2013)
10. S. Chung, Applications of smart technologies in logistics and transport: A review. *Transp. Res. E: Logist. Transp. Rev.* **153**, 17 (2021)
11. K.R. Choo, K. Gai, L. Chiaraviglio, et al., A multidisciplinary approach to Internet of Things (IoT) cybersecurity and risk management. *Comput. Secur.* **102**, 102136 (2021)
12. N.M. Thomasian, E.Y. Adashi, Cybersecurity in the internet of medical things. *Health Policy Technol.* **10**(3), 100549 (2021)
13. E. Siow, T. Tiropanis, W. Hall, Analytics for the Internet of Things: A survey. *ACM Comput. Surv.* **51**(4), 1–36 (2018)
14. R. Lohiya, A. Thakkar, Application domains, evaluation data sets, and research challenges of IoT: A Systematic Review. *IEEE Internet Things J.* **8**(11), 8774–8798 (2021)
15. A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, M. Ayyash, Internet of Things: A survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tutor.* **17**(4), 2347–2376 (2015)
16. A.C. Franco da Silva, P. Hirmer, Models for Internet of Things environments – A survey. *Information* **11**(10), 487 (2020)
17. Z. Yang, Y. Yue, Y. Yang, et al., Study and application on the architecture and key technologies for IOT, in *International Conference on Multimedia Technology*, 2011, pp. 747–751
18. A. Giri, S. Dutta, S. Neogy, et al., Internet of Things (IoT): A survey on architecture, enabling technologies, applications and challenges, in *International Conference on Internet of Things and Machine Learning (IML '17)*, Article 7, 2017, pp. 1–12
19. A. Farahzadi, P. Shams, J. Rezazadeh, et al., Middleware technologies for cloud of things: A survey. *Digit. Commun. Netw.* **4**(3), 176–188 (2018)
20. W. Kassab, K.A. Darabkh, A–Z survey of Internet of Things: Architectures, protocols, applications, recent advances, future directions and recommendations. *J. Netw. Comput. Appl.* **163**, 102663 (2020)
21. G. Fortino, C. Savaglio, G. Spezzano, et al., Internet of Things as system of systems: A review of methodologies, frameworks, platforms, and tools. *IEEE Trans. Syst. Man Cybern. Syst.* **51**(1), 223–236 (2021)
22. J. Zhang, M. Ma, P. Wang, et al., Middleware for the Internet of Things: A survey on requirements, enabling technologies, and solutions. *J. Syst. Architect.* **117**, 102098 (2021)
23. M. Sarnovský, P. Kostelník, J. Hreno, et al., Device description in HYDRA middleware. *Proceedings of the Workshop on Intelligent and Knowledge Oriented Technologies2*, 1–4 (2007)
24. V. Tsiatsis, A.T. Gluhak, T. Bauge, *The SENSEI Real World Internet Architecture* (IOS Press, Guildford, 2010), pp. 247–256
25. P.R. Pietzuch, Hermes: A scalable event-based middleware, in *UCAM-CL-TR-590 Computer Laboratory*, (Cambridge University Press, Cambridge, MA, 2004)
26. J. Soldatos, N. Kefalakis, M. Hauswirth, et al., OpenIoT: Open source internet of things in the cloud, in *Interoperability and Open-Source Solutions for the Internet of Things*, (2015), pp. 13–25
27. Z. Theodore, P. Andreas, A. Federico, et al., FIWARE lab: Managing resources and services in a cloud federation supporting future internet applications, in *7th International Conference on Utility and Cloud Computing*, 2014
28. J.R. Silva, F.C Delicato, L. Pirmez, et al., Prisma: A publish-subscribe and resource-oriented middleware for wireless sensor networks, in *ICT 2014*, pp. 87–97, 2014
29. H. Vincent, V. Issarny, N. Georgantas, et al., CHOReOS: Scaling choreographies for the Internet of the future, in *Middleware'10 Posters and Demos Track*, 1–8, 2010
30. Xively. Available: <https://xively.com/>. Accessed 04 Nov 2021
31. C. Perera, P.P. Jayaraman, A. Zaslavsky, et al., An internet of things middleware for resource constrained mobile devices, in *47th Hawaii International Conference on System Sciences*, 2014, pp. 1053–1062
32. J.P. Calbimonte, S. Sarni, J. Eberle, et al., XGSN: An open-source semantic sensing middleware for the web of things, in *Proceedings of the TC/SSN@ ISWC*, 2014, pp. 51–66
33. H. Rahman, M.L. Hussain, A comprehensive survey on semantic interoperability for Internet of Things: State-of-the-art and research challenges. *Trans. Emerg. Telecommun. Technol.* **31**(12), 1–25 (2020)
34. A. Gawanmeh, J.N. Al-Karaki, Disruptive technologies for disruptive innovations: Challenges and opportunities, in *ITNG 2021 18th International Conference on Information Technology-New Generations* (Springer 2021), pp. 427–434
35. M. Noura, M. Atiquzzaman, M. Gaedke, Interoperability in internet of things: Taxonomies and open challenges. *Mobile Netw. Appl.* **24**, 796–809 (2018)
36. W3C. Semantic Integration & Interoperability Using RDF and OWL. Available: <https://www.w3.org/2001/sw/BestPractices/OEP/SemInt/>. Accessed 29 Nov 2021

IoT Machine Learning Based Parking Management System with Anticipated Prediction of Available Parking Spots

Grzegorz Chmaj and Michael Lazeroff

Abstract

Machine Learning based designs provide an extensive means to recognize patterns and do various kinds of predictions. In this paper we apply machine learning to the Internet of Things architecture to optimize access to smart parking. We assume that the system will detect which driver parked at which spot, and also will recognize driver's habit of returning to the car. This way we predict that the spot of the driver walking towards the parking lot or garage will soon be available for other drivers. Drivers looking for a parking spot will receive such information in advance, in a form of "Spot N will be available in X minutes". The design operates over the features offered by smartphone devices: to determine the parking spot, determine the walking driver's position and also serving as a base for a mobile application, so the system is convenient to use and doesn't require additional infrastructure.

Keywords

Machine learning · Internet of Things · Smart cities · Smart parking · GPS · Trajectory · Time estimation · ETA · REST API · Coordinates

G. Chmaj (✉)

Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, USA
e-mail: grzegorz.chmaj@unlv.edu

M. Lazeroff

Department of Computer Science, University of Nevada, Las Vegas, USA
e-mail: lazerm1@unlv.nevada.edu

41.1 Introduction

Smart cities implementing Internet of Things design include multiple components such as traffic light control, air quality monitoring, emergency response, smart buildings and smart parkings. Intense traffic and concentration of high number of cars in the centers of the cities led to development of various solution to minimize the time of finding an empty spot; this way also reducing the emission and in-parking traffic. Many solutions implemented already are based on sensors that are mounted at the parking spot (one sensor per one spot). Sensors are monitoring the presence of a vehicle at the parking spot and then either locally control the red/green indicator light, or are wired to a central control unit, that can count number of available spots and display this information to the drivers entering the parking garage. Smart parking approaches can also feature more advanced options like counting vehicles entering given floor, so number of available spots could be updated before these vehicles reached their spots. Other features are smart routing of exiting vehicles (for garages having multiple exits) and balancing the distribution of vehicles over the floors by the criterion of vicinity of free spots to the elevators (so the more occupancy, the farther available free parking spots are from elevators, regardless the floor).

We propose the live prediction of free parking spots ahead of time – for the optimization of traffic and parking time. In our approach, each driver participating in the system has a smartphone with our parking application installed (safe to assume it's obligatory for company parking garages, but not only). Manual spot entering is used to recognize that a particular driver has just parked in a spot of certain identifier (identifier describes the location on the floor but also the floor number). On the other hand, our design recognizes driver's habit of returning back to the car, so we can predict ahead of time, that the certain driver will be freeing the parking spot

within some estimated time. This information is presented to drivers that are currently looking for a parking spot. Our idea is flexible and can have multiple factors included, such as different types of parking spots, different day times when a certain spot is available, parking fees and many other. We consider our system as easy to implement, as no other infrastructure like sensors, wiring etc. is required – smartphones, application and active users is sufficient.

41.2 Related Work

Internet of Things, specifically its implementation to cities – resulting in a concepts called Smart Cities faces multiple challenges, at the same time solving many old problems and providing numerous optimizations of city-related factors. The trends for IoT-enabled smart cities were surveyed in [1] – and included the confrontation of user approach and technological approach, along with the overview of components and their then presence on the market (in numbers) and projections for the following years. Authors also included the overview of IoT-Smart City applications, such as transportation and mobility, smart homes, smart infrastructure, IoT solutions for retail, healthcare, energy optimization and brief discussion of other areas related with smart cities. The technological look at the smart cities – in a perspective of services being delivered – were described in [2]. Authors specified the following services: building management, automation and home automation (for buildings), remote parking management, business fleet management and vehicle telematics (for transportation), home security and people protection (for security area), smart healthcare and hospitals (for healthcare) and optimization of distribution and usage of electrical energy (for electricity). The paper includes the results of survey done among IoT experts regarding specified IoT/Smart Cities aspects.

Aside of the research and engineering work being done for IoT Smart Cities in general, a lot of attention is brought to the transportation aspect of smart cities that is closely related to our work presented in this paper. The systematic analysis of transportation in smart cities was described in [3]. The work includes the taxonomy and results of review of 199 smart cities projects – geographically located around the entire globe. Authors of [4] analyze the IoT technologies selection to be applied into smart transportation. The analysis includes considering specific requirements of the application. Most of the technologies that are considered are IoT-specific or closely related. The aspect of data propagation for the smart city transportation was described in [5], where authors analyze the communications between multiple participants of Intelligent Transport System. Several factors are consid-

ered: latency, heterogeneous connectivity, mesh architecture, cloud-based approach and many others. Similarly, as in our work, authors included machine learning in the solutions – for the goal of optimization and independent decision making. Also the scenarios for smart cities and smart roads are included. The mobile object authentication, being important in the transportation area was de-scribed in [6]. Authors propose multilayer and hierarchical architecture for mobile objects, including: perception layer (mainly sensors, but also other typical IoT devices), network layer, mobility support layer and application layer. The mobility support layer provides the authentication of the mobile objects that are relocating between different environments and systems.

The convergence of Internet of Things and Machine Learning, being applied to transportation in smart cities was described in [7]. Authors describe multiple solutions in which Intelligent Transportation Systems operate using ML, also pointing out the trends and directions that possibly need more work (also mentioning smart parking). Extensive information is provided about ML techniques applied to the smart transportation. Applications considered in that work are: Route-Optimization-Navigation, Parking, Lights, Accident Detection, Road Anomalies and Infrastructure. Each application is analyzed against the ML solutions applied, including the ML-related details.

41.3 Proposed System

The proposed system predicts the time when a parking space is to become available. This system is most useful in places with high parking density, for example, a university parking lot or garage. The prototype of this system aims to leverage a user's smartphone as much as possible and limit the amount of physical hardware needed in the actual parking structures. Example execution:

- (i) An individual is returning to the parking lot where their car is parked.
- (ii) The machine learning model predicts that the person is indeed returning to the parking lot – and a time prediction to return is generated.
- (iii) Patrons seeking a parking space can view the estimated time for a vacant spot through a mobile application.

Architecture consists of: Frontend Client (Mobile Application), Backend Server, REST API, Machine Learning Pipeline and Database. The Backend Server is the main focus of this paper. However, details about the role of the mobile application and some potential features it could include are discussed.

41.3.1 Mobile App

The main interface for users of the system would be through a mobile application. There are a few different purposes that the mobile application would serve:

- *Provide GPS coordinates to the Backend Server:* The Backend Server utilizes GPS coordinates in the Machine Learning Pipeline. Using the GPS API provided with many smartphone application frameworks, the individual's current location can be sent to the server and processed as needed
- *Allow users to pick their spot and checkout:* One barrier to using GPS coordinates to track a user's location is the loss of precision when inside buildings, including parking garages. Because of the limitation, precise GPS information could not determine which spot a user chooses. Since the system uses as minimal extra hardware as possible, users manually select the parking spot and "check out" when they leave. From a user experience standpoint, this area would benefit most from having an automatic system in place.
- *Display helpful information to those looking for a parking spot:* The mobile app can display beneficial information – such as a map of the parking lot, which spots are currently open, and the times for spots that are estimated to be available.

Specific implementation details and features to be included can be tweaked to suit individual needs. For example, in a university setting where there are many different parking spots for varying roles, it would be beneficial to have additional permissions for each user and potentially only display locations they have access to. Furthermore, existing systems such as parking passes and metered pay could be integrated into the mobile app for a comprehensive experience for the user. Depending on the parking lot provider's exact needs, a network of different parking lots could also be connected and integrated into the application. However, in this system prototype, the three features outlined above are essential to the application (Fig. 41.1).

41.3.2 Database

Google's Firebase databases were chosen for this system. This database service contains many benefits such as development speed, client syncing, and ease of use [8]. Furthermore, with Firebase's two different types of databases – the Real-Time Database and Firestore, and the tight integration with Google Cloud Platform services like data analytics, it can be quite valuable with IoT systems [9].



Fig. 41.1 Example mobile UI

- *Real-Time Database:* The Real-Time Database contains the state of the parking lot. Individual spots are mapped in the database with information such as an identifier, type of spot, and relative location. The RTDB also contains the time predictions for spots to open – which can be queried.
- *Firestore:* The Cloud Firestore database contains all other data for the system. This includes information such as users and their permissions and useful information such as coordinate information for system users.

41.3.3 Machine Learning Model

41.3.3.1 Purpose

The goal of the machine learning model is to determine whether a set of coordinates is returning to the parking lot. The model receives a list of GPS coordinates as input and outputs a binary classification of 0 (does not return) or 1 (does return).

41.3.3.2 Dataset

The dataset used consists of GPS coordinates (latitude and longitude pairs) that simulate individuals' trajectories when traveling near the parking garage. The parking lot chosen for the system is a heavily trafficked garage at the University of Nevada, Las Vegas [10]. Entrances to the different locations were labeled and are shown in Fig. 41.2.

Fig. 41.2 Garage entrance labels



Fig. 41.3 Trajectories



The trajectories were created using Google Earth's path creation tool. The dataset consists of 776 simulated trajectories that are each four coordinates long. Each trajectory was labeled as entering the garage or missing the garage (Fig. 41.3).

The decision to use four long coordinate trajectories was not arbitrary. If the trajectories are too long – the predictions

may not be made as frequently as needed. If the trajectories are too short – say, every one or two coordinates, it would be more challenging to train the model to make proper predictions. Four coordinates is a happy medium for achieving frequency in polling and accuracy.

The distributions of the different trajectories are shown in Table 41.1. A unique coloring is given to each category

Table 41.1 Trajectory distributions

Location	Entered garage	Missed garage
A	136 (Red)	131 (Blue)
B	149 (Green)	130 (Yellow)
C	82 (White)	74 (Dark Orange)
D	38 (Light Orange)	36 (Purple)
Total	405	371

of trajectories to distinguish them visually. It is essential to note the decision-making process behind the distribution of trajectories. For this specific garage, entrances A and B are more heavily trafficked than entrances C and D due to their location on campus. This bias towards locations A and B is reflected in the data – however, simulated data can never capture the actual traffic of the entrances to this garage. In a production environment, collecting actual traffic distributions to the garage entries would be crucial for accurate time prediction results.

41.3.3.3 Model Parameters

Each input to the model consists of four pairs of latitude and longitude coordinates (eight inputs total). The dataset is partitioned to a 70/30 training and testing split. Once partitioned, the data is scaled using the *MinMaxScaler* found in the Sklearn Python library. The *MinMaxScaler* is fit transformed to the training data, and then the testing data is subsequently transformed. Once the data is scaled, it is ready to be fed into the model for training.

41.3.3.4 Model Architecture and Training

The model is sequential with three dense layers: an input, a hidden, and an output layer. Eight neurons in the first layer correspond to the four pairs of latitude and longitude coordinates. The input layer uses the Rectified Linear Unit (ReLU) activation function [11]. The single hidden layer contains six neurons and as well uses the ReLU activation function. A single neuron and the sigmoid activation function are used for the output layer. The goal of the output layer is the binary classification of whether the coordinates enter the garage or miss the garage.

Loss for the model is calculated using the Binary Crossentropy function [12]. A learning rate of 0.01 is specified for the Adam optimizer, and the model is trained using 200 epochs as the limit. An early stopping callback function is declared, which monitors the validation loss for the model. It has a 15 epoch patience parameter and aims to prevent potential overfitting of the model.

41.3.3.5 Model Results

Valuable insights can be extracted from the results of the model training. The model accuracy steadily increases

throughout 198 epochs, concluding at an average of about 95% accuracy on the training and validation data. A large gap between the training and validation data curves is absent, which suggests no significant amount of overfitting in the accuracy plot. For the model loss plot, the gradual decrease in the curve of both the training and validation data indicates that no considerable overfitting of the model occurs. The decline rate suggests that the model's learning rate is appropriate and does not learn too slow or too fast. The training and validation loss concludes at approximately a 15% loss. Although the accuracy and loss plots suggest that the model is not overfitting to the data during the training phase, testing on unseen data is essential for gauging the model's actual performance (Figs. 41.4, 41.5 and 41.6).

A separate testing dataset was created for testing the trained model. It contains 27 trajectories that miss the garage and 31 trajectories that enter the garage. In Fig. 41.7, the trajectories entering the garage are colored purple, and those missing are blue (Fig. 41.8).

We get precision results of 96% when the model predicts coordinates are not returning to the garage and precision of 97% when the model predicts they will. This is about on par with the results of the training data seen in the accuracy plot. These are promising results – the model is performing just as well with unseen data as it did with the training data. A relatively simple model (recall this model only has three layers) can accurately predict whether a set of coordinates is returning to a location. This finding could extend to other systems trying to determine some type of user intent where coordinates are the primary input. In this system, it is being used for predicting user intent to return to a location, however, this approach could be shaped for other types of intent depending on the specific goals of that system. Comparisons to similar models will be completed in future work.

41.3.4 Backend Server

The Backend Server is where the main work of the system takes place. The goal of the Backend Server is to have a scalable API that the front end (mobile application) can efficiently reference. The API follows the REST architecture and uses the FLASK web framework [13]. The benefits of using FLASK for the REST API were the development speeds, ease of use, and future scalability for features and traffic (Fig. 41.9).

41.3.5 Geofences

A geofence acts as a virtual perimeter around a physical location [14]. To create the geofences necessary for the system, polygons are made using Google Earth by defining

Fig. 41.4 Early stopping callback triggered

```

Epoch 194/200
17/17 [=====] - 0s 4ms/step - loss: 0.1411
Epoch 195/200
17/17 [=====] - 0s 3ms/step - loss: 0.1292
Epoch 196/200
17/17 [=====] - 0s 4ms/step - loss: 0.1269
Epoch 197/200
17/17 [=====] - 0s 3ms/step - loss: 0.1250
Epoch 198/200
17/17 [=====] - 0s 3ms/step - loss: 0.1297
Epoch 00198: early stopping

```

Fig. 41.5 Model accuracy

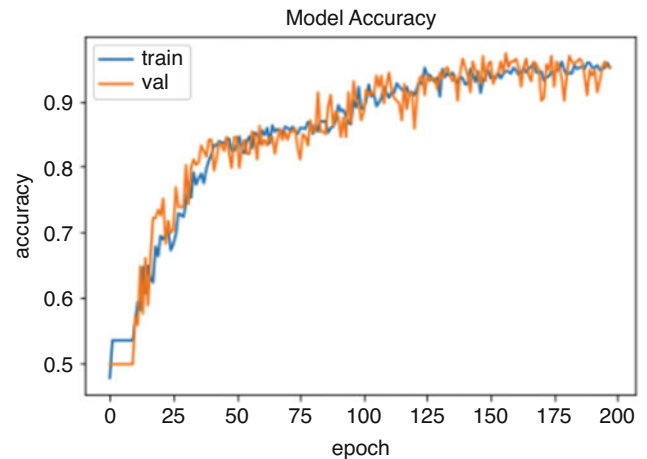
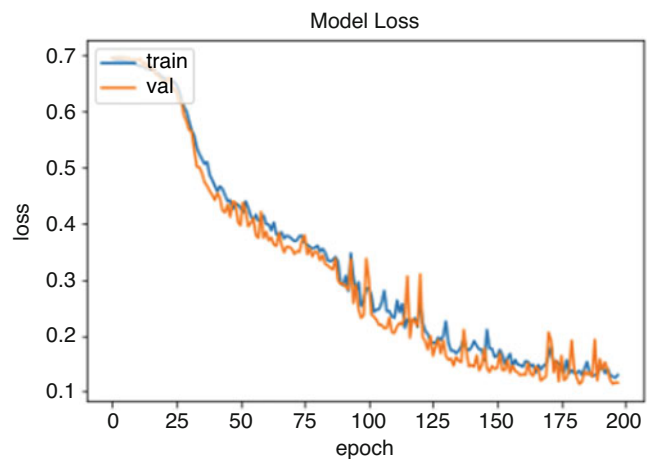


Fig. 41.6 Model loss



bounds around a region. The goal of the geofences is to limit unnecessary processing (Fig. 41.10).

Two geofences play critical roles in the pipeline.

- **Inner Geofence** – The inner geofence is a polygon defined by coordinates that encompass the garage. If coordinates are found within this geofence, we know that the person has returned to the garage.
- **Outer Geofence** – The outer geofence denotes the region around the garage where the model can reasonably predict. The perimeter defined by the outer geofence is based on several factors, such as the paths leading back to the lot,

the traffic, and the buildings or structures surrounding the lot.

It is simple to determine whether a coordinate is inside the bounds of a geofence. Since a geofence is just a polygon defined by a set of coordinates (latitude and longitude pairs), a coordinate is inside the geofence if it exists inside the polygon. Using the Python math library Shapely, a polygon is defined using the coordinates from Google Earth [15]. Once the polygon object is created, coordinates can then be tested for if they are contained within the polygon.

Fig. 41.7 Testing trajectories



Fig. 41.8 Test set results

	precision	recall	f1-score	support
0	0.96	0.96	0.96	27
1	0.97	0.97	0.97	31
accuracy			0.97	58
macro avg	0.97	0.97	0.97	58
weighted avg	0.97	0.97	0.97	58

With the pipeline, the last coordinate provided in the request is first used to determine whether it is inside the inner geofence. If it is, the user is in the garage, so quit. If the user is not in the inner geofence and is not in the outer geofence, quit. The machine learning model runs when the user is within the outer geofence but outside of the inner geofence. The purpose of this again is to maintain reasonable predictions by the model. It is only desirable to run the model when there is a likely chance they are returning to the garage instead of just polling continuously.

41.3.6 Model Prediction Consequences

As the model is a binary classification model, there are only two prediction possibilities – a prediction that the coordinates

are going to the garage or not. If the model predicts the coordinates will not return to the garage, then there is no more processing that needs to be done, and the backend will wait for another POST request with more coordinates. If the model does predict that the coordinates will return to the garage, however, then the time prediction part of the pipeline can occur.

41.3.7 Time Prediction

Time predictions will only occur when the model predicts a return to the garage. The approach to predicting time is made as simple as possible in the first iteration of the system. Placemarks are created in Google Earth that define arbitrary “regions” in the garage. Spots are assigned to these different

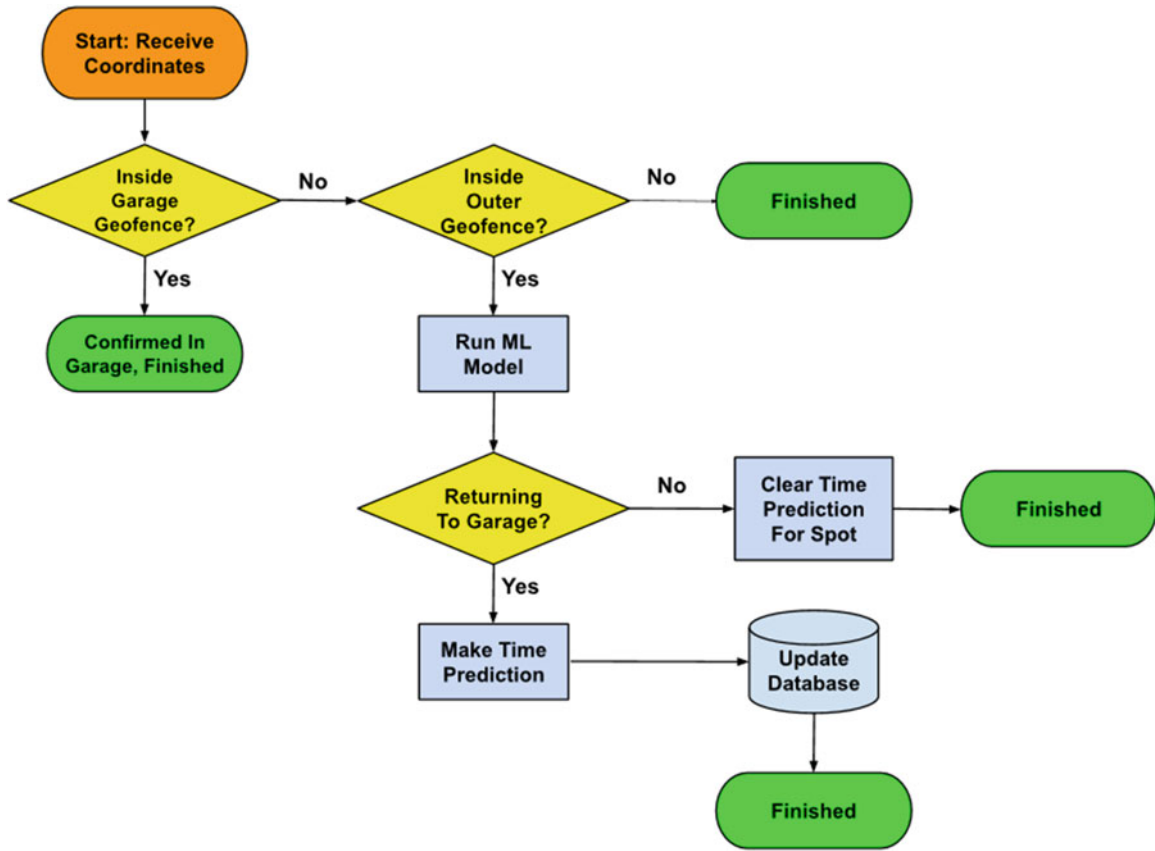
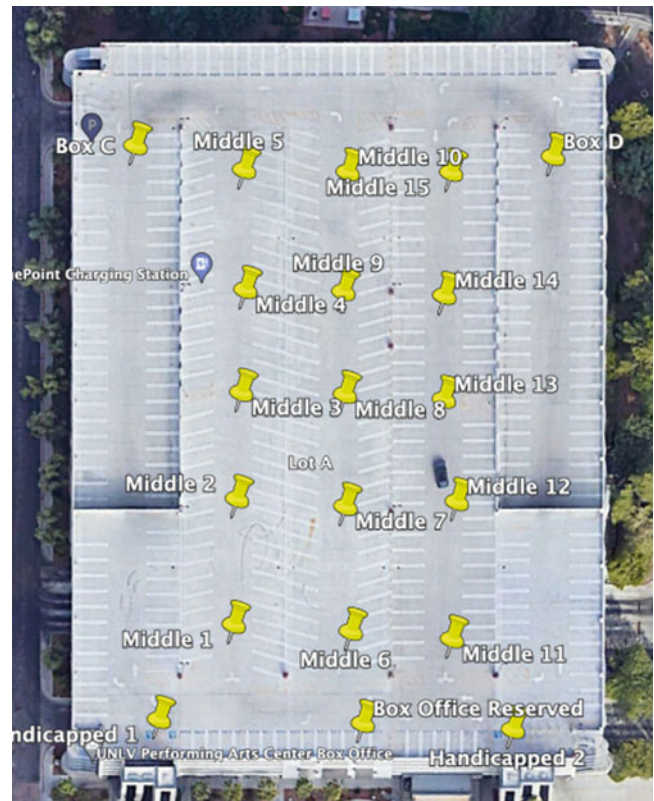


Fig. 41.9 Backend pipeline diagram

Fig. 41.10 Outer geofence



Fig. 41.11 Regional placemarks

regions based on their location. These placemarks are used to calculate the distance from a user's current location and their parking spot location.

$$Time = \frac{Distance}{Speed}$$

Description:

- Take the user's current location (provided in the POST request) and the regional location of their parking spot and calculate the distance.
- Set speed as a constant – to an average walking speed of 1.5 m/s
- Calculate the time using distance and speed

After calculating the time prediction, the database is then updated.

41.3.7.1 Multi-floor Garages

Although this system focused only on the first floor of the parking garage, other floors can be easily scaled using this approach by including a constant travel time to move between floors.

41.3.7.2 Dynamically Calculating Speed

It is not the case that everyone travels at the same speed – especially at a university setting where there are many modes of

transportation used to get around. Instead of setting speed as a constant, an individual's speed could be calculated using the coordinate data and time stamps. With this information, we can calculate the movement speed unique to every individual to create more precise time predictions.

41.3.7.3 Accounting for Uncertainties

To account for uncertainties when patrons arrive back at their car and leave the parking lot – a buffer time can be added on top of the prediction time. A buffer time would help account for individuals' time to settle and leave their parking spot (Fig. 41.11).

41.4 Conclusions

The work presented in this paper addresses the problem of congestion in parking garages. It describes the machine learning based approach to predict the availability of parking spot ahead of time – using driver's geographical position, walking path, geofences and other aspects. The proposed system delivered high accuracy of predictions. Once the system detects that a driver walks towards the parked car, it notifies driver(s) currently looking for available spots – that the parking spot with a specified symbol will be available in a certain amount of time. In this preliminary work, the time at which the spot is going to be available is estimated independently of a certain driver, but in the future work the personal

data is going to be included to get more precise individual time predictions. Other elements planned are parking spot selection preferences, different categories of parking spots and consideration of reserved parking spots with flexible time schedule.

Acknowledgment This material is based upon work supported by the National Science Foundation under Grant No. 1950872

References

1. A. Alavi, P. Jiao, W. Buttlar, N. Lajnef, Internet of Things-enabled smart cities: State-of-the-art and future trends. *Measurement* **129**, 589–606 (2018). <https://doi.org/10.1016/j.measurement.2018.07.067>. ISSN 0263-2241
2. E. Park, A. del Pobil, S. Kwon, The role of Internet of Things (IoT) in Smart Cities: Technology roadmap-oriented approaches. *Sustainability* **10**(5), 1388 (2018)
3. T.G. Crainic, G. Perboli, M. Rosano, Q. Wei, Transportation for smart cities: A systematic review, in *The Eleventh International Conference on City Logistics*, June, 12–14, 2019
4. W. Ayoub, A. Ellatif Samhat, M. Mroue, H. Joumaa, F. Nouvel, J.C. Prévotet, Technology selection for IoT-based smart transportation systems, in *Vehicular Ad-Hoc Networks for Smart Cities*, Advances in Intelligent Systems and Computing, ed. by A. Laouiti, A. Qayyum, M. Mohamad Saad, vol. 1144, (Springer, Singapore). https://doi.org/10.1007/978-981-15-3750-9_2
5. A.A. Brincat, et al., The internet of things for intelligent transportation systems in real smart cities scenarios, in *IEEE 5th World Forum on Internet of Things (WF-IoT)*, 2019
6. M. Saadeh, A. Sleit, K.E. Sabri, W. Almobaideen, Hierarchical architecture and protocol for mobile object authentication in the context of iot smart cities. *J. Netw. Comput. Appl.* **121**, 1–19 (2018). <https://doi.org/10.1016/j.jnca.2018.07.009>
7. F. Zantalis, G. Koulouras, S. Karabetsos, D. Kandris, A review of machine learning and IoT in smart transportation. *Future Internet* **11**(4), 94 (2019)
8. <https://firebase.google.com/docs/database>
9. B. Padmaja, E. Patro, S. Mahurkar, G. Akhila, Google firebase based modern IoT system architecture. *Int. J. Eng. Res. Technol.* **9**(8), 107–110 (2021) Accessed 7 July 2021
10. <https://www.unlv.edu/maps/pkg-1>
11. A.F. Agarap, Deep learning using rectified linear units (ReLU), *arXiv:1803.08375*, 2018
12. https://www.tensorflow.org/api_docs/python/tf/keras/losses/BinaryCrossentropy
13. <https://flask-restful.readthedocs.io/en/latest/>
14. <https://www.verizonconnect.com/glossary/what-is-a-geofence/>
15. <https://shapely.readthedocs.io/en/stable/>

Channel State Information Spectrum Gap Filling Using Shallow Neural Networks

42

Avishek Mukherjee, Beata Hejno, and Manish Osti

Abstract

We propose CSIFill, a novel system to predict the Channel State Information (CSI) in indoor wireless networks. CSIFill can estimate the CSI on different frequency sub-carriers by using the CSI measurements from neighboring frequencies. CSIFill is different from traditional estimation techniques which attempt to recreate the wireless channel and instead relies on already collected CSI data, to predict the CSI on different wireless frequencies. This is especially useful in indoor wireless networks where an Access Point (AP) needs to periodically measure the CSI on other frequency channels to find better data rates. CSIFill can be used to automatically determine when to switch to another channel to obtain better service without any additional probing or overhead. Our initial results with CSIFill have been very encouraging. CSIFill was evaluated using real world experimental CSI data and was found to accurately estimate CSI data for up to 7.5 MHz channel bandwidth using a shallow neural network.

Keywords

Channel state information · Deep learning · Wireless networks · Channel estimation · Channel prediction · Neural networks · Shallow neural networks · Wi-Fi networks · Quality of service · CSI collection

A. Mukherjee (✉) · B. Hejno · M. Osti
 Computer Science and Information Systems, Saginaw Valley State
 University, University Center, MI, USA
 e-mail: amukher1@svsu.edu

42.1 Introduction

There has been an exponential increase in internet capable devices in recent years. According to [1], the average US household has 25 connected devices as of March 2020. Most of these devices use Wi-Fi to connect to the AP/router on either the 2.4 or 5 GHz frequency bands. The 2.4 GHz frequency spectrum has roughly 80 MHz of bandwidth available for use in the United States, and is divided into several overlapping channels, each assigned about 20 MHz of frequency bandwidth. When a device connects to the wireless router, it is usually assigned a channel (e.g. 1, 6, 11 etc.), and all communications take place over the frequency spectrum associated with that particular channel. Since there can be several devices communicating over the same wireless channel, devices coordinate the transmission of data among themselves according to the protocols specified in the Wi-Fi standard [2, 3] to avoid falling into collision domain.

Wi-Fi networks typically use Orthogonal Frequency Division Multiplexing (OFDM) and Multiple Input Multiple Output (MIMO) for modulation. In Wi-Fi networks, the Channel State Information is a vector of complex numbers that can be used to indicate the quality of the wireless channel. More specifically, it refers to the channel coefficients from the transmitting antennas to the receiving antennas for each OFDM subcarrier. In practice CSI is a representation of the actual multi-path components of the signal that undergo several physical phenomena like reflection, diffraction etc. from the transmitter to the receiver. The CSI between any transmitter and receiver antenna pair can be measured at the receiver and sent back to the transmitter to make decisions like rate and channel selection among others. This paper investigates the accuracy of predicting the CSI in certain frequencies without actually probing the channel, but rather by using already existing CSI measurements from nearby wireless channels. Since the CSI data on a particular fre-

quency represents the actual physical paths taken by the signal, in theory it should be possible to estimate these channel coefficients on other channel. This is very useful, as it frees up APs from continuously monitoring channel quality on other channels. In addition, with wireless devices utilizing increasingly larger antenna arrays and occupying larger frequency spectrum, measuring the CSI between all transmitting and receiving pairs on all wireless channels can be time consuming. Approximating some of these measurements will allow network access points to quickly decide on modulation parameters and achieve better service with higher speeds.

CSIFill is a data-centric approach to solve the above problem. It uses shallow neural networks to predict the CSI between neighboring sets of subcarriers with high accuracy. We conducted real world experiments to measure the CSI under varying channel conditions and evaluated the performance of CSIFill against the actual measured CSI. Our results show that CSIFill can be used to accurately measure up to 24 channel coefficients or up to 7.5 MHz of frequency spectrum, in a 20 MHz channel by learning the channel coefficients from neighboring wireless channels. The performance of CSIFill was also evaluated using different learning algorithms as well as varying the number of measurements it uses to train the network. Overall, CSIFill performs very well and can accurately predict the missing channel coefficients with a very low error rate.

The rest of the paper is organized as follows. Section 42.2 discusses related work. Section 42.3 provides a high level overview of the problem and some theoretical background. Section 42.4 discusses the details of CSIFill. Section 42.5 evaluates the performance of CSIFill on real world data. Section 42.6 concludes the paper.

42.2 Related Work

CSI estimation has remained a popular area of research to improve the performance of WiFi networks. This section explores some of the related work in this area. Some of the earlier work involving CSI estimation is outlined in [4] which uses a probabilistic approach to predicting the CSI based on prior observed values. More recently, some research has been conducted using machine learning techniques to predict the CSI. In [5], the authors use a deep learning approach to develop a low computational complexity CSI predictor for 5G networks. While CSIFill also uses a machine learning approach, it is widely different from the methods described above and is also aimed at indoor wireless networks, which poses a different set of challenges towards estimating the CSI. The authors of [6] attempt to solve a similar prob-

lem by developing a channel prediction framework for LTE networks and attempt to design an accurate channel model for predicting the CSI. Crepaldi et al. [7] and Shirani-Mehr et al. [8] is also different from our solutions as it predicts the channel quality for larger antenna configurations by combining measurements in smaller configurations. CSIFill instead attempts to estimate the CSI in the frequency domain from measurements in neighboring channels. There has also been some recent studies which use the sparsity in the space domain to improve the efficiency of CSI generation. For example, in [9], the authors propose solutions using temporal and spatial correlation to reduce the training overhead. Other proposed methods include estimation in LTE networks like optimal training signal design for one flat channel and using the sparse nature of the channel [10–12] to estimate the CSI. These problems are similar in nature to the proposed CSI prediction problem, however these are primarily geared towards LTE and cellular networks where the channel conditions and spectrum availability are very different from indoor wireless networks. Finally CSI estimation is a different application than CSI localization methods described in [13, 14] or security [15, 16]. We also note that CSI compression methods such as those published in [17] are complementary to CSIFill and can be used to actually improve their performance.

42.3 Overview of CSIFill

This section provides a high level overview of CSIFill.

The CSI measured by the receiver can be represented as the summation of the multiple signal propagation paths from the sender antenna to the receiver antenna. This can be seen in Fig. 42.1 which shows the absolute values of some measured CSI data. Thus, the CSI data on each OFDM subcarrier can be approximated as

$$H = \sum_{p=1}^P \alpha_p e^{jf\delta_p}.$$

where α_p and δ_p denotes the amplitude and delay of path p and P is the total number of multipath components from the sender to the receiver.

While the fundamental channel model is straightforward, in practice there are several factors such as noise, linear and non-linear phase errors [18] that make it difficult to utilize the model directly to predict the CSI. CSIFill uses a different approach. Instead of trying to recreate the exact paths taken by the signal, it uses neural networks to learn the wireless channel by training the network on a wide variety of CSI data captured on different environments.

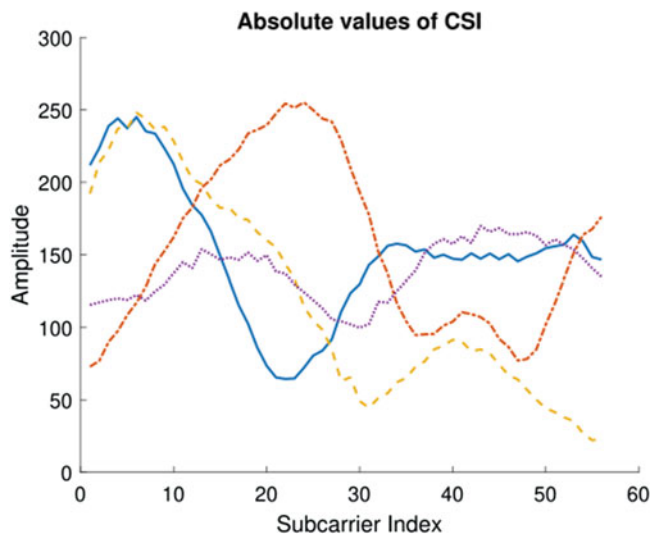


Fig. 42.1 Measured CSI in a high mobility environment

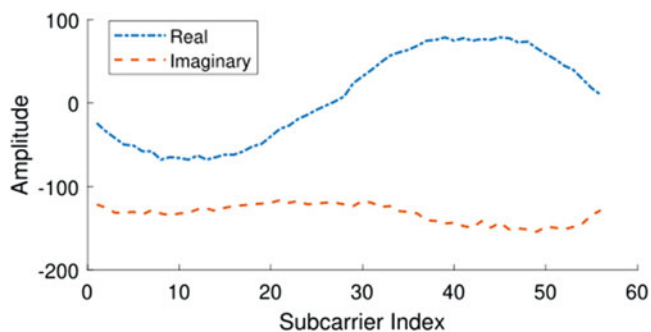


Fig. 42.2 A typical example of CSI data

A typical example of CSI coefficients is illustrated in Fig. 42.2. The figure shows the real and imaginary parts of the channel coefficients on 56 subcarriers measured by the receiving antenna. It can be seen that the data is continuous and sinusoidal in nature. CSIFill uses a shallow neural network to make two types of predictions. First, it uses the measurements from the beginning and end of the CSI data, to estimate the channel coefficients in the middle subcarriers. This is primarily the main the focus of this paper and is referred to as Bounded Estimation (BE). The paper also does some preliminary investigation into the performance of CSIFill in the event of an unbounded estimation (UBE), by only using the subcarrier measurements at the beginning of the channel to predict the remaining subcarriers.

CSIFill uses a shallow neural network with a single hidden layer of size 16 to perform Bounded Estimation. The choice of neural network is explained in later sections. CSIFill also defines 2 different variations of network by using different backpropagation algorithms. In the first case, it uses Levenberg-Marquardt backpropagation (LM) for training the network. In the second case, it uses Bayesian Regularization

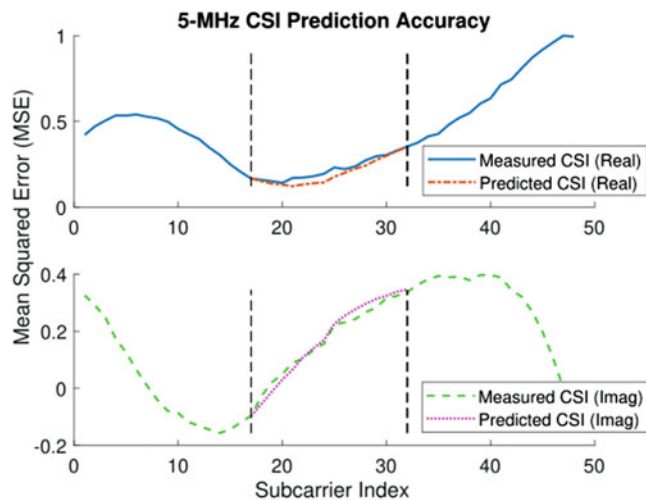


Fig. 42.3 Example bounded prediction by CSIFill

(BR) to update the network weights. This was done primarily for 2 reasons. First, regularization avoids overfitting the training data. Secondly, LM is computationally faster and can be used on devices with limited computing power. Ultimately, it is a trade-off between speed and performance, and we have included an evaluation for both algorithms.

As mentioned earlier, CSIFill was also used to perform unbounded estimations. In this case, a deeper neural network was trained with 3 hidden layers, each with 32 neurons. While this part of the paper was only a preliminary study, the accuracy of CSIFill was very encouraging.

Figure 42.3 shows a typical estimation of the Channel State Information performed by CSIFill. It can be seen that the real and imaginary parts of the estimated CSI follow very closely to the actual data. This was a bounded estimation of 16 subcarriers or 5 MHz on the 2.412 GHz channel using a Bayesian Regularization (BR) backpropagation trained neural network.

42.4 Details of CSIFill

This section outlines the details of the learning algorithm as well as the different configurations used in CSIFill.

42.4.1 Training Algorithms

As mentioned earlier, two versions of CSIFill was developed. Both versions implement a feed forward neural network or multilayer perceptron (MLP) with different backpropagation algorithms. The first version uses Levenberg-Marquardt backpropagation to adjust the weights of the neural network. The second version uses Bayesian Regularization. A high level overview of the two algorithms are outlined below.

- Levenberg-Marquardt(LM) backpropagation: The LM algorithm is a fast method for training medium sized networks with a single hidden layer. LM is used to solve non-linear least squares problems by combining features from Gauss-Newton's method and gradient descent. Basically, LM finds the best weights by minimizing the total squared error between the target and predicted values. It then updates the weights using a Taylor series approximation of the Jacobian. The key idea is LM dynamically dampens the update with a parameter, to avoid overshooting or getting stuck in a local minima. CSIFill uses a pre-existing implementation of the Levenberg-Marquardt algorithm available in MATLAB [19].
- Bayesian Regularization (BR): The BR algorithm is used in conjunction with Levenberg Marquardt back propagation and is primarily used to avoid overfitting the training data. It has been shown [20] to have better performance than the LM algorithm. The key idea is that instead of minimizing the mean squared error, BR uses a probabilistic distribution of network weights to minimize the sum of squared errors in the model.

42.4.2 Network Configuration

We developed a heuristic to find the best configuration for the number of hidden layers in the network as well as the size of each layer. A small subset of the collected CSI was randomly chosen and supplied to the neural network to evaluate its performance across different configurations. The network was trained to estimate 5 MHz of CSI data using neighboring subcarriers. This was repeated for both bounded and unbounded estimations.

42.4.2.1 Configuration for Bounded Estimation

For bounded estimations, the network was trained using 3 hidden layer configurations. The first configuration consisted of a single hidden layer and an additional hidden layer was added for every subsequent configuration. In addition, the network was also trained by varying the size of each hidden layer using 8, 16 and 32 neurons. Finally, the above process was repeated for both LM and BR algorithms.

Figure 42.4 shows the mean squared error (MSE) performance for each configuration for both algorithms when estimating 5 MHz of CSI data. It can be seen that increasing the number of hidden layers beyond 1 produces very little improvement in the prediction. In addition, increasing the size of each hidden layer beyond 16 also does not improve the

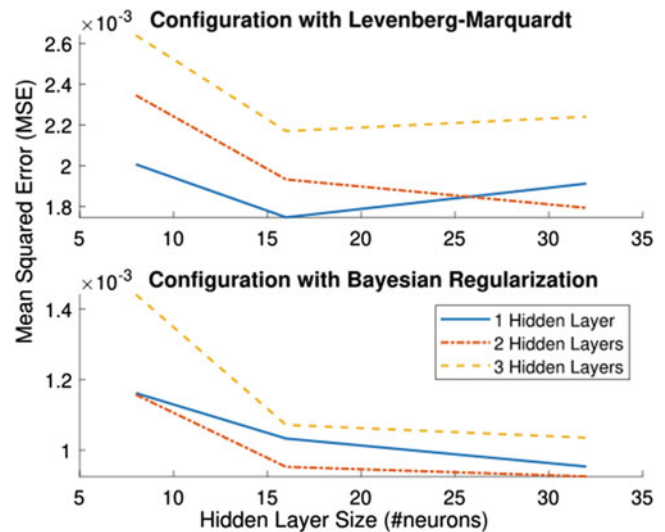


Fig. 42.4 Network configuration selection for bounded predictions

performance significantly. Figure 42.5 shows the selected architecture for bounded estimations, a shallow neural network with a single hidden layer with 16 neurons that was used to train the experimental CSI data.

42.4.2.2 Configuration for Unbounded Estimation

A similar process was followed for selecting the best configuration across both algorithms for unbounded estimations. Using the same CSI data subset, a total of 9 different configurations was tested, with 1, 2 and 3 hidden layers along with 8, 16 and 32 neurons for each layer. The performance can be seen below in Fig. 42.6.

We note that since unbounded estimations are a much harder scenario to train, it is likely that moving to higher configurations would have further improved the performance. However, due to complexity constraints, we selected a configuration of 3 hidden layer networks along with 32 neurons for each layer as shown in Fig. 42.7.

42.4.3 Training the Network

The training process was similar for both training algorithms. The general process is described below.

- Suppose CSIFill is used to estimate the channel coefficients for x subcarriers. Then the feature set was built using the channel coefficients from the remaining $[N - x]$

Fig. 42.5 Network architecture for bounded predictions

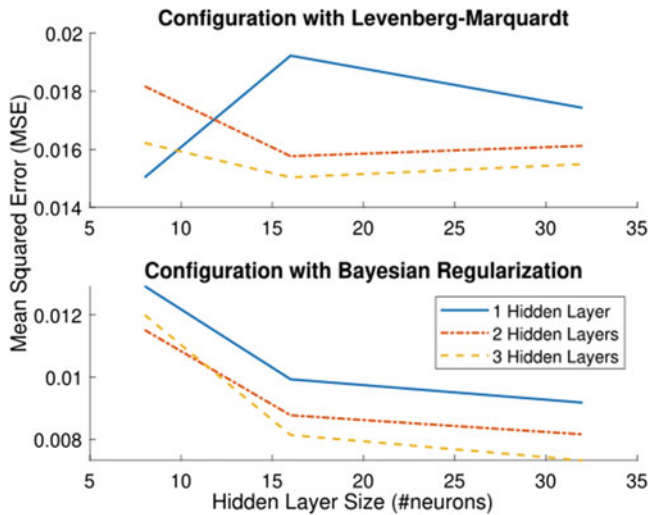
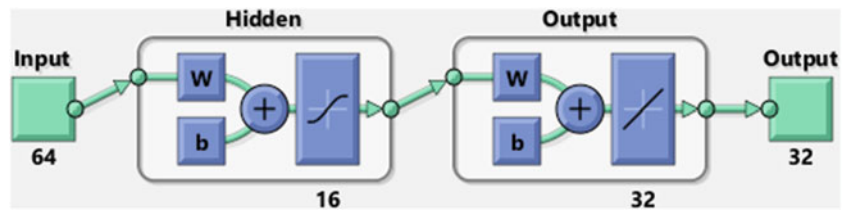


Fig. 42.6 Network configuration selection for unbounded predictions

subcarriers where N represents the total number of subcarriers. In addition, since the channel coefficients are complex, the real and imaginary components were extracted from the coefficient and added separately to the feature set. Thus the input layer for the network consists of $2 * [N - x]$ neurons.

- The target set consists of the channel coefficients for the x subcarriers that are to be estimated. These were also split into real and imaginary parts and linearly concatenated to form $2 * x$ neurons in the output layer for the network.
- The data set was divided into training, validation and test sets with ratios of 0.50, 0.15 and 0.35 respectively.
- The data division was done in blocks to intentionally keep some test locations completely isolated from the training phase. This would pose a greater challenge to CSIFill.
- Finally, to avoid overfitting, 10 identical networks are trained, with the stopping criteria set to the earlier of 500 epochs or 6 consecutive failed validation checks. The rationale for this is that each network starts of with different randomized initial parameters and may converge at different points, this will result in a generalized approach to approximating the target CSI data.

Figures 42.8 and 42.9 shows some statistics measured during a typical training phase. In this scenario, the network was trained to predict 5 MHz of CSI data using LM backprop-

agation. It can be seen that the network converges at around 27 epochs, with a MSE performance of around 0.002.

42.4.4 Testing the Network

CSIFill uses a standard testing process by measuring the performance of the trained network using the Mean Squared Error value to compute the difference between the predicted output versus the actual data. The prediction was computed from all 10 trained networks described above and the average of all the predictions was calculated to represent the overall prediction. As mentioned earlier, this was also done to prevent the model from fitting the noise in the CSI data. For example, Fig. 42.10 shows the distribution of the Mean Squared Error values for every predicted subcarrier across all test cases in a 5 MHz bounded estimation trained network. It can be clearly seen that most of the errors are very small and mostly exist due to the noisiness of the original data.

42.5 Evaluation

CSIFill was evaluated using real world experimental data collected over the course of one week at different locations in an university setting. The performance of CSIFill was measured using the test set described above. The evaluation process is described below.

42.5.1 Data Collection and Pre-processing

We ran several experiments using Atheros CSITool [21] on two laptops that were configured with AR9462 wireless cards. We enabled only a single antenna for both the transmitter and receiver and so each CSI vector had dimensions of $[1 * 1 * 56]$ representing the 56 subcarrier values on a wireless channel across each receiver transmitter pair. The experiments were conducted by continuously moving the receiver and the transmitter to introduce significant variance in the wireless channel. This allowed us to simulate different types CSI data representative of real world wireless channels. We collected over 90,000 CSI data across 10 different locations inside a classroom, hallway and other locations.

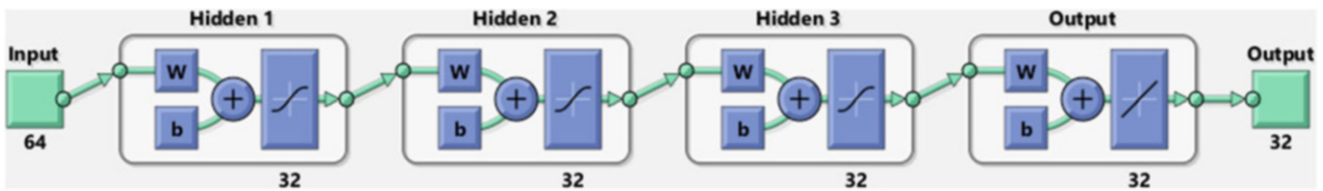


Fig. 42.7 Network architecture for unbounded predictions

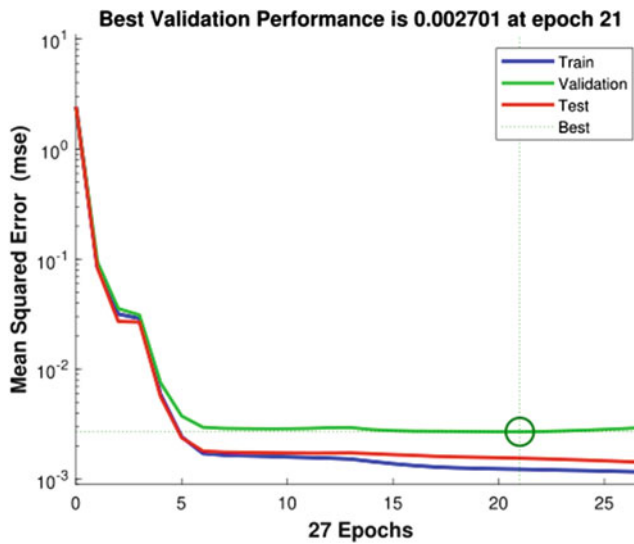
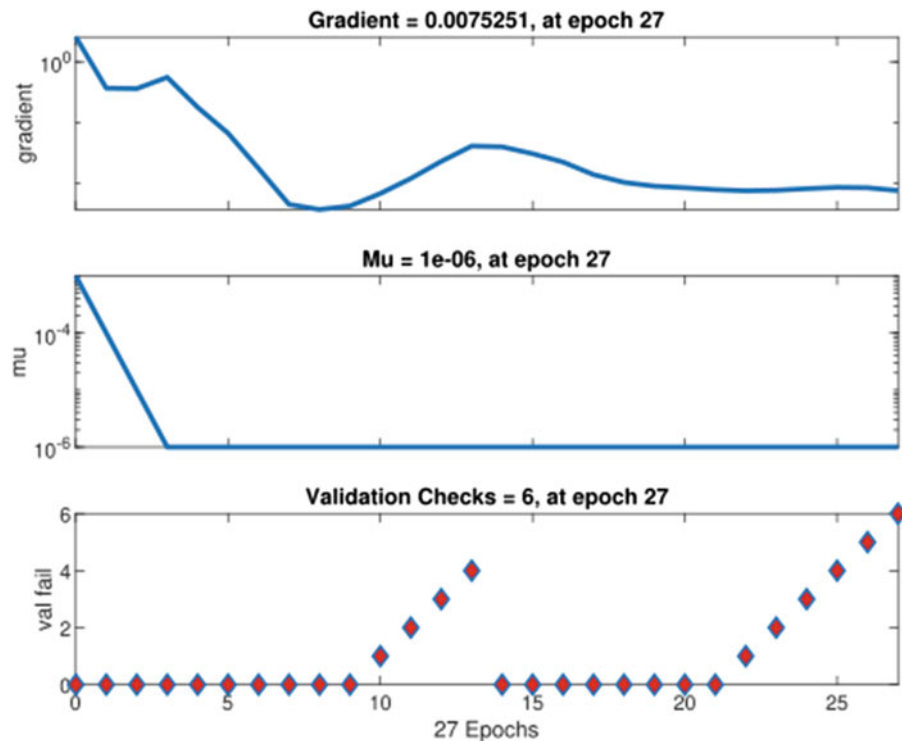


Fig. 42.8 Training phase performance

We ran through several pre-processing steps before training our network with the collected data. These are outlined below.

- First, we observed that there seems to be some signal attenuation at both end of the spectrum. This was likely caused by the hardware and thus we removed 4 subcarriers on either end of the measured CSI, keeping only the middle 48 subcarriers.
- Second, for the purposes of training our network, we discarded very weak signals (RSSI < 30 dB) and used signals that were relatively strong.
- The data set was trimmed down to use the CSI measurements roughly 1ms apart to ensure there was sufficient variance in the CSI data.
- Finally, each CSI vector was normalized by dividing each subcarrier with the maximum amplitude of the CSI measurement.

Fig. 42.9 Training phase convergence



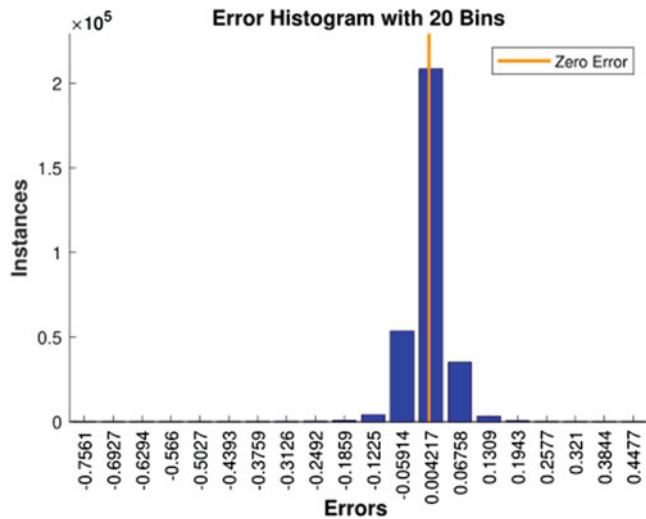
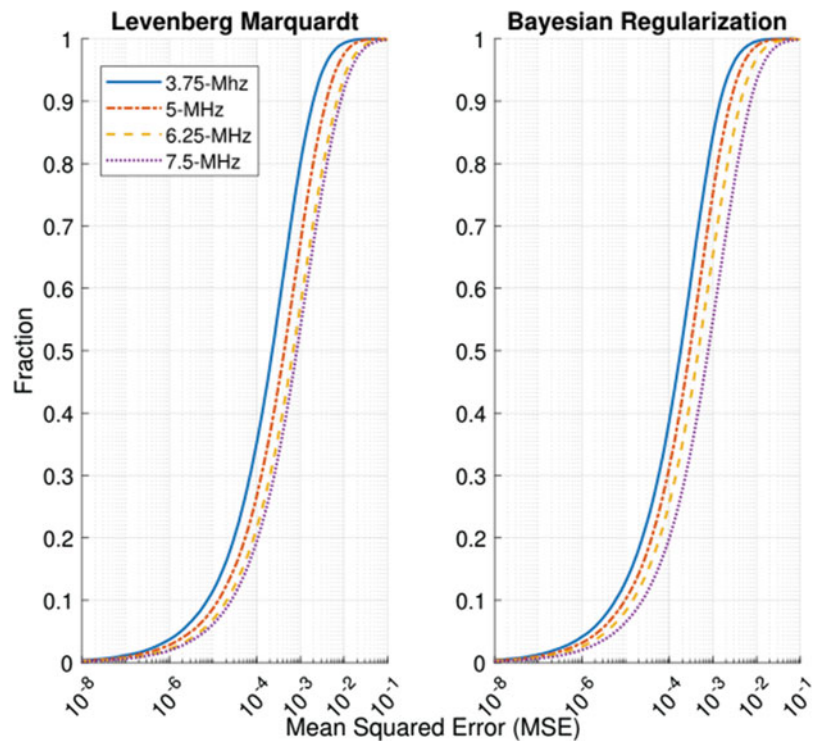


Fig. 42.10 Sample performance of a network used in bounded estimations

42.5.2 Evaluation Process

CSIFill was tested in a number of different scenarios. First, all evaluation scenarios were repeated using both the LM and BR backpropagation models. In addition, 4 variations of the network described above were trained to predict the CSI data at 12, 16, 20 and 24 subcarriers. Finally, CSIFill was evaluated on both bounded and unbounded estimation performance.

Fig. 42.11 Overall performance of CSIFill in bounded estimations



42.5.3 MSE Performance

This section describes the estimation accuracy of CSIFill.

42.5.3.1 Bounded Estimation Accuracy

Figure 42.11 shows in log scale the cumulative density plot of the MSE for bounded estimations. The plot on the left shows the overall performance of CSIFill using LM backpropagation. It can be seen that the average estimation error is below 0.002 in almost 90% of cases. Naturally, the accuracy of the network does degrade as the bandwidth of the target frequencies become larger, which is expected, however the error is still relatively low even at 7.5 MHz at around 0.01 in around 90% of cases. The accuracy of CSIFill can be slightly improved when using BR backpropagation as can be seen in the right side plot. However, this comes at the cost of higher computational complexity.

It should be noted that the mean squared error values are mostly due to the noisiness of the CSI data. These could have been reduced further by moving to a higher configuration of the network, which would have resulted in overfitting the data along with noise.

42.5.3.2 Unbounded Estimation Accuracy

While CSIFill was not intended to be used for unbounded estimations, it can be seen that in its current state it can already provide good predictions for 16 subcarriers or 5 MHz CSI data. An example of this can be seen in Fig. 42.12 where

Fig. 42.12 Typical example of unbounded estimation

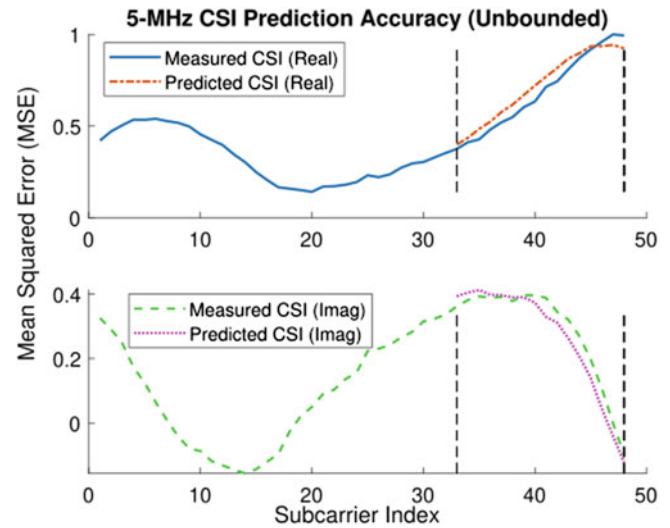
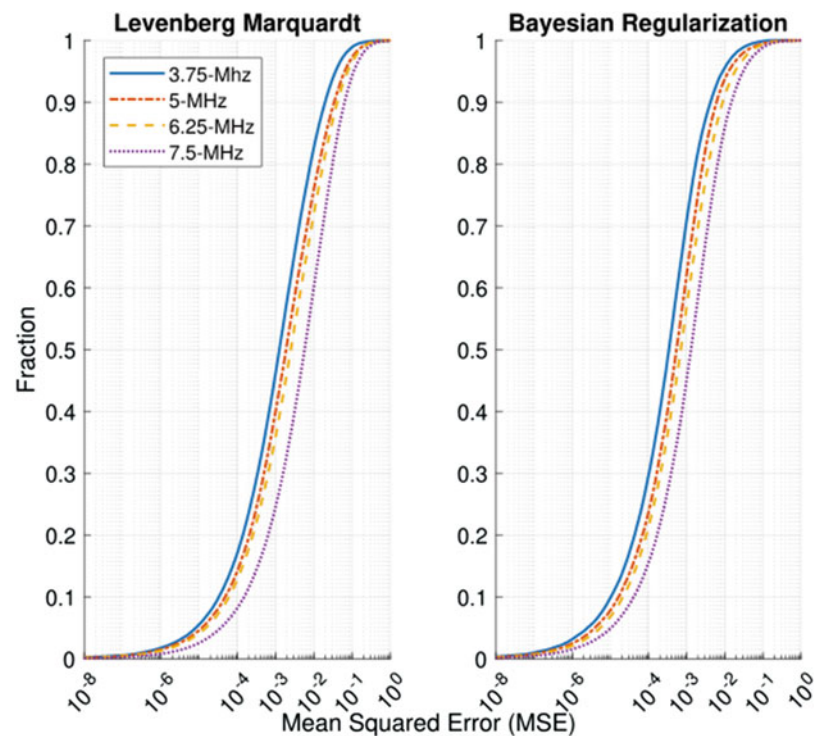


Fig. 42.13 Overall performance of CSIFill in unbounded estimations



CSIFill can accurately predict the shape of CSI data without a pivot subcarrier on the other end.

It can be seen from Fig. 42.13 that the performance degrades as we aim for larger predictions, however the overall fit accuracy is still acceptable without making further modifications to the network.

42.6 Conclusion and Future Work

We proposed CSIFill, a novel CSI spectrum gap filling solution, to predict the CSI in indoor wireless networks on subcarriers, by using measurements from neighboring subcarriers.

By using a shallow neural network, CSIFill enjoys accurate predictions of up to 7.5 MHz of channel bandwidth when the measurements on either end of the spectrum are known. While CSIFill is not optimized for unbounded predictions, our preliminary testing shows relatively good performance for up to 5 MHz of predicted bandwidth. We can infer that a more complex neural network could be trained to achieve even higher accuracy and should even be possible to infer the CSI of the entire spectrum using the CSI data from non-overlapping channels. This application of CSIFill can be very useful as it can potentially be used to measure the CSI on other wireless channels without any probing, but rather by simply extrapolating the wireless network channel from

the CSI measured on the current channel. This will allow faster optimal channel selection and improve the overall network performance. In addition, while our proposed system is focused on improving the performance of indoor wireless networks, it can potentially be extended to other wireless technologies like 5G and LTE.




References

1. 2021 connectivity and mobile trends survey. <https://www2.deloitte.com/us/en/insights/industry/telecommunications/connectivity-mobile-trends-survey.html>
2. I. C. S. L. M. S. Committee, Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 5: enhancement for higher throughput (2009)
3. I. P. T. G. AC, Status of project IEEE 802.11ac (2013). http://www.ieee802.org/11/Reports/tgac_update.htm
4. M. Morelli, U. Mengali, A comparison of pilot-aided channel estimation methods for OFDM systems. *IEEE Trans. Signal Process.* **49**(12), 3065–3073 (2001)
5. C. Luo, J. Ji, Q. Wang, X. Chen, P. Li, Channel state information prediction for 5g wireless communications: a deep learning approach. *IEEE Trans. Netw. Sci. Eng.* **7**(1), 227–236 (2020)
6. L. Liu, H. Feng, T. Yang, B. Hu, MIMO-OFDM wireless channel prediction by exploiting spatial-temporal correlation. *IEEE Trans. Wireless Commun.* **13**(1), 310–319 (2014)
7. R. Crepaldi, J. Lee, R. Etkin, S. Lee, R. Kravets, CSI-SF: estimating wireless channel state using CSI sampling & fusion, in *IEEE INFOCOM* (2012), pp. 154–162
8. H. Shirani-Mehr, D.N. Liu, G. Caire, Channel state prediction, feedback and scheduling for a multiuser MIMO-OFDM downlink. *CoRR*, vol. abs/0811.4630 (2008)
9. M. Biguesh, A. Gershman, Training-based MIMO channel estimation: a study of estimator tradeoffs and optimal training signals. *IEEE Trans. Signal Process.* **54**, 884–893 (2006)
10. W. Bajwa, J. Haupt, A. Sayeed, R. Nowak, Compressed channel sensing: a new approach to estimating sparse multipath channels. *Proc. IEEE* **98**, 1058–1076 (2010)
11. X. Rao, V. Lau, Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems. *IEEE Trans. Signal Process.* **62**, 3261–3271 (2014)
12. J.-C. Shen, J. Zhang, E. Alsusa, K.B. Letaief, Compressed CSI acquisition in FDD massive MIMO with partial support information, in *2015 IEEE International Conference on Communications (ICC)* (2015)
13. S. Sen, R. Choudhury, B. Radunovic, T. Minka, Precise indoor localization using PHY layer information, in *Proceedings of the 10th ACM Workshop on Hot Topics in Networks* (2011), p. 18
14. M. Kotaru, K. Joshi, D. Bharadia, S. Katti, SpotFi: decimeter level localization using WiFi, in *ACM Sigcomm* (2015)
15. Z. Jiang, J. Zhao, X. Li, J. Han, W. Xi, Rejecting the attack: source authentication for wi-fi management frames using csi information, in *IEEE INFOCOM* (2013), pp. 2544–2552
16. H. Liu, Y. Wang, J. Liu, J. Yang, Y. Chen, Practical user authentication leveraging channel state information, in *the 9th ACM Symposium on Information, Computer and Communications Security (ASIACCS)* (2014)
17. A. Mukherjee, Z. Zhang, Fast compression of OFDM channel state information with constant frequency sinusoidal approximation, in *IEEE GLOBECOM* (2017)
18. H. Zhu, Y. Zhuo, Q. Liu, S. Chang, Pi-splicer: perceiving accurate CSI phases with commodity wifi devices. *IEEE Trans. Mobile Comput.* **17**(9), 2155–2165 (2018)
19. Levenberg-marquardt backpropagation. <https://www.mathworks.com/help/deeplearning/ref/trainlm.html>
20. M. Kayri, Predictive abilities of bayesian regularization and levenberg–marquardt algorithms in artificial neural networks: a comparative empirical study on social data. *Math. Comput. Appl.* **21**, 20 (2016)
21. Y. Xie, Z. Li, M. Li, Precise power delay profiling with commodity WiFi, in *ACM Mobicom* (2015), pp. 53–64

Part IX

Potpourri

Unveiling a Novel Corporate Structure in World-Class Business, Merging Digital-Physical Environment in Hyper Famili Incorporation

Mohammad Khakzadeh , Fatemeh Saghafi , Seyed Milad Seyed Javadein , Mohammad Hossein Asmaie, and Masoud Matbou Saleh

Abstract

In this paper we aim to design a brand new corporate structure which merges Physical and Digital Technology in World-Class Business, whilst utilizing new technologies based on digitalization drivers. We propose a method in which we have introduced a state of the art Organizational structure that surveys our world-class Business Model case studies and all of the mandatory technologies which shape the fundamental framework of such innovative business ecosystems that end up as the most important input for the decision making process in management levels. In the upcoming research we will be comparing this model with other models in different aspects like comparability in utilizing brand new technologies to building up the dominant share of global market.

Keywords

Digital-physical environment · Organizational structure · Training · Human resource · Artificial intelligence · Machine learning

M. Khakzadeh
Operations Research Management, University of Tehran, Tehran, Iran
e-mail: m_khakzadeh@ut.ac.ir

F. Saghafi (✉)
Faculty of Management, University of Tehran, Tehran, Iran
e-mail: fsaghafi@ut.ac.ir

S. M. S. Javadein
Operations Research Management, University of Tehran, Tehran, Iran
e-mail: m.s.javadein65@alumni.ut.ac.ir

M. H. Asmaie
Family Modern Incorporation, Golrang Industrial Group, Tehran, Iran
e-mail: Asmaei.MohamadHoseyn@HyperFamili.com

M. M. Saleh
Family Modern Incorporation, Tehran, Iran
e-mail: Masood_Msaleh@yahoo.com

43.1 Introduction

A short review of all the successful business tell us the organizational structure and the business elements play a huge role in their prosperity. These organisms influence each other, and their environment constantly; they compete and collaborate, share and create resources, and coevolve; and they are unavoidably subjected to external disruptions, to which they inevitably adapt together [1]. Some observers, notably John Hagel, have suggested that such ecosystems will prove most enduring and influential, and provide the most sustained and important benefits to those businesses that create, lead, and participate in them [1]. Hyper Famili is a young Business incorporation which pursues the ambitious goal of creating its own business ecosystem because of the exceptional profit margin which has enabled the company to a successfully implement an economic umbrella effect throughout all sub modules. In order to help the company to expand its business life cycle, we have designed our structured investigation around the whole system including organizational structure, processes, business intelligence, decision support systems (DSS), in three level of Organizational, regional and international business environment. In order to expand to the International business scope, we have considered our rivals to be exceptionally proactive in terms of technology forecasting, futures studies and future analysis of business elements and business relationship and business trends. Our model is based on international trends, thus the definition of group borders, primary relationship between firms, sources of transformation and changes and applicability are to some extent processed [2]. Our surveys around the peer rivals in worldwide ranges show that not only accelerating the digitalization is enough to overcome this problem, but also can smoothen the way, helping the company to conquer the apex of retail in its region. In the aftermath of the Covid-19 pandemic, the company was faced

with numerous changes in the business and strategized to digitalize its structures to get ahead of the competition. The next step was responding to the change in their customers' habits, which had affected both e-commerce and in-store experiences known as Digital-Physical environment. Also focusing on agility meant, choosing an efficient team for digitalizing [3].

43.2 Literature

43.2.1 World-Class Business Ecosystem Model

A business ecosystem is the network of organizations including suppliers, distributors, customers, competitors, government agencies, and etc., who are involved in the delivery of a specific product or service through both competition and cooperation. A Finland based ecosystem also featured strongly in the RDI¹ roadmap and previous policy statement of Finland [4]. The idea is that, each entity in the ecosystem affects and is affected by the others, creating a constantly evolving relationship in which each entity must be flexible and adaptable in order to survive in a biological ecosystem. The theory of business ecosystems was developed by business strategist James Moore in 1993. This concept also can be depicted as four clear stages including Creation, Development, Maturity and Decline [5]. In this paper we consider world class business ecosystems corporations such as Walmart, Carrefour and Kroger of in order to reach the best practice of our case studies [6].

43.2.2 Intelligent Seamless Ecosystem

Digital technologies are changing the way companies do business. New technologies are connecting buyers and suppliers across more locations and activities, creating immense opportunities for some but putting others under constant pressure. Nevertheless, one thing is clear for all businesses: those who do not adjust will find it much harder to thrive in the digital age. Ecosystems should focus on using digital technologies for internationalization, they must be connected to ICT and transport infrastructure [7] which has been proven to be an enormous challenge, particularly in remote areas. We call this phenomenon an Intelligent Seamless Ecosystem.

43.2.3 Hyper Famili Incorporation

Hyper Famili is a new business incorporation focused on retail industry which up to now has followed the traditional

business framework. Since in the international market trends of retail fields are diverse, it is prominent for the company to change its path toward brand new technology based frameworks and to participate in a higher level of the ecosystem. In this research we follow the incumbent parts of ecosystem, encountering the real future risks and fluctuations against expanding international market share.

43.3 The Physical-Digital Boundary

Digitalization began influencing the physical Economy 50 years ago, with information technology automating many business processes. Advent of the Internet increased the pace, scope, and scale of that process, with some commentators initially distinguishing between an "old" physical and "new" digital economy: "E-commerce" was different from "commerce," "bricks and mortar" separate from "online". That boundary, however, quickly blurred, with terms such as "clicks and mortar" and "Omni-channel" emerging in retail, for example, to describe a much more blended and integrated reality [1]. However many Asian groceries insist on traditional way of doing business indifferent to technology trends [24] and that is what Hyper Famili is trying to change. The networked digital structure of today's computing environment demands that we develop an understanding of how networks and key firms within them support or inhibit digital innovation, how they enhance or damage business productivity, and how they provide a healthy environment for the creation of new businesses and products which would thrive whilst in this new environment [25].

43.3.1 Training; the Immersive Learning as a Mandatory Technology in New Ecosystems

With the Industry 4.0 revolution and the uptake of increased digitization, there is a shift in skill requirements for the future workforce. According to ABI Research, the enterprise Virtual Reality (VR) training market will generate US\$216 million in 2018 and grow to US\$6.3 billion in 2022. For example, XR education-focused company zSpacesaw a 128 percent CAGR² during 2014 to 2016. Experiential learning has long been argued as the most effective way to learn, and studies have shown that learning through experience increases the learning quality and improves retention by up to 75% [26]. Using immersive learning based on virtual reality would also provide 12% higher accuracy and 17% decrease in the

¹The National Research, Development and Innovation.

²Compound Annual Growth Rate: is a business and investing term for the geometric progression ratio that provides a constant rate of return over the time period.

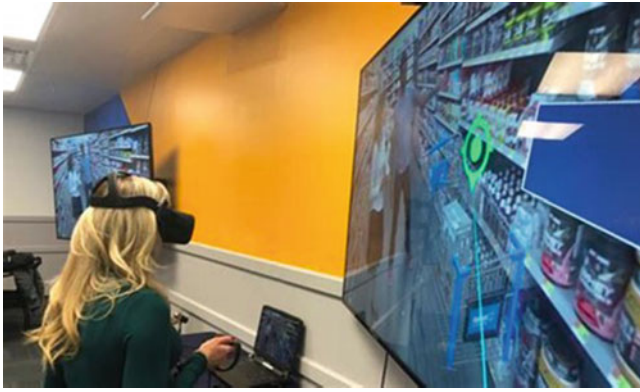


Fig. 43.1 Walmart virtual reality training system in retail field (Roche, 2018) [41]

amount of time needed to complete the instructional video amongst participants (Fig. 43.1).

43.3.2 Benefits of Immersive Learning

- **Mirror real-life situation:**

Immersive learning is effective because of the process of emphasizing things through visualization. By providing environments that mimic real-life situations more accurately.

- **End of distance:**

Results from the Accenture Technology Vision 2018 survey 9, indicates that 36% of executives believe that removing distance barriers between people and information is a key driver for their adoption of XR solutions. Through immersive experiences, businesses can tap onto the expertise of thousands of skilled workers from anywhere in the world. XR can also provide remote guided tours and remote collaboration. The greater the expertise the less time they need to get acquainted with XR.

- **Reduced operational costs:**

Organizations that adopt immersive learning can cut costs on employee travels and transporting equipment to training locations and even save space on real estate. Figures 43.2 and 43.3 demonstrate the benefits of VR³ systems in comparison to older LMS⁴ systems in terms of time and accuracy.

³Virtual Reality.

⁴Learning Management Systems.

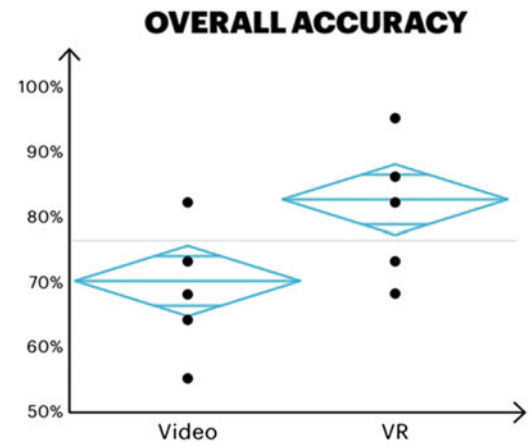


Fig. 43.2 Comparing Traditional Training V.S modern immersive learning over accuracy [26]

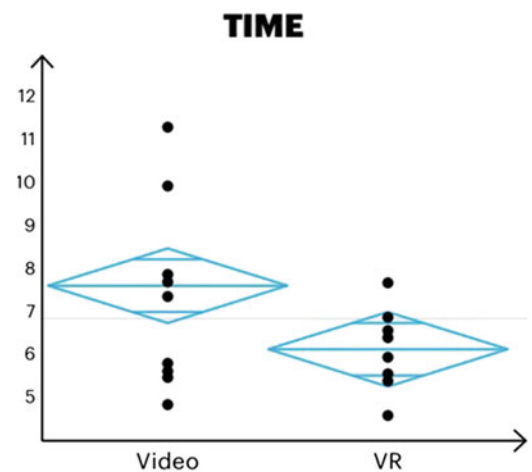


Fig. 43.3 Comparing Traditional Training V.S modern immersive learning over accuracy [26]

- **Learning through mistakes:**

One of the most compelling advantages of immersive learning is that people do not have to worry about making mistakes, which can be excessive in the real world both in terms of machinery and safety.

- **Increased engagement:**

With the ability to build-in gamification,⁵ immersive learning can be fun. When trainees are engaged and interested, it leads to better retention.

⁵Gamification is the strategic attempt to enhance systems, services, organizations, and activities in order to create similar experiences to those experienced when playing games in order to motivate and engage users.

Table 43.1 Comparison of top retail giant of business ecosystem

Company name	Structure kind	Country	Robotic warehouse	Virtual training	Virtual technology	Artificial intelligence	Web address
Walmart	HIERARCHICAL	USA	*	*	*	*	walmart.com
Carrefour	Semi-Matrix	France	*	*	*	*	carrefour.com
Kroger	Semi-Matrix	USA	*	*	–	–	Kroger.com
ALDI	HIERARCHICAL	USA	–	–	–	–	aldi.com [8]
COSTCO	Semi-Matrix	USA	–	–	–	–	costco.com [9]
Walgreens	HIERARCHICAL	USA	–	–	–	–	walgreens.com [10]
Home Depot	HIERARCHICAL	USA	–	–	–	–	homedepot.com [11]
Giant Eagle	HIERARCHICAL	USA	–	–	–	–	gianteagle.com [12]
SHERWIN WILLIAMS	HIERARCHICAL	USA	–	–	–	–	sherwin-wiliams.com [13]
BURLINGTON	HIERARCHICAL	USA	–	–	–	–	burlington.com [14]
SPAR	Semi-Matrix	Germany	–	–	–	–	spar-international.com [15]
IKEA	Semi-Matrix	USA	–	–	–	–	ikea.com [16]
Amazon	Semi-Matrix	USA	–	–	–	–	amazon.com [17]
Ahold delhaize	Semi-Matrix	Netherland	–	–	–	–	aholddelhaize.com [18]
Star Bazaar	Semi-Matrix	India	–	–	–	–	starbazaarindia.com [19]
Ulta	HIERARCHICAL	USA	–	–	–	–	ulta.com [20]
Rite Aid	HIERARCHICAL	USA	–	–	–	–	riteaid.com [21]
Meijer	HIERARCHICAL	USA	–	–	–	–	meijer.com [22]
Wumart	HIERARCHICAL	China	–	–	–	–	wumart.com [23]

- **Advanced analytics:**

XR captures enriched user data behavioral, eye tracking, heat maps and gesture tracking. Management can review immersive learning experiences and test results through automated reports that can be helpful to employees for future growth [26]. Organizational structure is described as the anatomy of an organization that critically influences how an organization functions [27] (Table 43.1).

43.4 Our Nearest Mega-Retail Rivals Strategic Ecosystem Structure Plan

- **Walmart Express store** [28]

They believe their multi-format portfolio will fuel the next generation of retail, enable the convergence of digital and physical store locations through e-commerce and unlock value, giving customers anytime, anywhere access to Walmart Express store. Its o targets are as follow:

- Streamlining the organizational structure to make it more agile, without losing sight of the end customers
- Achieving more gains in productivity and competitiveness
- Creating a leading Omni-channel ecosystem uniting stores and e-commerce
- Overhauling the merchandise offering with emphasis placed on greater quality [29]

Walmart brand new solutions:

- Autonomous delivery with Driver-Free cars. They operate retail formats which include grocery stores, discount warehouse clubs and a combination of general merchandise store [30].
- Walmart intends to install an AI -powered system in its stores that will prevent theft.
- Company has launched its own mobile payment system [31]
- Using VR technology for training human resources
- Walmart has concentrated its Training policies toward virtual shopping.

Immersive Learning program based on VR, run through hyper-real practice scenarios. An emerging intuition for supporting innovation of VR training [32]. In 2017, the VR headsets first rolled out in 30 Walmart academies where associates were trained to handle different situations from the everyday tasks, like managing the produce section, to the rare events, like Black Friday madness. Next, these situations can be virtually replicated and standardized for hundreds of employees, eliminating bias while placing employees in positions that best fit their skill sets [33]. And there is the great impact to remove the bias factor in human resources by new technologies based on artificial intelligence and FIS⁶ or

⁶Fuzzy Inference System.

ANFIS⁷ systems. Also it can be followed by forecasting the human resources factors by using ML systems.

- **Carrefour** [34]

Carrefour deploys a simplified and open organization transforming path which is simultaneously undertaking several projects: rationalizing processes, prioritizing them, developing their partnerships with specialists from the digital or retail sectors, and familiarizing their employees with the digital transformation using internal training program. To achieve this aim, they are preplanning e-commerce websites, logistics systems and a branded store network as part of an Omni channel distribution approach. Specifically, E-commerce websites per country, opening 200 additional drive pick-up points and in 2018, the expansion of home delivery to new towns as well as the opening of convenience stores (3000 by 2022) [35].

- **Kroger** [36]

According to their fact book they've aimed for a Seamless Ecosystem globally due to face the future survival risks, because it will have predicted the prominence of a digital strategy. Seamless has been a huge help in the current environment, enabling customers to shop in the way they prefer and feel most comfortable. They will continue to invest to make a world-class seamless experience available to their customers. They are well-positioned because several of their grocery competitors are not taking these steps today [37].

43.5 Brand New Warehouse and Inventory Management Model for Hyper Famili Ecosystem

In order to compete with their competitors in future, the newest technologies to supply elements such as: Time saving, Quality of inventory system, Acquiring precise warehouse system, Stock settlement affairs and Reorder point polices and order strategies should be studied. According to the newest retail business plan, the policy of the central warehouse is the most important item supporting digital shipping and virtual shopping. Building a superior end to end distribution network which utilizes robotic technologies and artificial intelligence in the warehouse will provide customers with a cost effective and convenient seamless shopping experience and adds value to the company. Through this partnership in the ecosystem and a diverse ecosystem of fulfillment channels, it will have the capability to support customer delivery, pickup fulfillments and store replenishments [37]. Which implies subsystems as below:

- Annual warehouse buffer supporting seamless shopping
- Packing material storage
- Robotic pallet handling area
- Global warehouse sizing and dimensioning
- Dock dimensioning [13]

43.6 Organizational Structure

The new ecosystem design introduces a new organization structure, supporting goals and strategies for thriving in the ecosystem, especially in emerging technologies such as artificial intelligence that the human resources jobs should be included in the new corporate structure (Figs. 43.4, 43.5, and 43.6).

The new organizational goals should support them in future. In Fig. 43.7 our new design for new pathways including artificial intelligence, coordination offices to investing more in new business subfields like virtual shopping, which support our green policies in order to decrease our greenhouses gases use and utilizing renewable or sustainable technologies such as solar panels or Biomass energies techniques for energy maintaining. The company needs to finalize and then actualize their decisions in regards to artificial intelligence, robotic and training (Fig. 43.7). The main advantage of this model is delegating more decisions and roles to local stores instead of establishing control committees to monitor the quality of the policies that have been put into application which can be supervised by coordinator vice president.



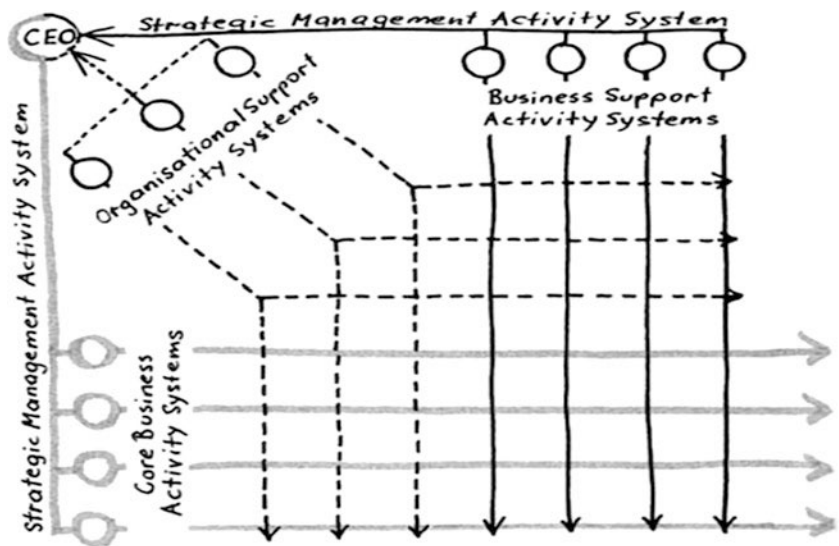
Fig. 43.4 Kroger's first customer fulfillment center – an automated warehouse facility with digital and robotic capabilities [42]

⁷Adaptive Neuro-Fuzzy Inference System.



Fig. 43.5 Business ecosystem requirements [1]

Fig. 43.6 Biomatrix model [40]



43.7 Comparative Study in Hyper Famili

As Walmart is the nearest company to Hyper Famili in terms of Structure, business model and the type of customers, the best ideas for digitalization and changing measures and

strategies for future growth come from repurposing its plan. There are several Artificial Intelligence Projects in Human resources including hiring and sale forecasting, concentrated in emerging AI.⁸ An independent department in hyper famili

⁸Artificial Intelligence.

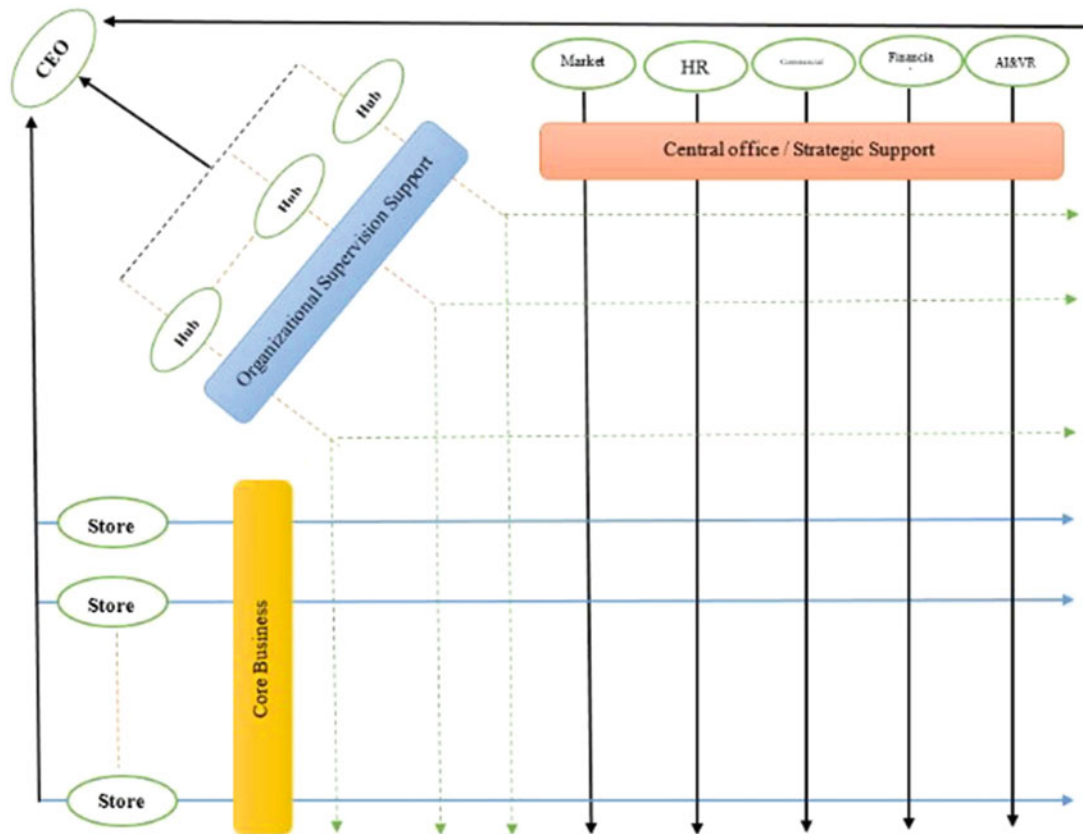


Fig. 43.7 New organizational structure, Hyper Famili Incorporation

which would work as a matrix structure. This department will consist of computer coders, system analyzers, digital data processors and they would be working with other experienced consultants who themselves would collaborate with many high educated human resources experts on AI and ML.⁹ Also Hyper Famili has started to implement strategies in Storage policies in centralized warehouses to smooth the path of digitalizing the stock of goods. Also it has started to collaborate with Kourosh Holdings (a sub holding of Golrang Industrial Group) in the field of e-commerce to help the task of supporting the digital infrastructure of e-shopping. Next, in training sector, Hyper Famili started a new Fuzzy ROI¹⁰ for efficiency studies to prove the importance of training roles and define the preliminary requirements of virtual training. Also there has been new studies published on gamification used as a brand new page in human resources development. In part of the structure, a new biomatrix level in sale planning and supervising in staff part to increase the level on organizational coordination has been formed and they hope these changes would contribute to measurable increase in efficiency and reallocating resources, funding

new opportunities and avoiding potential pitfalls, helping at critical decision points and vital relationships and connections, and any future needs that require further investigation, and present problems that require further investigation [38]. In terms of automation, it perceives environments instrumentally, as raw materials, waste sinks, and eco-services for objectified control. In contrast, widening access to low-cost, low-power sensors, micro-controllers, platform-based data-storage, analytical visualization, and Wi-Fi or mobile connectivity, enables more communities to sense, map and relate to a growing range of environmental phenomena for the Fourth Industrial Revolution [39].

43.8 Biomatrix Corporate Structure

In this section we introduce one of the most flexible extreme modern organizational chart which somehow would play a different role in order to solve the potential problems of enterprise. Naturally this point of view will decrease the central power which we have called semi-decentralized structure here. Another aspect of this organizational surgery is that it will help the CEO to delegate confidently and arrange the new generation of talented staff to implement the latest

⁹Machine Learning.

¹⁰Return On Investment.

strategies according to futures studies and business ecosystem. It's a permanent rule to monitor all peers and rivals to maintain technological advantage over them and to penetrate the innovation walls and barriers. These mechanisms should conclude to not to acquire the most market shares, but to take over all the niche markets. Otherwise this structure will improve the agility of implementing human resource strategies all around the company. Imposing fast hierarchical orders to react to market fluctuation, will empower the operational power and saving the time and energy and reserve for reaction scenarios at business ecosystem level. This is the marvelous core of this kind of structure that will differentiate it from other competitors, (see Fig. 43.6) to address their needs, especially in human resources and commercial affairs. Additionally, the new department of Artificial intelligence has emerged as a new sector in Hyper Famili organization. The part that they call Hub originally is generated from two convergent paths in order to maintain supervision on functional sectors and to make the highest coordination to central staff part.

This new arrangement will also bring functional and administration sectors under CEO control. In the next stage every stages will be independent and have their staff to consecrate on more important details of each hub stores. Hub managers can arrange all stores of their region to be kept under control and to monitor all operational affairs in detail. Central staff will facilitate him to be successful in fulfilling his role in leading every part of the sale.

43.9 Conclusion

In this paper we proposed a brand new organizational Structure based on business ecosystem and comparative study in order to first, guaranty the entirety and survival of Hyper Famili Retail Corporation and second, to stream line the mutation of the whole enterprise. In this way they would be inspired by their international counterparts such as Walmart and Carrefour in utilizing new tools such as futures studies, international retail business benchmark and technology trends due to select or create our prerequisites to be able to complete our business ecosystem. Consequently, it should be stated that our design for corporate structure based on ecosystem futures studies, trends and also the best practices of global giants of retail industry.

References

1. K.E. Deloitte, *Business Trends*. Deloitte Business Trends Series (2015), pp. 3–16
2. M. Mäntymäki, H. Salmela, In search for the core of the business ecosystem concept: A conceptual comparison of business ecosystem, industry, cluster, and inter organizational network (2017)
3. Retail in a post-COVID-19 world. Orange Business Services (2020)
4. K.A. Piirainen, V. Salminen, J. Kettinen, A. Reid, S. Zegel, *Impact Study: World-Class Ecosystems in the Finnish Economy, Part A – A New HoPE* (Helsinki, 2020)
5. G.I. Khotinskaya, *Business Ecosystem and Its Features in the Real Sector of the Economy 2020*, pp. 367–73. <https://doi.org/10.15405/epsbs.2020.03.53>
6. A. Hayes, M.J. Boyle, *Business Ecosystem* (Investopedia, 2021)
7. A. González, *SME Competitiveness Outlook, Executive summary, Business Ecosystems for the Digital Age* (International Trade Centre, 2018)
8. www.aldi.com (n.d.)
9. www.costco.com (n.d.)
10. www.walgreens.com (n.d.)
11. www.homedepot.com (n.d.)
12. www.gianteagle.com (n.d.)
13. www.sherwin-williams.com (n.d.)
14. www.burlington.com (n.d.)
15. www.spar-international.com (n.d.)
16. www.ikea.com (n.d.)
17. www.amazon.com (n.d.)
18. www.starbazaarindia.com (n.d.)
19. www.ultra.com (n.d.)
20. www.hy-vee.com (n.d.)
21. www.riteaid.com (n.d.)
22. www.meijer.com (n.d.)
23. www.wumart.com (n.d.)
24. F.S. Ahmad, A. Ihtiyar, R. Omar, A comparative study on service quality in the grocery retailing: Evidence from Malaysia and Turkey. *Procedia Soc. Behav. Sci.* **109** (2014)
25. M.R.L. Iansiti, *The New Operational Dynamics of Business Ecosystems: Implications for Policy, Operations and Technology Strategy* (Harvard Business School, 2002)
26. V. Roy, Immersive learning for the future workforce (2019), p. 2
27. M. Good, B. Soppe, M. Knockaert, A typology of technology transfer ecosystems: How structure affects interactions at the science–market divide. *J. Technol. Transf.* (2019)
28. www.walmart.com (n.d.)
29. Walmart, Walmart Inc 2020 Annual Report. WalmartCom (2021), pp. 1–88
30. L. Guruprasad, Dynamism in international business ecosystem a study on WALMART (2019)
31. Walmart: retail & payments ecosystem. PA7Space (2020)
32. B. McKeel, Senior Director of Digital Operations W. In the footsteps of trailblazers: How Walmart embraces Immersive Learning. STRIVER (2021)
33. R. Tuchscherer Walmart uses virtual reality to test new store managers. USA Today (2019), pp. 1–2
34. www.carrefour.com (n.d.)
35. Carrefour Annual Report 2020
36. www.kroger.com. No Title (n.d.)
37. Kroger Fact-Book 2019
38. M.J.B. Lauren Withycombe Keeler, The future of aging in smart environments: Four Scenarios of the United States in 2050. *Futures* (2021)
39. M.F. Adrian Smith, Post-automation. *Futures* (2021) 2021.102778.
40. A.R. Wigger, Managing organizational change: Application of the Biomatrix theory to the transformation of a non-profit organization (Leibniz, 2008)
41. <https://ca.movies.yahoo.com/tried-new-vr-training-1-million-walmart-associates-will-go-year-151310912.html>
42. <https://www.fastcompany.com/90279838/amazon-and-walmart-add-more-robots-but-insist-they-wont-terminate-jobs>

Harrison Ridley, Stuart Cunningham, and Richard Picking

Abstract

This research aims to understand how soundtrack elements can stimulate an emotional response in viewers whilst watching film and how sound professionals utilise sound for this purpose. The article reviews research on affective audio in film and informs ten semi-structured interviews with film sound professionals. These interviews cover a variety of topics, all with a focus on sound design. The interview transcripts were thematically analysed, and six themes were drawn from them and are discussed in detail and explored through case-study examples. A workflow for sound design is synthesised using each theme as a step in the post-production process, intended as a way of visualising how this research can aid the film-sound profession and inform future research in the field. It is concluded that sound is currently used to bring about affective responses in specific instances, when other emotional devices are not already in use. The use of sound as a narrative aid is found to be more prevalent, and indeed currently seen as more important within the industry than for specific affective use. Further research is suggested to enhance understanding of affective audio and to develop a framework for its widespread effective implementation in film.

Keywords

Affective audio · Emotion elicitation · Sound design · Sound practice · Affect · Audio post-production ·

H. Ridley (✉) · S. Cunningham
Centre for Advanced Computational Science (CfACS), Manchester Metropolitan University, Manchester, UK
e-mail: harrison.ridley@stu.mmu.ac.uk; s.cunningham@mmu.ac.uk

R. Picking
Faculty of Arts, Science and Engineering, Wrexham Glyndŵr University, Wrexham, UK
e-mail: r.picking@glyndwr.ac.uk

Interviews · Data collection · Toolbox · User requirements

44.1 Introduction

George Lucas has been quoted as saying that ‘sound is half the experience’ when watching a film [1]. While this is certainly not false, it begs the question as to how much sound (disregarding music) has an emotional impact on its audience when watching a film, as well as on their perception of many aspects of the film such as settings/locations, characters and key objects, the plot, and themes? While research into the subject of affective audio within film exists, much of it consists of research into the effect of music and score, or attempts to break soundtracks down into quantifiable sub-mixes for particular affect [2, 3]. The intention of the research reported here is to delve further into the sonic make-up of film soundtracks and uncover how the audience is affected whilst watching cinema.

This work is a component part of a wider piece of research aiming to develop a toolbox for film audio professionals to work toward creating more emotionally impactful sound design through the development and training of machine learning algorithms. The trained algorithm/s shall be used in a software toolkit for sound editors to identify the affective qualities of sounds in their libraries in order to utilise them to bring about more emotive responses in audiences. The research described here is an investigation into the techniques that sound professionals utilise to aid the viewers emotion and perception whilst watching film. It is the intention that this will be used to better understand the sound design and editing processes currently used, as well as film practitioners’ subjective options and whether or not they specifically consider affect in the production process. The questions that

this research attempts to answer are: Do sound professionals currently utilise sound(s) to bring about an affective response in the audience? How aware of sound's affective qualities is the film industry? Is there a need for more affective use of sound, or could there be more affective uses of sound in films?

The paper is organised into sections that describe: the background and reasoning for conducting this and future research; the methods in which the research was conducted, including the use of thematic analysis; simply and efficiently describes the results of the interview conducted; details analysis of the themes discovered through the results, and shows example responses; gives examples of themes identified in modern cinema; organisation of the themes into a workflow for affective audio-post production; describe how this research will aid and inform further research in developing an affective audio toolkit utilising machine learning techniques to assist sound editors in audio post-production.

44.2 Background

For the purposes of this study, the use of music in film is deliberately disregarded as the effects of music, particularly on emotion, are already well documented [4–7]. One example details the differences between emotion perception and induction in music, showing that perception is more common than induction [8]. In cases such as this, emotion perception indicated that an audience understands a performer to be portraying an emotion: emotion induction would be to bring about an emotion in an audience – this may not be the same as the perceived emotion.

Research within emotional audio for film has so far been limited. Ildirar et al. [9] have undertaken research into continuity and time perception within film, and audio's effect on it, and found that first-time viewers could successfully perceive continuity more frequently when diegetic sound was present than when it was not. Research has also investigated the effect of sound on audience response [10] and perception [11], but either only on a film-by-film basis (former), or to underline sound's importance in film (latter), as opposed to how it may be used to explicitly alter or inform one's perceptions and emotion.

There has been discussion about the role of sound effects in perception changing and emotion elicitation. Chion [12] speaks of the 'audio-visual contract', in which the viewer agrees to think of the sound and visual elements of a film as one entity, a suspension of disbelief, in which the viewer believes the sound to be real, or from within the world shown on screen. Blumstein et al. [13] conducted research in which the non-linear vocal attributes of a soundtrack were

studied. The study found that there were some instances of abrupt changes in 'noise' and frequency being used with the intention of bringing about fearful and 'dramatic emotional' responses respectively.

Donaldson's article 'Feeling and Filmmaking' [14], a study of the consideration of sound and affect, combines theory and practice in the filmmaking process. Donaldson draws on interviews with sound practitioners, some of which were conducted herself for the article. The article exposes how sound practitioners describe their work aesthetically as well as how the filmmaking process involves '*physical activity . . . play and experimentation*' and a '*kind of tactile analogy*'. Donaldson describes the affective quality of sound as being in the '*labour of making the film, as well as our watching of it*'.

Research on implementing affective sound design has been conducted [2, 3]. A model is described for sound practitioners to follow when mixing audio for affect. The 'Four Sound Areas' framework consists of categories of sound: Logical; Abstract; Temporal; and Spatial. The model created was used to measure how much each category was utilised in sections of film intending to evoke a particular emotion. It was found that the majority of negative emotions tended to have similar make-up of these four categories.

Whilst the published research described makes good progress toward understanding how to implement affective audio into a film's soundtrack, it does little to suggest what *actually* evokes a response in an audience. The 'Four Sound Areas' framework is aimed more toward attempting to evoke a general feeling or emotion over the course of several scenes. But how do film practitioners currently utilise audio's affective capabilities? This study aims to uncover how practitioners consider sound in their films and use this to inform further research into the topic.

44.3 Methodology

The research described herein is a qualitative study of film professionals' views and experiences of working with sound in various ways to aid an audience's perceptions and emotions. Ten semi-structured interviews were conducted with audio professionals about their use or awareness of the audience's perception, emotion, and how they take advantage of it. The use of interviews in works discussing the use of sound in film [15] and its affective qualities [14] is typical and use of semi-structured interviews allows for divergence and more in-depth exploration of topics that interviewees may bring up [16]. Interviews are also commonly used to elicit user requirements for interactive requirements as part of a user-centred design approach [17].

Interviewees were selected based on their role/s as either professional film practitioners (such as a director, re-recording mixer, or sound editor) or teachers of filmmaking and/or film studies and were based in a range of countries including the USA, UK, Serbia and Australia. By selecting sound professionals and practitioners to be interviewed and not cinema goers, data collected will be more relevant [16, 18] to professional aspects of filmmaking.

The interviewees varied throughout a range of career levels and backgrounds and included: an Academy Award winning sound designer; award-winning sound editors and mixers; independent filmmakers; and film studies teachers. The range of skill levels and industry experience provided by this sample of professionals allows for a wide-ranging overview of the industry as a whole. Interviews were audio recorded and transcribed for thematic analysis using NVivo 12 [19]. Questions were asked about multiple aspects of sound in filmmaking. Topics included, but were not limited to: thematic elements and/or subjects that directors wanted to convey; intended emotional responses during scenes and how sound was considered in this; how the sound design team implemented audio for emotive effect. Some questions were altered to suit the individual's background and position in the industry.

Thematic analysis was conducted using the interview transcripts. It aims to identify patterns in qualitative data [20]. In this, case word clouds were generated to give a visual guide to the most frequently used words and themes. After the word clouds were generated, in-depth analysis was conducted using the four-step approach described by Kumar [16].

From analysis of the themes and interviewee responses, the authors discussed examples of the themes in modern cinema and synthesise a workflow for affective audio implementation in film post-production.

44.4 Results

A summary of results is presented in this section and they shall be discussed fully in following sections.

The themes identified and selected for coding were: Sound as an Affective Device; Sound as a Narrative Device; Sound Design Technique; Mixing Technique; Role of a Sound Editor/Mixer; and Director's input to Sound. These were purposefully selected so that the research could differentiate film professionals' use of sound as narrative and affective devices, as well as to gain an insight into the role of sound teams, their communications with directors and how they go about creating sound for affect (Table 44.1).

44.5 Analysis and Discussion

44.5.1 Sound as an Affective Device

When talking about sound's use as an affective device there were two clear schools of thought in the interviewees. Those who see sound as a necessity to aid the visual and nothing more, and those who see sound as a means to enhance the audience's experience through its creative implementation – with many of these uses either purposefully or coincidentally having affective results. For example, the producer/director said that they want to 'get it [sound] right and make it very standard and let the dialogue be the main thing that people think about' (p 06) and that the use of sound to fill out a scene depends on 'how much do you want the people to feel like they're actually there' (p 02). Whilst an award-winning sound editor/supervisor said that: '*sounds that are not music, have the ability to impart some sense of emotion*' and that (when talking about the use of ambiences in film) '*the sound of birds twittering is generally a peaceful sound . . . it's not a sound that causes agitation or anxiety*' (p 10) and that such sounds may be used to imply the feeling or mood for the scene ahead. This was validated by other sound editors/mixers who, when describing the how they may establish mood and emotion in a scene, said:

in some kind of very dodgy neighbourhoods, you could have certain sounds . . . the pipes and the heating or, or the lights, the electricity being more noticeable, things that are kind of not very nice sounds highlighting them inside the space.
- Participant 09

It is worth noting that there were instances where interviewees fell in the middle of these two groups. For example, an independent sound editor said that whilst sound design can be used for an '*emotional boost or manipulating emotions. I just don't think it's used very often, Not least as often as a film score*' (p. 04).

Typical responses when asked about how sound can be used for affective response included describing: the use of point-of-view sound design to help the audience understand what a particular character is feeling; the use of atmospheric sounds to convey the emotion of a character (such as exhaustion or tiredness), as well as a way to create feeling about that space, such as oppression, anxiety or relaxation; using tones to create a sense of danger, or to unsettle an audience.

Responses such as this show that film professionals are actively using sound effects to convey and/or elicit emotion. It is worth noting that many of the affective devices described by the interviewees could also be used as narrative ones, with an affective side-effect. For example, the birds twittering

Table 44.1 Numbers of theme instances

Theme	No of instances	Example response
Sound as an Affective Device	43	'I think we can unequivocally state that sounds that are not music, have the ability to impart some sense of emotion' (p 10)
Sound as a Narrative Device	25	'we have a squeak of a bird if it's a romantic thing . . . ' (p 10)
Sound Design Technique	23	'you have this kind of increasing crescendo that leads to the impacts . . . so the spectators can feel emotionally that an explosion is arriving' (p 03)
Mixing Technique	11	'So those recordings that were underwater were then . . . mixed to add aggression to highlight their anger or struggle' (p 09)
Role of a Sound Editor/Mixer	14	'You're like frauds. You're, misleading your audience and lying to them . . . yeah, that's our job!' (p 07)
Director's Input to Sound	24	'The director[s] say that this is what they [want the] audience to feel or they'll say, this is what the character is feeling, which are two totally different things. And then, and then you choose [which] you're going to be representing' (p 04)
Total	140	

described as an affective device by one sound editor were similarly described as a narrative tool by another – the primary use of sounds such as these is determined by the need for each scene, and the sound team's interpretation of what is necessary. A great example of this was brought up when talking about the sound of *Mad Max: Fury Road* [21]

... in a very operatic sense, the death of the War Rig to me was the death of a creature. It wasn't a truck crashing; it was the death of one of the protagonists in the film. And so [omitted], created sounds of dying animals of whales and bears and something else for that scene... All those [sounds] of it flipping and turning her – all dying animal sounds because it was its, as we say in Hollywood, its swan song, its death rattle.
- Participant 10

44.5.2 Sound as a Narrative Device

Interviewees described the use of sounds for narrative aid as useful and common. One interviewee stated that sound and image 'work simultaneously, and they're only effective because they're running simultaneously' (p 01). Others (p 03, 04, 05, 07) told of using 'sweetened' sounds to reinforce a character or object's importance or power. For example, a technique was highlighted in a scene from *Raiders of the Lost Ark* [22] in which the interviewee stated that the sound team inserted a tiger roaring over the swerves and revs when Indiana Jones takes control of the Nazi truck, to give a sense of ferocity when he is in control. Uses such as this seem obvious and explicit, but when watching without knowing this trick was used, an audience may not notice the absurdity of the sound underneath.

Whilst uses such as this give exaggerated perspectives of important plot points, more subtle uses were also discussed. An interviewee described using sound to convey narrative efficiently, saying:

Think of how long it takes a writer to write the words for an actor to speak to tell the audience, 'I feel very calm right now.'

'I feel good.' – The minute you hear that bird chirping in the background of the scene, you've already assumed it. You've said it so economically with a sound, without having to use exposition.
- Participant 10

This technique was described as an efficient method to convey a character's emotion or mood. Other participants tended to agree with this statement and described similar uses such as echoic sounds to provide 'a sense of loneliness' (p 05) that was necessary for the narrative in their scenario.

44.5.3 Sound Design Technique

When talking about their approaches to, and understanding of sound design, the interviewees gave a wide array of responses ranging from simple use of silence before impact sounds to create the perception of greater loudness to creating unique sound effects for particular scenes or objects.

Notable techniques brought up by interviewees when talking about this subject included:

- Finding a link between the colour palette of the film and the timbre of the sound design: when colours on film are cooler, use sounds that are cooler in timbre, and vice versa. This is used to complement the visual element of the film and reinforce an aesthetic.
- Creating bespoke sounds for fantastical objects: marrying organic and synthesised sounds together to create new, believable and relatable sounds. This helps with the narrative in some cases, but mostly aids audience immersion.
- Using sound in an expressionistic sense, for example: detaching the sound from the realism of the image allows the sound editor to amplify or tone down elements of the soundtrack and guide the viewer through a scene. This could be used to make mundane sounds/scenes more interesting, comical, or impactful, or to highlight elements of a scene that would otherwise go unnoticed.

- ‘Painting’ with sound: one interviewee spoke of generating lots of ambiences for a sci-fi feature without any context. These sounds were then assigned ‘colours’ based on properties such as overall timbre, mood and sonic make-up, and then ‘splashed’ onto scenes. This technique was described as a way of trying to both categorise ambiences and find out which sounds could work with each other harmoniously. The sound design team found they could eventually call up a red, green, blue, or other ‘colour’ sound as required, to create a unique soundscape that fitted the mood.

Through conversations with these film professionals, it is clear there is not one single ‘right’ way to create sound for use in film, nor is there a universal approach. A variety of methods should be used to create the most cohesive, believable, and enjoyable soundtrack that complements the film, its themes, and emotions that the audience should feel throughout their viewing.

44.5.4 Mixing Technique

On the topic of mixing sounds, only a few interviewees spoke of its use for affect. Of particular note was a technique in which a mixer would reduce or remove sounds that were played over a long period of time. The idea being that people begin to ‘tune out’ (p 09) sounds such as traffic, fans and crowds, but when these sounds disappear the audience notices their absence. This was described as a way of generating impact in a scene where something shocking was told or occurred.

Other mixing techniques involved blending sounds together to apply a certain quality to a character or scene, for example, blending dangerous animal sounds together to use when a killer is about to strike, to indicate to the audience that although they are not seen, they are present.

44.5.5 Role of a Sound Editor/Mixer

When speaking about roles as sound editors/mixers the respondents gave varying answers. Some described the job as ‘telling another part of the story . . . to move a plot forward . . . when words don’t suffice’ (p 10) and also as the ‘emotional link between the director and the audience’ (p 08), whilst others described the job quite colourfully as being ‘frauds’, explaining that their job is to lie to the audience, in order to make the film more believable and therefore better – ‘the bigger the fraud, the bigger the lie; the better the movie, the better the soundtrack’ (p 07). It is interesting to see how sound editors are beginning to view their role as being similar to a composer:

I’m a composer. Musicians are discovering that they don’t need to be bound by melody. Music does not necessarily imply a harmonic structure. And of course the polar opposite of melodic music is sound design, music, concrete or sound. What I do is no different than what a composer does. Every sound that you hear in a motion picture is chosen, designed, timed, considered, everything that a composer might take into consideration about voicing metre and tempo, I do, but I just don’t follow a grid and I don’t constrain myself to 12 notes.

- Participant 10

It seems that the sound editors/mixers generally see their function in a film’s production as more than that of a traditionally technical, rigid role where they simply fill out the film with appropriate sounds, but instead as a creative, nuanced one. These professionals seem to be understanding how their input to a film’s soundtrack can help guide an audience through a film, whether narratively or emotionally.

44.5.6 Director’s Input to Sound

Whilst most of the interviewees described their relationships with directors as a collaborative one, there were varying accounts of direction in relation to sound.

One interviewee described three types of director: a director who knows what they want, but allows some freedom in sound design; a director who doesn’t think about sound design and gives a completely free reign over its development; and a director who wants something very specific, and watches over everything. It was notable that the director who gives total freedom was considered the hardest to work with as they ‘often don’t understand what we [Sound Editors] do . . . and they want something else [after the work is done]’ (p 07).

Other interviewees spoke of directors giving an indication of what the audience should feel at certain points in the film and ask the sound team to help generate or accentuate this. One said that as a sound editor ‘you choose whether you’re going to be representing the character and what they’re feeling or what they’re seeing or representing . . . , even though they [the audience] might be seeing and feeling something the character isn’t or doesn’t.’ (p 09).

In general, all agreed that directors give enough creative freedom, whilst reining in that freedom only in specific scenes (often key narrative points).

44.6 Themes in Action

Whilst discussing the themes within this research is an effective and useful means to describe the findings, their implementation in real-world films act as a means of validation of their practicality. This section highlights the themes’ uses

in some examples of modern cinema, wherever a theme is identified or referenced, it is highlighted in bold.

44.6.1 *Blade Runner* (1982) & *Blade Runner 2049* (2017)

Ridley Scott [23] and Denis Villeneuve's [24] sci-fi films make considerable use of both musical score in the form of synthesised sounds and drones, as well as elaborate and well-considered soundscapes to create and reinforce mood and tone throughout scenes. Soundscapes and sound design were used to aid the narrative: the opening scene with the Voight-Kampf machine uses beeps, ticks and other familiar electromechanical sounds to set the scene as a realistic futuristic reality. These sounds help build the feeling of tension and foreboding in the audience, altering their affective state in a scene that culminates in the death of a character.

The role of the sound designer in the production of *Blade Runner 2049* was discussed with participants in the interviews that made up the basis for this research. An interview participant described their role as similar to a composer that works to build on the visuals of the film world, utilising sound design techniques such as keeping timbres matched to the colours of the visual elements. This technique affords the viewer an immediate and clear distinction of feeling and mood for a scene, without the necessity for dialogue or facial expression.

44.6.2 *Once Upon a Time . . . in Hollywood* (2019)

Tarantino's [25] film utilised sound as part of the characterisations throughout. The sound design team describe the recreation of vehicles and ambiances to create not only a cohesive world that aids the narrative, but also bring about affect in the audience: nostalgia at times, joy throughout, and momentary unease and fear [26]. The Manson's Ford Galaxie was mentioned by the sound team as a 'junker' that pops, bangs, and stumbles its way down the road. Its sound was purposefully created to give a sense of unease, as its arrival signifies the sinister nature of the scene ahead. They made sure that each car was its own character, at the behest of the director, all reflections of their owners' characters, brought to life through affective sound design that evokes emotion, and aids the narrative with subtle nuance.

44.7 Working with Themes

The themes identified in this research are all elements of the post-production process of filmmaking. We now organise the themes into a hierarchical workflow (Fig. 44.1). The flowchart was constructed using the information from interviewees about the typical hierarchy of post-production, alongside their responses relating to the importance of narrative and affective sound, as well as our own knowledge of the industry and process, having experience in audio post-production.

The themes are organised into three sections, shown in the circles and colours, which analogise the sound design process to that of painting or drawing.

The first two themes, 'Director's Input to Sound' and 'Role of a Sound Editor/Mixer' are used as high-level outline of sound direction and the themes that are to be presented in a given film or scene. The Director usually gives a brief to the sound design team, and as such is placed higher than the 'Role of a Sound Editor/Mixer'. The arrows pointing in both directions here depict the conversation between director and editor/mixer, and the constant exchange of ideas and implementation throughout the post-production process.

The next two, 'Sound Design Technique' and 'Mixing Technique' are of equal importance, although rarely occur at the same time. In a typical post-production process, the

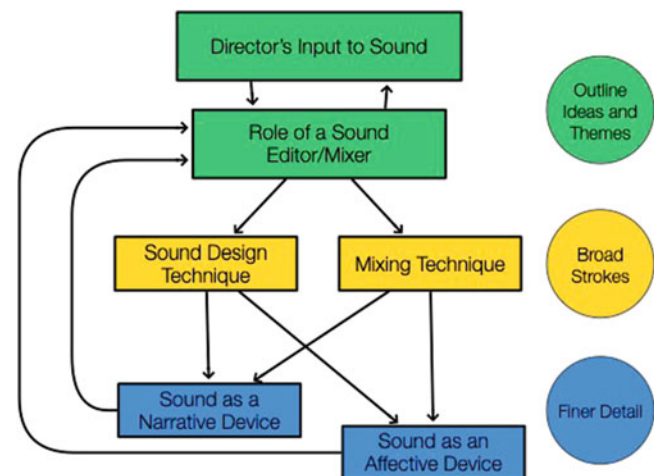


Fig. 44.1 Flowchart depicting the research themes in a hierarchical workflow

sound design (or editing) is completed and handed to the mixing (or re-recording) team. From here it is best to look at the flowchart as a parallel timeline. This section is where the broad strokes of the ‘painting’ are filled in. Sounds are filled into scenes on necessity, in addition to some more stylistic sound choices.

The final section is where the ‘painting’ comes together, with sound now finalising each scene with finer details. As interviewees indicated that the narrative is usually more important than bringing out emotion in an audience, the ‘Sound as a Narrative Device’ theme is placed higher in the hierarchy than ‘Sound as an Affective Device’. Note that both Sound Design Technique and Mixing Technique have influence on the outcome of both of these themes.

The third section feeds back into the first at the ‘Role of a Sound Editor/Mixer’ theme for two reasons. Firstly, because of the parallel timeline of post-production workflow depicted in the flowchart, the sound editing is typically done before the mixing, but would often go through the same supervising team. Secondly, this is done because the ‘finer details’ would be refined through feedback from the supervisory and production teams. This is shown with the two-way communicative arrows between the two top level themes.

44.8 Conclusion and Future Work

The results of this research show that there is great consideration of audio’s affective use in filmmaking, particularly in post-production. In general, Sound Professionals consider the effect that sound can have on an audience, or how the sound can bring further depth to aspects of a film in their everyday work.

During one interview, the emotional quality of breathing sounds in horror films were discussed. The conclusion of this conversation was that, without these sounds, scenes where a killer is stalking their victims would not have as much as a physical effect on the audience than it would without:

That whenever we have his breathing, it’s the overriding force of that scene. It’s the main affective thing in that scene... So we get a feeling of that claustrophobia of what that breathing is like when you’re trying to breathe but you’ve got something hovering. So it’s not just the breathing isn’t just saying: ‘Look something bad is about to happen’ ‘Oh there’s a baddie – watch out!’ It’s effectively making us feel claustrophobic or asking us to feel claustrophobic. It’s designed ‘affectively’ to make us feel quite claustrophobic, make us maybe feel a little bit sweaty... so a sound isn’t just – It’s not just a signifier... it works on us physically.

- Participant 01

This confirmed that audio, and not music, can have a strong affective use, when utilised with care and consideration.

The interviewees did not suggest that they found it difficult to find sounds that explicitly caused affect; however,

many did mention that they purposefully use sounds to do this. Interviewees generally agreed that the sound editing process allowed for more creative uses of sound and utilisation of its affective qualities than the mixing process, where balancing the music, dialogue and effects becomes more of a narrative choice.

An interesting and useful note from analysis of responses was the use of colour in sound editing. Many participants spoke of matching timbres with on-screen colours to match the mood of the general film, a scene, character, or other element. Participant 10 also described a categorisation method for sound editing wherein they developed a ‘colour palette’ of sounds, particularly ambiances, that could be ‘painted’ onto the editing session to create specific ambiances based around what kind of mood was desirable for each scene.

Finally, to answer the research questions outlined in the introduction: Yes, sound designers do utilise sound to bring about an affective response in the audience, although this is *usually* done in a subtle manner, and this is most used to bring about feelings of fear or unease. Film sound professionals are *mostly* aware of sound’s affective qualities, some more than others. Certain situations demand a use for affective audio, but many do not. There could be more uses of affective audio in film, and more research into its implementation is suggested.

This research helps understand the post-production workflow with an emphasis on emotion. It informs further research is now being undertaken as part of the first author’s PhD studies investigating what characteristics of sound make them evoke emotional responses in listeners. A database of affective sounds that could be used in film sound design and used to aid research in this emerging field is being created. It is planned to then lead on to training machine learning algorithms in order to develop a toolkit for sound professionals to aid their workflow by suggesting useable sounds that are known to have desirable affective qualities.

References

1. L. Blake, George Lucas: Technology and the art of filmmaking [Magazine] (2004, November 1). Retrieved 16 January 2019, from Mixonline website: <https://www.mixonline.com/recording/george-lucas-365460>
2. N. Hillman, S. Pauletto, The Craftsman: The use of sound design to elicit emotions. *Soundtrack* 7(1), 5–23 (2014). https://doi.org/10.1286/st.7.1.5_1
3. N. Hillman, S. Pauletto, Audio imagineering: Utilising the four sound areas framework for emotive sound design within contemporary audio post-production. *New Soundtrack* 6(1), 77–107 (2016). <https://doi.org/10.3366/sound.2016.0084>
4. K. Kalinak (ed.), *Sound: Dialogue, Music and Effects*, 1st edn. (I.B. Tauris & Co. Ltd, New Brunswick, 2015)
5. A.J. Cohen, Music as a source of emotion in film, in *Series in Affective Science. Music and emotion: Theory and research*, (Oxford University Press, New York, 2001), pp. 249–272

6. P. Ivanka, M. Slobodan, The effect of music background on the emotional appraisal of film sequences | Directory of Open Access Journals. *Psihologija* **44**(1), 71–91 (2011). <https://doi.org/10.2298/PSI1101071P>
7. B. Langkjær, Making fictions sound real – On film sound, perceptual realism and genre. *MedieKultur J. Media Commun. Res.* **26**(48), 5–17 (2010). <https://doi.org/10.7146/mediekultur.v26i48.2115>
8. P.N. Juslin, P. Laukka, Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *J. New Music Res.* **33**(3), 217–238 (2004). <https://doi.org/10.1080/0929821042000317813>
9. S. Ildirar, D.T. Levin, S. Schwan, T.J. Smith, Audio facilitates the perception of cinematic continuity by first-time viewers. *Perception* **47**(3), 276–295 (2018). <https://doi.org/10.1177/0301006617745782>
10. D. Candusso, J. Thompson, How sound design shapes the audience’s response in Baz Luhrmann’s ‘Australia’. *Int. J. Image* **5**(1), 25–31 (2014). <https://doi.org/10.18848/2154-8560/CGP/v05i01/44117>
11. J.R. Drake, The importance of sound design and its affect on perception. MA Thesis, State University of New York (2013). Retrieved from <https://dspace.sunyconnect.suny.edu/handle/1951/58385>
12. M. Chion, *Audio-Vision: Sound on Screen*, ed. by C. Gorbman, 2nd edn. (Columbia University Press, 1994).
13. D.T. Blumstein, R. Davitian, P.D. Kaye, Do film soundtracks contain nonlinear analogues to influence emotion? *Biol. Lett.* **6**(6), 751–754 (2010). <https://doi.org/10.1098/rsbl.2010.0333>
14. L.F. Donaldson, Feeling and filmmaking: The design and affect of film sound. *New Soundtrack* **7**(1), 31–46 (2017). <https://doi.org/10.3366/sound.2017.0095>
15. V. LoBrutto, *Sound-on-film: Interviews with Creators of Film Sound* (Greenwood Publishing Group, Westport, 1994)
16. R. Kumar, *Research Methodology: A Step-by-Step Guide for Beginners*, 4th edn. (SAGE, Thousand Oaks, 2014)
17. A. Dix, R. Beale, G.D. Abowd, *Human-Computer Interaction* (Pearson Education UK, London, 2003)
18. J. Francombe, J.S. Woodcock, R. Hughes, R. Mason, A. Franck, C. Pike, et al., Qualitative evaluation of media device orchestration for immersive spatial audio reproduction. *J. Audio Eng. Soc.* **66**, 414–429 (2018). <https://doi.org/10.17743/jaes.2018.0027>
19. QSR International, Qualitative Data Analysis Software | NVivo [Information/Product] (2020, December 8). <https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home>
20. V. Braun, V. Clarke, Using thematic analysis in psychology. *Qual. Res. Psychol.* **3**(2), 77–101 (2006). <https://doi.org/10.1191/1478088706qp063oa>
21. Miller, G. (2015). Mad Max: Fury Road [Action, Adventure, Sci-Fi]. Retrieved from <http://www.imdb.com/title/tt1392190/>
22. S. Spielberg, Raiders of the Lost Ark [Action, Adventure]. Paramount Pictures, Lucasfilm (1981, June 12)
23. R. Scott, Blade Runner [Action, Sci-Fi, Thriller]. The Ladd Company, Shaw Brothers, Warner Bros (1982, June 25)
24. D. Villeneuve, Blade Runner 2049 [Action, Drama, Mystery, Sci-Fi, Thriller]. Alcon Entertainment, Columbia Pictures, Sony (2017, October 4)
25. Q. Tarantino, Once Upon a Time... In Hollywood [Comedy, Drama]. Columbia Pictures, Bona Film Group, Heyday Films (2019, July 24)
26. T. Muirhead, 111 – Once upon a time in Hollywood, vol. 111 (2019, September 10). Retrieved 26 October 2020, from <https://soundcloud.com/tonebenders-podcast/111-once-upon-a-time-in-hollywood>

Laxmi Gewali and Samridhi Jha

Abstract

Algorithms for covering and simplifying a 1.5D terrain have been extensively investigated. Minimally covering a 1.5D terrain is an intractable problem. We present a critical review of existing approximation algorithms for simplifying a 1.5 D terrain by a fewer number of vertices. We introduce the problem of simplifying a 1.5D chain with a fewer number of vertices subject to visibility requirement. We then present the development of an efficient algorithm for chain approximation with minimal effect on external visibility.

Keywords

Chain approximation · Visibility · Terrain illumination · Monotone chain · Covert penetration · Visibility algorithm

Euclidean metric, Hausdorf metric and Frechet metric. It is generally required that each vertex in the approximated chain is within ε distance from some point on the original boundary.

Simplifying a complex polygonal chain with simpler ones has been extensively used in many application areas such as cartography, geographical information system (GIS) [5], computer graphics, medical imaging, transportation research, data transmission, and pattern recognition [2, 8]. Input polygonal chain Ch_1 is represented by listing coordinates of its vertices $p_1, p_2, p_3, \dots, p_n$ in the order they occur along the boundary. When the first and the last vertices are the same we have the boundary of a simple polygon. If the first and the last vertices are not same we have an open polygonal chain. Polygonal chains are approximated in term of error ε . The goal is to approximate the given input chain with a new one Ch_2 which has fewer number of vertices. Furthermore, each vertex of the approximated chain Ch_2 is required to be within ε distance from some point on the boundary of original chain Ch_1 . The vertices of the approximated chain are usually a subset of the input chain.

In this paper, we make an attempt to explore the change in the visibility properties of a given polygonal chain when it is approximated with a fewer number of vertices. We show that most of the well-known chain approximation algorithms generate approximated solutions that do not retain the visibility properties of the original chain.

The paper is organized as follows. In Sect. 45.2, we present an overview of important algorithms reported in applied computational geometry literature, dealing with polygon chain approximation. In Sect. 45.3, we examine the effect of boundary approximation on the visibility properties of the input polygon. In particular, we show that the widely used polygon simplification algorithms do not retain visibility properties. We then present an algorithm for simplifying polygonal boundary so that the approximated boundary tends to retain visibility properties. In Sect. 45.4, we discuss applications and further generalizations of the proposed problem.

45.1 Introduction

Approximating a polygonal boundary with a fewer number of vertices is a well investigated problem by researchers from computational geometry, computer graphics and big data analysis communities. The main objective is to approximate a given chain with fewer number of vertices/edges without losing the structural and feature properties. While most investigators require the vertices in the approximated chain to be a sub-set of the original chain, a few have allowed new vertices to be present in the approximated solution. Popular distance measures used for approximating a polygon chain include

L. Gewali (✉) · S. Jha
 Department of Computer Science, University of Nevada Las Vegas,
 Las Vegas, NV, USA
 e-mail: laxmi.gewali@unlv.edu

45.2 Review of Polygonal Approximation

An interesting polygonal approximation algorithm based on the divide and conquer paradigm was reported in Douglas and Peucker Algorithm [3] (DP Algorithm, in short), which is very intuitive and straightforward to understand. The DP algorithm takes (i) an open polygonal boundary chain $Ch = \langle v_1, v_2, \dots, v_n \rangle$ and (ii) predetermined error level ϵ as input and outputs a reduced size chain Ch' with a fewer number of vertices. It first compares the line segment s_l connecting the first and the last vertices v_1 and v_n with the chain Ch . If all vertices of Ch are within distance ϵ from s_l then s_l is the approximation for Ch . On the other hand, if some vertices of Ch are at distance larger than ϵ then the DP algorithm partitions the chain Ch into two parts: left chain Ch_l and right chain Ch_r by using the vertex v_r of Ch that has the largest distance from segment s_l . Ties are resolved arbitrarily. The algorithm then proceeds recursively in the left chain Ch_l and right chain Ch_r . Figure 45.1 shows an example of the execution trace of DP algorithm for an input chain with 36 vertices and error level equal to the length of the last edge in the input chain. The bottom part of Fig. 45.1 shows the approximated chain (drawn as dashed) with 8 vertices.

Another wisely cited boundary approximation algorithm was reported by Imai and Iri [8]. We refer to this algorithm as **I-I Algorithm** in short. For a given error level ϵ the I-I algorithm uses the concept of **ϵ -Rectangle** to capture the sub-boundary that can be approximated that satisfies the error level requirement. In simple term an **ϵ -Rectangle** is the smallest rectangle of width ϵ that covers the maximum number of consecutive vertices on the input boundary. The algorithm scans the boundary and identifies the sequence of ϵ -Rectangles to cover the entire boundary. The edges corresponding to the constructed ϵ -Rectangles give the approximated boundary.

The main ingredients of I-I algorithm are depicted in Fig. 45.2. The top part of Fig. 45.2 is an input monotone polygon with 68 vertices. The error level is as indicated in the top left corner. I-I algorithm covers the boundary with twelve ϵ -Rectangles as indicated in the middle part of Fig. 45.2. The approximated boundary with 13 edges is shown on the bottom part of Fig. 45.2. The segments approximating the chains corresponding to ϵ -Rectangles are drawn with dashed edges.

45.3 Visibility and Boundary Approximation

We begin with the definitions of visibility inside or outside of a simple polygon as practiced by researchers in computational geometry [1, 2, 4, 6, 7, 9, 10]. Two points p

and q inside a simple polygon are said to be **visible** if the line segment $s_{p,q}$ connecting them lies completely inside the polygon. This essentially means that $s_{p,q}$ does not intersect with the boundary of the polygon. Similarly, two points p and q outside a simple polygon are said to be **visible** if the line segment $s_{p,q}$ connecting them lies completely outside the polygon. This also means that $s_{p,q}$ does not intersect with the boundary of the polygon.

One of the key concepts in the study of visibility properties of polygons is the notion of the **visibility polygon** from a point in the presence of the polygon. The net area visible from a point p is the visibility polygon. Visibility polygons in the interior and exterior of a simple polygon are illustrated in Fig. 45.3. While the visibility polygon in the interior of a polygon is bounded, the visibility polygon in the exterior could be unbounded.

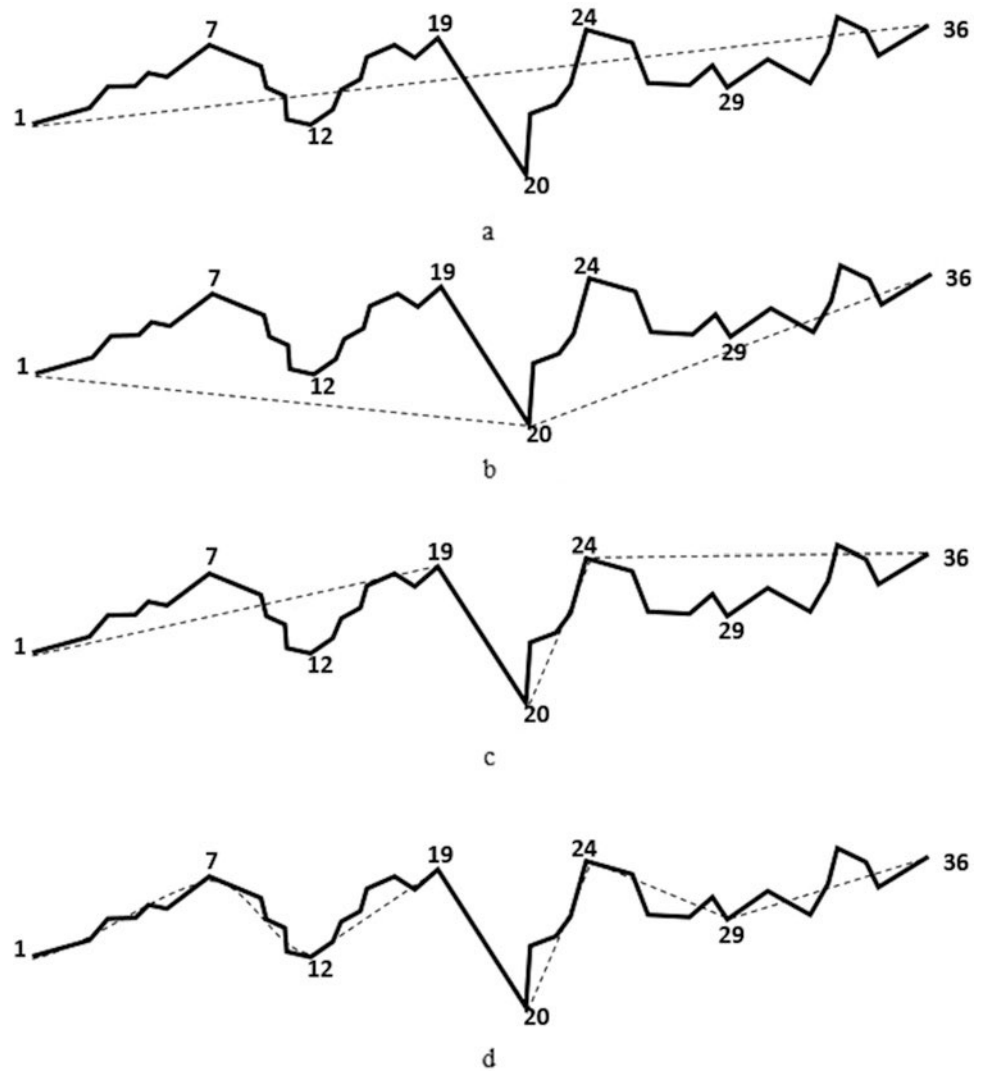
We now look at the change in the visibility properties of polygons when the boundary of the input polygon P is approximated with a polygon Q with fewer number vertices. Additionally, the approximation should satisfy **ϵ -tolerance** property. The notion of **ϵ -tolerance** is formally defined as follows.

Definition 1 Let Q denote a polygon with a fewer number of vertices obtained by approximating the boundary of polygon P . Polygon Q is said to satisfy “ **ϵ -tolerance** property” if any vertex of Q is within distance ϵ from some point on the boundary P .

The visibility properties of polygons are usually investigated in two modes: *internal mode* and *external mode*. While in the internal mode, the visibility properties are considered in the region enclosed by the boundary, in the external mode, the visibility properties are examined in the region outside of the polygon. Effect of boundary approximation on the internal visibility of simple polygons are examined in [5]. Visibility properties in the interior of the polygon can change significantly when the boundary is approximated within **ϵ -tolerance**. Specifically, it is observed in [5] that visibility inside the polygon can change arbitrarily, when a sub-chain is approximated by a single line segment, even though each point on the approximated line segment is within ϵ distance from some point in the original chain.

Our interest here is to explore the effect on the external visibility when the boundary of a 1.5D terrain is approximated subject to error tolerance level ϵ . A 1.5D terrain can be formally defined in term of monotone polygons. Monotone polygons have been extensively studied by investigators in computational geometry community [9]. Specifically, a simple polygon is called **monotone** with respect to a given direction \mathbf{d} if its boundary consists of two disjoint chains each of which are monotone with respect to \mathbf{d} . It may be noted that if one of the chains of a monotone polygon is just one

Fig. 45.1 Illustrating the progress of DP algorithm



horizontal line segment then it is called a **monotone mountain** [9]. Quite a few computational geometry researchers refer to monotone mountain as 1.5D terrain. The use of this term is becoming popular due to the fact that the sky line of the terrain is exactly a monotone chain. One can view 1.5D terrain to be a cross-section formed between a vertical plane and a 2D terrain.

The specific problem we propose to investigate can be stated as follows.

External V-Aware Approximation Problem (EAP)

Given: (a) A 1.5D polygon P with vertices p_1, p_1, \dots, p_n ,
 (b) Error tolerance level ϵ .

Question Construct a smaller 1.5D polygon Q with a fewer number of vertices so that (a) each vertex of P is within distance ϵ from some vertex in Q , and (b) the external regions visible from any approximated line segment in Q is identical to the corresponding chain in the original polygon P .

Definition 2 A sub-chain $Ch_{i,j} = p_i, p_{i+1}, p_{i+2}, \dots, p_j$ of a 1.5D terrain T is called a **concave chain** if external angles at vertices $p_i, p_{i+1}, p_{i+2}, \dots, p_j$ are all greater than 180 degrees.

Definition 3 The line segment $\langle p_i, p_j \rangle$ connecting the first and the last vertex of a concave chain is called its **lid**. In Fig. 45.4, the chain $\langle p_{16}, p_{17}, p_{18}, p_{19}, p_{20} \rangle$ is a maximal concave chain and the line segment connecting p_{16} to p_{20} is its lid.

Maximal concave chains play a key role in understanding the external visibility of 1.5D terrain T . It turns out that the external area visible from the concave chain (outside the lid) is exactly the area visible from its **lid**. This is stated in the following lemma.

Lemma 1 Let $Ch_{i,j} = p_i, p_{i+1}, p_{i+2}, \dots, p_j$ be concave chain of a 1.5D terrain. The external visibility polygon VP_2 (outside the lid) from the chain $Ch_{i,j}$ is identical to the external visibility polygon VP_1 from the **lid** $e_{i,j}$ of the chain.

Fig. 45.2 Illustrating the progress of I-I algorithm

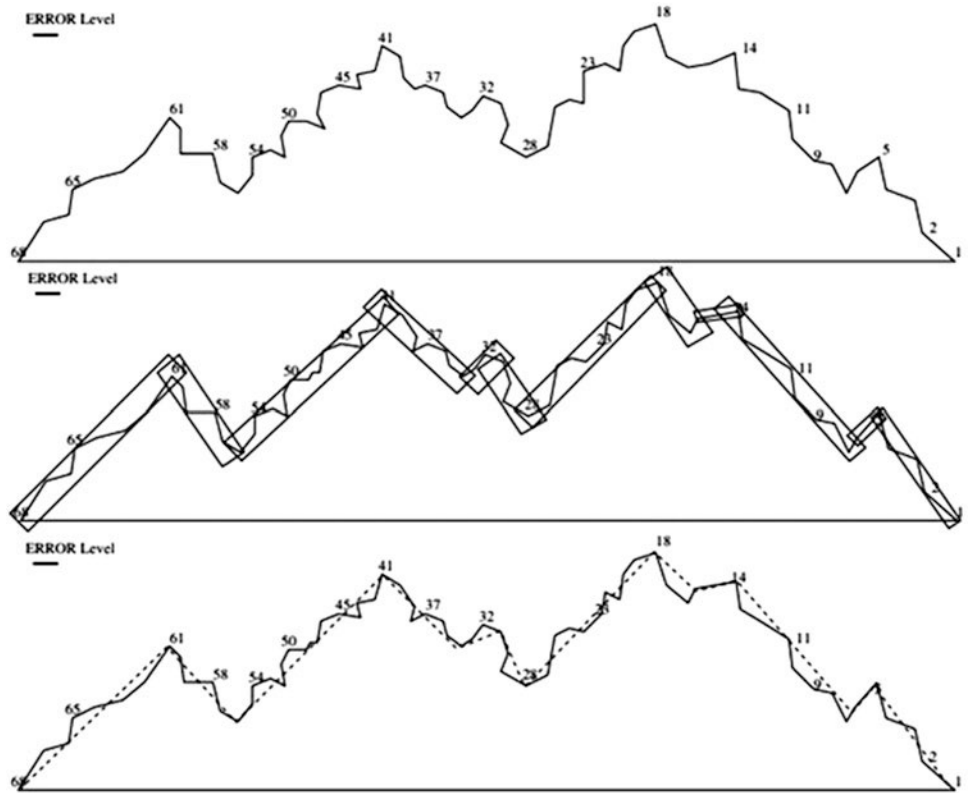


Fig. 45.3 Illustrating internal and external visibility polygons

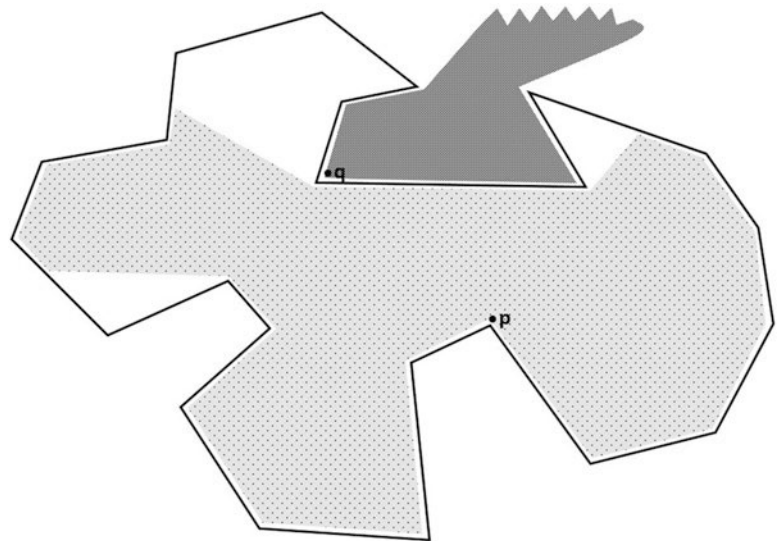
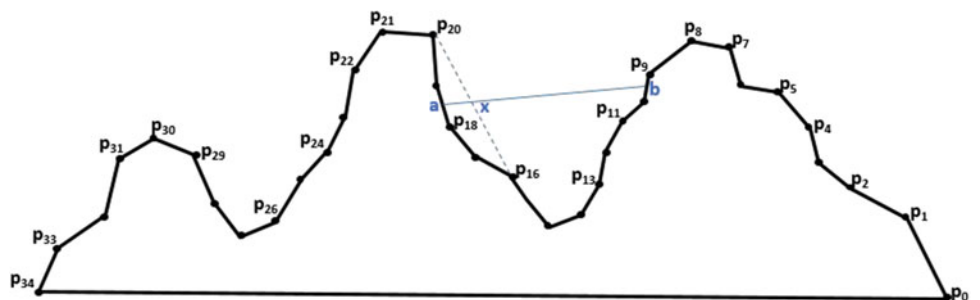


Fig. 45.4 Proof of Lemma 1



Proof We prove the lemma by arguing that for any visibility ray r_1 originating from a point on chain $Ch_{i,j}$, there is a visibility ray r_2 originating from a point in $e_{i,j}$ and coinciding with r_1 and vice versa. Consider a 1.5D terrain T as shown in Fig. 45.4.

Let us examine the visibility polygon VP_2 produced by chain $Ch_{i,j}$. Any visibility ray r_i originating from a point a in $Ch_{i,j}$ and hitting the boundary of the polygon at a point b must intersect the edge e_i at a point x . The existence of such an intersection point x follows from the convexity of the region enclosed by $Ch_{i,j}$ and its lid. This means for every visibility ray $r_1 = [a,b]$, emanating from a point in Ch_i , there is a corresponding visibility ray $r_2 = [x,b]$, emanating from a point x in the lid and coinciding with r_1 . This implies that the part of VP_2 outside the lid is identical to VP_1 . \square

C-Chain Approximation

The approximation algorithm we present is based on identifying concave chains of ϵ -width (referred now on simply as ϵ -C-Chains) on the boundary.

We can now describe the C-Chain Approximation algorithm. The inputs to the algorithm are (i) error tolerance level ϵ , and (ii) a 1.5D terrain T . The algorithm outputs approximated terrain T' whose vertices are within ϵ distance from some point on the boundary of T . The algorithm scans the boundary of T and identifies ϵ -C-Chains one at a time such that the identified ϵ -C-Chains are disjoint from each other. For capturing ϵ -C-Chains only those chains are considered that form concave a chain from outside.

The boundary of 1.5D Terrain is partitioned into maximal concave chains as shown in Fig. 45.5, where the lids of the maximal concave chains are drawn red. For the purpose of describing the algorithm we take the length of the edge

$\langle p_{33}, p_{34} \rangle$ (Fig. 45.5a) as the value of the error tolerance level ϵ . The widths of some maximal concave chains are larger than ϵ . For example, the width of maximal concave chain $\langle p_{25}, p_{26}, p_{27}, p_{28}, p_{29} \rangle$ is far larger than ϵ . Similarly, the width of the maximal concave chain $\langle p_{16}, p_{17}, p_{18}, p_{19}, p_{20} \rangle$ is bigger than ϵ . Large width maximal concave chains are partitioned into the minimum number of concave chains whose widths are less or equal to ϵ we obtain the intended sequence of concave chains. The approximation of the terrain's boundary is achieved by replacing each concave chain with their lids which is shown in Fig. 45.5c. In Fig. 45.5c, the approximated boundary is drawn with dashed edges, where the original boundary with 35 vertices is approximated by a boundary with 24 vertices.

We now proceed to develop a formal sketch of the C-Chain Approximation algorithm. The proposed C-Chain Approximation algorithm makes use of the function for extracting ϵ -C-Chains from a given concave chain $CH = \langle v_i, v_{i+1}, v_{i+2} \dots v_j \rangle$. The function examines the width of sub

chains $CH_{i,j} = \langle v_i, v_{i+1}, v_{i+2} \dots v_j \rangle$ of lengths 3, 4, 5, ... one at a time to find the consecutive sub chains $CH_{i,j}$ and $CH_{i,j+1}$ such that the conditions listed below are satisfied. In these conditions, $WD(CH_{i,j})$ denotes the width of the chain $CH_{i,j} = \langle v_i, v_{i+1}, v_{i+2} \dots v_j \rangle$.

$$WD(CH_{i,j}) < \epsilon \quad (45.1)$$

$$WD(CH_{i,j+1}) > \epsilon \quad (45.2)$$

We refer to the first vertex of the sub-chain v_i as first leg and the end vertex v_j as last leg. The algorithm initially starts with sub-chain $CH_{i,i+2}$ (the triangle corresponding to the first three vertices) as the candidate ϵ -C-Chain. If the width of $CH_{i,j} = CH_{i,i+2}$ is greater than ϵ then the input chain has no ϵ -C-Chain and hence $CH_{i,j}$ cannot be approximated. On the other hand, if the width of $CH_{i,i+2}$ is smaller than ϵ then the second leg (j) is incremented and the width of $CH_{i,j}$ is checked. This process of moving the second leg continues until the end of the input chain is reached or conditions (3.1) and (3.2) are satisfied. If the end of the input chain is reached, then the last ϵ -C-Chain is found. If conditions (3.1) and (3.2) are satisfied without reaching the end of the input chain, then a ϵ -C-Chain is found and the algorithm proceeds to identify the next ϵ -C-Chain by resetting the first leg to the running last leg and last leg equal to new first leg plus 2.

A formal description of the algorithm for extracting ϵ -C-Chain components for a given input maximal concave chain is listed as Algorithm 45.1.

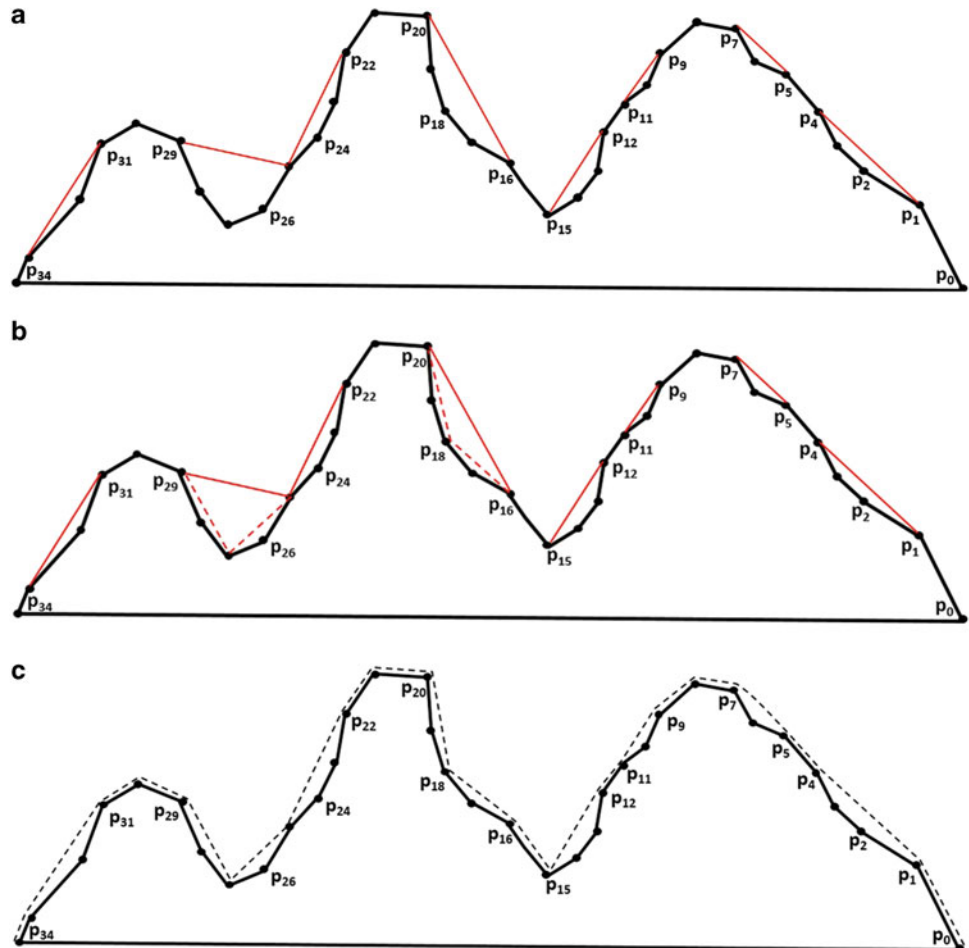
Algorithm 45.1: Extract ϵ -C-Chains

```

1:  Input: (i) Concave chain  $ch_1 = \langle v_1, v_2$ 
      , . . . .  $v_n \rangle$ 
2:  Input: (ii) Tolerance threshold  $\epsilon$ 
3:  Output: Disjoint  $\epsilon$ -C-Chains components
4:   $i=1$ ;  $j = 3$ ;
5:  while ( $j \leq n$ ) do {
6:      if (  $wd(ch_1, i, j) \geq \epsilon$  )
7:           $i++$ ;  $j = j + 2$ ;
8:      else if ( $j \geq n$ )
9:          Output segment ( $p_i, p_j$ );
10:     else if ( $wd(ch_1, i, j) \geq \epsilon$ ) {
11:         Output segment ( $p_i, p_j$ );
12:          $i = j$ ;  $j = j + 2$ ;
13:     }
14:     else  $j++$ ;
15: }
```

The actual C-Chain approximation algorithm works by first marking maximal concave chains on the boundary of 1.5D terrain T . The consecutive sequence of vertices with external angle greater than 180 degrees are precisely the maximum concave chains. All such chains can be marked

Fig. 45.5 Alternate sequence of concave/convex chains



in one scan of the boundary by checking external angles. In the second pass around the boundary of T , Algorithm 1 is invoked for each maximal chain. Replacing each ϵ -C-Chain with corresponding lids gives the approximated boundary T' .

Time Complexity The time complexity of the proposed algorithm can be analyzed in straightforward manner. Algorithm 1 checks each vertex constant number of times. The shape formed by the maximal concave chain and its lid is a convex polygon such that the projection of edges of the concave chain on the lid are disjoint. Because of convexity and the disjoint property, the width of the shape corresponding each ϵ -C-Chain can be computed in time proportional to its size. Hence the total time to build the approximated chain is $O(n)$ where n is the number of vertices on the boundary of 1.5D terrain T .

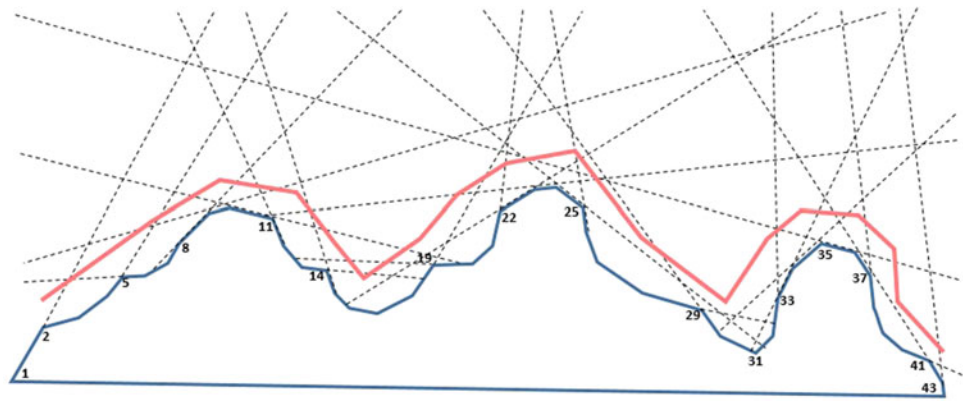
45.4 Discussions

The problem of boundary approximation subject to external visibility has rich applications in (i) the placement of cellular tower and (ii) covert path planning for surveillance on terrain

surface. If the input 1.5D terrain is of very large size, then the proposed approximation algorithm can be applied to obtain a reduced size 1.5D terrain. Furthermore, in situations where we need to transmit the terrain data we could apply the proposed approximation algorithm to compress the input and transmit efficiently.

For applications in covert path planning the objective is to construct a collision-free path over the 1.5D terrain such that the constructed path is least exposed to observation tower placed on the surface. A covert path P should be such that its exposure to points on terrain surface should be reduced and at the same time the path should have acceptable clearance from the terrain. Somehow we need to extrapolate visibility exposure of regions above the terrain surface to construct such a path. A path that penetrates the regions of low exposure can be taken as a covert path. One approach would be to evaluate external visibility profile of region outside the terrain surface by extending the visibility rays induced by the edges of 1.5D terrain. If we extend the rays emanating from the edges of the terrain we get an arrangement of visibility rays as shown in Fig. 45.6, where visibility rays are drawn with dashed lines. A path above the ground that crosses the minimum number of visibility edges could be taken as a covert path. A path

Fig. 45.6 Illustrating a possible covert penetration



consisting of line segments as shown by red colored edges in Fig. 45.6 could be taken as the initial candidate for covert path. Some of the turn angles implied by the initial path may be too sharp for an aerial vehicle used for reconnaissance mission. To address such problems sections of the path (sub-paths) near the sharp turns need to be patched. For example, in Fig. 45.6, the part of the path above the terrain vertices v_{15} and v_{16} has a very sharp turn which is clearly forbidden for aerial vehicles. Such portions of the paths could be patched by a detour path that does not have sharp turn and its visibility to other parts of the terrain is reduced.

Details of constructing covert path above a terrain are in progress and will be reported in the future.

For computing ε -C-Chains (in Sect. 45.3) we required that the external visibility from the C-Chain is identical to external visibility from the corresponding approximating segment. We can relax this requirement to develop another version of approximation algorithm described in Sect. 45.3. The detail of this version of the algorithm would be worth investigating.

References

1. W.S. Chan, F. Chin, Approximation of polygonal curves with minimum number of line segments or minimum error. *Int. J. Comput. Geom. Appl.* **6**, 59–77 (1996)
2. M. de Burg, M. van Kreveld, M. Overmars, O. Schwarzkopf, *Computational Geometry: Algorithms and Applications*, 2nd edn. (Springer, 2000)
3. D. Douglas, T. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica Int. J. Geogr. Inf. Geovisualiz.* **10**, 112–122 (1973)
4. S. Eidenbenz, Approximation algorithms for terrain guarding. *Inf. Process. Lett.* **82**(2), 99–105 (2002)
5. L. Gewali, S. Jha, Effect of boundary approximation on visibility, (in ITNG 2021). *Adv. Intell. Syst. Comput.* **1346**, 247–253 (2021)
6. S.K. Ghosh, D.M. Mount, An output sensitive algorithm for computing visibility graphs, in *28th Annual Symposium on Foundations of Computer Science (sfcs 1987)*, pp. 11–19
7. M. Hacar, T. Gokgoz, A new score-based multi stage matching approach for road network conflation in different road patterns. *ISPRS Int. J. Geo Inf.* **8**(2), 81 (2019)
8. H. Imai, M. Iri, Computational-geometric methods for polygonal approximations of a curve. *Comp. Vis. Graph. Image Process* **36**(1), 31–41 (1986)
9. J. O'Rourke, *Computational Geometry in C*, 2nd edn. (Cambridge University Press, Cambridge, 1998)
10. Z. Xie, Z. Ye, L. Wu, Research on building polygon map generalization algorithm, in *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, vol. 3 (SNPD, 2007), pp. 786–791

Detection of Strictly L3-Live Structures by Structural Analysis of General Petri Net Using SAT-Solver

Yuta Yoshizawa and Katsumi Wasaki

Abstract

One of the dynamic properties of Petri nets is liveness, which ranges from $L0$ to $L4$ depending on the severity of the condition. The structure required for the existence of strictly $L3$ -live transitions is referred to as the strictly $L3$ -live structure. The strictly $L3$ -live structure consists of three elements: a repeating closed circuit ($L3$ -circuit), a transition that cancels the liveness of the $L3$ -circuit ($CircuitBreaker$), and a place that receives a token supplied by the $L3$ -circuit (k -place). To detect these elements, we use the partially-conservative, partially-repetitive, and bounded structural properties of general Petri nets, as well as solving the matrix inequalities, which are necessary and sufficient conditions for the three structural properties, using the SAT solver to detect the three elements that constitute the strictly $L3$ -live structure.

Keywords

Petri nets · Model check · Liveness check · SAT solver · HiPS tool · Structural properties · Strictly L3-live · Strictly L2-live · Structural analysis · Liveness level

46.1 Introduction

Recently, information technology-based systems have been used in various aspects of our lives, and the scale and num-

Y. Yoshizawa
Graduate School of Science and Technology, Shinshu University,
Nagano, Japan
e-mail: 21w2081h@shinshu-u.ac.jp

K. Wasaki (✉)
Faculty of Engineering, Shinshu University, Nagano, Japan
e-mail: wasaki@cs.shinshu-u.ac.jp

ber of these systems have been increasing. Because these technologies serve as the foundations for social systems, ensuring not only the performance but also the reliability of these systems has become critical. However, guaranteeing the design and operation of parallel systems, such as asynchronous circuits and asynchronous communication protocols, is difficult because errors do not easily appear during verification. One way to verify the reliability of these parallel systems is to model them with Petri nets. Petri nets are mathematical models that show the behavior of discrete event systems with parallelism, asynchrony, and non-determinism of event occurrence [1, 2]. Petri nets can perform three functions simultaneously: graphical tools, simulation tools, and mathematical methodologies. This describes the system structure in a visual representation as a graphic tool, and it can simulate parallel and dynamic events in the system using tokens in Petri nets. However, as a mathematical tool, it can build equations of state, algebraic equations, and other models that describe the behavior of the system.

A Petri net support tool, HiPS (Hierarchical Petri Net Simulator), has been developed at Shinshu University to describe and analyze Petri nets [3–5]. HiPS has an intuitive and common GUI and various functions such as structural and dynamic property analysis. Petri nets represent a system's simulation by manipulating the firing of transitions. The possibility of a transition firing is referred to as activity level, and it is closely related to the system being deadlocked or not happening at all. Being active is an ideal property of many systems, but it is a demanding property for large systems; thus, activity levels have been defined, with the activity condition relaxed at each level. Understanding the behavior of a system requires knowing the activity level a transition in a Petri net belongs to.

In this study, we describe the structure and analysis required for the existence of activity level-3 transitions. Here, we discuss the structural properties of general Petri nets and how to SAT-solver to efficiently analyze them. Section 46.2

describes the concept and properties of Petri nets. Section 46.3 defines the strictly $L3$ -live structure. Section 46.4 describes the structure detection method. Section 46.5 presents examples of structural analysis.

46.2 Petri Nets

46.2.1 Place/Transition Net

Carl Adam Petri proposed the Petri net model to graphically represent a discrete event system consisting of multiple processes. It can be used to describe and study information processing systems with asynchronous, distributed, parallel, nondeterministic, and stochastic behavior. The model can be used to describe and study information processing systems with linear, asynchronous, distributed, parallel, nondeterministic, and stochastic behavior. It can be used to simulate system behavior using tokens in Petri nets, additionally to be used as a visual representation of the system structure. A Petri net is a directed bipartite graph whose vertices are places and transitions, and whose four components are arcs and tokens.

In a Petri net, places are drawn as circles, and transitions are drawn as squares or bars. Arcs connect places to transitions and transitions to places. They are represented by arrows according to their connection directions.

A P/T-net is a 5-tuple, $PN = (P, T, F, W, M_0)$, where:
 $P = \{p_1, p_2, \dots, p_m\}$ is a finite set of places,
 $T = \{t_1, t_2, \dots, t_n\}$ is a finite set of transitions,
 $F \subseteq (P \times T) \cup (T \times P)$ is a set of arcs (flow relations),
 $W : F \rightarrow \{1, 2, 3, \dots\}$ is a weight function,
 $M_0 : P \rightarrow \{0, 1, 2, 3, \dots\}$ is the initial marking,
 $P \cap T = \emptyset$, and $P \cup T \neq \emptyset$.

A Petri net structure $N = (P, T, F, W)$ without any specific initial marking is denoted by N .

A Petri net structure $N = (P, T, F, W)$ without any specific initial marking is denoted by N . For a Petri net (N, M_0) , the set of all possible markings reachable from the initial marking M_0 is denoted by $R(N, M_0)$ or simply $R(M_0)$. The set of all possible firing sequences from M_0 in a net (N, M_0) is denoted by $L(N, M_0)$ or simply $L(M_0)$.

46.2.2 Incidence Matrix

In a Petri net (N, M_0) such that the number of transitions is n and the number of places is m , the incidence matrix $A = [a_{ij}]$ is an n -by- m matrix that represents the connection between transitions and places. Each component of the incidence matrix is given by $a_{ij} = a_{ij}^+ - a_{ij}^-$. Here, $A_{ij}^+ = w(i, j)$ is the weight of the arc from transition i to its output place

j , and $A^+ = [a_{ij}^+]$ is called the positive incidence matrix. Furthermore, $a_{ij}^- = w(j, i)$ is the weight of the arc from the input place j of transition i to transition i , and $A^- = [a_{ij}^-]$ is known as the backward connection matrix.

46.2.3 Firing Condition

The transition t can fire if there are more tokens in the input place p of the transition t than the weight of the arc from the input place p to the transition t . The transition t is fireable if the input place p has at least one token of the weight of the arc from the input place p to the transition t . A fireable transition can either fire or not fire. This is because, while an event may be in a state where it can occur, it may or may not occur. When the transition t fires, a token of the arc weight is removed from each input place. When the transition t fires, a token of the arc weight is removed from each input place. Then, a token for the weight of the arc is added to each output place of t .

46.2.4 Structural Properties

Structural properties are Petri net properties that are either independent of the initial marking or associated with the presence of a particular firing sequence from some initial marking. These properties can be described by the incidence matrix A and its associated homogeneous equations and inequalities because they depend on the structure of the Petri net. In the necessary and sufficient condition equation, x denotes the firing frequency vector, and y denotes the weighted sum of the total number of tokens.

1. *Structurally Bounded (SB)*: A Petri net N is said to be structurally bounded if it is bounded for any finite initial marking M_0 . The necessary and sufficient conditions are

$$\exists y > 0, Ay \leq 0 \quad (46.1)$$

2. *Partially Repetitive (PRP)*: A Petri net N is said to be partially repetitive if there exists a marking M_0 and a firing sequence σ from M_0 such that some transitions occur infinitely. The necessary and sufficient conditions are

$$\exists x \geq 0 (x \neq 0), A^T x \geq 0 \quad (46.2)$$

3. *Partially Conservative (PCS)*: A Petri net N is said to be partially conservative if there exists a positive integer

Table 46.1 Liveness level

L1-live
The transition t can fire at least once in the firing sequence with $L(M_0)$.
L2-live
For any positive integer k , the transition t is firable at least k times in some firing sequence of $L(M_0)$.
L3-live
The transition t appears infinitely many times in the firing sequence with $L(M_0)$.
L4-live
Transition t is <i>L1-live</i> for all markings in $R(M_0)$.

solution $y(p)$ for some place p , such that the weighted total number of tokens ($M^T y = M_0^T y$) is constant for all $M \in R(M_0)$ and any fixed initial marking M_0 . The necessary and sufficient conditions for partially conservative are

$$\exists y \geq 0 (y \neq 0), Ay = 0 \quad (46.3)$$

46.2.5 Liveness

One of the dynamic properties of Petri nets is a property known as Liveness. We say a Petri net (N, M_0) is Liveness if any transition in the net can be fired from any marking M_0 through any firing sequence, regardless of which marking is reached from the initial marking M_0 . This means that an active Petri net ensures deadlock-free operation regardless of the firing sequence used. Therefore, while liveness is an ideal property for many systems, verifying this stringent property is prohibitively expensive for large systems. Therefore, the levels of liveness shown in Table 46.1 are defined by relaxing the liveness condition [7, 8]. A Petri net (N, M_0) is said to be *Lk-live* if all transitions in the net are *Lk-live* ($k = 1, 2, 3, 4$). The *L4-live* is the strongest activity level and corresponds to the liveness defined. A transition is said to be strictly *Lk-live* if it is *Lk-live* and not *L(k + 1)-live*, where ($k = 1, 2, 3$). In the following, when we say *Lk-live* or *Lk-live*, we mean strictly *Lk-live*.

46.3 Proposal of Strictly L3-Live Structure

The strictly *L3-live* structure is the structure required for the existence of strictly *L3-live* transition. t_3 is strictly *L3-live* in the net of Fig. 46.1 because it can fire an infinite number of times if t_1 does not fire. After t_1 has fired, t_2 can fire as many times as t_3 has fired (the number of tokens in p_2). When t_2

can fire, t_1 has already fired and t_3 is no longer active, so the upper limit of the number of tokens in p_2 is bounded. Therefore, while t_2 cannot fire an infinite number of times in any firing sequence, it can fire as many times as t_3 fires, making it strictly *L2-live*. Therefore, the existence of strictly *L2-live* transitions depends on the existence of strictly *L3-live* transitions.

In order for strictly *L3-live* transitions to exist, there must be a repeating closed circuit (*L3-circuit*) such as the one composed of t_3 and p_1 , and a transition (*CircuitBreaker*) such as t_1 that removes the liveness of the circuit. The strictly *L3-live* structure is composed of these three elements plus a place (*k-place*) that receives a token from a *L3-circuit* transition such as p_3 . The existence of *L3-circuit* and *CircuitBreaker* is required for the strictly *L3-live* structure to exist.

- L3-circuit*: A Petri net circuit has a structure in which all transitions on the circuit are iterative and deprive them of activity. In the net in Fig. 46.1, the circuit consisting of t_3 and p_1 is repetitive and its liveness is taken away by the firing of t_1 , resulting in a *L3-circuit*.
- CircuitBreaker*: A transition that removes the liveness of a transition in a *L3-circuit* that has a place in the *L3-circuit* as input; the input place in the *L3-circuit* must be a conservative place, and the transition must be strictly *L1-live*. This transition is represented by t_1 in the net in Fig. 46.1.
- k-place*: A place outside the *L3-circuit* receives tokens from transitions inside the *L3-circuit* and is infinitely supplied with tokens only by transitions inside the *L3-circuit*. An extended *k-place* is a place that receives tokens from a *k-place* and has the same properties as a *k-place*. p_2 is a *k-place* in the net in Fig. 46.1.

46.4 Structural Analysis

The flow of the structural analysis is shown in Fig. 46.2, where the structural properties are determined using the SAT-solver, and each structure is analyzed by net exploration using the solution. The analysis method is explained in Fig. 46.3.

46.4.1 Solving structural properties using SAT-solver

SAT-solver is used to obtain the necessary and sufficient conditions for structural properties. Google OR-Tools [6] is used as the solver. When solving for structural partially

Fig. 46.1 Contain L3 transition net

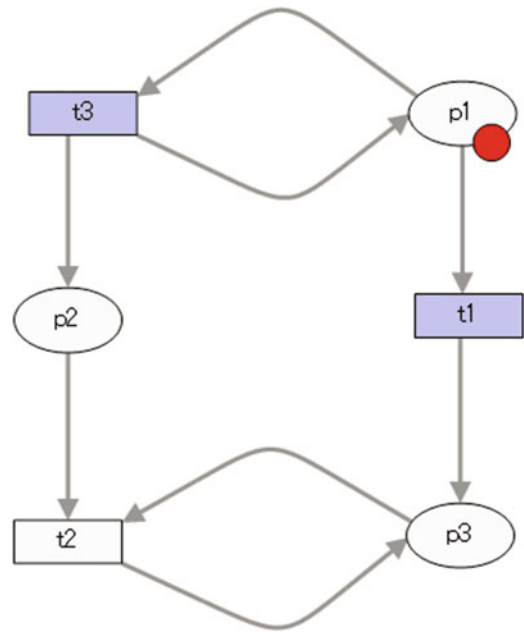
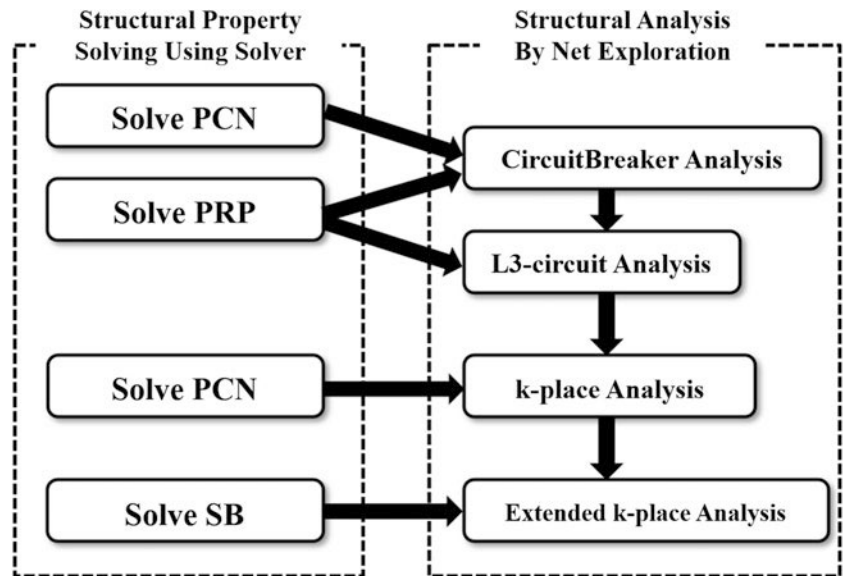


Fig. 46.2 Flow of structural analysis



repetitive, the elements of the array of solutions obtained are converted to bool values, and a conditional expression is added to exclude solutions with the same bool values as those already obtained when solving for the next solution.

$\{t_2, t_3\}$. Additionally, t_4 , a non-repetitive transition with a conservative p_2 as input, is determined to be a *CircuitBreaker*, and the set of conservative places $\{p_0, p_1, p_2\}$, as well as its input place p_2 is saved as *CircuitBreakerSource*.

46.4.2 Detection of CircuitBreaker

We can determine which places are conservative by finding a solution to structural partially conservative. Determine that the non-iterative transition whose input is a place on the iterative circuit is a *CircuitBreaker* by finding an iterative circuit such that the places on the iterative transition are conservative. In the net of Fig. 46.3, there are two iterative circuits: one that iterates at $\{t_0, t_1\}$ and one that iterates at

46.4.3 Detection of L3-circuit

We can find a pair of iterating transitions by solving for structural partially repetitive. It is determined to be an *L3-circuit* if there is *CircuitBreakerSource* on the iteration of that transition. In the net of Fig. 46.3, there are two iterative circuits, one iterating at $\{t_0, t_1\}$ and the other iterating at $\{t_2, t_3\}$. Because each of them has places *CircuitBreakerSource*

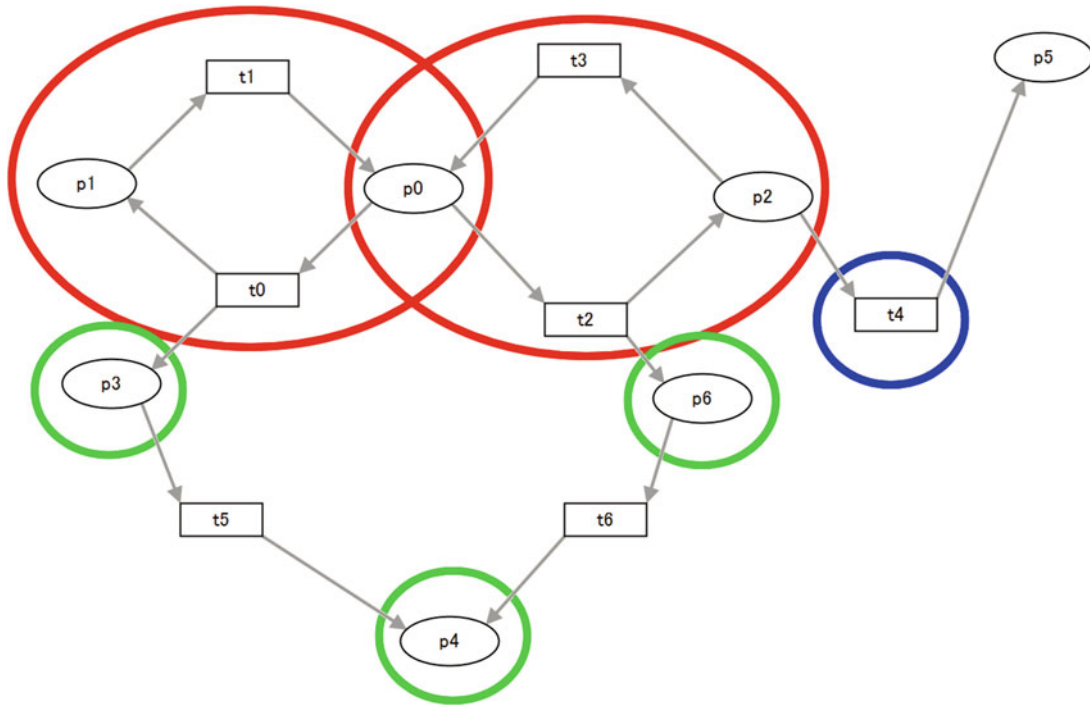


Fig. 46.3 A net containing strictly L3-live transitions (a)

lace on the circuit, these iterative circuits are determined as *L3-circuit*.

46.4.4 Detection of *k*-place

k-place is defined as a place outside the *L3-circuit* that receives tokens from transitions on the *L3-circuit* but does not receive an infinite supply of tokens from transitions on the *L3-circuit*. An *Extended k*-place is a place that receives a token from a *k*-place and has the same properties as a *k*-place. An *Extended k*-place is a place that receives a token from a *k*-place and has the same properties as a *k*-place. An *Extended k*-place is a place that receives a token from a *k*-place and has the same properties as a *k*-place. In the net of Fig. 46.3, p_3 is a *k*-place because it has as input only t_0 , a transition on the *L3-circuit*. Similarly, p_6 is a *k*-place. Although p_4 is originally an unbounded place, it becomes a bounded place in the subnet when all the *L3-circuit* that deprive the liveness of t_4 , which is the *CircuitBreaker* of the *L3-circuit* of p_3 , are deleted, and thus it becomes an *Extended k*-place of p_3 . Similarly, p_4 is an *Extended k*-place of p_6 (Table 46.2).

Table 46.2 Strictly L3-live structures of Fig. 46.3

L3LS	<u>L3-circuit</u>	<u>CircuitBreaker</u>	<u>k-place</u>
L3LS_1	$\{t_0, t_1\}$	$\{t_4\}$	$\{p_3, p_4\}$
L3LS_2	$\{t_2, t_3\}$	$\{t_4\}$	$\{p_4, p_6\}$

46.5 Example of Structural Analysis

Figure 4 shows an example of structural analysis. The structural partially conservative problem is solved, obtaining the result that $p_0, p_1, p_2, p_3, p_4, p_5, p_6$ are simultaneously conservative. We can also solve for structural partially repetitive to find iterative transitions, and then find a pair of transitions such that the places on the iterations are conservative, which are candidates for *L3-circuit*. For example, because $\{t_0, t_1\}$ is iterative and the places on the iterations, p_0, p_1 are conservative, $\{t_0, t_1\}$ is a candidate for *L3-circuit*. There is also a structural partially repetitive solution, $\{t_0, t_1, t_{12}\}$, but the place on this iteration, p_{14} , is not conservative, so it is not a candidate for *L3-circuit*. Here, t_{18} is a transition does not have iterative properties, and because it has an input p_0 , which is conservative on $\{t_0, t_1\}$ and *L3-circuit* candidate, it becomes a *CircuitBreaker*, and thus $\{t_0, t_1\}$ becomes an *L3-circuit*. Additionally, t_{18} is a *CircuitBreaker* for all other *L3-circuit* because it deprives the *L3-circuit* that have conservative places on the circuit of liveness simultaneously as p_0 . A *k*-place is a place outside the *L3-circuit* that receives a token from a transition on each *L3-circuit*. Although p_{20}

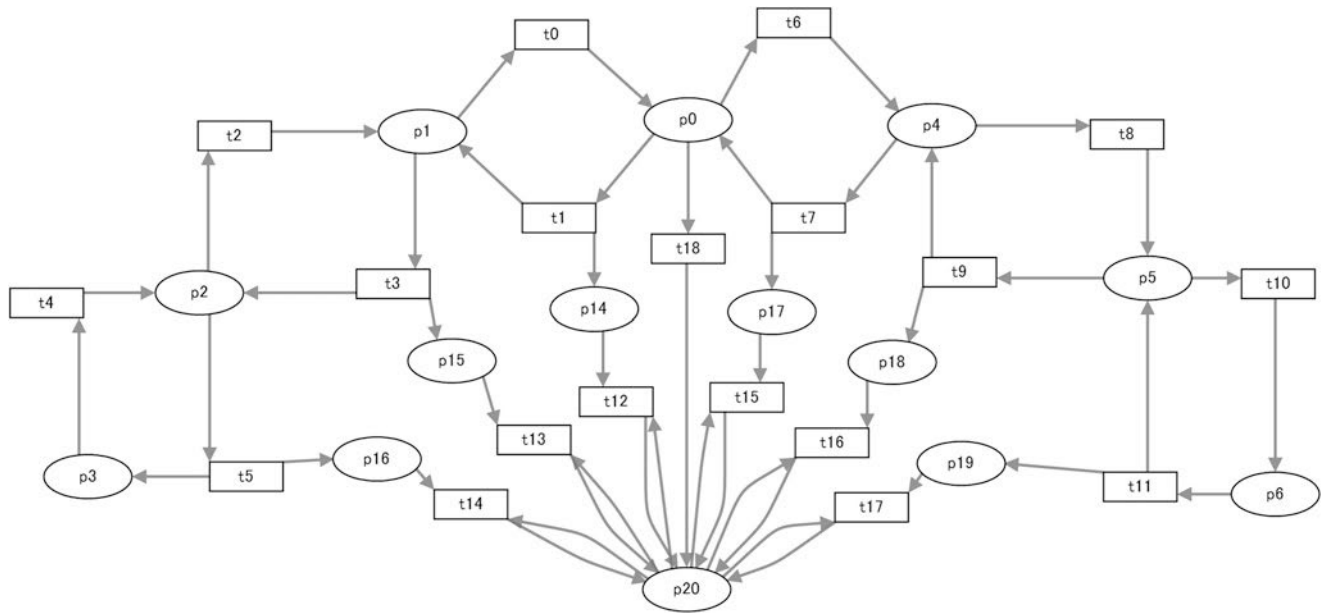


Fig. 46.4 A net containing strictly $L3$ -live transitions (b)

Table 46.3 Strictly $L3$ -live structures of Fig. 46.4

L3LS	L3-circuit	CircuitBreaker	k-place
L3LS_1	$\{t_0, t_1\}$	$\{t_{18}\}$	$\{p_{14}\}$
L3LS_2	$\{t_2, t_3\}$	$\{t_{18}\}$	$\{p_{15}\}$
L3LS_3	$\{t_4, t_5\}$	$\{t_{18}\}$	$\{p_{16}\}$
L3LS_4	$\{t_6, t_7\}$	$\{t_{18}\}$	$\{p_{17}\}$
L3LS_5	$\{t_8, t_9\}$	$\{t_{18}\}$	$\{p_{18}\}$
L3LS_6	$\{t_{10}, t_{11}\}$	$\{t_{18}\}$	$\{p_{19}\}$

is structured to receive tokens from all k -place, it is not an *Extended k -place* because it is originally a bounded place. Thus, the strictly $L3$ -live structure in Fig. 46.4 becomes Table 46.3.

46.6 Conclusions and Future Work

In this study, we discovered the strictly $L3$ -live structure, which is required for the existence of strictly $L3$ -live transitions. The SAT-solver was used to efficiently analyze the structural properties.

Future work includes the detection of the structures required for the existence of strictly $L2$ -live transitions. The strictly $L3$ -live structure is required for the existence of the strictly $L2$ -live transition. The strictly $L2$ -live transition can only fire after the *CircuitBreaker* removes the liveness of the

L3-circuit, as shown in t_2 in Fig. 46.1, and it can fire as many tokens as p_2 and k -place.

References

1. T. Murata, Petri nets: Properties, analysis and applications. Proc. IEEE **77**(4), 541–580 (1989)
2. C.A. Petri, W. Reisig, Petri net. Scholarpedia **3**(4), 6477 (2008)
3. Y. Harie, Y. Mitsui, K. Fujimori, A. Batajoo, K. Wasaki, HiPS: Hierarchical Petri Net design, simulation, verification and model checking tool, in *Proceedings of the 6th IEEE Global Conference on Consumer Electronics (GCCE)* (2017), pp.686–690
4. HiPS: Hierarchical Petri net Simulator, Shinshu University. Available at <https://sourceforge.net/projects/hips-tools/>
5. Y. Harie, K. Wasaki, A Petri Net design and verification platform based on the scalable and parallel architecture: HiPS, in *Proceedings of the 14th International Conference on Information Technology – New Generations (ITNG2017)*, Advances in Intelligent Systems and Computing, vol. 558 (Springer, 2017), pp. 265–273
6. GoogleOR-Tools: <https://developers.google.com/optimization>
7. F. Commoner, Deadlocks in Petri nets, Wakefield, Applied Data Research, Inc., Report #CA-7206-2311 (1972)
8. K. Lautenbach, Liveness in Petri nets, St. Augustin, Gesellschaft fur Mathematik und Datenverarbeitung Bonn, Interner Bericht ISF-75-02.1 (1975)

Space Abstraction and of PetriNets Using the Submarking Method Quasi-home States

47

Tomoki Miura and Katsumi Wasaki

Abstract

The home state is a property of Petri nets. The home state is a state that can be returned to from all markings, and is the stable state in the system. In this study, we define a quasi-home state as a state in which the conditions of the home state are relaxed. The quasi-home state is the home state in submarkings, which are abstracted markings. The submarking method can compress the state space while giving meaning to the marking. In this study, we defined four submarking methods. The user specifies the place to be abstracted and the submarking method, and submarking is obtained. The home state is obtained through the submarking's dynamic analysis. This study aims to improve the net analysis' efficiency using the submarking method to determine the quasi-home state.

Keywords

Petri nets · Reachability graph · HiPS tool · Coverability graph · Dynamic analysis · Home state · Quasi-home states · Concurrent models · Submarking · Closed circuit

47.1 Introduction

Recently, information technology-based systems have been employed in various applications, and their scale and number are increasing. Since these technologies support social sys-

T. Miura
Graduate School of Science and Technology, Shinshu University,
Nagano, Japan
e-mail: 21w2066d@shiinshu-u.ac.jp

K. Wasaki (✉)
Faculty of Engineering, Shinshu University, Nagano, Japan
e-mail: wasaki@cs.shinshu-u.ac.jp

tems' infrastructure, ensuring the performance and reliability of these systems is crucial. However, it is difficult to ensure the design and operation of parallel systems, such as asynchronous circuits and communication protocols that operate asynchronously, because errors do not easily occur during verification. Modeling these parallel systems with Petri nets [1, 2] is an approach for verifying their reliability. A Petri net is a mathematical model that shows the behavior of a discrete event system with parallel, asynchronous, and nonde-terministic event occurrences. Simultaneously, Petri nets can function as graphical tools, simulation tools, and mathematical approaches.

As a graphic tool, it describes the system structure in a visual representation, and by using tokens that represent the state of the Petri net, it can simulate parallel and dynamic events of the system. However, as a mathematical tool, it is used to develop equations of state, algebraic equations, and other models that describe the system's behavior.

When using Petri nets to model a large and complex system, the model is difficult to handle and understand from the creator's perspective. In this case, using a Petri net tool facilitates the descriptiveness, simulation, and behavior analysis of Petri nets. To solve the problems of descriptiveness, usability, and reusability of existing Petri net tools, a hierarchical Petri net design tool called Hierarchical Petri net Simulator (HiPS) was developed at Shinshu university [3–5]. HiPS is a Petri net design tool with an intuitive and common GUI. HiPS has a cover graph generator as an analysis function, and it can also generate graphs using the extended cover analysis (ECG) method, which is an extension of the regular cover analysis method.

When analyzing the dynamic behavior of Petri nets, it is important to examine all possible states (markings) from the initial state. However, counting the markings one by one while analyzing a large Petri net, is time-consuming due to numerous markings. Therefore, we reduce expenses by

abstracting the markings (submarking). This study aims to improve the net analysis' efficiency by determining the quasi-home state using the submarking method and compressing the state space by abstracting the markings without changing the system's essence. The home state is stable in the system because it can return from all markings. A quasi-home state is one in which the home state's conditions are relaxed. In submarking, the quasi-home state is the home state. Submarking is a marking abstraction that can compress the state space and give meaning to the marking. In this research, we defined four submarking methods. The user specifies the place and method of submarking, and the marking is abstracted. The submarking is abstracted by the user by specifying the place and method of submarking, and dynamic analysis is used to obtain the quasi-home state.

In Sect. 47.1, we describe the background, purpose, and overview of our study. In Sect. 47.2, Petri nets and their properties are discussed. In Sect. 47.3, submarking and quasi-home states are explained. In Sect. 47.4, an example of an analysis using the submarking method is presented, along with the results. In Sect. 47.5, we summarize and discuss future work.

47.2 Petri Nets

47.2.1 Overview of Petri Nets

In 1966, Carl Adam Petri proposed the Petri net model to graphically represent discrete event systems. It is a useful tool for describing and investigating information processing systems characterized by concurrent, asynchronous, distributed, parallel, nondeterministic, and stochastic behavior. In addition to visualizing the system structure, the behavior of the system can be simulated using tokens in a Petri net. A Petri net is a weighted directed bipartite graph with two types of nodes: place and transition, and an arc that connects the nodes. The following is a definition of a Petri net.

A P/T-net is a 5-tuple, $PN = (P, T, F, W, M_0)$, where:

$P = \{p_1, p_2, \dots, p_m\}$ is a finite set of places,

$T = \{t_1, t_2, \dots, t_n\}$ is a finite set of transitions,

$F \subseteq (P \times T) \cup (T \times P)$ is a set of arcs (flow relations),

$W : F \rightarrow \{1, 2, 3, \dots\}$ is a weight function,

$M_0 : P \rightarrow \{0, 1, 2, 3, \dots\}$ is the initial marking,

$P \cap T = \emptyset$, and $P \cup T \neq \emptyset$.

A Petri net structure $N = (P, T, F, W)$ without any specific initial marking is denoted by N .

In Petri nets, places are represented as circles, and transitions are represented as squares or bars. The place connected to the arc exiting the transition is called the output place and the place connected to the arc entering the transition is called

the input place. An arc with a weight of k is called a k -weight arc (multiple arc) and can be interpreted as a set of k parallel arcs. When k is 1, the weights are omitted. A Petri net in which all the arcs have a weight of one is said to be normal (ordinary). When k tokens are assigned to a place, k points or k are marked on the place. A place is marked with k points or k when it has k tokens assigned to it. k is a non-negative integer, and a place is said to be marked if it has tokens assigned to it. The placement of a token at a place in the entire Petri net is called marking M , and the initial state is called initial marking M_0 . Marking M is a m -dimensional vector of m places, with the p -th component denoted by M_p , which represents the number of tokens in place p . Many systems' behavior can be described by their states and transitions. There are two types of transitions: source and sink transitions. Source transitions are transitions without an input place and sink transitions are transitions without an output place. A source transition can fire unconditionally, and the firing of a sink transition consumes tokens. When place p is the input and output place of transition t , the pair of p and t is called a self-loop. A Petri net with no self-loop is called pure. A Petri net in which all arcs have a weight of one is called ordinary.

47.2.2 Reachability Analysis and Reachability Coverability Analysis

Reachability Analysis Reachability is a fundamental concept in analyzing a model's behavior of a model. When a firing sequence of transitions causes a transition from marking M_n to marking M_m , we say that M_m is reachable from M_n . The reachability analysis in Petri nets verifies that M_m is reachable from M_n . There are two main methods of verification: generating a state transition diagram and determining whether the reachability condition is satisfied.

Reachability Coverability Analysis The covering property is one of the Petri nets' dynamic properties. A marking M in a Petri net (N, M_0) is coverable if there is a marking M_1 in $R(M_0)$ for every place p in the net, such that $M_0(p) \geq M(p)$. $M(p)$ represents the number of markings. By firing a fireable transition from the initial M_0 in the Petri net (N, M_0) , we can obtain the same number of "new" achievable markings as the fireable transitions. From each new marking, another new reachable marking can be obtained. This process results in a tree representation of the markings. The nodes represent the initial marking M_0 (root) and the markings generated from it (descendants), and each arc represents the firing of a transition. However, if the net is unbounded, the traditional covered tree representation grows infinitely large. To keep the tree finite, we introduce a special symbol, ω . "infinity" is represented by ω . Here, ω has the properties $\omega > n$,

$\omega \pm n = \omega$, and $\omega \geq \omega$ for all integers n . Reachability analysis is a method for obtaining reachability set by sequentially generating markings. However, if we perform reachability analysis on an unbounded net, the reachable tree becomes infinite in size. In the covered analysis, the size is made finite even for unbounded nets by introducing the symbol ω . The algorithm for constructing covered trees was presented by Karp and Miller [6].

To solve the information gap in the covering analysis method (ω covering), the ECG method is proposed [7]. To determine the increase or decrease of the covered area, the infinite state ω used in the covering graph is extended to three values of the number of tokens in a place: increasing (N_u), constant (N_c), and decreasing (N_d). However, the step width of the increase and decrease in the number of tokens is 1. While the conventional covering graph uses N , the extended covering graph uses N and is denoted as $N *$ (or $N *$ transition). If there is an arc that transitions to N with a decreasing number of tokens, then Generate a branching arc with a condition. The transition destination of the branched arc is the bounded part in contrast to the unbounded $N *$. In conventional covering graphs, if more than one covering is possible, ω is substituted. To clarify the boundary between bounded and unbounded parts covered by $N *$ in the extended covering graph, a token with 0, 1 is considered bounded if it can be covered by $N *$ and two or more tokens

47.3 Submarking and Qusai-home State Proposals

47.3.1 Qusai-home State

Home State Sometimes it is not necessary to return to the initial state as long as it is possible to return to a (home) state. The home state is defined as follows after relaxing the reversibility condition. For all markings M in (N, M_0) , if the marking M' is reachable from M , then M' is called the home state. A net with home state M' is said to be reversible if M' is M_0 under the limited condition that M' is M_0 . The home state is a state in which one can return from all markings. By determining the home state, we can check the system's stable state.

Qusai-home State When performing dynamic analysis on a Petri net with $k \in \{1, 2, \dots, n\}$ places, the marking is obtained every time a transition is fired. In some systems, it is unnecessary to closely monitor the increase or decrease of tokens for the number of tokens $M(p_k)$ that a place p_k has.

There are also cases where it is unnecessary to consider the number of tokens held by a place p_k . The quasi-home state is defined as the marking M' is reachable from M'' for a marking M'' that abstracts the number of tokens in a place. The number of tokens a place has We define four abstractions, which are shown in the next section.

47.3.2 Submarking

The submarking method is a method for abstracting marking. Submarking helps to compress the state space by removing unnecessary information during analysis. The abstracted markings can be given meaning and considered from various perspectives when analyzing process graphs. In this study, we define four submarking methods based on the abstraction of marking presented above.

Submarking method 1: Remove the marking $M(p_i)$ of place p_i from all markings $M(p_i)$, $k \in \{1, 2, \dots, n\}$. Given a set of places P and a set of places to be removed P' , generate a state space for the set of places p_d (the set of places from P minus P') such that $p_d = P - P'$.

Submarking method 2: When there exists $\mu(p_i)$ such that $\mu(p_i)$ satisfies the closed interval $[a, b]$ in the token number $\mu(p_i)$ of place p_i . When $\mu(p_i) = D_i$, where a and b are natural numbers including 0, $a \leq b$. Since the tokens are non-negative integers, the union set of all ranges must be $[0, \infty)$. Therefore, the lower end of the next range of the closed interval $[a, b]$ is $b + 1$. In the case of cover analysis or extended cover analysis, the upper end includes ∞ .

Submarking method 3: When there exists $\mu(p_i)$ such that $\log 10^n \leq \mu(p_i) < \log 10^{n+1}$, ($n \geq 0$) in the number of tokens $\mu(p_i)$ that place p_i has. If there exists $\mu(p_i)$ such that ($n \geq 0$) is satisfied, then $\mu(p_i) = D'_i$. However, when $\mu(p_i) = 0$, no state value transformation is performed. In the case of cover analysis or extended cover analysis, the upper end includes ∞ .

Submarking method 4: Exclude from all markings $M(p_k)$ the markings $M(p_i)$ of the unbounded place p_i . Generate a state space for the place set (P minus P'') such that $p_{ub} = P - P''$ when P is the set of places and P'' is the set of places to be removed.

47.4 Application Examples

An application example is illustrated below. The vending machine's model diagram and state space are depicted in Fig. 47.1 and Table 47.1, respectively. In Fig. 47.1, p_4 represents the inventory, where $\mu(p_4) = 1000$.

Fig. 47.1 Petri Net model of vending machine (Inventory 1000)

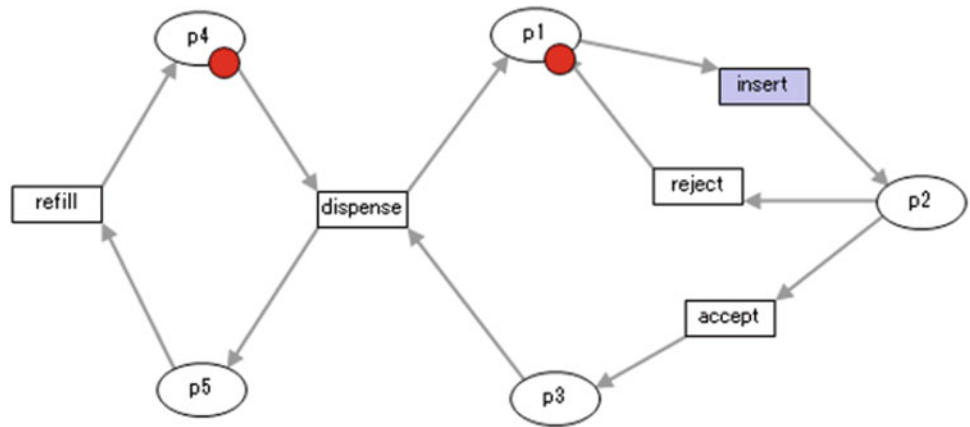


Table 47.1 State space in Fig. 47.1

Marking	Place (p_1, p_3, p_4, p_5)
M_0	(1, 0, 1000, 0)
M_1	(0, 1, 1000, 0)
M_2	(1, 0, 999, 1)
...	
M_{100}	(1, 0, 645, 355)
M_{101}	(0, 1, 645, 355)
M_{102}	(1, 0, 644, 356)
...	
M_{1999}	(0, 1, 1, 999)
M_{2000}	(1, 0, 0, 1000)
M_{2001}	(0, 1, 0, 1000)

Table 47.2 State space when submarking method 1 is applied to Fig. 47.1

Marking	Place (p_1, p_3, p_4, p_5)
M_0	(1, 0, 1000, 0)
M_1	(0, 1, 1000, 0)
M_2	(1, 0, 999, 1)
...	
M_{100}	(1, 0, 645, 355)
M_{101}	(0, 1, 645, 355)
M_{102}	(1, 0, 644, 356)
...	
M_{1999}	(0, 1, 1, 999)
M_{2000}	(1, 0, 0, 1000)
M_{2001}	(0, 1, 0, 1000)

47.4.1 Application Examples 1 (Removing the Insert-Reject Loop)

Table 47.2 shows the state space when submarking method 1 is applied to the place p_2 in the net of Fig. 47.1. By excluding p_2 from the net diagram in Fig. 47.1, p_2 is considered to be the same regardless of the number of tokens, and thus the state space is the state space of the four places excluding p_2 . There are 3003 states without submarking and 2002 states with submarking. The insert-reject closure in the net of Fig. 47.1 is a livelock. If all markings are considered to be in the home state, markings in the livelock state are also considered to be in the home state. However, in the analysis of the home state of a vending machine system, the inventory and sales information are the most important, so by excluding p_2 , the insert-reject closed circuit that is a live lock is not represented, and the number of states is reduced. In this way, the state space can be compressed by excluding unnecessary information from the system during analysis (Table 47.3).

Table 47.3 Compare of the number of states with the net applying the submarking 1 method 1 in Fig. 47.1

	Vending machine	Vending machine with submarking method 1 applied
Number of states	3003	2002

47.4.2 Application Examples 1 (Abstracting by Inventory Count)

Table 47.4 shows the state space when submarking method 2 is applied to the place p_4 in the net of Fig. 47.1. The range of the number of tokens that p_4 has is A:0, B:1–50, C:51–100, D:101–500, and E:501–1000. The model diagram of Fig. 47.1 has a closed circuit consisting of p_4 and p_5 , and the sum of the number of tokens in the closed path is retained. The sum of the number of tokens in the closed circuit is retained. Therefore, when p_4 is abstracted, p_5 can also be abstracted. In the vending machine model (Fig. 47.1), p_4 and p_5 represent the inventory and sales of the vending machine,

Table 47.4 State space when submarking method 1 is applied to Fig. 47.1

Marking	Place (p_1, p_2, p_3, p_4, p_5)
M_0	(1, 0, 0, E, 1000-E)
M_1	(0, 1, 0, E, 1000-E)
M_2	(0, 0, 1, E, 1000-E)
M_3	(1, 0, 0, D, 1000-D)
M_4	(0, 1, 0, D, 1000-D)
M_5	(0, 0, 1, D, 1000-D)
M_6	(1, 0, 0, C, 1000-C)
M_7	(0, 1, 0, C, 1000-C)
M_8	(0, 0, 1, C, 1000-C)
M_9	(1, 0, 0, B, 1000-B)
M_{10}	(0, 1, 0, B, 1000-B)
M_{11}	(0, 0, 1, B, 1000-B)
M_{12}	(1, 0, 0, A, 1000-A)
M_{13}	(0, 1, 0, A, 1000-A)
M_{14}	(0, 0, 1, A, 1000-A)

Table 47.5 Compare of the number of states with the net applying the submarking 1 method 2 in Fig. 47.1

	Vending machine	Vending machine with submarking method 2 applied
Number of states	3003	15

respectively. By giving a range to the number of tokens in p_4 , it is possible to classify the state as A: no inventory, B: needs replenishment, C: does not need replenishment. The number of states decreases with the number of state values given, compared to the conventional state space. In the vending machine's model in Fig. 47.1, it is better to know roughly whether the machine is sold out or has more than half of its stock remaining, rather than how many are in stock. Submarking method 2 allows us to comprehend the system's general state, rather than the transition of each token (Table 47.5).

47.4.3 Application Examples 3 (Non-bounded Place Abstraction)

Below is an example of the Petri net model (Fig. 47.2) and the state space for Fig. 47.2 (Table 47.6). The state space (Table 47.7) for this net when the submarking method 4 is applied is presented below. As a preprocessing step for the quasi-home state determination, a boundedness determination is conducted. If the net is unbounded, the unbounded place p_i is obtained. In submarking method 4, all markings of the obtained place p_i in the state space are excluded.

The net in Fig. 47.2 is a model diagram of a producer/-consumer system. The $[p_0, p_1]$ closed circuit represent producers. The $[p_2, p_3]$ closed circuit, $[p_4, p_5]$ closed circuit, and

Table 47.6 State space pf the Fig. 47.2

Marking	Place ($p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8$)
M_0	(0, 1, 0, 1, 0, 1, 0, 1, 0)
M_1	(1, 0, 0, 1, 0, 1, 0, 1, 0)
M_2	(0, 1, 0, 1, 0, 1, 0, 1, ω)
M_3	(1, 0, 0, 1, 0, 1, 0, 1, ω)
M_4	(1, 0, 1, 0, 0, 1, 0, 1, ω)
M_5	(0, 1, 1, 0, 0, 1, 0, 1, ω)
M_6	(0, 1, 1, 0, 1, 0, 0, 1, ω)
M_7	(1, 0, 1, 0, 1, 0, 0, 1, ω)
M_8	(1, 0, 0, 1, 1, 0, 0, 1, ω)
M_9	(0, 1, 0, 1, 1, 0, 0, 1, ω)
M_{10}	(0, 1, 0, 1, 1, 0, 1, 0, ω)
M_{11}	(1, 0, 0, 1, 1, 0, 1, 0, ω)
M_{12}	(1, 0, 0, 1, 0, 1, 1, 0, ω)
M_{13}	(0, 1, 0, 1, 0, 1, 1, 0, ω)
M_{14}	(0, 1, 1, 0, 0, 1, 1, 0, ω)
M_{15}	(1, 0, 1, 0, 0, 1, 1, 0, ω)

Table 47.7 State space when submarking method 4 is applied to Fig. 47.2

Marking	Place ($p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7$)
M_0	(0, 1, 0, 1, 0, 1, 0, 1)
M_1	(1, 0, 0, 1, 0, 1, 0, 1)
M_2	(1, 0, 1, 0, 0, 1, 0, 1)
M_3	(0, 1, 1, 0, 0, 1, 0, 1)
M_4	(0, 1, 1, 0, 1, 0, 0, 1)
M_5	(1, 0, 1, 0, 1, 0, 0, 1)
M_6	(1, 0, 0, 1, 1, 0, 0, 1)
M_7	(0, 1, 0, 1, 1, 0, 0, 1)
M_8	(0, 1, 0, 1, 1, 0, 1, 0)
M_9	(1, 0, 0, 1, 1, 0, 1, 0)
M_{10}	(1, 0, 0, 1, 0, 1, 1, 0)
M_{11}	(0, 1, 0, 1, 0, 1, 1, 0)
M_{12}	(0, 1, 1, 0, 0, 1, 1, 0)
M_{13}	(1, 0, 1, 0, 0, 1, 1, 0)

$[p_6, p_7]$ closed circuit represent consumers and p_8 represents a buffer, respectively. Figure 47.2 shows that the place p_8 is unbounded. Excluding all markings of p_8 in the state space (Table 47.6) generates a state space of eight places ($p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7$). By performing a covering analysis on this state space, identical markings are covered and the state space shown in Table 47.7 is obtained. It is possible to perform home state analysis in the bounded part for an unbounded net.

47.5 Performance Comparison

In this study, we defined the submarking method along with the quasi-home state determination. We confirmed that the

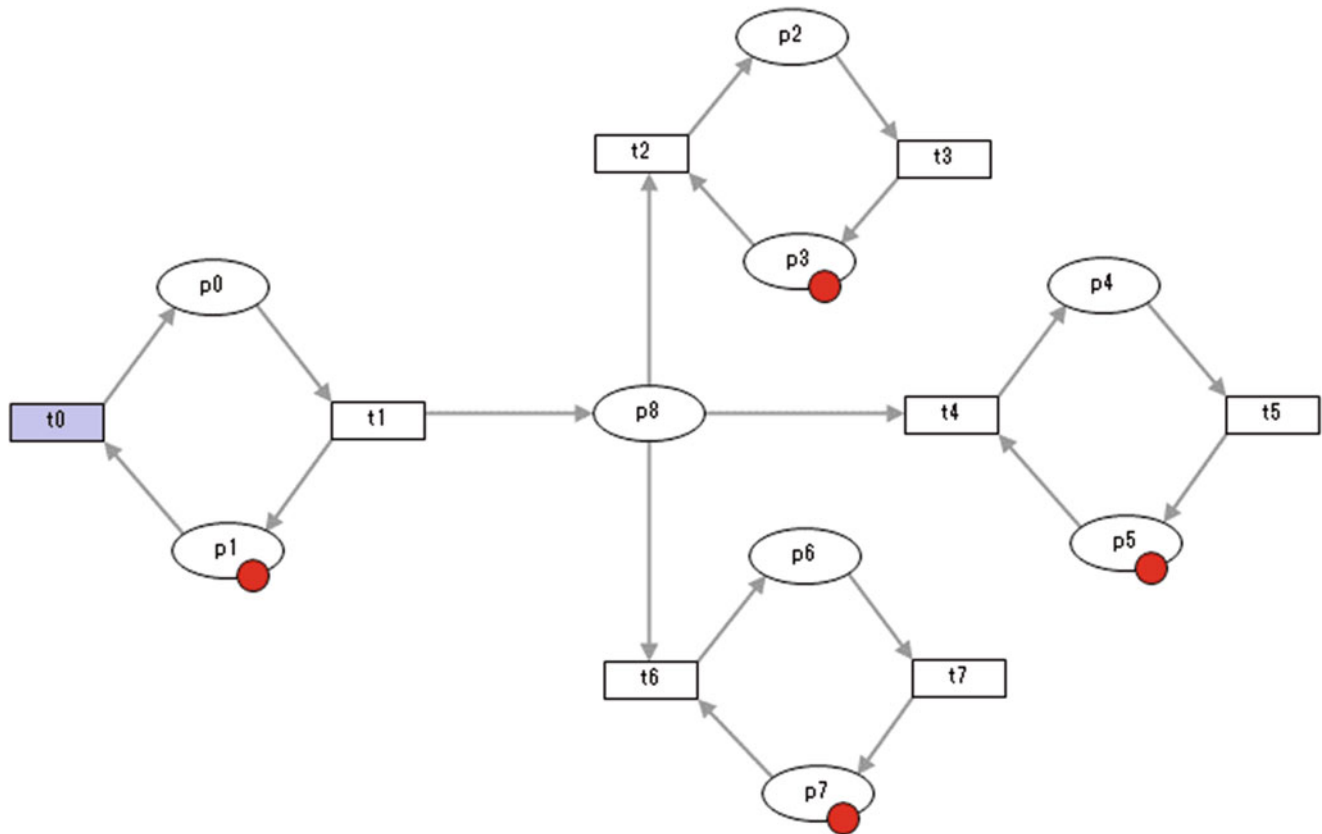


Fig. 47.2 Model diagram of a producer/consumer system

submarking method compresses the state space. Not removable at the point in time represented by Fig. 47.1.

In future research, we will implement a quasi-home state determination function. There are many home states in a single net. Therefore, it is necessary to consider an efficient algorithm to determine the home state and its representation approach. As a preprocessing step for determining the home state, we perform structural analysis to identify possible candidates for the home state and then perform dynamic analysis to improve the analysis' efficiency.

Another issue is the compression of state space. Currently, only markings are shown in the state space. The state space can be made more understandable by adding a representation of the home state and information on the firing of transitions on the state space.

Acknowledgment This work was supported by JSPS KAKENHI Grant Number 19K11821.

References

1. T. Murata, Petri nets: Properties, analysis and applications. *Proc. IEEE* **77**(4), 541–580 (1989)
2. C.A. Petri, W. Reisig, Petri net. *Scholarpedia* **3**(4), 6477 (2008)
3. HiPS: Hierarchical Petri net Simulator, Shinshu University. Available at <https://sourceforge.net/projects/hips-tools/>
4. Y. Harie, K. Wasaki, A Petri Net design and verification platform based on the scalable and parallel architecture: HiPS, in *Proceedings of the 14th International Conference on Information Technology – New Generations (ITNG2017)*, Advances in Intelligent Systems and Computing, vol. 558 (Springer, 2017), pp. 265–273
5. Y. Harie, Y. Mitsui, K. Fujimori, A. Batajoo, K. Wasaki, HiPS: Hierarchical Petri Net design, simulation, verification and model checking tool, in *Proceedings of the 6th IEEE Global Conference on Consumer Electronics (GCCE 2017)*, pp. 686–690
6. R.M. Karp, R.E. Miller, Parallel program schemata. *J. Comput. Syst. Sci.* **3**(2), 147–195 (1969)
7. Y. Mitsui, Y. Harie, K. Wasaki, Implementation of L2/L3-liveness analyzer using the extended coverability graph to Petri Net Tool: HiPS, in *Proceedings of the 16th IPSJ Forum on Information Technology Conference (FIT 2017)*, pp. 91–94

Index

- A**
- Abstractive text summarization
 - algorithm implementation, 274
 - bi-directional LSTM, 273
 - data cleaning, 271
 - GloVe word embeddings, 270
 - using NLTK, 272, 273
 - overview, 277
 - ProductReviewHeader, 271
 - ProductReviewText, 271
 - Sequence-to-Sequence modeling (Seq2Seq model), 270, 272, 274
 - training decoder, 272, 273
 - training encoder, 272, 273
 - Abstract syntax tree (AST), 6, 30
 - Accelerometer, 141, 144
 - Access control, 176–177
 - ACCESS/CPN, 63
 - ACCU-CHEK Spirit, 58, 60, 61
 - Advanced Encryption Standard (AES) algorithm
 - AES-NI, 151
 - algorithm, 149
 - definition, 148
 - vs. Rijndael256, 149
 - SSO, 196
 - vector AES-NI, 151
 - AESDEC, 148
 - AESDECLAST, 148
 - AESENC, 148–153
 - AESENCLAST, 148, 151
 - AESIMC, 148
 - AESKEYGENASSIST, 148
 - AES-NI, 148
 - Affective device, 373–374
 - Amalthea model, 47, 48
 - analyzeNode function, 8
 - Android Debug Bridge (ADB) tool, 138
 - Android's Sensor Manager, 138
 - Android WearOS, 138
 - Anti-Phishing protection system, 165
 - Anti-Phishing tools, 162, 165, 167
 - API-First Design, 73–74
 - anticipation of, 77
 - industry and grey literature, 76–77
 - infancy of, 77
 - modern API, 74
 - modern architectures survey, 78
 - state of academia, 75–76
 - Application Programming Interface (API), 74
 - APP4MC, 48
 - Apps from recent online mobile stores, 129–130
 - ArthroSim, 18
 - Artificial intelligence (AI), 156, 165
 - ASK-CTL model, 60
 - Assurance Case Exchange Standard (ACES), 58
 - Audio-visual contract, 372
 - Authentication system
 - biometric-based authentication, 195–196, 199
 - description, 195
 - proof of identity, 195
 - Automated Individual White-List, 163
 - Automation systems, 11
 - conceptual communication infrastructure, 13–14
 - framework, technologies comprising
 - OPC UA, 12–13
 - software development, languages and models, 13
 - time-sensitive networking, 12
 - holistic modeling approach, 14–15
 - Avatar eyeball movement, 22
 - Avatar lip tracking, 22
 - Avatars, 20–21
 - Avira, 165
- B**
- Backend as a Service (BaaS) platform, 76
 - Backpropagation Through Time algorithm (BPTT), 258, 259
 - Backup of General Product Owner (Bkp of GPO), 220
 - Backup of Team Scrum Master (Bkp of TSM), 220
 - Bagging, 163
 - BaitAlarm, 163
 - Bar chart, 297, 298
 - Behavior change detection, 121
 - Bidirectional Long-Term Short-Term Memory (Bidirectional LSTM), 240, 242
 - test data accuracy graph, 243
 - test data los-loss graph, 243
 - BigchainDB, 223
 - Big Data, 219, 220, 223, 224, 227
 - Big data analysis communities, 379
 - Biomatrix corporate structure, 369–370
 - Biometric-based authentication, 195–196, 199
 - Biometrics, 137
 - data, 196, 197
 - trait, 195, 196

- BitDefender, 165–166
- Blended graphical–textual modeling, 40
- Blended modeling, 39–46
 - framework
 - editor generation, 42
 - graphical and textual syntaxes, generation of, 41
 - mapping and synchronization, 41
- Blender software, 212
- Blockchain, 66
- Blockchain Based Trust Management (BBTS), 172
- Blockchain technology
 - application, 183
 - blocks, 180
 - BurstIQ, 182
 - EHR (*see* Electronic Health Record (EHR))
 - fogs, 175, 176
 - hash value of parent block, 172
 - in healthcare sector, 183
 - HealthNet, 182
 - ID management system, 173
 - IoT (*see* Internet of Things (IoT))
 - IoT passport, 174
 - manufacturing zones, 173
 - RIM, 172
 - RSUs, 173
 - security enhancement and privacy protection in healthcare (*see* Systematic literature review (SLR))
 - software industry, 184
 - SSO (*see* Single sign-on (SSO) fingerprint authentication)
 - structure of, 176
 - systematic mapping study, 187
 - TERMs, 173
 - trust approaches in IoT
 - assessment, 175–176
 - features, challenges, and limitations of, 177
 - trusted zones, 173
- Blocking, 108, 110
- Bluetooth, 138
- Boolean operators, 190
- Bootstrap technology, 223
- Boundary approximation, 379–385
- Bounded Estimation (BE), 358
- Burndown Charts metrics, 220
- BurstIQ, 182

- C**
- CalibratedClassifierCV, 231
- Campus Virtual, 130
- Cancer Genomic Atlas, 230
- Carrefour, 367
- C-chain approximation, 383
- CC-MR transitive closure using, 112
- CD-Adapco, 36
- CE-240 Database Systems Project, 226
- CE-245 Information Technologies, 226
- Centers of Academic Excellence in Cyber Defense Education (CAE-CDE) designation program, 156
- CE-229 Software Testing, 219
- Change Data Capture (CDC) techniques, 96
- Change detection, 119, 124
- Chrome, 25, 26
- C# & JSON-LD symantic tools, 176
- CLEAN MX, 165
- Clipboard, 21
- CliqueRank algorithm, 109
- Clustering process, 120
- CNN architecture, 164
- Codebases, 32
- Code of Ethics Canons, 156
- Collaborative Intrusion Detection system (CIDS), 163
- Coloured Petri Nets (CPN), 57
- Command-line tool, 30
- Component-Based Software Engineering (CBSE), 13
- Computational fluid dynamics (CFD), 35, 36
 - free energy, 37
- Computational geometry, 381
- Computer graphics, 379
- Concave chain, 381
- Conflict management, 324–326
 - functionality, 326
 - microservice, 323
- Consistency, 86
- Constrained Probabilistic Matrix Factorization (CPMF), 264, 292
- Containerization, 323
- Continuous user verification, 137
- Conventional Neural Network Phishing Detection (CNNPD), 164
- Corel Health, 182
- Corporate environment, software engineering and
 - proof of concept, 69–70
 - quantitative outcomes, 70
 - students' feedback, 70–71
 - studies, 75
 - tailoring scrum and IPBL, 66–69
 - scrum artifacts academically used, 67–68
 - scrum roles academically used, 67
 - STEPES-BD scrum ceremonies, 68–69
- Course learning outcomes, 158
- Course topics, 158
- COVID-19 pandemic, 166
 - information management system, 220, 224–226
- Credibility, 86
- Credit-based shaper (CBS) algorithm, 12
- Cross-Blockchain platform, 181
- Cross-platform blended modelling, 3
 - bridging MPS and EMF, 5–6
 - automatic export, 6–8
 - EcoreLanguage, 6
 - evaluation, 9
 - motivation and contribution, 4
 - studies, 4–5
- CSIFill
 - Bounded Estimation (BE), 358
 - configuration for bounded estimation, 354
 - configuration for unbounded estimation, 354
 - data collection and pre-processing, 355–356
 - fundamental channel model, 352
 - Levenberg-Marquardt backpropagation (LM), 353
 - LTE and cellular networks, 352
 - MSE performance
 - bounded estimation accuracy, 355, 357
 - unbounded estimation accuracy, 357–358
 - network configuration, 354
 - shallow neural networks, 353, 354
 - testing network, 355
 - training algorithms, 353–354
 - training process, 354–355
- CSS3 stack, 26
- C-style pseudocode, 149, 150
- Cumulative Sum Approach (CUSUM), 121

- Curriculum proposal
 - assessment methods, 158
 - course description, 158
 - course learning outcomes, 158
 - course topics, 158
 - rationale, 157
- Cybersecurity ethics education
 - curriculum proposal, 157–158
 - ethical challenges, 155, 157
 - ethical dilemmas, 156–158
 - ethical issues, 156, 158
 - mentoring, 156–158
- Cybersecurity Ethics (CSE) knowledge unit, 158

- D**
- Data classification, 121
- Data interoperability, 183–182, 184
- Data modeling, 180–182, 184
- Data propagation, 342
- Data security, 182
- Data storage techniques, 182, 184
- Data streams, 119–121, 124, 296, 300
- Data Visualization microservice, 323
- Data warehouse (DW), 91
- Data Washing Machine (DWM), 108
- Decentralized software applications (DApps), 182, 183
- Decision support systems (DSS), 363
- Deep learning neural network models
 - BERT Augmentation, 240
 - Bi-LSTM, 240
 - dropout function, 239
 - text augmentation, 240
- Defect discovery, 27
- Defect injection, 27
- Department of Homeland Security (DHS), 156
- Digital assets, 223
- Digitalization, 363
- Digital-Physical environment, 364
- DIMMER, 322, 327
- Directional selectivity, 205, 206
- Distance learning, 129
- Distributed databases, 181
- Distributed ledger technology, 196, 198
- DNS phishing attack, 164
- Document Object Model (DOM), 25
- Domain-Driven Design (DDD), 75
- Domain Specific Language (DSL), 40
- Domain-specific modeling languages (DSMLs), 3, 39
- Drift Detection Method (DDM), 121
- Dual-tree complex wavelet packet transform, 206–207
- Dynamic Security Skin, 163
- Dynamic Time Warping, 138

- preprocessing for endoscopic image
 - block division for endoscopic image, 210
 - image after preprocessing, 208
- proposed detection method, 208
- 2D-CWPT (*see* 2D complex discrete wavelet packet transform (2D-CWPT))
- Eclipse Modeling Framework (EMF), 4
 - bridging MPS and, 5–6
 - automatic export, 6–8
 - EcoreLanguage, 6
- E-commerce, 366, 367
- EcoreEMF* metamodels, 5
- Ecore language, 4, 6, 9
- Educational systems
 - e-learning (*see* E-learning systems)
 - inclusion and exclusion criteria, 132
 - m-learning, 129–132
- eHealth, *see* Electronic Health Record (EHR)
- E-learning systems
 - accessibility assessment, 129
 - apps from recent online mobile stores, 129–130
 - standalone operating system apps, 130
 - use of, 137
- Electronic CodeBook (ECB), 148, 151, 152
- Electronic Control Units (ECUs), 47
- Electronic Death Certificate, 225, 226
- Electronic Health Record (EHR)
 - academia and software industry, 183, 184
 - for administrative reasons, 180
 - analysis of investigated works by area, 183
 - challenges, 183
 - Corel Health, 182
 - COVID-19 pandemic, 220, 224
 - DApps, 182, 183
 - DBMS, 183, 184
 - development and validation, 184
 - e-patient, 180
 - field of research, 183
 - futurework, 184
 - HL7 standard, 181
 - Medicalchain, 182
 - MediLedger, 182
 - OpenEHR standard, 181
 - PoB, 180
 - PoS, 180
 - PoW, 180
- Electronic shopping, 165
- Email spoofing, 163
- Embedding layer network
 - augmented text model, 244
 - deep learning neural network models, 239
 - Keras' to_categorical, 242
 - Panda's get_dummies, 242
 - test data accuracy graph, 243
 - test data los-loss graph, 243
- Emotion perception, 372
- Endoscopy, 207, 208
- e-patient, 180
- Ethereum network, 196–198
- Ethical challenges, 155, 157
- Ethical dilemmas, 156–158
- Ethical issues, 156, 158
- Ethical mentoring, 158
- Express technology, 223
- Extended Backus–Naur form (EBNF) grammar, 40

Extended cover analysis (ECG), 395
 External V-aware approximation problem (EAP), 381
 Extraction (E), 92, 93
 Extractive text summarization, 269, 270
 algorithm implementation, 271
 data cleaning, 271
 TextRank algorithm, 269–275
 Extract, Transform, and Load (ETL) process, 91–96
 extraction (E), 93
 loading (L), 93
 methodology, 93
 studies, 92–93
 transformation (T), 93

F

Facebook, 130
 Face-to-face attendance, 224, 225
 False Negative (FN), 123
 False Positive (FP), 123
 Federal High-way Administration (FHWA),
 312
 Fictitious Death Certificate, 225, 226
 Fictitious Medical Record, 225
 FileNotFound, 31
 Film audio professionals, 371
 Filmmaking process, 372
 Fine-motor gestures, 137
 Fingerprints, 196–198
 Firefox, 25
 First system refinement, 58–59
 Fogs, 175–176
 Fortinet, 165
 Fossil Gen 4 Carlyle, 138
 Free energy, 37
 Function Block Diagram (FBD), 13

G

Galvanic Vestibular Stimulation system, 18
 Gen EcoreEMF, 6
 General Product Owner (GPO), 220
 Geographical information system (GIS), 379
 Gesture-based recognition systems, 137, 138
 GitHub, 31, 153
 Goal Question Metric (GQM) approach, 84
 Google Classroom, 84
 Google Cloud platform, 222
 Google Collaboratory environment, 164
 Google Meet, 84
 Google Safe Browsing API, 163
 Grading GRADE (Grading of Recommendations Assessment
 Development and Evaluation), 192
 Grafana application, 224
 Graph-based hierarchical record clustering, 110–114
 adapted CC-MR, transitive closure using, 111–112
 adapted Louvain, modularity optimization using, 112–113
 dataset, 114
 evaluation, 115–116
 experimental setup, 114–115
 Graph Database Management Systems (GDBMS), 98
 Graphical Modeling Framework (GMF), 40
 Graphical notations, 42
 Gueron's implementation, 149
 Gyroscope, 138, 140–141, 144

H

Handlebars technology, 223
 Haptic feedback, 287
 Hardware module, 59
 Head mounted displays (HMD), 18
 Health applications, 182
 HealthNet, 182
 Health science, 211–215
 HERIDAL dataset, 278
 Heroku, 222
 Heterogeneity, 338
 Heuristic evaluation criteria, 131
 Hierarchical Petri net Simulator (HiPS), 393
 Higher Order Transformations (HOTs), 45
 High performance computing (HPC) resources, 322
 Highway Economic Requirements System (HERS) model, 311
 Holistic modeling approach, 14–15
 Horizontal scalability, 97
 HTML5-JavaScript-CSS3 stack, 28
 Hybrid graph-based entity resolution, 109
 Hyper Famili, 363
 comparative study in, 368–369
 inventory management model, 367
 organizational structure, 367–368
 Hyper Famili Incorporation, 364
 Hyperledger Fabric tool, 176

I

IceLake (ICL), 150
 Identity provider (IdP), 196
 idleSlope, 12
 I-I algorithm, 380
 Illegal values, 86
 Imagery microservice, 324
 Immersive learning
 advanced analytics, 366
 increased engagement, 365–366
 learning through mistakes, 365
 mirror real-life situation, 365
 reduced operational costs, 365
 Immunotherapy, 230
 Industrial Internet of Things (IIoT) paradigm, 11
 Industry 4.0 revolution, 364–365
 InfluxDB, 222, 223
 Information and communication technologies (ICT), 156, 179, 364
 Information Systems Audit and Control Association (ISACA), 156,
 157
 Information Technology (IT) domain, 13
 Initial dependency model, 28
 Instruction List (IL), 13
 Insulin Infusion Pump Systems (IIPS), 57
 empirical evaluation
 analysis, 61–63
 procedure and measures, 61
 scoping, modelers and variables, 60
 model-based approach, 58
 first system refinement, 58–59
 hardware and software decomposition, 58
 second system refinement, 59–60
 quality assessment scenarios, 60
 first refinement, verification of, 60
 second refinement, verification of, 60
 web-based application, 63–64
 Integrated Development Environment (IDE), 30

- Integrity, 84
 - Intelligent Transportation Systems, 342
 - Intel® Turbo Boost Technology, 151
 - Interdisciplinary Problem-Based Learning (IPBL) approach, 65–67, 219
 - scrum artifacts academically used, 67–68
 - scrum roles academically used, 67
 - STEPES-BD scrum ceremonies, 68–69
 - International Electrotechnical Commission (IEC), 13
 - International Information System Security Certification Consortium (ISC), 156
 - Internet Explorer (IE) Test Suite, 31
 - Internet of Things (IoT)
 - aim of, 171
 - application domains, 337
 - applications, 339
 - architecture for, 173
 - blockchain based trust approaches in
 - assessment of, 175–176
 - features, challenges, and limitations of, 177
 - digital interconnection, 333
 - heterogeneity levels, 338
 - interoperability in, 338, 339
 - intrinsic features of, 171
 - MASs, 173
 - ontology, 339
 - platforms, 337
 - public blockchain, 175
 - semantic models for, 339
 - taxonomy scheme
 - for classifications and analysis of existing solutions, 174
 - feasibility, 175
 - features, 175
 - limitations, 175
 - security function, 175
 - suitability, 175
 - trust, 171, 172
 - ubiquitous computing, 333
 - web of things (WoT), 333
 - Internet of Vehicles (IoV), 172–173
 - Interoperability, 190, 335
 - InterPlanetary File system (IPFS), 196
 - Interpretability, 84
 - Interviews, 372
 - IoT passport, 174
 - Ivy Bridge (IVB), 150–151
- J**
- Jar file, 197
 - JavaScript, 25
 - JavaScript Object Notation (JSON), 28, 30, 221
 - Java Tomcat REST Server, 197
 - Java web API, 197
 - JetBrains MPS, 4
 - and EMF, 5–6
 - automatic export, 6–8
 - EcoreLanguage, 6
 - Jupyter Framework Architecture, 164
 - Jupyter Notebook, 231
- K**
- Kahoot quiz, 213
 - Kaspersky, 165
 - Kaspersky Internet Security (KIS), 165
 - Kernelized Probabilistic Matrix Factorization (KPMF), 264–266
 - Keyless Signature Infrastructure (KSI), 181
 - Keystroke dynamics (KD), 137
 - K-Nearest-Neighbor (KNN), 122–124, 126, 230, 232
 - Knowledge, skills, and abilities (KSAs), 156
 - Kroger, 367
- L**
- Ladder Diagram (LD), 13
 - Language engineering, 3
 - Language workbenches, 3, 4
 - Large-scale teaching methods, 129
 - Latest processor generation (ICL), 153
 - LDAT home page, 297
 - Leap motion controllers, 137
 - Learning Management Systems (LMS)
 - effective use, 131
 - evaluation, 130
 - Facebook, 130
 - usability features, 130, 131
 - Lenovo T430s Laptop, 150
 - Levenberg-Marquardt backpropagation (LM), 353
 - LiDAR sensors, 293
 - Light Detection and Ranging (LiDAR), 293
 - box plot, 300, 301
 - data analysis, 293, 294
 - definition of, 293
 - prototype
 - bar chart, 297, 298
 - data visualization dashboard, 297
 - line graph, 297, 298
 - 3D point cloud mesh, 297–299
 - user interface, 296–297
 - software design
 - data workflow, 295–296
 - design application, 296
 - high level diagram, 295
 - LiDAR sensors, 294
 - system architecture, 294–295
 - system level diagram, 295
 - website map, 296
 - terrestrial LiDAR, 294
 - 3DSYSTEM viewer, 294
 - user performance, 300
 - user study, 299–300
 - change color, 299
 - change point size, 299
 - change sensor location, 299
 - fill out form, 299
 - telecommunication option, 299
 - William Pennington Engineering Building (WPEB), 299
 - visualization, 294
 - Line graph, 297, 298
 - Linux OS, 151
 - Loading (L), 91, 93
 - Louvain method, 112–113, 116
 - Lung adenocarcinoma (LUAD), 230
 - Lung cancer, 230
 - Lung Squamous Cell Carcinoma (LUSC), 230
 - Lymphoma cancer, 230

M

- Machine learning algorithms
 - Adam optimizer, 260
 - backpropagation algorithm, 258
 - bidirectional LSTM network architecture, 260
 - Exploratory Data Analysis, 240
 - GloVe.840.300D embeddings, 260
 - logistic regression, 257
 - Naïve Bayes classifier, 257
 - NLP techniques, 240
 - proposed technique, 242
 - Random Forest classifier model, 241, 243
 - recurrent neural networks (RNNs), 258
 - Stacking classifier model, 241
 - support vector machine (SVM), 258
 - viable approach, 256
 - word count vector, 257
 - word embeddings, 257
 - Machine Learning and High Performance Computing, 321
 - Machine learning, cancer dataset for gene mutation based treatment
 - comparison of log-loss performance, 233
 - future work, 232
 - gene types, 231
 - KNN, 232
 - LUAD, 230
 - lung cancer diagnosis, 229, 230
 - LUSC, 230
 - Naïve Bayes classifier, 232
 - one-hot encoding, 231
 - Random Forest Classifier, 232
 - response encoded table, 231
 - Stacking Classifier, 232
 - SVM, 232
 - system architecture, 231
 - Malicious website detection, 163, 165, 166
 - Manson's Ford Galaxie, 376
 - Manual spot entering, 341
 - Manufacturing execution systems (MES), 11
 - MapReduce, 111
 - Markov Chain (MC), 120
 - Massive Online Analysis (MOA), 122
 - MATLAB, 141
 - Medicalchain, 182
 - Medical education, 211, 212
 - Medical training, 211–215
 - MediLedger, 182
 - Memorial Sloan Kettering Cancer Center (MSKCC), 230, 238
 - Mentoring, 156–158
 - Meta-modeling, 40
 - Micro-Clustering DBScan (M-DBScan)
 - behavior change detection, 120–121
 - novelty detection, 120
 - Microservice Architecture (MSA), 75, 78
 - Microservice-based Envirosensing Support Applications (MESA), 321
 - conflict management, 324–326
 - DIMMER smart city platform, 322
 - functionality, 321
 - generic service infrastructure for PEIS, 322
 - microservice architecture, 322
 - near real-time autonomous quality control, 326
 - Nevada Research Data Center, 322
 - NRDC quality assurance application, 323–324
 - OceanTEA, 323
 - SEPHAS Lysimeter Visualization, 323
 - software specification
 - high level design, 323
 - technology utilized, 323
 - Microsoft SQL Server (MSSQL), 323
 - Micro-Web-Services, 322
 - Missing attribute value, 84–86
 - MixColumns* transformations, 149
 - Mixing technique, 375, 376
 - Mix Probabilistic Matrix Factorization (MixPMF) model
 - architecture of, 264
 - collaborative filtering, 263
 - frequency distribution, 267
 - picture representation, 267
 - proposed model, 266
 - recommendation system, 263
 - training and evaluation, 266–268
 - Mobile learning (m-learning), 129–132
 - Model-based approach (MBA), 58
 - first system refinement, 58–59
 - hardware and software decomposition, 58
 - second system refinement, 59–60
 - Model-based engineering (MBE), 13
 - Model-based software, 47
 - Model-driven design (MDD), 75, 98
 - Model-driven engineering (MDE), 3, 39
 - Modeling editor, 41, 42
 - Modeling languages
 - AmaltheaModel and APP4MC, 48
 - classification framework, 49
 - timing model, 49–51
 - timing verification, 51
 - evaluation
 - timing properties, modeling of, 49–50
 - timing requirements, modeling of, 50–51
 - timing verification, support for, 53
 - research, 48–49
 - rubus component model, 47, 48
 - Modularity, 112–113
 - optimization method, 116
 - Molecular dynamics (MD), 35
 - comparison of, 35–36
 - free energy, 37
 - MongoDB database, 221
 - MongoDB technology (NoSQL), 223
 - Monotone, 380–381
 - Moodle, 130, 131
 - Mouse dynamics, 137
 - Multiagents systems (MASs), 173
 - Multilayer Perceptron (MLP), 122–124, 126
 - Multiplayer VR simulators, 18–19
 - MySQL, 222
- N**
- Naive Bayes algorithm, 164
 - Naïve Bayes classifier, 232
 - Narrative device, 374
 - National Institute for Standards and Technology (NIST), 148
 - National Security Agency (NSA), 156
 - Natural language processing (NLP), 231, 239
 - Natural Language Toolkit (NLTK), 257
 - Near Real-time Autonomous Quality Control (NRAQC) System, 326
 - Nebula Genomics, 182
 - .NET Framework, 323
 - Nevada Research Data Center (NRDC), 321
 - NICE Cybersecurity Workforce Framework (NCWF), 156

- NLPAug library, 238, 240, 242
 - NodeJS technology, 223
 - NoSQL (Not Only SQL) databases, 97
 - Novelty detection, 120
 - NRDC Quality Assurance (QA) Application, 323
 - NSF Track 1 Solar Energy-Water-Environmental Nexus project, 321
 - NVIVO software, 192
- O**
- Object coordination and interaction, 288
 - OceanTEA, 322, 326, 327
 - One-Hot encoding, 231, 232
 - One-time authentication, 137
 - Online checkers, 166
 - “O3” optimization level, 151
 - OpenEHR, 181
 - Open government data, 83
 - attribute, distribution of error quantity by, 87
 - data dictionary, 89
 - data quality dimensions, 83–84
 - error type, distribution of error quantity by, 85, 86
 - frequent quality mistakes in, 85, 86
 - illegal values, 86, 90
 - improvement suggestions, 85, 86
 - incorrect values, 86
 - methodology, 84–85
 - execution, 85
 - planning, 84
 - questionnaire, 88
 - Open Platform Communication-Unified Architecture (OPC UA), 12–13
 - Open-source web-application, 196, 198
 - OpenSSL, 150
 - Open Web Application Security Project (OWASP), 161
 - Operation Technology (OT) domain, 13
 - OramaVR, 18
 - Order of execution, virtual environment, 289
 - Orthogonal Frequency Di-vision Multiplexing (OFDM), 351
 - Overhead, 173–176
- P**
- Pairs trading, 245–253
 - Pairwise matching, 110
 - Pairwise testing technique, 223
 - ParseError, 30
 - Parsing, 109–110
 - Partially conservative (PCS), 388–389
 - Partially repetitive (PRP), 388
 - Participatory modeling, 97
 - entities’ incorporation degree, calculation of, 99
 - method flow, 99
 - Movies* database, modeling, 100–105
 - phases, 98
 - relationship’s incorporation degree II, calculation of, 99, 101
 - studies, 97–98
 - Patient handoffs, 18
 - Pattern-based authentication, 137
 - Peer-to-peer multi-node blockchain, 196, 198
 - Petri net
 - application, 395–397
 - detection of circuitbreaker, 390–391
 - detection of L3-circuit, 390–391
 - firing condition, 388
 - incidence matrix, 388
 - k -place, 391
 - liveness, 389
 - overview of, 394
 - performance comparison, 397–398
 - place/transition net, 388
 - Qusai-home state, 395
 - reachability analysis, 394–395
 - reachability coverability analysis, 394–395
 - SAT-solver, 389–390
 - strictly L3-live structure, 389
 - structural analysis, 389–391
 - structural properties, 388–389
 - submarking, 395
 - support tool, 387
 - PhishChecker, 163
 - Phishidentity, 164
 - PhishGuard, 163
 - Phishing attacks
 - anti-phishing tools, 162, 165, 167
 - auto-updated whitelist, 164
 - bagging, 163
 - classification, 163
 - CNNPD, 164
 - credentials from users, 163
 - detection tools
 - Avira, 165
 - BitDefender, 165–166
 - CLEAN MX, 165
 - Fortinet, 165
 - Kaspersky, 165
 - Sophos, 165
 - Virus Total (VT), 165
 - evaluation procedure
 - dataset description, 166
 - evaluation metrics, 166
 - forward credentials to attacker, 163
 - future work, 167
 - hybrid methodology, 164
 - legitimate behavior, 163
 - malicious links, 163
 - online detection tools, 162
 - performance evaluation results for website checking tools, 167
 - Phishidentity, 164
 - phishing emails to users, 162
 - phishy behavior, 163
 - scenario of, 162
 - smishing, 162
 - suspicious behavior, 163
 - vishing, 162
 - PhishingCorpus, 164
 - PhishNet, 163
 - PhishTank, 165
 - Photogrammetry, 20
 - Physical-Digital Boundary, 364
 - Pilot searches, 189
 - Platform as a Service (PaaS), 222
 - Plug-in Development Environment (PDE), 31
 - Polygonal approximation, 380
 - Portable test and Stimulus Standard (PSS) language, 40
 - implementation and use case, 40
 - editor generation, 44–45
 - graphical and textual notations, generation of, 42
 - input format, 42
 - mapping and synchronization, 42–44
 - selection of, 41
 - Predictive Blacklisting, 163

- Preprocessing, 107–108
 - Principal component analysis (PCA), 205
 - PRISMA-P checklist, 188
 - Privacy, 171–173, 175–178
 - Probabilistic Matrix Factorization (PMF), 264, 265
 - Probabilistic Neural Network (PNN), 164
 - Probabilistic token-based methods, 107
 - Product Owner (PO), 66
 - Programmable logic controllers (PLC), 11, 13
 - Proof-of-Burn (PoB), 180
 - Proof of Concept (PoC), 221
 - BigChainDB tool, 224
 - remote monitoring functionality, 224
 - Sprint 1, 223
 - Sprint 2, 224
 - Sprint 3, 224–225
 - Sprint zero, 223
 - Proof-of-Stake (PoS), 180
 - Proof-of-Work (PoW), 180
 - Public Environmental Information System (PEIS), 322
 - Python, 222, 223
 - PyTorch, 260
- Q**
- QA application, 323
 - Quality assessments, 58
 - Quora, 255
- R**
- Radiation therapy, 230
 - Random Forest Classifier, 232
 - Random radial basis function generator, 122
 - RDTSCP instruction, 151
 - Real-time ETL process, 91–92
 - studies, 92–96
 - applied purposes, 94–95
 - features, 94–95
 - general studies, 93–94
 - Record-record similarity-based graph entity resolution, 109
 - Recurrent neural networks (RNNs), 258
 - unidirectional vs. bidirectional, 259, 260
 - ReferenceError, 30
 - Reference integrity Metric (RIM), 172
 - Relational databases, 97
 - Reliability, 337
 - Resource Description Framework (RDF), 98
 - Response coding, 231, 232
 - REST API, 197
 - RESTful API, 222
 - REST-ful API Modeling Language (RAML), 78
 - REST service add_user, 197
 - REST service verify_user, 197–198
 - Rich Client Application (RCP), 31
 - Rijndael, 147
 - encryption and decryption, 149
 - ICL, 150
 - implementation, 149
 - IVB, 150–151
 - measurements methodology, 151
 - optimized code, 151
 - performance in ECB and CTR modes, 152
 - SKL, 150
 - Road Side Units (RSUs), 173
 - RP inclusion criteria, 334–335
 - Rubus component model (RCM), 47, 48
 - Rubus real-time operating system (RTOS), 53
 - Rule-based analysis, 30
- S**
- Safe online banking, 165
 - Sars-CoV-2 virus, *see* COVID-19
 - Scalability, 45, 175
 - Scikit-learn library, 231
 - Scrum Agile method, 66
 - Scrum Framework (SF), 65, 71, 219
 - Scrum model, 220
 - Scrum of Scrums (SoS) meetings, 66
 - Scrum Teams (STs), 220
 - Seamless, 367
 - Search and rescue unmanned aerial vehicles (SARUAV), 277, 278, 286
 - See also* Unmanned aerial vehicles (UAV)
 - Second system refinement, 59–60
 - Security, 337
 - Semantic assignment, 123
 - Semantic interoperability, 334, 337
 - Semantic-MDBScan
 - algorithm, 121, 122
 - flowchart, 122
 - future works, 124, 126
 - KNN, 124
 - MLP, 124
 - spatial entropy parameters, 123
 - test datasets, 122–123
 - abrupt/recurrent changes, 124, 125
 - gradual/recurrent changes, 124, 126
 - Semantic value, 123
 - Semantic Web of Things (SWoT), 333
 - Sensors, 341
 - SEPHAS Lysimeter Visualization, 323
 - Sequelize technology, 223
 - Sequence-to-Sequence modeling (Seq2Seq model), 272, 274, 275
 - Service Discovery, 323
 - Service Oriented Architecture (SOA), 73, 74
 - Shannon's entropy-based algorithm, 112
 - ShiftRows*, 149
 - Simple graph modeling, 98
 - Simulation, 59, 60
 - Single Instruction Multiple Data (SIMD), 150
 - Single Shot Detection (SSD) binary, 278
 - Single sign-on (SSO) fingerprint authentication
 - AES, 196
 - back-end modules, 197
 - backend of our proposed web app
 - biometric-based user authentication, 199
 - Java Tomcat REST server, 197
 - MVC architecture, 197
 - REST service add_user, 197
 - REST service verify_user, 197–198
 - blockchain advantages, 196
 - flow, 196
 - frontend of our proposed web app
 - API for user authentication does not recognize the user, 202
 - API for user authentication is successful, 201
 - denied registration of an existing user, 200
 - successful user registration, 200
 - user registration screen, 199
 - webform to authenticate user, 201
 - future work, 198–199
 - identity provider, 196

- IPFS, 196
 - 2-factor authentication, 196
 - web app development, 197
- Sirius platform, 44
- SkyLake (SKL), 150
- Smart Cities, 342
- Smart contracts, 181–183, 190
- Smart parking
 - approaches, 341
 - database, 343
 - data propagation, 342
 - free parking spots, 341
 - machine learning model
 - account for uncertainties, 349
 - backend server, 345
 - dataset, 343–345
 - dynamically calculating speed, 349
 - geofences, 345–347
 - model architecture and training, 345
 - model parameters, 345
 - model prediction consequences, 347
 - purpose, 343
 - time prediction, 347–349
 - manual spot entering, 341
 - mobile app, 343
 - mobility support layer, 342
 - sensors, 341
- Smartphones, 131, 137, 138
- Smart watch
 - additional synchronization, 141–142
 - data collection, 138–139
 - error computation, 140–141
 - error rates of accelerometer and gyroscope, 141, 144
 - future works, 144
 - inter-pattern comparison, 141
 - intra-pattern comparison, 141
 - overall error for inter and intra-pattern comparisons, 143, 144
 - pre-processing of data, 139–140
 - signal data comparison after applying native alignment for same pattern, 141, 142
 - signal data comparison after applying native alignment for two different patterns, 142, 143
 - signal data comparison after applying original alignment for same patterns, 141, 142
 - signal data comparison after applying original alignment for two different patterns, 141, 143
- Smishing, 162
- Social media user
 - Chrome browser, 306
 - collected data, 306
 - data analysis, 306–308
 - Facebook, 303
 - human computer interaction (HCI), 303
 - image and text times, 306
 - Instagram, 303
 - methodology
 - apparatus, 304–305
 - design, 305–306
 - participants, 304
 - procedure, 305
 - tasks, 305
 - one-way ANOVA, 308
 - participants' TikTok time, 306
 - questionnaire analysis, 308
 - social media sites, 304
 - TikTok, 303
 - Twitter, 303
 - video and text times, 306
- Social networks, 173, 177
- Soft skills, 70
- Software architecture, 74
- Software engineering
 - proof of concept, 69–70
 - quantitative outcomes, 70
 - students' feedback, 70
 - studies, 66
 - tailoring scrum and IPBL, 66–69
 - scrum artifacts academically used, 67–68
 - scrum roles academically used, 67
 - STEPES-BD scrum ceremonies, 68–69
- Software module, 59
- Solar Energy-Water-Environmental Nexus project, 322
- Sophos Endpoint Protection, 165
- Sound design technique, 374–376
- Sound editor/mixer, 375
- Sound Professionals, 377
- Soundscapes and sound design, 376
- SpamAssassin public datasets, 164
- Spark tool, 225
- Spatial entropy, 120, 123
- SpoofGuard tool, 163
- SpringerLink, 189, 190
- Stacking Classifier, 232
- Standalone operating system apps, 130
- State Transition Test, 223
- STEPES-BD project, 66, 70
 - BigchainDB, 221
 - business layer, 221
 - data interface, 223
 - data interoperability, 222
 - face-to-face care process, 224, 225
 - infrastructure, 221–222
 - interface layer, 221
 - persistence layer, 221
 - PoC (*see* Proof of Concept (PoC))
 - software application, 222–223
 - testing the application, 223
 - 3-tier architecture of, 222
- Stock Backtesting engine, pair trading
 - algorithmic trading, 246
 - analysis module, 249
 - automated trading, 245
 - backtesting, 246
 - Backtest module, 248, 249
 - Bayesian Kalman filter, 252
 - cointegration tests, 247
 - daily profit and loss chart, 253
 - database DB module, 248, 249
 - design diagram, 249
 - Engle-Granger method, 247
 - high frequency trading (HFT), 246
 - MetaTrader5, 246
 - MS Backtesting, 246
 - pair finding module, 249
 - Python's matplotlib library, 250
 - retail sector stocks, 250–252
 - statistical arbitrage, 245, 246
 - stock trends of Pepsi and Coca Cola, 247, 248
 - Tradingview, 246
 - Trend Spider backtester, 246
 - Yahho finance data, 250–251
- Storytelling, 66

- Stream reservation (SR) classes, 12
 - Structurally bounded (SB), 388
 - Structured Text (ST), 13
 - Student Financing Fund (FIES), 84
 - SubBytes*, 149
 - Submarking, 394, 395
 - Supervisory control and data acquisition (SCADA) systems, 11
 - Support Vector Machine (SVM), 164
 - Switched Ethernet, 13
 - Synchronizer, 43, 44
 - Syntactic dependency analysis
 - research questions, 27–28
 - tool design, 28
 - tool implementation, 28–30
 - tool presentation, 30–31
 - validation, 31
 - “Syntax” errors, 28
 - Systematic literature review (SLR), 188, 191
 - automated search, 188
 - confidence in cumulative estimate and meta-bias, 192
 - data collection process, 191
 - data items, 191
 - data management, 190
 - data synthesis, 192
 - future work, 192–193
 - inclusion/exclusion criteria, 188, 189
 - information sources, 188–189
 - outcomes and prioritization
 - primary outcomes, 191–192
 - secondary outcomes, 192
 - proposed methodology, 189
 - reporting, 335
 - research goals, 188
 - research protocol (RP), 334
 - research questions, 188, 334
 - results, 336–339
 - risk of bias individual studies, 192
 - search strategy, 189–190
 - selection process, 190–191
 - snowballing search, 188
 - System Usability Scale (SUS), 131
- T**
- Taxonomy, 174–175
 - Team Scrum Masters (TSMs), 66, 220
 - Telemedicine, 179, 190
 - Terrestrial LiDAR, 294
 - Test datasets, 122–123
 - Test Driven Development (TDD), 223
 - Text data analysis, 239
 - TextRank algorithm, 271–273, 275
 - Textual modelling, 4
 - Textual notations, 41
 - Thematic analysis, 373
 - Three dimensional (3D) visualization of human organs
 - average confidence level of students in health science I, 213, 214
 - average confidence level of students in health science II, 213, 215
 - Blender software, 212
 - experiment implementation, 212–213
 - future work, 215
 - of human heart, 212
 - improved model of human heart, 212
 - models used in survey, 215
 - recognition accuracy for health science I class, 213, 214
 - recognition accuracy for health science II class, 214
 - training technique, 211, 212
 - 3D point cloud mesh, 297–299
 - 3DSYSTEM viewer, 294
 - Time complexity, 384
 - Time-sensitive networking (TSN), 12, 15
 - Timestamp, 123
 - Token-based graph entity resolution, 108–109
 - Totally integrated scrum (TIS) method, 220
 - Touchscreen interaction, 137
 - TrainingDataSet, 122, 123
 - Transmission gates, 12
 - True negative (TN), 123–124
 - True positive (TP), 123
 - TrustBar, 163
 - Tuckman Maturity Model (TMM), 220
 - Tuckman Model Maturity Level (TMML), 66
 - 2D complex discrete wavelet packet transform (2D-CWPT)
 - dual-tree complex wavelet packet transform, 206–207
 - dual-tree complex wavelet transform, 206
 - flowchart, 207
 - index of each frequency component at levels 1 and 2, 207
 - preprocessing for endoscopic image, 207–208
 - Two dimensional (2D) organ images, 212, 213, 215, 216
 - Two Factor Authentication (2FA) property, 162
- U**
- Ubiquitous object interaction, 288
 - UI interaction, 288
 - UltraLeap’s Leap Motion Controller, 287
 - Unified modeling language (UML), 75
 - United States Naval Academy, 157
 - Universal verification methodology (UVM), 41
 - Unmanned aerial vehicles (UAV)
 - communication process flowchart, 280
 - machine learning techniques, 278
 - methodology, 278–279
 - object detection algorithms, 281
 - processing speed, 279
 - process outline, 280
 - PyTorch codes, 281
 - You Only Look Once (YOLO), 278
 - Unsupervised entity resolution
 - dataset, 114
 - evaluation, 115–116
 - experimental setup, 114–115
 - hybrid graph-based entity resolution, 109
 - method
 - blocking, 110
 - canonicalization, 113
 - graph-based hierarchical record clustering, 110–114
 - merged data file, 109
 - pairwise matching, 110
 - parsing, 109–110
 - probabilistic token-based methods, 107
 - record-record similarity-based graph entity resolution, 109
 - studies, 108
 - token-based graph entity resolution, 108–109
 - “Unused Ref”, 28
 - Usability, 129–132
 - User authentication, 137, 138
 - User-Centered Design (UCD), 130
 - U.S. National Initiative for Cybersecurity Education (NICE), 156

- V**
- vaes—Code, 151
 - vaesx*N*—Code, 151
 - Vector AES-NI, 148
 - Vehicle operating costs (VOCs)
 - application and challenges, 313
 - data flow diagram, 313, 314
 - driving cycles, 312
 - entry questionnaire, 315–316
 - Excel interface, 313, 318
 - exit questionnaire, 318
 - fuel consumption interface, 315, 316
 - fuel economy, 313, 314
 - mileage-related vehicle depreciation, 313
 - performance, 312
 - physics-based vehicle models, 312
 - repair and maintenance modules, 313
 - statistical difference, 319
 - task completion time, 317
 - tasks performed by the participants, 316–318
 - user study
 - design, 315
 - participants, 313
 - procedure, 314
 - tasks, 314–315
 - technology, 313–314
 - website interface, 313
 - Virtual Reality (VR), 17
 - avatar eyeball movement, 22
 - avatar lip tracking, 22
 - framework
 - interactables, 21
 - voice, 21–22
 - implementation
 - eye tracking, 19
 - photogrammetry, 20
 - process of execution, 19
 - recording, 19–20
 - room creation, 20
 - room objects, 20
 - interaction
 - avatars, 20–21
 - clipboard, 21
 - voice and audio, 21
 - medical training simulators, 18
 - in medicine, 18
 - multiplayer VR simulators, 18–19
 - notepad improvements, 22
 - object interaction, 22
 - Virus Total (VT), 165
 - Vishing, 162
 - Visibility and boundary approximation, 380–384
 - Visual Similarity Based Phishing Detection, 163
 - Vive Pro Eye, 19
 - Voice communications, 21–22
 - Voight-Kampf machine, 376
 - Volume renderings, 211
 - VR controllers
 - detailed use cases
 - common object usage, 288
 - locomotion, 288
 - physics rigidbody object, 288
 - virtual input, 288
 - virtual menu usage, 288
 - gloves, 287
 - HCI VE considerations, 286
 - implementation
 - menu interaction, 289
 - rigidbody physics object interactions, 290–291
 - virtualized everyday objects, 289–290
 - input usability factors
 - ease-of-use, 287
 - performance, 287
 - interaction design, 286–287
 - laser pointing system, 287
 - remapping, 286
 - key component, 285
 - ready-made VR controller, 286
 - virtual input devices, 287
 - VR cross-compatibility, 22
- W**
- Walmart Express store, 366
 - Walmart virtual reality training system, 365
 - Warnings, 31
 - WearOS operating system, 138
 - Web-based application, 63–64
 - Web3j, 197
 - Web of things (WoT), 333
 - Weka program, 163
 - Wi-Fi connectivity, 138
 - Wi-Fi networks, 351
 - Wilcoxon test, 124
 - William Pennington Engineering Building (WPEB), 299
 - Windows Communication Foundation (WCF), 323
 - World-Class Business Ecosystem Model, 364
- X**
- x86-64C intrinsics, 151
 - XOR operations, 148, 153
- Y**
- You Only Look Once (YOLO), 278
- Z**
- Zero-hour phishing attack, 164
 - Zotero, 190