ComSIS
Computer Science and Information Systems
Volume 16, Number 1, January 2019

Contents

ISSN: 1820-0214 (Print) 2406-1018 (Online)

Computer Science and Information Systems
Vol 16, No 1, January 2019

ComSIS

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 16, Number 1
January 2019

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Editorial Office:
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Informatics
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
Phone: +381 21 458 888; **Fax:** +381 21 6350 458
www.comsis.org; Email: comsis@uns.ac.rs

Volume 16, Number 1, 2019
Novi Sad

Computer Science and Information Systems

ISSN: 1820-0214 (Print) 2406-1018 (Online)

The ComSIS journal is sponsored by:

Ministry of Education, Science and Technological Development of the Republic of Serbia
http://www.mpn.gov.rs/

Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2017 two-year impact factor 0.613,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Summon (Serials Solutions),
- EBSCO bibliographic databases,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (MS Word or LaTeX), or three hard-copies of the manuscript written in English, intended for publication and prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be sent to official journal e-mail.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 20 (using the appropriate Word or LaTeX template).

Manuscripts will be refereed in the manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

Computer Science and Information Systems

Volume 16, Number 1, January 2019

CONTENTS

NOTE ON (SELF-)PLAGIARISM

Before commencing with the editorial and the contents of this issue, we have an important announcement regarding (self-)plagiarism.

During a routine check of articles published in Computer Science and Information Systems using plagiarism-detection software, two borderline cases of self-plagiarism were detected. The articles in question amply used excerpts from other papers by the same authors, to the point of obscuring the contributions of said articles.

Although the Editorial Board determined that the articles contain just enough novelty not to warrant retraction, we emphasize that we will implement a much stricter anti-plagiarism policy in the future. If a borderline case of self-plagiarism "slips through" and is published, upon later detection we will at the very least publicly label the article as such. Needless to say, as always, clear cases of plagiarism and self-plagiarism will be retracted immediately upon detection.

In the end, we hope not to have to deal with such issues again, and implore all prospective authors to carefully check the manuscripts intended for submission to our journal.

Editor-in-Chief
Mirjana Ivanović

Managing Editor
Miloš Radovanović

EDITORIAL

This issue, starting Volume 16 of the Computer Science and Information Systems journal, consists of one invited and 13 regular articles. As always, we acknowledge the hard work and enthusiasm of our authors and reviewers, who were invaluable to producing the current issue.

The current issue begins with the invited paper "Cross-layer Design and Optimization Techniques in Wireless Multimedia Sensor Networks for Smart Cities" by Hasan Ali Khattak et al., which surveys the state-of-the-art in cross-layer optimization approaches for the structural design of wireless sensor networks, in order to solve issues such as energy conservation, variable channel capacity, and the resource-constrained nature of such networks. Techniques are evaluated and discussed through the prism of six criteria: energy efficiency, quality of service, communication reliability, security, error correction, and network resource management. Challenges and future directions are highlighted in detail, emphasizing the use of multimedia data and applications on the internet of things.

In the first regular article, "SOCA-DSEM: a Well-Structured SOCA Development Systems Engineering Methodology" Laura C. Rodriguez-Martinez et al., tackle the theoretical and practical deficiencies of service-oriented computing application (SOCA) development methodologies. The authors first argue for the use of agile approaches (more specifically agility-rigor methodologies), and then propose a new SOCA development systems engineering methodology (DSEM) giving its description, theoretical foundations, illustration of use on a prototype running example, and two pilot empirical evaluations on usability metrics.

"Building a flexible BPM Application in an SOA based Legacy Environment," by Mladen Matejaš and Krešimir Fertalj proposes an integration model for building a business process management application (BPMA) and connecting the BPMA with legacy systems based on service-oriented architecture (SOA), with notable characteristics being simple co-dependence of BPMA and existing systems, minimal changes to legacy applications and maximal use of existing functionalities. Feasibility of the approach is demonstrated on a real-life business use case scenario.

Marko Janković et al., in "Reconstructing De-facto Software Development Methods," present an approach through which companies can document and reconstruct the software development methods that they truly use. Contrary to existing approaches, the proposed method does not place great burden on company staff, but rather extracts information on development practice directly from software repositories. The approach was developed through studying five companies, and evaluated on a real software repository shared by an additional company, demonstrating its effectiveness on various aspects of the development process.

"CrocodileAgent 2018: Robust Agent-based Mechanisms for Power Trading in Competitive Environments," authored by Demijan Grgic et al., addresses the problem of power brokerage in sustainable energy systems enabled by smart grids. The proposed agent-based approach fared very well (3rd overall) in the 2018 Power Trading Agent Competition, thanks to its focus on the creation of smart time-of-use tariffs to reduce peak-demand charges.

In their article entitled “Estimating Point-of-Interest Rating Based on Visitors Geospatial Behaviour,” Matej Senožetnik et al. deal with the problem of sparse ratings by extracting points-of-interest from various sources and proposing an approach to estimating point-of-interest ratings based on geospatial data of their visitors. Ratings are estimated using various machine-learning techniques (including support vector machines, linear regression, random forests and decision trees), and the approach is experimentally evaluated on the problem of motorhome users visiting campsites in European countries.

“A Study of Sampling with Imbalanced Data Sets – Case of Credit Risk Assessment” by Kristina Andrić et al. investigates the role of sample size and class distribution in credit risk assessments, performing a large-scale experimental evaluation of real-life data sets of different characteristics, using several classification algorithms and performance measures. Results indicate performance measure, classification algorithm and data set characteristics all play a role in determining the optimal class distribution, offering insight on how to design the training sample and select the classification algorithm to maximize prediction performance.

The article “Towards Understandable Personalized Recommendations: Hybrid Explanations,” by Martin Svrcek et al. addresses the problem of trust in recommender systems, i.e. recommendations, through proposing an approach for making recommendations transparent and understandable to users. This is achieved with a hybrid method of personalized explanation of recommendations, independent of recommendation technique, which combines basic explanation styles to provide the appropriate type of personalized explanation to each user. Experimental evaluation in the news domain demonstrated not only improvements to users’ attitudes towards recommendations, but also to recommendation precision.

George Lagogiannis et al. in their article “On the Randomness that Generates Biased Samples: The Limited Randomness Approach,” tackle the problem of generating exponentially biased samples from data streams by proposing an algorithm using O(1) random bits per stream element, as opposed to existing algorithms and the theoretical limit of O(n), where n is the sample size. This is achieved in a specific setting where survival probabilities are assigned to the stream elements before they start to arrive.

In “Retinal Blood Vessel Segmentation Based on Heuristic Image Analysis,” Maja Braović et al. improve automatic analysis of retinal fundus images for disease diagnosis by proposing an approach for retinal blood vessel segmentation that is simpler than existing approaches (since it is based on a cascade of very simple image processing methods). The method emphasizes specificity over sensitivity, and is demonstrated to achieve high average accuracy on two publicly available data sets.

“Data Visualization Techniques in Medical Image Compression Evaluation – An Empirical Study,” by Dinu Dragan et al. presents an empirical study of multidimensional visualization techniques, considering tables (control), scatterplots, parallel coordinates, and star plots in the problem of decision making in picture archiving and communications system (PACS) design. In the experiments, using a specially developed tool three sets of subjects were presented with visualizations, recording their decisions in order to determine which visualization technique produces the best decisions. Conclusions include visualizations producing better results than tables, and 2D parallel coordinates being best on high-dimensional data.

Francisco Jurado and Pilar Rodriguez, in “An Experience on Automatically Building Lexicons for Affective Computing in Multiple Target Languages,” tackle the problem of lexicon building for affective computing on texts written in languages other than English. The presented approach starts with initial seeds of English words in that define the emotions of interest, and then expands these with related words in a bootstrapping process, finally obtaining a lexicon by processing the context sentences from parallel translated text where the terms have been used. Exploratory analysis shows promising results, with similar affective fingerprints observed in different translations of the same books.

In their article “Improving Sentiment Analysis for Twitter Data by Handling Negation Rules,” Adela Ljajić and Ulfeta Marovac explore how the explicit treatment of negation by using grammatical rules that influence the change of polarity impacts the effectiveness of sentiment analysis of tweets in the Serbian language. Experimental evaluation showed statistically significant improvement of sentiment analysis when negation was processed using the proposed rules with the lexicon-based approach or machine learning methods.

Finally, “Rejecting the Death of Passwords: Advice for the Future,” by Leon Bošnjak and Boštjan Brumen discusses the problems of using textual passwords for authentication, as well as their alternatives. An interesting historical perspective is given, showing that real user-generated passwords used today are no better than the ones used four decades ago. The article gives recommendations how to improve password security, arguments against replacements of textual passwords, and discusses conditions that need to be met in order for passwords to finally be replaced.

Editor-in-Chief
Mirjana Ivanović

Managing Editor
Miloš Radovanović

Computer Science and Information Systems 16(1):1–17 https://doi.org/10.2298/CSIS181115004K

Cross-layer Design and Optimization Techniques in Wireless Multimedia Sensor Networks for Smart Cities ⋆

Hasan Ali Khattak[1], Zoobia Ameer[2], Ikram Ud Din[3], and Muhammad Khurram Khan*[4]

[1] COMSATS University Islamabad, 44550 Islamabad, Pakistan.
hasan.alikhattak@ieee.org
[2] Shaheed Benazir Bhutto Women University Peshawar, 25500 Peshawar, Pakistan.
zoobia.ameer@sbbwu.edu.pk
[3] Department of Information Technology, The University of Haripur, Haripur 22620, Pakistan.
ikramuddin205@yahoo.com
[4] Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh 11451, Kingdom of Saudi Arabia.
mkhurram@ksu.edu.sa

Abstract. The future smart cities vision can be developed through leveraging the potentials of Internet of Things (IoT) and wireless sensor network (WSN) technologies. WSN is a resource constrained network where network nodes are tiny devices that are run on battery power. Diverse types of applications such as environmental and habitual monitoring, detection, and tracking, use WSNs. The invention of new network protocols, the establishment of new models for communications, and testing the available solutions in real world environment are some of the current research issues in WSNs. Main challenges in such networks include energy conservation in an efficient way, dealing with variable channel capacity, and the resource constrained nature of such networks. The design of architecture for such networks has a vital role in solving the issues to some extent, i.e., the cross layer design approach is an architectural technique that offers the interaction of different layers together to enhance the performance, minimize the energy consumption, enhance the network life time, and provide Quality of Service (QoS) in real time communications. These are some of the current areas where cross-layer design approaches are being used. This paper presents different types of cross-layer design techniques in wireless multimedia sensor networks. Using such architectural techniques, different state of the art cross-layer optimization approaches are discussed while giving the reader an insight on prominent challenges and issues along with future directions.

Keywords: Optimization, Multimedia, Wireless Sensor Networks, Smart Cities, Internet of Things.

1. Introduction

Wireless Sensor Networks (WSNs) include a large number of low-power devices normally known as sensor nodes. Sensor devices are resource-constraint nodes with respect to computational power, storage capacity, energy, bandwidth, and communication range [26]. Depending on the application scenario, sensor nodes are either precisely placed in the area of interest (e.g., structure monitoring of a building) or randomly deployed in the

⋆ Corresponding Author: Muhammad Khurram Khan (mkhurram@ksu.edu.sa)

field (e.g., environmental monitoring). Furthermore, these nodes are responsible for the sensing parameter of interest and/or detect an event from their environment and report it to a base station that is often called sink node or gateway node. The sink is more resourceful node compared to other nodes in the WSN.

Recent advancement in the sensor technology and micro-miniaturization has enabled the emergence of a more specialized field of WSN that is known as Wireless Multimedia Sensor Network (WMSN), which has applications in various domains in futuristic smart cities such as health-care monitoring [27], home automation [16], transportation & vehicular networks [18,11]. In WMSN, sensor nodes are equipped with audio and video capturing devices, such as microphones and inexpensive CMOS cameras. Sensor nodes are also responsible for capturing and delivering multimedia content, such as (a) audio, (b) video, and (c) still images to the gateway node [2]. Some applications of the WMSN are surveillance [10,16], traffic management [29,36], health-care services [2,19], environmental monitoring [35], and industrial process management [41], as shown in Fig. 1.

The rapid and persistent evolution in tools and technologies along with the explosive growth in the market proliferation of IoT devices has made WSN a ubiquitous medium towards the realization of smart cities [12,13]. WSNs have presented various applications and have given space for developing several services to benefit smart city infrastructures. There have been several issues and challenges which hinder the full scale deployment of the future generation of IoT enabled smart cities [14].

In resource constraint IoT based network infrastructures, for example, WMSN, achieving (a) low frame loss, (b) reduced delay, and (c) better end-to-end video quality are a challenging task. In order to ensure higher quality of service with constrained resources, one has to think out of the box. Traditional smart city design approaches might not be suitable and efficient for taking full benefit from the WMSN. One such approach is the cross-layer design for cross layer optimization, being an optimal approach, which is used not only to improve but also optimize the overall network performance. The cross layer optimization has made it possible for the smart cities research community.

”Cross-layer design techniques with regards to the reference layered model is the design of protocols, architectures, or algorithms, that not only exploit but also offer a set of inter-layer interaction which is a super set of the standard interfaces provided by the overall reference layered architecture” [20]. According to [38], the cross-layer design is a back-and-forth information flow that merges different layers and couples those layers that have no common interface. In contrast to the layered reference model, there are explicit interfaces among layers and some time the concept of shared database. In a layered structure, protocols communicate with adjacent layers via function calls, while a cross-layer design violates such limitations of communications [6].

The focus of this paper is to explore cross-layer optimization techniques in the context of WMSN for future smart cities. The rest of the paper is organized into the following sections: Section 2 covers the basic concepts of cross-layer design techniques. State of the art cross-layer design approaches in WMSNs are discussed in Section 3, and Challenges of cross-layer designs in WMSNs are discussed in Section 4. At the end, we finalize the paper with the concluding remarks and future directions presented in Section 5.

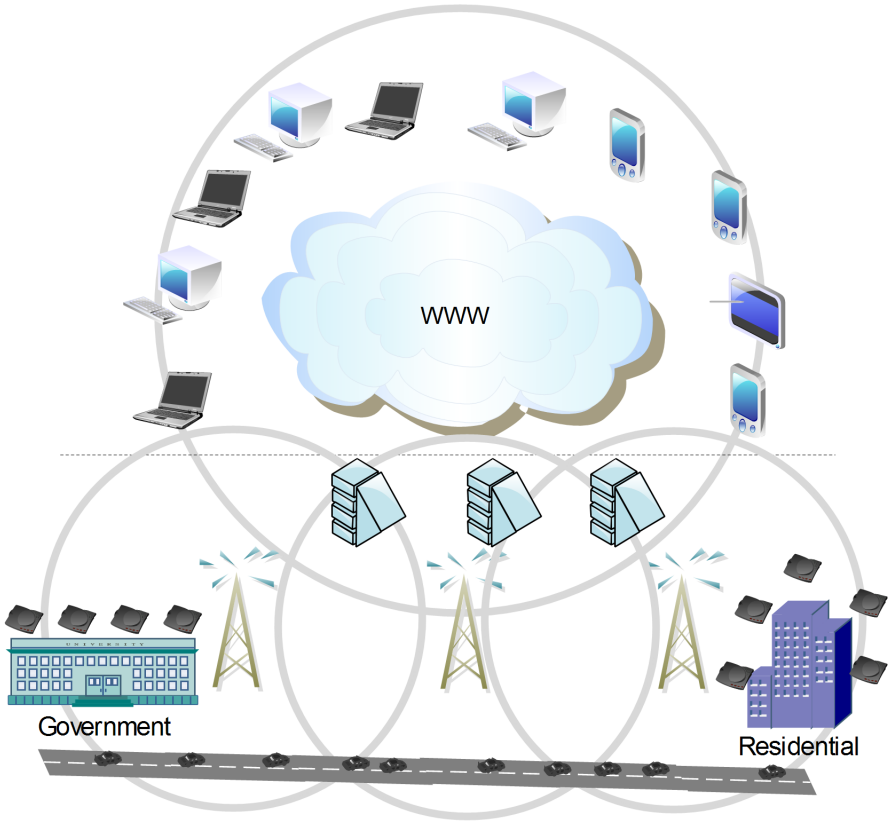


Fig. 1. Smart City Applications Home Automation, Building Monitoring, Health-care monitoring, and Vehicular Networks to name a few.

2. Cross Layer Design

A great deal of research is being carried out since last few years in the field of wireless technology to come up with better architectural approaches. Most of these efforts are focused towards reducing strong inter-dependency among layers, which is a typical characteristic of the layered architecture. The cross-layer design is an emerging technique for both wireless and wired networks. The theme behind this approach is optimizing the flow of data in such a way that any two or more layers can be crossed for upgrading the complete performance of the network [49]. The ultimate goal is to enhance the performance of WMSNs in terms of energy usage, routing, delivery of multimedia contents, and network management.

Depending on the user requirements, two, three, or four layers of the layered architecture can be crossed. For example, in order to accomplish in-network processing and better multi-path selection, the application layer and routing layer are combined. Service level differentiation, priority scheduling, and efficient routing techniques are achieved when the routing and MAC layers work together. Moreover, achieving QoS and reliability of multimedia content when multiple routing paths scheme is used, the routing and transport layer contribute in achieving these enhancements [3].

The cross-layer design approach is one of the preposition in this regard that enables the designers to develop protocols which enable communications across layers (in contrast to the layered architecture approach) [15].

3. Cross Layer Design Techniques

Every layer contributes different functionalities in the protocol stack of WSNs and WMSNs. Crossing these layers can achieve performance gain in different scenarios. The most common of these performance enhancements are discussed in this section. Energy efficiency is a broad category as the nature of WMSN is resource constrained with respect to energy. The ultimate focus of every alike contribution is energy preservation of such networks. Other areas such as QoS and communication reliability along with security are also discussed later in this section.

3.1. Energy Efficiency

Energy efficient communication is always the most important and critical aspect while conceiving WMSNs. Researchers and developers come up with new frameworks, new architectures, and new protocols whose main goal is achieving the energy efficiency. Generally, every layer is responsible for something like protocols at network layer construct routes and do resource reservations. While considering multi-hop communications, the MAC layer is responsible for efficient channel utilization. The combination of both Network and MAC layer achieves to provide bounded delay in the communication via multi-hop paths. Another example is the use of routing information at the network layer that can maximize the sleep duration of each network node [40], the information taken from network layer that the network layer is busy, the MAC layer differs the channel access for nodes, and thus the nodes remain in sleep mode for a longer time. Hence, in the long

run, the energy efficiency is provided for the whole network when the sleep duration is increased. Following are related works from the literature that contribute energy efficiency with the use of cross-layer design approaches.

The cross-layer technique, discussed in [48], deals with the life time optimization of WMSNs, where the correlation between sensing data and the redundancy due to overlap regions consume more energy and degrade the life time of the network. Three layers are contributing in the optimization in such a way that the application layer is sharing the global variables, the transport layer is controlling the source rate, and the responsibility of the network layer is to adjust the flow rates on links.

Similarly, Lee *et al.* [21] address the problem of end to end video delivery in resource constrained sensor networks and life time enhancement. Routing and power allocation are optimized by the cross layer optimization. This is a three step procedure, i.e., first, the physical layer does the energy management by announcing power allocations to upper layers. Second, the network layer performs optimal routing on the basis of channel state and hop count. Third, the MAC layer performs power allocation by finding the optimum power allocation matrix with regards to information from other layers. The data rate and network life time are increased that achieve better end to end video delivery.

Collision Aware Routing Protocol (CARP) [45] uses two parameters, i.e., collision degree and energy level, for routing and rout adjustment rather than traditional hop count. High collision in the network consumes more energy, while the cross layer design avoids high collisions. The energy level is calculated with the help of residual and initial energies. A logical control unit over the network layer attains parameters from the physical and link layers. The more residual energy nodes are chosen based on the parameter value from the physical layer as relay nodes to save the energy. The CARP's routing metric needs physical layer functions characteristics such as data rate, number of collisions, and state parameters so as to set up an autonomous logical control unit in the network layer that fetches the physical and link layer parameters, and computes the degrees of collision and energy for routing. The control overhead is avoided by using the route metric in the RREP and RREQ packet header.

The cross-layer technique presented in [8] optimizes the energy usage by balancing the traffic flow across multiple paths, thus, the life time of the network is increased. The contributions presented in this work for different layers are as follows: At the routing layer, the multiple paths are used for data transmissions. In addition, when the retry limit of transmissions is over, the MAC layer is controlled for each channel.

The cross-layer optimization design presented in [30] increases the life time of WMSNs in two stages. In the first one, the number of connected sensor nodes is increased while in the second case, the network life time is increased by controlling the connection of low priority nodes in the network. All layers are involved in this optimization design technique.

3.2. Quality of Service

Quality of Service (QoS) is a very generic term used in computer networks domain. The term QoS refers to the end-to-end data transmission without much delay. The term is more specific for real time communications where the basic requirement is on time delivery and reception. A few of cross-layer design techniques are discussed in the following sections.

Table 1. State of the art in Cross Layer Energy Optimization Techniques

Ref.	Layered Crossed	Research problem	Performance Matrix
[40]	MAC and Network Layers	Information of routing at the network layer is used for the MAC layer, thus, it can maximize the sleep duration of all nodes, which ultimately increase the energy efficiency.	Data rate
[48]	Application, Transport and Networks	Densely deployed WMSNs have redundancy and correlation in multimedia data; allocating source rate and flow rate can enhance the life time.	Data rate
[21]	MAC, Network and PHY Layers	Efficient power management can enhance the network life time based on optimal routing from the network layer.	End to end delay.
[45]	MAC, Network and PHY Layers	Control overhead create the energy of WMSNs nodes depletes more rapidly, controlling the unwanted traffic can increase the life time of WMSNs.	Hop count, /collision degree/energy level.
[8]	MAC and Network Layers	Balancing the traffic flow across multiple paths can enhance the energy preservations.	data rate/ retry limit
[30]	All Layers	The network life time is maximized by controlling the admission of low priority nodes in the network	Data rate.

Well-known areas like QoS achievement are discussed, e.g., QoS in terms of resource reservation, video quality, and packet loss etc.

QoS in terms of resource Reservations A model based channel capacity prediction scheme is presented in [50] for providing statistical QoS. The scheme perceives the state of a link from the MAC re-transmission information where statistical channel capacity is calculated when the traffic load is saturated. Based on the proposed model, the QoS routing optimization and resource reservations are carried out. The proposed scheme assigns network resources and forwards packet taking into consideration the interference among flows and state of a link.

The path reservation technique, VSN-Module, is presented in [9]. This protocol is based on UltraWideBand (UWB). The VSN-Module provides the QoS support while the path reservation is done in a distributed manner in the network.

Multi-hop routes are set from source to sink and different parameters such as jitter, end-to-end delay, and bandwidth are satisfied based on flow requirements. The physical layer is integrated with the upper layer functionalities, for example, MAC, network and transport layers. The UWB technology is set at the physical layer while the channel access and transmission parameters are coupled.

QoS in terms of communication protocols function optimization The framework presented in [34] provides the QoS imposed on each stream. The tasks of communication protocols are optimized in this framework through maximizing video stream requests.

The increase in video sources in WMSNs along with the QoS is hard to achieve. Thus, frames are classified by the traffic classifier module and the rout classifier module of the framework selects the appropriate path for each frame delivery. The MAC layer maps frames in access controllers (ACs) according to the priorities and the appropriate transmission mode is then chosen by the network layer.

QoS in terms of prioritization of streams The new framework introduced in [15] provides QoS for each stream by prioritizing them. The traffic in WMSNs is classified into six classes. The network is partitioned into hexagonal cells to minimize the interference. A shared database is used for the cross layer design.

A single node supports multiple applications in this framework via reserving some area in the memory. Each application requests for its protocol parameters and a middleware is used for the interaction of shared database and applications. The middleware sets suitable parameters at different layers on the basis of network conditions and application requirements from physical channel conditions and MAC layer feedback. Then the application is informed by the middleware about the status of the network so that its behaviors can be adapted by the applications accordingly.

QoS in terms of application requirements The communication paradigm on the basis of time-hoping impulse radio ultra-wide band technology is presented in [25]. A control unit called cross-layer control unit (XLCU) controls and configures the networking functionalities at three layers, i.e., physical, MAC, and network layer. The QoS is provided on the basis of unified logic that takes decisions. These decisions depend on two conditions. First is the application requirements at application layer and second is the status of the functional blocks that implement the networking functionalities. Reliability and flexibility are achieved when delivering contents to different types of applications.

QoS in terms of streaming video quality The cross-layer framework in [39] aims to deliver only the summaries of video instead of the video file over wireless channels. Three layers, i.e., physical, data link, and application are combined in a cross-layer fashion that provide the best quality of video across the WMSN. Each layer optimizes the parameters like coding scheme, adaptive modulation, source coding, and retransmission. Source coding is optimized at the application layer. Similarly, ARQ at the data link layer and other parameters, e.g., adaptive modulation and encoding schemes at the physical layer.

An integrated cross-layer optimization algorithm is presented in [4] for video streaming across multi-hop networks [4]. It optimizes different control parameters across the application layer, network layer, MAC layer, and physical layer to provide QoS in multipath transmission networks using IEEE 802.11a/e. At routing layer, the end-to-end optimization is done for path choosing. At MAC layer, the maximum retry limit is set and the modulation scheme is chosen at physical layer.

The proposed system presented in [23] consists of a video encoder module, MAC layer cognitive module, modulation and coding module, and cross-layer module. The cross-layer module adjusts all network functions by an appropriate selection of system parameters, hence, the network functions are jointly optimized in order to obtain best user

perceived video quality using cognitive radio. The dynamic channel coding MAC (DCC-MAC) technique is used in this system to allow the concurrent transmissions of distinct pairs of source and destination that reduce the interference among nodes.

A new architecture, called QoS MAC is presented in [46], where different contention windows for application specific QoS requirements are set thereby the traffic is categorized on the basis of delay requirements. The proposed MAC scheme efficiently shares channel and contention windows are updated dynamically to set it for different types of categories of traffic in the WMSN.

QoS in terms of packet loss A novel transport layer protocol, i.e., User Datagram Dispatcher Protocol (UDDP), minimizes packet loss ratio in WMSNs [1]. Information is taken from MAC and Application layers which send number of packets and frame types to the transport layer. Similarly, the MAC layer sends MTU to the transport layer for fragmenting a video frame into a corresponding number of packets.

3.3. Reliable communication

The technique presents two geographic forwarding schemes [17]. These schemes are based on the cross-layer design. The first scheme is called Load Balanced Reliable Forwarding (LBRF). The LBRF integrates network and MAC layers for efficient and reliable data delivery in WMSNs. The idea behind this scheme is considering the buffer occupancy level information of those sensor nodes that are used as relay nodes. Every sensor reports buffer occupancy conditions of relay sensors by passively monitoring and storing this information.

A sensor is simply piggybacking information to the MAC layer and neighbors store this information for future use by overhearing. Stored buffer occupancy information of candidate nodes is used for dynamic load balancing.

To enhance the video quality of every video stream in WMSNs and maintaining the balance in video quality among video streams, a new rate control strategy is introduced [31]. The end-to-end data rate is controlled through the cross-layer optimization control algorithm, while the strength of the channel coding and video quality at the physical layer. The SNR is calculated along the forward path at the physical layer and the lowest SNR value is recorded in the packet. This information is sent back to the sender after extracting from the packet at the network layer. Thus, the most recent forward path SNR value is used. Congestion is avoided by computing the end-to-end data rate.

The interaction of physical, network, and transport layers is a cross-layer design that dynamically adjusts the transmission radius and data generation rate for energy efficiency [44]. The optimum transmission radius of sensing devices is chosen at the physical layer. Multiple routing paths are discovered at the network layer. The qualified multiple paths are selected at the transport layer and the information creation rate adjustment is carried out at the Physical layer.

The cross-layer design that cumulatively regulates the video encoding rate, the channel coding rate, and the transmission rate is proposed in [32]. First, the sensed video is encoded to be transmitted over the channels. Second, a rate controller is used which balances the videos streams and maximizes the video quality. Third, the compressed video

Table 2. Cross Layer Optimization for Quality of Service

Author	Layers Crossed	Research Problem	Performance Matrix
[6]	MAC and Network Layers	Resource reservation and routing optimization decreases delay and improves the delivery ratio.	Packet delivery ratio.
[9]	Physical, MAC and Transport Layers	The video traffic is reached from source to destination across multiple path; single path creates delay; multiple path minimizes delay and achieves QoS for video traffic.	Delay / jitter / power consumption.
[34]	Application, Network and MAC	QoS can be achieved in WMSNs by improving the functionalities of communication protocols as the nature of WNS and WMSN is different due to heavy traffic in WMSNs.	Data rate/ SNR/ delay
[15]	Application, Network, MAC, and PHY layers	Traffic in WMSNs is affected by high interference level; minimizing the interference level achieves QoS in WMSNs by classifying the traffic classes.	Delay/ throughput
[25]	Application, Network, MAC, and PHY layers	QoS to heterogeneous applications	Bit error rate/ delay/ throughput
[39]	Physical, Data Link, and Application layers.	Transmissions consume more energy than computations; delivering only meta data can achieve better quality and minimize the end-to-end delay.	Video quality/ End-to-end delay.
[4]	Application, Network, MAC, and Physical Layers	The video streaming consumes more time; improving the video streaming achieve better QoS with respect to video quality.	Data rate/ delay
[24]	Application, Network, MAC, and PHY layers	Adjustment of all network functions by selecting appropriate system parameters so that the network functions are jointly optimized to obtain the best user perceived video quality of cognitive radio	Data rate.
[1]	Application, MAC, and Transport layers	The packet loss rate degrades the QoS in WMSNs; minimizing the packet loss rate improves the QoS.	Frame rate/ delay/ throughput

coding is controlled at the application layer. Last, the adaptive parity at the physical layer and the rate is controlled at the transport layer.

A technique [28] where all layers perform some controlling, for example, on the application layer, optimal routing is achieved by controlling video bit stream and adaptive FEC. The MAC utilization is controlled with the help of congestion aware scheme on the Network layer. The adaptive channel allocation on MAC layer controls channel information.

Some cross-layer techniques [42] combine all layers into a single one for energy efficiency. The objective of this model is the development of a highly reliable communication, avoiding congestion locally, and adaptability in communication decisions. The design principle is a completely unified cross-layered such that both functionalities and information of legacy communication layers are encapsulated in one particular protocol.

Congestion is always caused by the communications waste and the minimization in energy efficiency. Sensor fUzzy-based Image Transmission (SUIT) is a cross-layer based image transport protocol [37]. Congestion estimation is done based on fuzzy logic that

Table 3. Cross Layer Optimization for Communication Reliability

Ref.	Layered Crossed	Problem and Contribution	Performance Metrics
[17]	MAC and Network Layers	Load balancing locally improves the reliability of video across the network. The node dynamically selects the next hope to send data to the sink.	Frame rate/ delay/ latency
[31]	Physical and Transport layer	Controlling the data rate in the network achieves better end-to-end delay and video quality.	Packet error rate/ video rate/ distortion loss
[44]	Transport, Network, and PHY Layers	Reliable video Transmission	Data rate/ delay
[32]	Application, Transport, Physical	Encoding the video can achieve better video quality and minimize the congestion.	End-to-end delay/ video quality.
[28]	Application, Network, MAC, and PHY layers	Reliable video Transmission	MAC utilization is controlled with the help of congestion aware scheme on the network layer; adaptive channel allocation on the MAC layer controls channel information
[37]	Application, MAC, Transport, Network Layers.	Congestion Control	Application layers get the congestion level from the Transport layer for deciding the frame rate.
[42]	All Layers	Efficient and reliable communication	Completely unified cross-layer design in which both functionalities and information of legacy communication layers are encapsulated into a single unified mechanism

decreases the image quality on the fly, which can mitigate the congestion. The idea behind this is to drop some packets of frames that lower the quality but with acceptability. The application layer receives the congestion level from the transport layer while determining the frame rate. A table is generated by the MAC layer that caches the neighbors' buffer occupancy by piggybacking the completeness of the transport layer's buffer into every outgoing packet. The transport layer uses the elements of this table during the fuzzy logic-based congestion estimation. Moreover, at the time of congestion estimation, the transport layer invites the ID of the next hop sensing device from the network layer.

3.4. Security

Security is always the main concern of WSNs and WMSNs. For securing the communication among wireless sensor nodes, the algorithm is based on the information taken from the application layer wherein this information will trigger the algorithm of the provided solution [5], which is a routing algorithm at the network layer. Event notification message is broadcasted by a node that would probably be under attack in some time. The closer nodes will receive a lot of messages, which means that the node is closer with which the incident is going to happen. The counter for that event messages will be used to locate the node in the neighborhood of any node. The algorithm running on the network layer is based on an assumption that there is one or more backup receivers associated with the sink. In this case, the distance is calculated by Euclidean distance formula. Once it is detected that a receiver is under attack, information from the application layer is forwarded and the secondary sink is selected as a receiver. The proposed cross-layer optimization approach [43] is to secure the delivery of images via wireless channels. The parameters– BER, ARQ retry limit, and transmission rate are optimized across the three layers (i.e., PHY, MAC, and APP).

An Efficient Dynamic Selective Encryption Framework (EDES) to ensure security of multimedia traffic in WMSNs is proposed in [33]. Among three security levels, the selection depends upon the energy and QoS parameters, and for the EDES, a cross-layer approach is used to take different parameters from physical, MAC, and upper layers. The capacity metric is used for the evaluation to increase or decrease the level of security. The EDES is a two-step framework wherein in the first step, network performance parameters are assessed by combining the QoS parameters and residual energy to capacity function CAP. This needs parameters from the MAC, network, and upper layers. The second step selects the security level based on CAP.

3.5. Error Correction

The work in [47] presents the forward error control (FEC) scheme based on cross layering of physical layer, transport layer, and application layer. Three approaches which are based on FEC are jointly considered. RVLCs at the application layer and RSC codes at the physical layer are exploited. Redundant packets are generated at the transport layer for the FEC. Strong bit error correction is offered which also overcomes packet loss in multi-hop wireless communications. A cross layered scheme is proposed in [22] that joins the functionalities at different layers (i.e., network, link, and application layers). The scheme avoids the video quality degradation when burst error occurs in the communication. The scheduling mechanism is jointly exploited on these layers.

3.6. Network Resource Management

A cross-layer optimization across three layers, i.e., physical, link, and application layer, manages network resources [7]. This new approach is adapted for the Direct Sequence Code Division Multiple Access (DS-CDMA) using visual sensor networks. Source coding rate, channel coding rate, and power level are also assigned simultaneously to all nodes in the network. The network resources can be optimally allocated to nodes based on two criteria that maximize the video quality of the network as follows:

- Minimize the average end-to-end distortion among nodes.
- Minimize the maximum distortion of the network.

Based on the node density and video quality, the capacity of the WMSNs can be determined to allocate resources.

4. Challenges of Cross-layer Design

Cross-layer design approaches though present much needed improvements to the network and improve the overall Qos. However, some studies have suggested that these techniques

Table 4. Cross Layer Optimization for Security Provision

Ref.	Layers Crossed	Research Problem	Performance Metrics
[5]	Application and Network Layers	Event notification message is broadcast by a node that would probably be under attack in some time. The more closer nodes will receive a lot of messages which means that the node is closer on which the incident is going to happen	Radio range
[43]	Physical, MAC, and Application layer	The images traveling across the network need to be secured.	Retry limit/ transmission rate/ frame length
[33]	MAC, Network, Transport, and Application	Defining the security level for different scenarios can enhance the energy efficiency as well as QoS.	Residual energy/ capacity function

Table 5. Cross Layer Optimization for Error Corrections

Ref.	Layers Crossed	Problem and Contribution	Working of technique
[47]	Application, Transport, and PHY Layers	Strong bit error correction	3 FEC approaches are jointly utilized in the ISCFC, i.e., the RSC codes at the physical layer and the RVLCs at the application layer while generating redundant packets by different random interleaves corresponding to the FEC method of the transport layer.
[22]	Application, Network and Link layer	The burst error in the communication degrades the video quality.	Packet loss rate/ PSNR

violate strong dependencies in layered approaches. Such dependencies have significant role to play once we consider the abstraction and seamless integration they provide. Thus, there are several inherited challenges of wireless technology for multimedia sensor networks.The new challenges for cross-layer approaches make it a difficult task for the designers and developers alike.

While achieving the cross-layer design in order to optimize the overall interactions, the TCP/IP layered architecture is dismantled. This not only disturbs the encapsulation mechanism of a well structured model but also makes it rather difficult to make further improvements in the well structured and planned TCP/IP reference model. In the event of a modification, we must consider almost all layers that are interacting. This might be the most significant issue with the cross-layer design.

One of the prominent challenges is the unavailability of a uniform cross-layer design. There have been many designs proposed but what lacks is a unified approach. This is due to the fact that these designs are normally modelled for specific problems and therefore not easy to have two different designs which can achieve the desired goals. One solution for this problem is to minimize the interactions between cross-layers as much as possible. This way, multiple cross-layer designs may co-exist through a proper standardization.

Generally, cross-layer design tends to resolve a specific issue or target a particular scenario such as video surveillance or vehicular infotainment system. Rather than providing a uniform approach for all situations, it improves only specific aspects of the communication. For example, in vehicular networks, infotainment system requires a lower TCP retry value but in case of text based transmission regarding vehicle's situations, we may require a longer TCP retry value to cope with fragmented or bad networks. Developing a universal cross-layer design is also a significant open research issue. One approach with finding common schemes in various designs can be a good approach towards modeling a uniform design. In case of vehicular networks, we may need to find common grounds in video streaming designs and finally come up with a design which has advantages of each of these designs.

5. Conclusion and Future Directions

Future smart urban spaces are realized through a lot of IoT and WSN applications requiring millions of sensors to work together. These sensor nodes generate a large amount of data. Among these applications, a prominent one is where sensors are deployed for capturing multimedia data such as for video surveillance, patient health monitoring, and traffic management to name a few. In order to have an optimal network throughput, cross-layer design techniques have posed promising results.

Table 6. Cross Layer Optimization for Network Resource Management

Ref.	Layers Crossed	Problem and Contribution	Working of technique
[7]	Application, Link, and Physical Layers	Proper resource management in WMSNs can receive a better video quality.	Source coding rate/ channel coding rate/ transmission power level

In this paper, we have discussed in detail various cross-layer optimization techniques for Multimedia Wireless Sensor Networks (WMSN) on different layers. WMSNs present one of the promising applications for the IoT in Smart Cities. Though similar to generic WSN applications, WMSN leverages the underlying infrastructure to communicate audio visual feedback from a certain environment. Thus, enabling various applications and services in smart cities such as high definition surveillance, traffic control and management, health-care diagnosis and efficient drug delivery, industrial and building automation, surveillance and monitoring to name a few.

Multimedia applications tend to put a considerable burden over the WMSN. This kind of network load can be optimized through several cross-layer optimization techniques. In this paper, we have put forward several such techniques which not only improve the overall interactions but also improve the Quality of Service (QoS). Similarly, we consider that along with the cross-layer optimization, WMSNs also require raw data in-network processing. In future work, we intend to extend the work by developing a test-bed using state of the art Technologies. One of the important future directions would be to fully test the cross-layer optimization techniques through proper testing it under controlled experiments.

Acknowledgments. Authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group no (RG-1439-58)

References

1. Aghdam, S.M., Khansari, M., Rabiee, H.R., Salehi, M.: UDDP: A user datagram dispatcher protocol for wireless multimedia sensor networks. 2012 IEEE Consumer Communications and Networking Conference, CCNC'2012 pp. 765–770 (2012)
2. Akyildiz, I.F., Melodia, T., Chowdhury, K.R.: A survey on wireless multimedia sensor networks. Computer Networks 51(4), 921–960 (2007)
3. Almalkawi, I.T., Zapata, M.G., Al-Karaki, J.N., Morillo-Pozo, J.: Wireless multimedia sensor networks: Current trends and future directions. Sensors 10(7), 6662–6717 (2010)
4. Andreopoulos, Y., Mastronarde, N., Van Der Schaar, M.: Cross-Layer Optimized Video Streaming Over Wireless Multihop Mesh Networks. IEEE Journal on Selected Areas in Communications 24(11), 2104–2115 (2006), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1717624
5. Approach, C.l., Faghani, M.R., Nguyen, U.T.: Incident-Driven Routing in Wireless Sensor. Computing pp. 402–406 (2011)
6. Aune, F.: Cross-Layer Design Tutorial. Norwegian University of Science and Technology, p. 35 pp. (2004)
7. Bentley, E.S., Matyjas, J.D., Medley, M.J., Kondi, L.P.: Wireless visual sensor network resource allocation using cross-layer optimization. Proceedings of SPIE 7257(0704), 72570G–72570G–10 (2009), http://link.aip.org/link/PSISDG/v7257/i1/p72570G/s1{&}Agg=doi
8. Bouabdallah, F., Bouabdallah, N., Boutaba, R.: Cross-layer design for energy conservation in wireless sensor networks. In: IEEE International Conference on Communications. pp. 1–6 (2009)
9. Campelli, L., Akyildiz, I.F., Fratta, L., Cesana, M.: A cross-layer solution for ultrawideband based wireless video sensor networks. In: GLOBECOM - IEEE Global Telecommunications Conference. pp. 240–245 (2008)

10. Cucchiara, R.: Multimedia surveillance systems. In: Proceedings of the third ACM international workshop on Video surveillance & sensor networks - VSSN '05. p. 3 (2005), http://portal.acm.org/citation.cfm?doid=1099396.1099399
11. Din, I., Kim, B.S., Hassan, S., Guizani, M., Atiquzzaman, M., Rodrigues, J.: Information-centric network-based vehicular communications: overview and research opportunities. Sensors 18(11), 3957 (2018)
12. Din, I.U., Asmat, H., Guizani, M.: A review of information centric network-based internet of things: communication architectures, design issues, and research opportunities. Multimedia Tools and Applications pp. 1–16 (2018)
13. Din, I.U., Guizani, M., Hassan, S., Kim, B.S., Khan, M.K., Atiquzzaman, M., Ahmed, S.H.: The internet of things: A review of enabled technologies and future challenges. IEEE Access (2018)
14. Din, I.U., Guizani, M., Kim, B.S., Hassan, S., Khan, M.K.: Trust management techniques for the internet of things: a survey. IEEE Access (2018)
15. Farooq, M.O., Kunz, T., St-Hilaire, M.: Cross layer architecture for supporting multiple applications in Wireless Multimedia Sensor Networks. In: IWCMC 2011 - 7th International Wireless Communications and Mobile Computing Conference. pp. 388–393 (2011)
16. Iqbal, J., Khan, M., Talha, M., Farman, H., Jan, B., Muhammad, A., Khattak, H.A.: A generic internet of things architecture for controlling electrical energy consumption in smart homes. Sustainable Cities and Society 43, 443 – 450 (2018), http://www.sciencedirect.com/science/article/pii/S2210670718300489
17. Isik, S., Donmez, M.Y., Ersoy, C.: Cross layer load balanced forwarding schemes for video sensor networks. Ad Hoc Networks 9(3), 265–284 (may 2011)
18. Khan, O.A., Shah, M.A., Din, I.U., Kim, B.S., Khattak, H.A., Rodrigues, J.J., Farman, H., Jan, B.: Leveraging named data networking for fragmented networks in smart metropolitan cities. IEEE Access (2018)
19. Khattak, H.A., Ruta, M., Di Sciascio, E.: Coap-based healthcare sensor networks: a survey. In: 11th International Bhurban Conference on Applied Sciences & Technology, Islamabad. (2014)
20. Kumar, K., Yang, E.H.: On the growth mode of two-lobed curvilinear graphene domains at atmospheric pressure. Scientific Reports 3 (2013), http://www.springer.com/engineering/signals/book/978-0-387-39022-2
21. Lee, K., Lee, H., Lee, S.: A cross-layer scheduling technique for maximizing end-to-end video quality. In: International Conference on Information Networking 2011, ICOIN 2011. pp. 410–413 (2011)
22. Liu, H., Zhao, S., Qian, J.: Quality-aware cross-layer scheduling for robust video streaming over wireless channels. Journal of Electronics 28(3), 303–312 (2011)
23. Luo, H., Ci, S., Wu, D.: A Cross-Layer Design for the Performance Improvement of Real-Time Video Transmission of. IEEE Transactions on Circuits and Systems 21(8), 1040–1048 (2011)
24. Luo, H., Ci, S., Wu, D.: A cross-layer design for the performance improvement of real-time video transmission of secondary users over cognitive radio networks. IEEE Transactions on Circuits and Systems for Video Technology 21(8), 1040–1048 (2011)
25. Melodia, T., Akyildiz, I.: Cross-layer QoS-aware communication for ultra wide band wireless multimedia sensor networks. Selected Areas in Communications, ... 28(5), 653–663 (2010)
26. Murthy, J.K., Kumar, S., Srinivas, A.: Energy Efficient Scheduling in Cross Layer Optimized Clustered Wireless Sensor Networks. International Journal of Computer Science and Communication 3(1), 149–153 (2012)
27. Muzammal, S.M., Shah, M.A., Khattak, H.A., Jabbar, S., Ahmed, G., Khalid, S., Hussain, S., Han, K.: Counter measuring conceivable security threats on smart healthcare devices. IEEE Access 6, 20722–20733 (2018)
28. Oh, B., Chen, C.: A cross-layer approach to multichannel MAC protocol design for video streaming over wireless ad hoc networks. Multimedia, IEEE Transactions on 11(6), 1052–1061 (2009)

29. Pascale, A., Nicoli, M., Deflorio, F., Dalla Chiara, B., Spagnolini, U.: Wireless sensor networks for traffic management and road safety. IET Intelligent Transport Systems 6(1), 67 (2012), http://digital-library.theiet.org/content/journals/10.1049/iet-its.2010.0129
30. Phan, K.T., Fan, R., Jiang, H., Vorobyov, S.A., Tellambura, C.: Network lifetime maximization with node admission in wireless multimedia sensor networks. IEEE Transactions on Vehicular Technology 58(7), 3640–3646 (2009)
31. Pudlewski, S., Melodia, T.: DMRC: distortion-minimizing rate control for wireless multimedia sensor networks. Mobile Adhoc and Sensor Systems, ... (2009)
32. Pudlewski, S., Prasanna, A., Melodia, T.: Compressed-sensing-enabled video streaming for wireless multimedia sensor networks. IEEE Transactions on Mobile Computing 11(6), 1060–1072 (2012)
33. Rachedi, A., Kaddar, L., Mehaoua, A.: EDES - Efficient dynamic selective encryption framework to secure multimedia traffic in Wireless Sensor Networks. IEEE International Conference on Communications (2012)
34. Shah, G.A., Liang, W., Akan, O.B.: Cross-layer framework for QoS support in wireless multimedia sensor networks. IEEE Transactions on Multimedia 14(5), 1442–1455 (2012)
35. Shi, H., Hou, K.M., Diao, X., Xing, L., Li, J.J., De Vaulx, C.: A wireless multimedia sensor network platform for environmental event detection dedicated to precision agriculture. arXiv preprint arXiv:1806.03237 (2018)
36. Siddiqa, A., Shah, M.A., Khattak, H.A., Akhunzada, A., Ali, I., Razak, Z.B., Gani, A.: Social internet of vehicles: Complexity, adaptivity, issues and beyond. IEEE Access 6, 62089–62106 (2018)
37. Sonmez, C., Isik, S., Donmez, M.Y., Incel, O.D., Ersoy, C.: SUIT: A cross layer image transport protocol with fuzzy logic based congestion control for wireless multimedia sensor networks. In: 2012 5th International Conference on New Technologies, Mobility and Security - Proceedings of NTMS 2012 Conference and Workshops. pp. 1–6 (2012)
38. Srivastava, V., Motani, M.: Cross-layer design: A survey and the road ahead. IEEE Communications Magazine 43(12), 112–119 (2005)
39. Studi, D., Ferrara, D.I., Di, D., In, R., Dell, S.: Cross-layer Optimization for Video Delivery over Wireless Networks. IEEE Access 25(4), 1–10 (2013)
40. Suh, C., Ko, Y.B.Y., Son, D.D.M., Salarian, H., Chin, K.W., Naghdy, F.: An energy-efficient mobile-sink path selection strategy for wireless sensor networks. Vehicular Technology, IEEE Transactions on 63(5), 2407–2419 (2014)
41. Tekin, N., Erdem, H.E., Gungor, V.C.: Analyzing lifetime of energy harvesting wireless multimedia sensor nodes in industrial environments. Computer Standards & Interfaces 58, 109–117 (2018)
42. Vuran, M.C., Akyildiz, I.F.: A Cross-Layer Protocol for Wireless Sensor Networks A Cross-Layer Protocol for Wireless Sensor Networks. Information Sciences (2006)
43. Wang, W., Peng, D., Wang, H., Sharif, H.: A cross layer resource allocation scheme for secure image delivery in wireless sensor networks. In: Proceedings of the 2007 international conference on Wireless communications and mobile computing. p. 157 (2007), http://portal.acm.org/citation.cfm?id=1280940.1280973
44. Wang, Y.: Cross Layer Optimization for Data Gathering in Wireless Multimedia Sensor Networks within Expected Network Lifetime. Journal of Universal Computer Science 16(10), 1343–1367 (2010)
45. Yan, J., Zhou, M., Ding, Z.: Recent advances in energy-efficient routing protocols for wireless sensor networks: A review. IEEE Access 4, 5673–5686 (2016)
46. Yan-liang Jin, Zhen Zhang, Hai-tao Liu: Contention window based QoS DCC-MAC for wireless multimedia sensor networks. In: IET International Communication Conference on Wireless Mobile & Computing (CCWMC 2009). pp. 201–204. IET

(2009), http://digital-library.theiet.org/content/conferences/10.1049/cp.2009.1925
47. Yang, Y., Chen, Y., Yi, W.: Cross-layer forward error control for reliable transfer in wireless multimedia sensor networks. Proceeding of The 7th IEEE Consumer Communications and Networking Conference (CCNC '10) pp. 1–2 (2010)
48. You, L., Liu, C.: Robust cross-layer design of wireless multimedia sensor networks with correlation and uncertainty. Journal of Networks 6(7), 1009–1016 (2011)
49. Zacharias, J.: Sensor Networks, Where Theory Meets Practice, vol. 10. Springer (2009), http://link.springer.com/10.1007/978-3-642-01341-6
50. Zhao, H.T., Zhang, R., Song, R.F., Zhang, H., Dong, Y.N., Tian, F.: A capacity-aware cross-layer QoS optimization scheme for wireless multimedia networks. In: 2011 International Conference on Wireless Communications and Signal Processing, WCSP 2011 (2011)

Hasan Ali Khattak received his PhD in Electrical and Computer Engineering degree from Politecnico di Bari, Bari, Italy in April 2015, Master's degree in Information Engineering from Politecnico di Torino, Torino, Italy, in 2011, and B.CS. degree in Computer Science from University of Peshawar, Peshawar, Pakistan in 2006. He is currently serving as Assistant Professor of Computer Science since January 2016. His current research interests focus on Distributed system, Web of Things and Vehicular Ad Hoc Networks, and Data and Social Engineering for Smart Cities. He is involved in a number of funded research projects in Internet of Things, Semantic Web and Fog Computing while exploring Ontologies, Web Technologies using Contiki OS, NS 2/3 and Omnet++ frameworks.

Zoobia Ameer received her Ph.D. in Bio-molecular Magnetic Materials from University of Salento, Lecce. Her Research work has been centered around characterization and fabrication of DNA based sensors which can be used for invasive diagnosis as well as drug delivery. She is currently work as Assistant Professor of Physics Department, Shaheed Benazir Bhutto Women University Peshawar. Where is involved in a number of research activity for optimization through various bio-inspired techniques and study of sensor network integration through in-body and on-body monitoring of patients.

Ikram Ud Din received his PhD degree in Computer Science from the School of Computing, Universiti Utara Malaysia (UUM). He received his MSc in Computer Science and MS in Computer Networking from the Department of Computer Science, University of Peshawar, Pakistan in 2006 and 2011, respectively. Currently, he is a Lecturer in the Department of Information Technology, The University of Haripur, Pakistan. He also served as the IEEE UUM Student Branch Professional Chair. His current research interests include resource management and traffic control in wired and wireless networks, mobility and cache management in Information-Centric Networking, and Internet of Things.

Muhammad Khurram Khan is currently working as a full professor at the Center of Excellence in Information Assurance, King Saud University, Saudi Arabia. He has published more than 300 papers in international journals and conferences. In addition, he is an inventor of 10 U.S./PCT patents. He is the Editor-in-Chief of a well reputed journal, Telecommunication Systems (Springer), and a member of several editorial boards. His research interests include cybersecurity, biometrics, multimedia security, and digital authentication. His profile can be visited at http://www.professorkhurram.com

Received: November 15, 2018; Accepted: January 5, 2019.

Computer Science and Information Systems 16(1):19–44 https://doi.org/10.2298/CSIS170703035R

SOCA-DSEM: a Well-Structured SOCA Development Systems Engineering Methodology

Laura C. Rodriguez-Martinez[1], Hector A. Duran-Limon[2], Manuel Mora[3], and Francisco Alvarez Rodriguez[3]

[1] Tecnológico Nacional de México/I.T. Aguascalientes, Av. Adolfo López Mateos 1801 Ote., Bona Gens, CP 20256, Aguascalientes, Ags., Mexico
lrodriguez@mail.ita.mx
[2]CUCEA, University of Guadalajara, Periférico Norte 799, Zapopan, Jalisco, Mexico
hduran@cucea.udg.mx
[3] Autonomous University of Aguascalientes, Av. Universidad 940, Ciudad Universitaria, CP 20131, Aguascalientes, Ags., Mexico
mmora@correo.uaa.mx, fjalvarez@correo.uaa.mx

Abstract. Service-oriented Software Engineering (SOSE) is a software engineering paradigm focused on Service-oriented Computing Applications (SOCAs), for what SOCA development methodologies are required. Recent studies on SOCA development methodologies revealed theoretical and practical deficiencies. Thus, academicians and practitioners must adapt development methodologies from other paradigms or use the available partial SOCA development methodologies. Also, since the high acceptance of agile approaches, we claim new well-structured and balanced agility-rigor methodologies are required. Then, this paper proposes a new ***SOCA Development Systems Engineering Methodology***, including its description, the explanation of its theoretical foundations and the illustration of its use with a prototype of a running example. Two pilot empirical evaluations on usability metrics are also reported. Findings support both theoretical adequacy and positive perceptions from the evaluators. While further empirical tests are required for gaining more conclusive evidences our preliminary results are encouraging.

Keywords: Service-oriented Software Engineering (SOSE), Service-oriented Computing Application (SOCA), Model-Driven Architecture (MDA), RUP-SE, MBASE, agility-rigor balance.

1. Introduction

This study is located in the research area of Software Engineering (SE), in the part referred to as the Service-oriented Software Engineering (SOSE) paradigm, which can be seen as an evolution of the previous Object-oriented (OOSE) and Component-based Software Engineering (CBSE) paradigms [25], but that addresses loosely coupled distributed systems. The SOSE paradigm aims at producing high-quality Service-oriented Computing Applications (SOCAs). These SOCAs can be defined as software systems built through a suite of software services. Software services are auto-contained

computational work units provided by internal or external parties (service providers) to other parties (service customers) [37]. Software services have been also defined as "*autonomous, platform-independent entities that can be described, published, discovered, and loosely coupled in novel ways*" [50]. Different from the object- and component-oriented paradigm, SOSE focuses on coarse-grained distributed services that are loosely coupled. Nevertheless, both the object- and component-oriented paradigms are still helpful to construct independent services, which are the building blocks of a SOCA.

The SOSE research stream investigates "*systematic, disciplined and quantifiable approaches to develop service-oriented systems*" [25]. For this aim, SOCA engineering methodologies are required [37], [25]. In particular, [37] indicates that one of the core research challenges in the SOSE paradigm is defining "*development processes and methodologies for service-oriented systems".* Some SOCA development methodologies have been proposed such as [39][40][51]. However, these methodologies have the following limitations. Current methodologies emphasize elaborating on the system architecture, the use of specific technology for constructing services -like BPEL (Business Process Executable Languages), WSDL (Web Service Definition Language), Web Services-, as well as aligning on business only in the requirements Phase (some of them include a market study and requirements engineering). However, such methodologies do not consider the analysis and design of services, composition, orchestration, and choreography –i.e. regarding the latter, they do not facilitate the transformation of the representation of the system in the business domain to the final implemented application-. These limitations make it difficult to narrow the conceptual gap between the problem (or business) domain and the system implementation (or application) domain [18][61].

We believe the reason there is a lack of more recent methodologies for SOCA development is that only until recently it has been recognized by academicians and Industry that Service-oriented Computing (SOC) [50][69] is not only a fashionable technology but a real need. For instance, only in recent years there have been some reviews of SOCA methodologies [25][5][34]. Also, the increasing need for business-IT alignment [24] reflects the fact that SOCA development is only recently gaining more attention.

In this paper, we propose a **SOCA D**evelopment **S**ystem **E**ngineering **M**ethodology, called ***SOCA- DSEM***, that considers: (i) concerns of business, architecture and system (application) along the whole development process; and (ii) the analysis and design of services, composition, orchestration, and choreography. In addition, our proposal is a well-structured methodology meaning that it covers all the software system development life-cycle in a well-structured manner i.e. our methodology includes phases, activities, models, and products as well as roles and an iterative development process. Our methodology focuses on defining the structural activities, the execution order of these activities and their associated products. In the case of complex systems, our methodology can employ specific methods. This is similar to other methodologies or models such as RUP and Spiral that enable the use of specific methods. For example, a specific method regarding graphical user interface design can be applied as well as others more specialized methods like [7] that considers the architectural principles to develop SOCAs.

The design of our development methodology is theoretically underpinned by four core design building blocks: 1) Service-oriented Computing (SOC) [50][69], 2) Model-

Driven Architecture (MDA) [42], 3) two popular development systems engineering methodologies (RUP-SE [10] and MBASE [12]), and 4) recommendations for agility-rigor balance [9]. Currently, our approach does not have tool support to automate model-to-model transformations, rather, these transformations need to be performed manually.

We illustrate the use of our methodology with a running example consisting of a service-oriented system that performs a quality-assessment of academic-courses. We evaluated our proposed methodology by employing two empirical evaluations with a group of 15 international academicians on Software Engineering, and a group of 32 MSc partial-time students from an MSc course on an IT program in Mexico. In the first evaluation, the subjects reviewed the methodology for rating the level of theoretical validity. In the second evaluation, the subjects rated the levels of usefulness, ease of use, compatibility, result demonstrability, and behavioral intention of use. These results show both theoretical adequacy of our methodology and adequate scores on the perceived metrics collected from the MSc course on IT.

The remainder of this paper is structured as follows. In Section 2, the theoretical basis for the design of our methodology is reported. In Section 3, the methodology is described and its utilization is illustrated with a running example in Section 4. In Section 5, the results of the two empirical evaluations are reported. Finally, in Section 6, both the conclusions and recommendations for further research are presented.

2. Review of Theoretical Basis

This section describes the aforementioned theoretical basis-core design building blocks- of our proposed methodology namely, Service-Oriented Computing (SOC), Model-Driven Architecture (MDA), two popular development methodologies -RUP-SE and MBASE-, and recommendations for achieving agility-rigor balance.

2.1 Service-Oriented Computing (SOC). The implementation of the Service-Oriented paradigm is addressed by Service-oriented Computing (SOC), which can employ the Object-oriented and Component-based paradigms [69], [50] to build individual coarse-grained services. SOC approaches target loosely distributed systems and aims at building applications that integrate the interoperability of such coarse-grained services. Thus, SOC approaches involve the study and advances referring to building services in a way that multiple SOCAs can simultaneously use the same set of services [69]. The set of services are platform-independent for increasing the SOCA inter-operability [69]. The SOC paradigm also advocates providing services in different heterogeneous and external computational platforms [25]. Moreover, some SOCAs can dynamically re-configure their contracted services (e.g. a service provider can be replaced by an alternative service provider at runtime) [25]. Furthermore, the utilization of newer SOC programming technologies can be used for constructing SOCAs -e.g. we use the Service Component Architecture (SCA) hosted in a Java interface development environment for the running example presented in this paper [6].

A SOCA can be defined as a distributed and loosely coupled software system represented and constructed as a main customer control code calling a suite of computational services [3]. According to [16], services -as a broad concept- can be classified in *Business Services* and *Computing Services*. In turn, *Computing Services*

can be classified as a *Business Computing Service* when it implements a Business Service or as an *Information and Communication Technology Computing Service* when it provides a service to another software system. Based on these concepts, a SOCA can be also conceptualized as a composition of *Business Computing Services* [56].

2.2 Model-Driven Architecture (MDA). MDA [42] and Model-Driven Development (MDD) [18] are part of Model-Driven Engineering (MDE) [18]. They address the conceptual gap between the problem domain and the implementation domain [18] in the context of the transformation of several models that are constructed during the process of software development [61]. One of the main differences among MDA and MDD is that the former advocates the use of a general-purpose modeling language like UML and the Meta Object Facility (MOF) [49] for meta-modeling whereas MDD advocates employing domain specific languages instead [54]. In addition, both MDD and MDA study models at different abstraction-levels, but the MDA approach defines four core types of models based on four similar related views: 1) Computing Independent Model (CIM); 2) Platform Independent Model (PIM); 3) Platform Specific Model (PSM); and 4) Executable Platform Model (executable PM). The CIM focuses on the system requirements and its environment without any technical consideration of the target computing platform. The PIM focuses on detailed system operational specifications without involving a specific platform. The PSM focuses on the specific computational implementation and how the system uses it. The executable PM focuses on both the final runtime model and the specific parts and provided capabilities of the computational platform. Based on an analysis of the main MDA literature [42] - regarding service-orientation- the following design recommendations can be derived. A software product/service can be developed by employing several transformations and refinements from one model to other models. The models must foster the portability, interoperability, and reusability of the final software product/service. Moreover, a software product/service, should be developed by phased and iterative development engineering methodologies (by applying successive model transformations and refinements). Finally, a software product/service must be designed by using several types of models (i.e. it is unlikely we are able to design a software product/service by using only a single or few models).

This paper is based on our previous work where we defined a generic MDA-based development engineering methodology [56] that is derived from the main core SOCA literature. Such a core SOCA literature involves MDA principles [42], a Service-oriented Analysis and a Design approach [69], a set of software service conceptualizations [2], and a three-phased software development engineering macro model [55]. This paper extends and adapts such a generic MDA-based development engineering approach [56] in two dimensions. The first dimension regards the levels of abstraction of *Business, Architecture* and *Application*. The second dimension involves the phases *Requirements, Design, Construction* and *Operation* of a generic SOCA development engineering methodology [55]. Where, the Business level matches the CIM, the Architecture level matches the PIM whereas the Application level matches both the PSM and the executable PM.

Hence, our previous work provides theoretical insights for helping on the design of specific SOCA development engineering methodologies. For the practical purposes of this paper, we found that SOCA development activities are implicitly related one-by-one to a product that we can name with the same name of the activity, e.g. the activity A1 “Business Process Modeling” produces a “Business Process Model”.

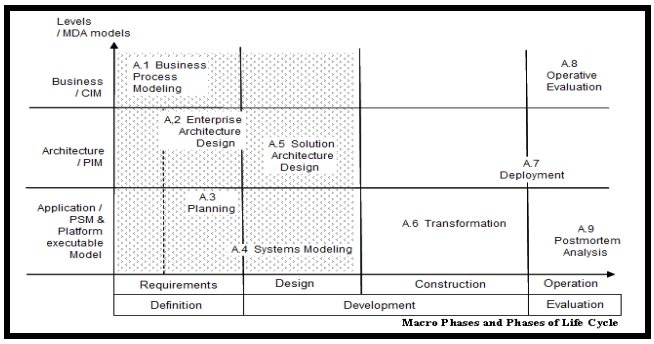


Fig. 1. Generic MDA-based Software Development Engineering Methodology for SOCAs (Quoted by [56])

Then, in order to define coarse-grained blocks of the structure of our SOCA methodology, from the MDA approach [56], we took the activities labeled as *A1..A9* (see Fig. 1) each one associated with a specific product. We also took from [56] the macro-phases *Definition*, *Development and Evaluation*, and the phases *Requirements*, *Design*, *Construction*, and *Operation*. At the same time, such an MDA approach [56] integrates some knowledge about software system development processes for SOCA, including RUP processes and promoting MBASE well-structured processes.

2.3 Software development engineering methodologies. Development methodologies combine the best practices from their predecessors [4], [55]. Two exemplary cases of these methodologies are the RUP for Systems Engineering (RUP SE) [10] and the Model-based System Architecting and Software Engineering (MBASE) [12]. They both include the flexibility of the iterative and incremental process and relevant activities for software projects like Risk Management and Business Modeling activities. RUP SE [10] is a methodology that is an enhancement of the well-known object-oriented or component-based RUP methods, and incorporates additional diagrams and techniques that are not reported in the standard RUP such as: context diagrams, system use cases with services, business process, process interaction diagrams, and system and sub-system architectures. RUP SE combines activities (that can be repeated in an iterative way), within phases that end with specific control milestones. The phases (Inception, Elaboration, Construction, and Transition) and the activities (Business Modeling, Requirements, Analysis and Design, Implementation, Test, Deployment, Configuration and Change Management, Project Management, and Environment Management) in the RUP SE are the same ones reported in standard RUP. In particular, the Business Modeling activity is mandatory and not optional as in normal RUP, and it is augmented with additional analysis diagrams [10]. Regarding MBASE, this methodology provides a system-based approach for developing integrated software systems [12]. MBASE is based on the Boehm's Win-Win Spiral Model [8] which proposes a system's specification with six elements: Operational Concept Description, System and Software Requirements Definition, System and Software Architecture Description, Life Cycle Plan, Feasibility Rationale Description, and Risk-driven prototypes. MBASE combines an incremental approach (with Inception, Elaboration, Construction, and Transition phases) with an iterative strategy (for developing the six core elements previously reported) for developing software systems. Similar to RUP SE, each phase ends with specific control milestones. In MBASE there are three core control milestones: Life-

Cycle Objectives, Life-Cycle Architecture and Plan, and an Initial Product ready for Operational use. In MBASE, the Business Modeling and Risk Analysis activities are also considered core activities and not optional ones as reported in other methodologies.

Hence, these two methodologies can be considered well-defined -according to the expected structure of phases, activities, products, roles and tools- [48], and well-balanced on agility-rigor issues [56]. Both methodologies are considered suitable for general purpose development. Therefore, for their use in SOCA development, they must be adapted. Nevertheless, we consider these two methodologies useful as a basis for generating new methodologies for SOCAs. Finally, we adopted the incremental and iterative philosophies of RUP and MBASE for favoring agility in the process. Specifically, from RUP we took the loops to control the development project, in particular inception, elaboration, transition, and deployment. Moreover, from RUP and MBASE we took a similar idea to milestones to implicitly control the project progress where the documents give evidence of the end of a stage or loop.

2.4 Agility-Rigor balance recommendations. Boehm and Turner [9] recommend employing an Agility-Rigor balance to the software development processes. These authors define a comparison framework that enables or facilitates the evaluation of software development processes. This framework defines three factors and divides each one in some characteristics, which can be measured for a software-system development life-cycle. Each characteristic can be assessed with the following values: non-existent, medium, and high. This indicating that a characteristic is not covered, partially covered, and fully covered, respectively. For an agility-rigor balance it is expected the assessed values to be in average medium (i.e. partially covered). Such factors and characteristics are the following:

*Factor 1. Level of Concerns***.** This factor identifies the organizational scope of use for which a development life-cycle proposes specific guides for its use (not assumptions). Five possible concerns are the following. The *Business Enterprise concern* indicates us whether the methodology can be used within the organizational boundaries (not necessarily with co-located teams) and whether the methodology provides guidance to potentiate the inter-team communication by suggesting information sharing strategies, use of communication tools, and ways to gradually introducing rigor in project work in order to support distributed development. The *Business System concern* indicates if the methodology is easier to be applied to business systems than to engineering or scientific applications. The *Multiteam Project concern* indicates if the methodology can be used by co-located teams. The *Single-team* concern indicates if the methodology can be used by a single team located in a single office building. Lastly, the *Individual Project concern* indicates whether the methodology can be used by a single individual. The less restrictions we have the more agility we obtain.

Factor 2. Life-Cycle Coverture. This factor identifies the life-cycle (structural) activities for which the methodology proposes specific guidance. This factor evaluates the life-cycle coverture. The activities to evaluate it are: Concept Development, Requirements, Design, Development, and Maintenance. For each one it is verified if the life-cycle activities are covered by the development process of the methodology under evaluation. Here a more complete coverture is considered to have more rigor, i.e. less agility.

Factor 3. Sources of Constraints. Each development life-cycle proposes constraints to the implementer. When such a life-cycle is less constrained, it is considered more agile. Five possible sources of project-areas constraints are considered by this factor:

management processes, technical practices, risk/opportunity considerations, measurement practices, and the customer interface.

Boehm and Turner [9] employed such a framework and evaluated RUP and other methodologies (that are either considered agile or disciplined), finding that RUP is a good example of agility-rigor balance, i.e. RUP is balanced on agility-rigor [9]. Because of this, we took RUP as a reference. Then, we included an agility-rigor balance in our methodology by considering the three factors - *level of concern, life cycle coverture*, and *source of constraints*- defined by [9]. Regarding the level of concern and life cycle coverture factors, our methodology, similar to RUP, partially covers all stages of the development process and proposes local work for both single and multiple teams. With respect to the source of constraints factor, RUP partially covers enterprise restrictions, technical-practices restrictions, and measurement-practices restrictions. Moreover, customer restrictions are not covered whereas risk/opportunity restrictions are fully covered. On the other side, since our methodology focuses on the development process, our methodology fully covers technical-practices and customer-interfaces restrictions. Nevertheless, in order to promote agility, our methodology does not cover risk/opportunities or measurement-practices restrictions (which mainly focus on the management process).

3. Structure of the proposed SOCA Methodology

Our methodology defines the structure of the development-process [62] i.e. such a methodology defines the components (phases, activities, artifacts, and roles) [48], and the way to iterate. The roles are responsibilities, assigned to agents (a person or teams), involving the activities of the processes. As said earlier, our methodology is based on the MDA approach, thus it is required a set of "models" as part of the structure of the methodology. The models in the methodology involve a set of products [10]. We use some specific notations to construct the artifacts (i.e. products) in our running example, however, the methodology enables the use of any other notations. In general, our methodology includes structural activities, but also, some umbrella activities [62] like planning and evaluation are considered.

The steps of our methodology are divided in four phases:

1. Requirements. This phase includes activities related to CIM modeling, CIM evaluation, and life cycle planning.

2. Design. This phase concerns PIM modeling, PIM evaluation, PSM modeling, and PSM-evaluation.

3. Construction. This phase includes the activities of construction planning, executable PM modeling, evaluation, and deployment.

4. Operation. The activities included in this phase are evaluation planning, evaluation of the executable PM operation, results specification, and analysis for evolving the executable PM.

Our methodology involves a number of model-to-model transformations, which currently have to be carried out manually. Importantly, our methodology is not constrained to specific tools to build the artifacts. Each one of the phases is described in turn below.

As said earlier, other methods can be employed within our methodology. For example, for developing the system architectural design we can apply the method in [7]. Other examples involve methods that target specific software engineering areas such as requirements engineering and human computer interactions. In any case, our methodology gives explicit guidance on the activities that need to be carried out as well as their associated artifacts, this independently of the particular methods being used.

3.1 Requirements. In the Requirements phase, the entire development process is planned, the general requirements and the scope of the system are defined, and the expected system delivery is also planned. In the Requirements phase, a requirement elicitation is carried out, the objectives of the software system are specified based on requirement agreements, and finally, both the software systems scope and the expected software systems delivery are planned. The activities involved in this phase are the following:

1. CIM Modeling. The CIM (according to our SOCA Development Systems Engineering Methodology) corresponds to the Business Process Model. In this activity, the elicitation and specification of general requirements are elaborated for creating the Computation Independent Model (CIM). This activity generates the following products: (i) Service Specification that includes the Process and work flows specification and the Specification of the Services and their conceptual definitions, and (ii) Data Semantic Model of the Enterprise architecture that aims at defining data to narrow the gap between the business and technical domain. The suggested artifacts for representing the CIM are: System Context Diagram, Work System Snapshot, System Responsibility Table, Business Process Diagram, and Enterprise data view.

2. CIM Evaluation. In this activity, several stakeholders participate for detecting and correcting errors in the previous activity.

3. Life Cycle Planning. In this activity, a Life Cycle Plan is elaborated. Such a plan must contain: (i) a system scope and delivery plan, and (ii) a development plan.

At the beginning of the execution of the CIM Modeling activity of our methodology, we can apply any specific method for requirements elicitation that we can take from the requirements engineering research area such as [62][64]. Next, we can construct the specific products of the CIM.

3.2 Design. In the Design phase, the software system is designed, which implies the transformation of the CIM into the PIM takes place, where the PIM realizes the Solution Architecture of the System. The transformation of the PIM into the PSM is also carried out in this phase. The activities of this phase are the following:

1. PIM Modeling. In this activity, the CIM into PIM transformation is executed as the design of the Solution System Architecture. This activity must generate the following products: (i) Service Provider Architecture Design that also contains the definition of services operations; and the specification of the communication among services (service orchestration); (ii) Service Consumer Architecture Design that also contains processes with a business semantic specification and the workflows specification (service choreography); and (iii) the Database Design. The suggested artifacts for representing the PIM are: Service model diagram, Join Realization Table that represents the Service-Choreography, Service-Orchestration Definition, and the E-R System data model.

2. PIM Evaluation. In this activity, the PIM is reviewed to identify and correct critical errors.

3. PSM Modeling. In this activity, the PIM into PSM transformation, for defining the System Model, is realized. The PSM is composed of the products: (i) Service Provider Design that contains the service specifications of implementations -definition of service components, service design and service interfaces-, (ii) Service Consumer Design that contains the definition of the technology (languages) to be used for the implementation of the executable PM, and their collaboration, (iii) design of the database scheme, and (iv) User interface design. The suggested artifacts for representing the PSM are: Creation script of the database scheme in a Data Definition Language (DDL), Data Access Definition (e.g. XML-based, JDBC), Service realization diagrams (interfaces, operations = UML interface stereotypes + component-based specification of realization), and User interface design.

4. PSM Evaluation. In this activity, the PSM is reviewed to identify and correct critical errors.

At the end of the PIM Modeling activity we can employ a specific architectural design method such as [7][13]. Regarding user interface design, which is part of the PSM Modeling activity, we can use guidelines or methods such as [15][22] proposed in the human computer interface knowledge area.

3.3 Construction. In the Construction phase, the software system is constructed, and released, this involves the transformation of the PSM into the Executable Platform Model (Executable PM). The activities of this phase are the following:

1. Construction Planning. In this activity, a plan for transforming the PSM to the executable PM is elaborated.

2. Executable PM Modeling. In this activity, the PSM is transformed into the executable PM, and each service along with its orchestration is implemented and tested. The artifacts suggested to use in this phase are: Services implemented and orchestrated, Services composed and published with a Service Component Architecture, and Choreography of services in work flows with HTML.

3. Executable PM Evaluation. In this activity, the executable PM is evaluated to detect and fix errors.

4. Deployment. In this activity, the executable PM is released.

3.4 Operation. In the Operation phase, the software system (the executable PM) is monitored and evaluated. Hence, monitoring and evaluation logs are analyzed for correcting unexpected problems and for functionality improvement or functionality increment. The activities of this phase are the following:

1. Evaluation Plan. In this activity, a plan is elaborated for evaluating the software system operation.

2. Evaluation of the executable PM operation. In this activity, the software system operation is monitored and evaluated.

3. Results Specification. In this activity, the evaluation of the software system is registered.

4. Analysis for evolving the executable PM. In this activity, the results are analyzed to make decisions for the purpose of system evolution.

3.5 Description of Roles. In our methodology, the roles are defined as development teams. Similar to RUP-SE's role teams, each one of them has a project administrator and a technical leader. Table 1 shows the roles of our methodology.

Table 1. Description of Roles of the SOCA Development Systems Engineering Methodology

Team Roles	Main Responsibility
Enterprise modeling	To conduct the Requirements phase activities. To elaborate the CIM MDA model.
System architecture	To conduct the high-level Design phase activities. To elaborate the PIM MDA model.
System and Services modeling	To conduct the low-level Design phase activities. To design the composition of services. To elaborate the PSM MDA model.
System and Services development	To conduct the Construction phase activities. To elaborate the composition of services. To elaborate the executable PM MDA model.
Deployment and Operation	To conduct the Operation phase activities. These activities must set the system in operational mode, perform required user training, and guarantee the expected functionality.
Evaluation and Evolution	To conduct the Support loop activities. These activities must guarantee the continuous operation of the system. Such activities are also responsible for monitoring, evaluating, and evolving the system.

Each role is also responsible of the administration of its own activities and phases. However, an administration team is posed for coordinating all the administration activities and ensuring their execution. The teams "System and Services development", "Deployment and Operation", and "Evaluation and Evolution" can constitute the development team. For large-scale systems it is necessary to have one team for each role. In the case of small/medium-scale systems, there can be a person per team, or even a few people performing multiple roles.

3.6 The incremental iterative approach. Our methodology performs iterations like in RUP, as shown in Fig. 2. This figure reports phases and loops and the stages -as in RUP- of Inception, Elaboration, Construction, Transition, and Support [62]. Also, similar to RUP, the colored shapes represent the possible amount of effort required for executing each phase of the process. Such a representation of the iterative way of our methodology implies that the construction of the models is built in an incremental way.

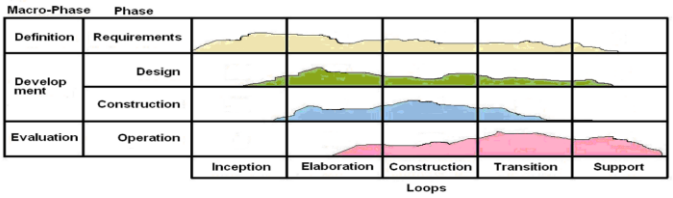


Fig. 2. The Iterative Model of the SOCA Development System Engineering Methodology

3.7 Model-to-model transformation. Our methodology defines three lines of model-to-model transformations. Such lines are represented in Fig. 3. The first artifacts constructed are System Context Diagram (D1) and the Work System Snapshot (D11) that are transformed into the artifacts System Responsibility Table (D12), Business Process Diagram (D7), and Enterprise data view (D13) each one in a line of transformation. Then the CIM is transformed into the PIM as follows. The System Responsibility Table (D12) is transformed into the artifacts Service model diagram (D2) and Service-Orchestration definition (D10). The artifact Business Process Diagram (D7) is transformed into the artifact Join Realization Table (D5), and the artifact Enterprise data view (D13) is transformed into the artifact ER System data model (D3). Next, PIM artifacts are transformed into PSM artifacts as follows. In the orchestration line, the artifacts Service model diagram (D2) and Service-Orchestration definition (D10) are

transformed into Service realization diagrams (D9). In the choreography line, the artifact Join Realization Table (D5) is transformed into the artifact User interface design (D8). In the data line, the artifact E-R System data model (D3) is transformed into the artifacts Creation script of the database scheme (D4) and Data access definition (D6). Finally, the PSM is transformed into the executable-PM as follows. In the orchestration line, the artifacts Service Realization Table (D9) and Service Orchestration Definition (D10) are transformed into the artifacts Service implemented and orchestrated in Java (D14) and Service composed and published with SCA (D15). In the choreography line, the artifact User interface design for HTML (D8) is transformed into Choreography of services in workflows with HTML (D16).

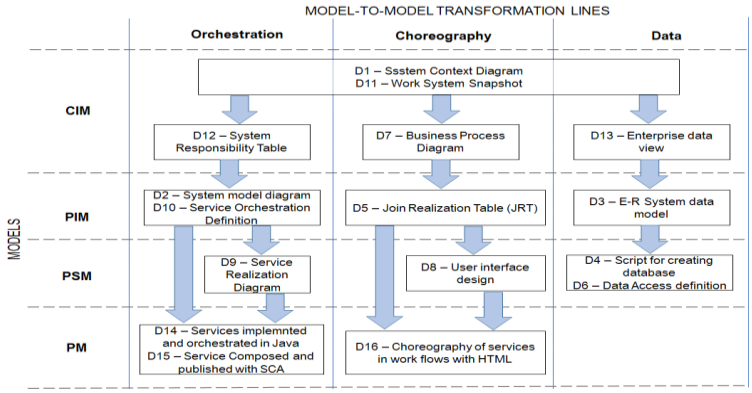


Fig. 3. Model-to-model transformation lines of the SOCA Development System Engineering Methodology

4. Motivating Example

We use a running example to illustrate the use of our methodology. Our running example involves a service–based system that supports the quality-assessment of biannual academic-courses. At the beginning, the system enables the teachers for elaborating two course-plans: "didactic instrumentation" that indicates the themes of the course and the didactic (teaching and learning) activities to be executed throughout the semester, and the "scheduled advance" that contains the planned dates to cover and to evaluate the course themes -or learning units of the course-. At three different audit trail moments, the teacher captures tracing data as audit trail of his/her work i.e. student-grades and real dates vs. planned dates. For each audit trail moment, the plans and tracing data are audited through the system by the department chair (chief) and the chief assistant. At the end of the semester, the system generates a final statistical report about the tracing data. In the running example, we illustrate the elaboration of the MDA-based models that conform to the artifacts of our methodology. More specifically, we report some of the artifacts that result from modeling the CIM, PIM, PSM and Executable PM.

4.1 Computation Independent Model (CIM). We present the artifacts System Context Diagram (see Fig. 4) -that represents the total context of the projected system-, and the

Business Process Diagram (see Fig. 5). The left part of the System Context Diagram artifact represents an overview of the business process, represented as a list of core elements: management forms, processes, and roles.

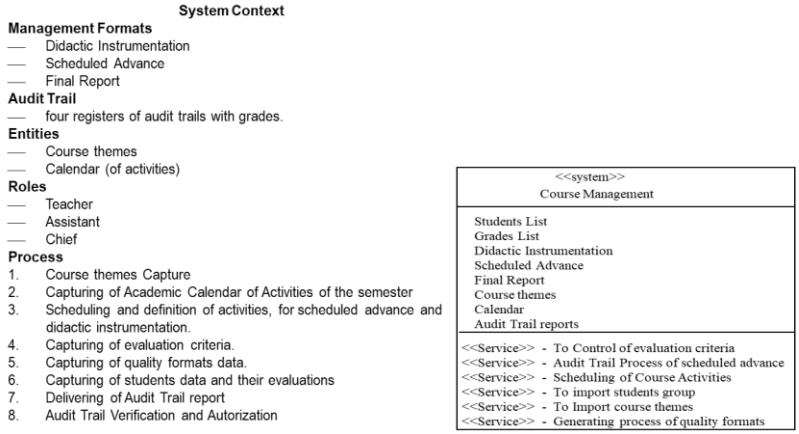


Fig. 4. CIM-D1 System Context diagram

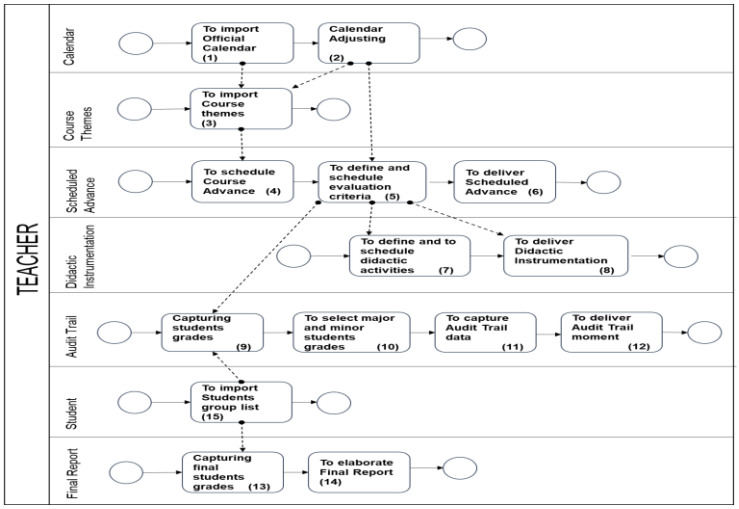


Fig. 5. CIM-D7 part c – A detailed Business Process Diagram

A Business-Process Diagram represents: (i) the activities of the private processes for each role -“CHIEF”, “TEACHER” and “ASSISTANT”-, where each private process is presented in a line; (ii) the interaction of public processes for the roles, that are presented at a high level of abstraction in a line per role; and (iii) diagrams of lower level of detail are constructed to show the specific workflows for each role for the specific cases where a more explanatory description is required. For the first increment of the development process (available in [57]), we only elaborated private processes of the role “Teacher”, the public processes collaborations of “Teacher” with the other roles, and the detailed workflow of the role “Teacher” (Fig. 5). Such diagrams present

the Business Processes from a low to a high level of detail based on the specific needs of the business requirements for any projected system. For example, the service "Calendar" in Fig. 6 has three operations: 2.1, 2.2, and 2.3 which are used for executing the sub-processes or activities 1, 2, 4, 5, and 7 of the detailed business process diagram (Fig. 5). The service "Audit Trail" has six operations labeled from 5.1 to 5.6.

4.2 Platform Independent Model (PIM). The PIM is the first model that is built during the first part of Design phase. The detailed Business Process Diagram presented in Fig. 5 is transformed into a Service model diagram (see Fig. 6), the service choreography -this one represented with Join Realization Tables, which are similar to use cases (available in [57])-, and the Service Orchestration Definition (see Fig. 7 a). The sub-processes or activities in the detailed Business Process Diagram are transformed into operations of the services in the Service model diagram. Also, these operations are grouped into components. Then, a Service orchestration definition (see Fig. 7 a) is elaborated with their implementations and interfaces. Interfaces (i.e. the operations the service component provides to other components) are represented on left side. Similarly, the component references (i.e. the services a component needs to call) are defined on the right side (see Fig. 7 b).

One component corresponds to one service (business service), and each service can have one or more interfaces. For each interface there is one implementation, and each implementation includes one or more operations of the corresponding service.

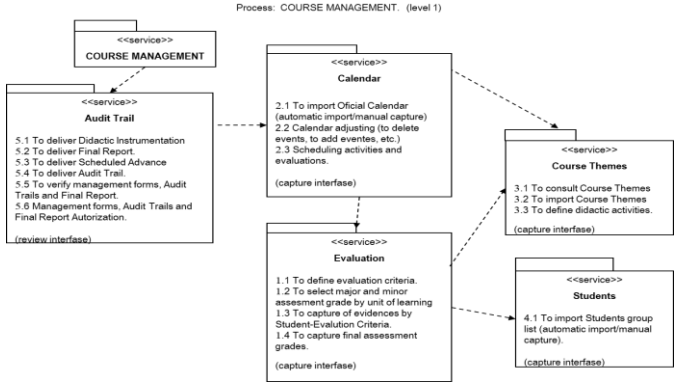


Fig. 6. PIM-D2 Service model diagram

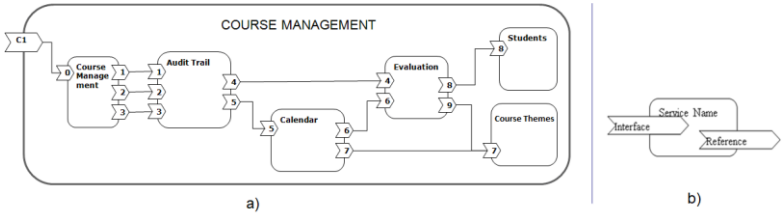
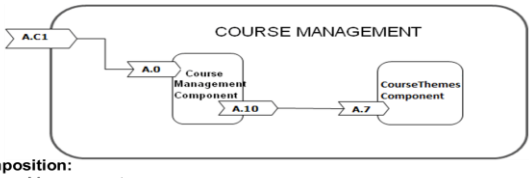


Fig. 7. a) PIM-D10 – Service Orchestration Definition-, **b)** Service Component with one Interface and one Reference

4.3 Platform Specific Model (PSM). The PSM is the second model constructed during the Design phase of our methodology. The Service orchestration definition and its

corresponding Service model diagram are transformed into the corresponding Service Realization diagrams (available in [57]), which describe the service and its realization with UML diagrams (i.e. one Service Realization Diagram is constructed for each service component in a Service Orchestration definition as shown in Fig. 7a). Also, JRTs are transformed into the user interface design (available in [57]).

4.4 Executable Platform Model (Executable PM). In our methodology, the Executable PM is built during the Construction phase. Here, Service orchestration definition and the Service orchestration diagrams were transformed into the artifacts of the Executable PM involving the programming code of the system prototype. Such a programming code is available in [57]. We implemented the main user interface of our running example in HTML. We also implemented the operation 3.1 "To consult Course Themes" (see the JRT in [57]) of the service "Course Themes" (see Fig. 6). Such an operation is used by the first step of the running example that regards the activity 3 "To import Course Themes". This activity is located in the private process (see Fig. 5). In the application domain, the prototype involves only a service composition of the subset of the Service Orchestration Definition presented in Fig. 7 a. Such a subset is shown in Fig. 8.



Service Composition:
(C1) Course Management

Service Components.
- CourseManagementComponet (just to publish the service composition, it does not have implementation)
- CourseThemesComponent

Service Interfaces = defines service implementation and interfaces.
(A.7) CourseThemesService + CourseThemesInterface .– Course Themes (this service interface is exactly how is presented in PIM-D10)

Service References
(A.10) CourseThemesReference .– Course Themes (this reference, in PIM-D10 is going to two service components. But, in this case it is going direct to CourseManagementComponent, because is the only operation to program in this first prototype).

Fig. 8. Fragment of Service-Orchestration definition of the implemented executable PM

We implemented the service "Course Themes" with a component (service). We defined code in SCA that orchestrates the service composition "Course Management", the service component "CourseThemes", and defined the reference "CourseThemeReference". The functionality of the services was implemented in Java. The service "CourseThemes" was implemented with both the class "CourseThemesService" and the interface "CourseThemesInterface", as shown in Fig. 8. The method "TemaItem[ ] get(String idMateria)" of the interface of the service "Tema" implements the operation 3.1 of the Service model diagram. The code of these programs can be obtained from [57]. Finally, some minor details regarding the implementation are the following. The service component "Course Management", the service component "CourseThemes", the reference "CourseThemesReference", the Java class "CourseThemesService", and the Java interface "CourseThemesInterface" correspond to "gestion", "Tema", "tema", "TemaImpl", and "Tema", respectively, in the implementation files.

5. Empirical Evaluation of the proposed SOCA Methodology from a Panel of SOCA Academicians and Practitioners

According to [44] a theoretical validation of a conceptual artifact can be also conducted with a similar *Face Validity* approach used in the simulation research stream [58]. Also, in other relevant disciplines a panel of experts is used for making decisions through a systematic process named Delphi method [47]. Furthermore, this evaluation method has started to be used for exploring a few relevant research problems in the discipline of software engineering [4], [26]. Thus, we applied this empirical approach as a preliminary evaluation of our work. The utilization of Likert 5- and 7-point scales have been reported as adequate regarding reliability and validity measures when the researcher wants to eliminate the neutral option [1]. In our research, we used for the first instrument a 5-point Likert scale because it is a new instrument and a moderate fine scale is recommended. In the second instrument we used a 7-point Likert scale because it has been extensively tested [31] and it has a finer scale.

5.1 Empirical Evaluation of proposed SOCA Development Systems Engineering Methodology from a Pilot Group of Software Engineering Academicians

A group of 15 software engineering academicians were contacted for evaluating the overall theoretical validity of the design of our methodology. Such evaluators were selected under the following criterion: to have a PhD level and at least 5 years of research-teaching experience on software engineering methodologies or software engineering professionals with at least a MSc student level and at least 5 years of professional expertise in software engineering methodologies. Most of them are from Latin America (LA) and two from Europe. This group was contacted through direct email contacts from the research team. Table 2 reports the specific demographic data of this panel of experts.

Table 2. Demographic Data of the Panel of Experts

Id. Eval	Academic Level of Evaluator	Expertise years in Software Engineering		Organization Type	Region
		Academic Settings	Industrial Settings		
1	PhD	5 – 10	< 5	Higher Education	LA
2	PhD	> 10	< 5	Higher Education	LA
3	PhD	> 10	------	Higher Education	LA
4	PhD	> 10	5 – 10	Government agency	LA
5	MSc Student	-----	5 – 10	Government agency	LA
6	PhD	> 10	< 5	Higher Education	LA
7	PhD	> 10	5 – 10	Higher Education	LA
8	PhD	5 – 10	< 5	Higher Education	LA
9	PhD Student	5 – 10	> 10	Government agency	LA
10	PhD	> 10	------	Higher Education	LA
11	PhD	> 10	< 5	Higher Education	LA
12	PhD	5 – 10	< 5	Higher Education	Europe
13	PhD	----	> 10	Higher Education	Europe
14	PhD	----	5 – 10	Higher Education	LA
15	MSc Student	5 – 10	> 10	Government agency	LA

The artifacts that were evaluated by the panel of experts were: 1) the user manual of our SOCA methodology, and 2) the documentation generated as a proof of concept for our running example. The 15 evaluators rated 6 items on theoretical validity (see Table

3). Each item was rated using a 5-point Likert Scale (from 1 for a total disagreement statement to 5 for a total agreement). The evaluation had a mean and standard deviation of **4.66** and **0.62,** respectively. Such results suggest that the panel of experts on software engineering methodologies considered our SOCA methodology as a theoretically supported development methodology.

Table 3. Evaluation of SOCA Development Systems Engineering Methodology from the Panel of Experts

Evaluated Item	Mean	Std. Dev.
P1. The conceptual model is supported by robust theoretical principles.	4.60	0.52
P2. The theoretical principles used to develop the conceptual model are relevant for the topic under study.	4.90	0.32
P3. The revised literature to develop the conceptual model does not present important omissions.	4.60	0.84
P4. The conceptual model is logically coherent.	4.70	0.67
P5. The conceptual model is adequate for the pursued purpose.	4.50	0.71
P6. The conceptual model provides a real contribution and it is not a duplicate of an existent model.	4.60	0.70
TOTAL	4.66	0.62

5.2 Empirical Evaluation of the SOCA Development Systems Engineering Methodology from Pilot Group of Software Engineering Practitioners

According to [27][35] a software development methodology can be evaluated through a *survey method* (also named field study). In particular in [35] it is reported a generic evaluation method named DESMET, where the survey method is one of the three ones suggested to be used. A survey method [35] "*is the collection and analysis of data from a wide variety of projects. The data collected in a survey are not as controlled as those obtained from a formal experiment but they can be analyzed statistically to identify important trends*".

Our methodology was evaluated through this *survey method*. In this survey took place a pilot group of 32 partial-time MSc students enrolled in a 2-year graduate program (23 enrolled in 2008-2009 period and 9 in the 2010-2011 period) in a Mexican public university located in the central region of Mexico. The first author taught a 4-weekeend graduate course on Software Oriented Architecture (SOA)/SOSE Development to this pilot group. These 32 partial-time MSc students (with a total of 64 class-hours per course) learned four main thematic units: 1) SOSE foundations, and its relationship with SOC and SOA; 2) SOC and SOA foundations, available technologies and the concept of SOCA or service oriented systems; 3) SOA/SOSE methodologies including a review of how service oriented systems can be implemented, and the methodology SOCA-DSEM; and 4) using SOCA-DSEM and SCA for constructing a SOCA example. For the End Term Project of this graduate course the students were asked to apply our SOCA methodology for developing a SOCA prototype of a real problem selected by themselves. The training to use the methodology included showing and explaining the complete running example that is illustrated in this paper.

The students were given a 2-week period for elaborating the project in teams of 4 people (to apply the methodology each team member took all team roles of the SOCA-DSEM). A total of eight teams were formed. Because this evaluation method does not qualify as experimental, the team formation was decided by the same MSc students. Each team selected the SOA technology to from the three technologies presented in the

second thematic unit of the course, namely Java Beans in J2EE and Servlets, C# in .NET using compositions in BizTalk Server with Business Activity Monitoring, and SCA for PHP platforms. The services were published as web services –i.e. remote component services-. Specifically, four of the teams used Java Beans with Servlets, two teams employed C# in .NET with BizTalk Server and Business Activity Monitoring, and two teams used SCA for PHP. The students only implemented Business Computing Services -also called "SOA Process Services"- as remote components, other kind of services such as Information and Communication Technology Computing Services -also called "SOA Infrastructure services"- were not implemented because SOCA -as we said earlier in subsection 2.1- is conceptualized as a composition of Business Computing Services [56].

The aforementioned technologies provide a way to publish the services in the web [59]. There are more sophisticated technologies and standards such as SOAP, WSDL, UDDI, RESTful Web Services among others that are useful to develop services [52][59][19]. However, these technologies were not used by the students as the technologies they employed are easier to learn and use in small projects. After this 4-weekend graduate course and the elaboration of the end term project, we measured the MSc perceptions on our SOCA development methodology by using the following constructs [31]: usefulness, ease of use, compatibility, result demonstrability, and intention of use. The questionnaires that were employed can be obtained from [57]. The demographic data of these 32 MSc part-time students is reported in Table 4. Finally, although the time of training involved only four weeks, the empirical evaluation was possible given that the students had already programming experience with the technologies they employed. On the other hand, our methodology is focused on the decomposition / composition of the services over distributed components, which can be implemented in three possible types of technologies, namely web services, Representational State Transfer (REST) approach, and messaging systems [7]. Despite the groups of students were small and there was limited time for developing the SOCA example, the students were able to build systems whose basic functionality worked correctly; this due to their previous programming experience with the employed technologies.

Table 4. Demographic Data of the Pilot Group of 32 MSc Part-Time Students

Demographic Variable	Highlights
IT background	All of them come from a BSc. in IT.
Age range	The sample contains similar groups (1/3) on the ranges of 26-30, 31-35, and 36-40 years.
Formal training in SwE	According to the MSc curricula, all of them (100%) were enrolled in at least 3 courses in SwE themes.
Main working role	Most of them (85%) are located in IT technical positions. Very few of them (15%) are located in managerial positions.
Scope of working organization	80% of them are working for organizations with a national scope and 10% are working for organizations with a worldwide scope.

According to [68], a construct is a theoretical concept, which cannot be measured directly, but through operational variables (named also items). We selected this set of constructs (*Usefulness, Ease of use, Compatibility, Result Demonstrability, and Behavioral Intention of Use*) because such constructs have been suggested elsewhere [53][46][65][45][36][66][30][41][11][21] as highly suitable for predicting the overall acceptance or rejection of a new designed artifact. Table 5 shows the constructs, their definitions, their operationalization, and their reported reliability measures from [31]

that reports the specific items for each construct C*i* where *i* goes from 1 to 5, thus, representing the 5 constructs we employed.

Table 5. Summary of Constructs used for Empirical Pilot Evaluation through a Survey Method

Construct	Definition	Operationalization	Reported Reliability
Usefulness	*the degree to which using IT innovation is perceived as being better than using the practice it supersedes.*	4-item 7-point Likert Scale	0.90
Ease of use	*the degree to which using a particular system is free of effort.*	3-item 7-point Likert Scale	0.90
Compatibility	*the degree to which adopting IT innovation is compatible with what people do*	3-item 7-point Likert Scale	0.88
Result Demonstrability	*the degree to which the results of adopting/using IT innovation are observable and communicable to others*	3-item 7-point Likert Scale	0.76
Behavioral Intention of Use	*the individual's intention to adopt or continue to use the IT innovation*	2-item 7-point Likert Scale	0.90

To assess the acceptance level from the pilot group of the 32 Part-Time MSc students, we planned initially to apply a one-tailed single-sample *t* Test [60] with an alpha value of 0.05 for each one of the five hypotheses (i.e. each one for each construct). First for supporting the normality assumption it was required to apply the single-sample t Test. Hence, we applied the Shapiro-Wilk test [43] to each 32 data sets for the 5 constructs. We found that the normality assumption cannot be supported, then we applied an alternative non-parametric test: the Wilcoxon signed-ranks test [60] for a single sample. There is another Wilcoxon test (Wilcoxon matched-pairs signed-ranks test) but it is only suitable for two dependent samples, and it was not our research case. Based on [60] a two-tailed bidirectional Wilcoxon signed-ranks test was used to evaluate the next null hypothesis "*does a sample of n subjects (or objects) come from a population in which the median value is equal to a specified value?*". In our research, we need to perform the test against a particular median value (fixed by the researchers) regarding whether it is less or equal to the fixed value. Thus, we use the one-tailed unidirectional test, and the null hypothesis was established as "*does a sample of n subjects (or objects) come from a population in which the median value is less or equal than a specified value?"*. This statistical non-parametric test has been used frequently in empirical software engineering research [38][17][29]. Thus, we used the one-tailed Wilcoxon signed-ranks test [60] to test the five null hypotheses statements which were stated as follows: "*H0.i The median of the construct Ci is <= 5.0*", where *i* goes from 1 to 5. We expected to reject all of the five null hypotheses with an alpha error of 0.05 as a maximum, and with it to obtain initial evidence on satisfactory perceptions for the SOCA methodology on usefulness, ease of use, compatibility, result demonstrability, and final behavioral intention of use in the pilot group of 32 evaluators. We considered that by using a Likert scale from 1 to 7, the fixed test value of 5.0 implies at least a moderately good score. For this task, we used the free statistical tool MaxStatLite (www.maxstatlite.com). Table 6 reports the associated statistical results.

Thus, in case the null hypotheses are rejected we obtain that our SOCA methodology fulfills the five metrics (i.e. it is perceived as useful, ease of use, compatible, with result demonstrability, and with a final behavioral intention to use in the near future) as the pilot group of evaluators (expert panel) previously suggested.

In this empirical statistical evaluation, we did not pursue to elaborate or test a predictive theory with a predicted construct of behavioral intension of use. Although this can be elaborated with specific statistical techniques like PLS [20], this is left for future research.

Table 6. Results of the Empirical Evaluation of the SOCA Development Systems Engineering Methodology with a Pilot Group of 32 Part-Time MSc Students

Null Hypothesis	Test Median	32 Sample Median	Test Statistics	Alpha value	P-value	Reject Null Hypothesis H0?
H0.1 "The median of the construct **usefulness** is less or equal to 5.00"	**4.000**	**5.625**	**63.000**	**0.0500**	**< 0.0001**	**YES**
H0.2 "The median of the construct **ease of use** is less or equal to 5.00"	**4.000**	**5.667**	**21.00**	**0.0500**	**< 0.0001**	**YES**
H0.3 "The median of the construct **compatibility** is less or equal to 5.00"	**4.000**	**4.833**	**94.500**	**0.0500**	**0.0038**	**YES**
H0.4 "The median of the construct **result demonstrability** is less or equal to 5.00"	**4.000**	**5.000**	**30.500**	**0.0500**	**< 0.0001**	**YES**
H0.5 "The median of the construct **behavioral intention of use** is less or equal to 5.00"	**4.000**	**5.250**	**22.00**	**0.0500**	**< 0.0001**	**YES**

5.3 Discussion

SOCA development methodologies are relatively new. Several SOCA methodologies have been reported in the last decade [25], but none of them has gained sufficient acceptance for SOCA academic and professional community, as it happened in the previous OOSE and CBSE paradigms. In [25] it is presented several generic process-related properties of SOCA development methodologies that allows us to compare SOCA engineering approaches. In particular, the objective/scope of the methodology is a relevant feature, and they distinguished five types: 1) focused on a full development process, 2) focused only on analysis and design, 3) focused only on service composition, 4) focused only on migration of SOA, and 5) focused only on project management. This categorization matches with different levels of abstraction of the methodology, from the management project, passing through the architecture perspective and process engineering, to the service composition. Our SOCA methodology is type 1.

SOCA development studies appear because of the novelty of the SOSE paradigm and the need of methodologies for service-oriented systems development. Although there are new proposals for SOCA development, they have several limitations. A desirable issue to be addressed by such methodologies is the ability to narrow the conceptual gap between the problem (or business) domain and the system implementation (or application) domain [18][61]. However, current proposals do not tackle this issue since they do not consider the analysis and design of services, composition, orchestration, and choreography all together. Furthermore, in general terms, these methodologies do not cover all the software development life-cycle [25]. In contrast, our SOCA methodology fills this gap by (i) including the product "Enterprise Architecture design" -although limited to business vocabulary-; (ii) considering some activities for planning, (iii) aligning Business with Information Technology since the beginning to the end of the development process -e.g. requirements analysis, a service oriented analysis and design, at architectural and system and the user interface design, and an operative evaluation at the business level-, (iv) suggesting activities to ensure an operative deployment at the architecture level; and (v) evaluating the operative system at the application level for correctness and for system evolution [56]. Also, our proposed methodology covers all the software system development life-cycle in a well-structured manner i.e. our

methodology includes phases, activities, models, and products as well as roles and an iterative development process.

The results of the empirical evaluation have shown that our methodology is practical and applicable for developing SOCAs. Specifically, the first evaluation involving a 5-point Likert scale was applied to an expert panel and obtained 4.50 or higher for each evaluated item, an average of 4.66, and an average standard deviation of 0.62. Such results suggest that the panel of experts on software engineering methodologies considered our SOCA methodology as having strong theoretical support (item P1) while considering that our proposal covers state of the art principles (item P2) without important omissions (item P3). The pilot group also considered our methodology as logically coherent (item P4) with a suitable conceptual model (item P5) as well as considering it as a real contribution that is not a duplicate of an existent solution (item P6). The second evaluation involving a 7-point Likert scale was applied to a group of software engineering practitioners and obtained a value higher than 4.0 for all the evaluated constructs with the directional hypothesis stating "H0.i The median of the construct C*i* is <= 4.0", where *i* goes from 1 to 5. We can interpret this result as all constructs were perceived as no negative or neutral ones. Reported reliability [31] for the constructs measured is higher than 0.70, which is the expected value for pilot studies. The results reported in Table 6, and supported by the Wilcoxon signed-ranks test, provide statistical-based evidence that SOCA methodology was perceived as useful, easy to use, compatible, with result demonstrability, and with a final intention to use it. Finally, Table 7 presents a comparative analysis of previous service oriented development methodologies with our work. (✓ indicates when the methodology gives guidance or defines activities related to the topic of the development processes/areas, while a blank square indicates no guidance is given for the topic). Although, all compared methodologies support at least one of three topics, namely Business Process Modeling, Requirements Engineering or Service Requirements, none of the revised methodologies provides a complete solution for all topics. However, our methodology covers all topics except requirements engineering and automated transformation.

Table 7. Comparative analysis of the SOCA Development Systems Engineering Methodology with other methodologies

Topic / area of the development processes.	M1 (Karastoyanova, 2003) [33]	M2 (Kotonya, 2004) [39]	M3 (Ivanyukovich, 2005) [28]	M4 (Kühne, 2005) [40]	M5 (Cox, 2005) [14]	M6 (Karakostas, 2006) [32]	M7 (Papazoglou, 2007) [51]	M8 (Gu, 2009) [23]	M9 (Cantor, 2003) [10]	M10 (Arsanjani, 2008) [3]	M11 SOCA Dev. Systems Eng. Methodology
Requirements Engineering		✓		✓				✓			
Business Process Modeling			✓		✓		✓	✓	✓	✓	✓
Enterprise architecture design										✓	✓
Service requirements and Service Design			✓			✓	✓		✓	✓	✓
Subsystems architectural design		✓			✓					✓	✓
SOA Design (service composition and orchestration)						✓	✓			✓	✓
Application modelling	✓		✓					✓	✓		✓
Automated transformation	✓			✓		✓					
Monitoring & Maintenance		✓			✓		✓	✓		✓	✓
Propose products / artifacts									✓	✓	✓
Propose to elaborate database									✓		✓

There are three methodologies whose focus is on the automatic transformation towards a set of services and compositions of executable services, specifically those proposed by Karakostas, Kühne and Kotonya's focuse mainly on requirements engineering, architecture design and monitoring and maintenance.

Unlike our methodology -except Karastoyanova, Ivanyukovich, Gu and Cantor (RUP-SE)-, the compared approaches do not support the topics of SOA design (i.e. composition and orchestration of the services). Therefore, current approaches do not have neither an implicit nor explicit vision of SOCA. In contrast, our methodology emphasizes -as part of the modeling of the application- the importance of the design of the user interface, and business alignment throughout the development process. Also, our methodology proposes both activities and products/artifacts to be built.

5.4 Limitations and Threats to Validity

There are a number of threats to validity regarding the statistical methods we employed. According to [63][67], when we use a survey research method, it must be considered the following validation issues: 1) *instrument validation*, 2) *internal validity*, 3) *statistical conclusion validity*, and 4) *external validity*. *Instrument validation* refers to using a reliable instrument that contains items (operational variables) strongly associated with the expected construct to be measured. The *instrument validation* must satisfy a *content validity*, a *construct validity*, and a *reliability of constructs*. We addressed threats to this type of validity by using well-tested instruments to measure the constructs of usefulness, ease of use, compatibility, result demonstrability and behavioral intention of use [31]. In [31] *content validity* was assessed by reusing a previous valid instrument. *Construct validity* was assessed through a satisfactory matrix of loadings factors for each construct (see pp. 210-211), where each value is greater than 0.70 and its maximum value belongs only to an associated construct. The *reliability of constructs* was assessed with the test of Cronbach's Alpha [31] and four of the five values were satisfactory obtaining higher or equal than 0.80 whereas only the construct Result Demonstrability exhibited a score of 0.70. Thus, the supported hypothesis H0.4 must be considered with caution.

Internal validity refers specifically to the extent to which an experimental variable is responsible for any variance in a dependent variable. This is also conceptualized as the reduction of alternative/competitive interpretations on the obtained results. In this research, it involves the fact that the pilot sample of evaluators reported satisfactory scores for the 5 constructs for other reasons different from the direct evaluation of our SOCA methodology. These scores can be biased for instance when there is pressure on the evaluators to assess with high scores, empathy of the evaluators with the research team, or lack of interest of the evaluators. The main threats for *internal validity* are: *history, maturation, testing, instrumentation, mortality,* and *selection* [67]. *History* refers to bias on scores when a critical event happened before conducting the evaluations. *Maturation* refers to a bias caused by time-passing effects on evaluators. *Testing* refers to bias on evaluators when they had previously evaluated the methodology. *Instrumentation* refers to changes with the instruments used. *Mortality* refers to the case when an evaluator dies before filling out all the questionnaires of the survey. Lastly, *selection* refers to the selection of evaluators to participate in one of two groups -experimental or control group-. The collection of data in this research avoids

the threats of *history* because no critical external event was registered during the evaluation process. *Maturation* was avoided because evaluations were transversal (no longitudinal) ones. *Testing* was avoided because a previous evaluation was not conducted. *Instrumentation* was avoided because we used the same valid instrument. *Mortality* was avoided because evaluators participated only once. Lastly, s*election* is only employed when both an experimental and a control group take place in the survey, this, in order to decide, which elements conform to each group.

Statistical conclusion validity refers to the correct utilization of statistical procedures. In this research, we avoided threats with the utilization of the adequate statistical test (a one-tailed Wilcoxon signed-ranks Test [60]) and by using an adequate alpha value (Type I error) of 0.05. Also, *external validity* refers to the extent to which the results can be generalized to other populations, settings or contexts. In this research, despite the pilot sample of evaluators is highly representative of the population, which is generically defined as "*population of software engineering practitioners working in medium and large enterprises in developing economies*", we cannot claim these results can be generalized to other similar populations because the pilot sample of evaluators were selected by a non-probabilistic procedure. Thus, the *external validity* must be limited to this type of sample.

Finally, given that small development teams and a small project were employed for evaluating this research, our claims must be considered as initial insights toward more definitive results. However, we believe our preliminary results are encouraging.

6 Conclusion

We have proposed a Service-oriented Computing Application Development Systems Engineering Methodology which was derived by employing four core design building blocks, namely Service-oriented Computing, the Model-Driven Architecture, two popular development methodologies -RUP-SE and MBASE-, and agility-rigor balance recommendations. Our methodology extends the existent and emergent body of knowledge in Service-Oriented Software Engineering by applying and enhancing the existent knowledge. Our methodology also includes features from the main SOCA development proposals and proposes the application of emergent techniques that are useful for SOCA development.

From a practitioner perspective, this research contributes with the following results. First, we defined a new agility-rigor balanced methodology. Second, we proposed a development methodology that provides a set of new techniques that are more suitable for SOCA development than those used in Object-Oriented Software Engineering or Component-Based Software Engineering approaches. Lastly, we used a running example that encourages practitioners to use and test our methodology.

Finally, we consider this study raises new interesting issues that require further research. Firstly, evaluating our methodology with empirical case studies involving a variety of business organizational problems will make it possible to capture more definitive quantitative usability metrics such as usefulness, ease of use, compatibility, and value of using our methodology for developing SOCAs. Secondly, experiments can be conducted with graduate part-time students to contrast the usability metrics obtained by using the most known software development methodologies such as RUP, MSF, and

RUP for SOA, versus our approach. Thirdly, an agile version of our methodology can be elaborated as well as identifying for what kinds of software projects are recommended the full methodology and for what kinds of projects can be suitable the agile version. Lastly, tool support can be developed for automating the model-to-model transformations.

References

1. Allen, E, Seaman, C.: Likert Scales and Data analyses, Quality Progress, Statistics Roundtable, 40(7), 64-65. (2007)
2. Amsden, J.: Modeling SOA IBM (parts 1,2,3,4,5), Level: Introductory, STSM, IBM, 2007
3. Arsanjani, J., Hailpern, B., Martin, J., Tarr, P.L.: Web Services: Promises and compromises, IBM Research Division, Thomas J. Watson Research Center, 1-18. (2008)
4. Avison, D., Fitzgerald, G.: Where now for development methodologies?, Communications of the ACM, 46(1), 78-82. (2003)
5. Baghdadi, Y.: A comparison framework for service-oriented software engineering approaches: Issues and solutions, International Journal of Web Information Systems, 9(4), 279-316. (2013)
6. Beisiegel, M., Blohm, H., Booz, D., Edwards, M., Hurley, O., et al.: SCA Service Component Architecture: Assembly Model Specification, SCA Version 1.00. (2007)
7. Bianco, P., Kotermanski, R., Merson, P.: Evaluating a Service-Oriented Architecture, Software Engineering Institute, Technical Report, 1-89. (2007)
8. Boehm, B., Egyed, A., Kwan, J., Port, D., Shah, A., Madachy, R.: Using the WinWin spiral model: a case study", IEEE Computer, 31(7), 33-44. (1998)
9. Boehm B. and R. Turner, Balancing Agility and Discipline: A Guide for the Perplexed, Addison Wesley, Boston, (2004).
10. Cantor, M.: Rational Unified Process for Systems Engineering, Rational Brand Services IBM Software Group. (2003)
11. Carter, L., Belanger, F.: The Influence of Perceived Characteristics of Innovating on e-Government Adoption, Electronic Journal of e-Government 2(1), 11-20. (2004)
12. Center for Software Engineering: Guidelines for Model-Based (System) Architecting and Software Engineering (MBASE), University of Southern California. (2000)
13. Clements, P., Bachmann, F., Bass, L., Garlan, D., Ivers, J., Little, R., Merson, P., Nord, R., Stanfford, J.: Documenting Software Architectures: Views and Beyond, 2nd Edition, Pearson, Addison-Wesley. (2011)
14. Cox, D.E., Kreger H.: Management of the service oriented-architecture life-cycle, IMB Systems Journal, 44(4), IBM Corporation. (2005)
15. Dix, A., Finlay, J., Abowd, G.D., Beale, R.: Human-Computer Interaction, 3rd Edition, Pearson, Prentice-Hall. (2004)
16. Dumas, M.: Towards a semantic framework for service description, Cooperative Information Systems Research Centre, Queensland University of Technology, Australia. (2000)
17. Ellis, B., Stylos, J., Myers, B.: The factory pattern in API design: A usability evaluation, In Proc. of the 29th international conf. on Soft. Eng., IEEE Computer Society, 302-312. (2007).
18. France, R., Rumpe, B.: Model driven development of Complex Software: A Research Roadmap, Department of Computer Science Colorado State University. (2007)
19. Garriga, M., Mateos, C., Flores, A., Cechich, A., Zunino, A.: RESTful service composition at a glance: A survey, Elsevier, Journal of Network and Computer Applications, 60(1) 32-53. (2016)
20. Gefen, D., Straub, D., Boudreau, M.: Structural equation modeling and regression: Guidelines for research practice, Communications of the Association for Information Systems, 4(1), 1-79. (2000)

21. Gefen, D., Karahanna, E., Straub, D.: Trust and TAM in online shopping: an integrated model, MIS Quarterly 27(1), 51-90. (2003)
22. Granollers, T., Lorés, J. & Perdrix, F.: Usability Engineering Process Model. Integration with Software Engineering. In Proceedings of HCI International 2003, 1-6. (2003)
23. Gu, Q., Lago, P.: A stakeholder-driven service life cycle model for SOA, ACM IW-SOSWE'07, Dubrovnik, Croatia. (2009)
24. Gu Q., Lago P.: Service Identification Methods: A Systematic Literature Review, In: Di Nitto E., Yahyapour R. (eds) Towards a Service-Based Internet. Service Wave, Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, 6481, 37-50. (2010)
25. Gu, Q., Lago, P.: Guiding the selection of service-oriented software engineering methodologies, Springer, SOCA 2011(5), 203-223. (2011)
26. Hassanzadeh, A., Namdarian, L.: Developing a framework for evaluating service oriented architecture governance (SOAG), Knowledge-Based Systems 24(5), 716-730. (2011)
27. Hevner, A.R.: Design Science in Information Systems Research, Information Systems and Decision Sciences, MIS Quarterly 28(1), 75-15. (2004)
28. Ivanyucovich A., Gandaharan, G.R., D'Andrea, V., Marchese, M.: Toward a service-oriented development methodology, University of Trento, Italy. (2005)
29. Jørgensen, M., Jøberg, D. I.: Impact of effort estimates on software project work, Information and software technology, 43(15), 939-948. (2001)
30. Kaba, B: Modeling information and communication technology use continuance behavior: Are there differences between users on basis of their status?, International Journal of Information Management, 38(1), 77–85. (2018)
31. Karahanna, E., Straub, D. W., Chervany, N. L.: Information Technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs, MIS Quarterly, 23(2), 183-213. (1999)
32. Karakostas, B., Zorgios, Y., Alevizos, C.C.: Automatic derivation of BPEL4WS from IDEF process models, Software & System Modeling, 5(2), 208-218. (2006)
33. Karastoyanova, D.: A methodology for development of Web Services-based Business Processes, Thechnische Universität Darmastadt. (2003)
34. Keith, M., Demirkan, H., Goul, M.: Service-oriented methodology for systems development, Journal of Management Information Systems, 30(1), 227-260. (2013)
35. Kitchenham, B., Linkman, S., Law, D.: Critical review of quantitative assessment. Software Engineering Journal, 9(2), 43-54. (1994)
36. Kiwanuka, A.: Acceptance Process: The Missing Link between UTAUT and Diffusion of Innovation Theory, American Journal of Information Systems 3(2), 40-44. (2015)
37. Kontogiannis, K., Lewis, G. A., Smith, D. B., Litoiu, M., Muller, H., Schuster, S., Stroulia, E.: The landscape of service-oriented systems: A research perspective, In Proceedings of the International Workshop on Systems Development in SOA Environments, IEEE Computer Society. (2007)
38. Kosar, T., Mernik, M., Carver, J. C.: Program comprehension of domain-specific and general-purpose languages: comparison using a family of experiments, Empirical software engineering, 17(3), 276-304. (2012)
39. Kotonya, G., Hutchinson, J., Bloin, B.: A method for formulating and architecting component and Service-oriented Systems, Computing Dept., Lancaster University. (2004)
40. Kühne, S., Thränert, M., Speck, A.: Towards a methodology for orchestration and validation of cooperative e-business components, Inst. of Cybernetics, Tallinn Technical Univ. (2005)
41. Legris, P., Inghamb, J., Collerettec, P.: Why do people use information technology? A critical review of the technology acceptance model, Information & Management 40(3), 191–204. (2003)
42. Miller, J., Mukerji, J.: MDA guide version 1.0.1, Object Management Group (OMG), (2003)
43. Mooi, E., Sarstedt, M.: Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics, Springer, Germany (2018)

44. Mora, M., Gelman, O., Paradice, D., Cervantes, F.: The case for conceptual research in information systems, In Proceedings CONF-IRM 2008, p. 52. (2008)
45. Mora, M., Rory, V., Rainsinghani, M., Gelman, O.: Impacts of electronic process guides by types of user: An experimental study, International Journal of Information Management, 36(1), 73-88. (2016)
46. Morris, M. G., Dillon, A.: How user perceptions influence software use, IEEE software, 14(4), 58-65. (1997)
47. Okoli, C., Pawlowski, S. D.: The Delphi method as a research tool: an example, design considerations and applications, Information & management 42(1), 15-29. (2004)
48. Oktaba, H., Ibargüengoitia, G.: Software process modeled with objects: Static view, Computación y Sistemas 1(4), 228-238. CIC-IPN ISSN 1405-5546. (1998)
49. OMG: OMG Meta Object Facility (MOF) Core Specification, v2.5.1. (2016) [Online]. Available: http://www.omg.org/spec/MOF (Current 2017)
50. Papazoglou, M.P., Traverso, P., Schahram, D., Leymann, F., Bernd, J.: Service-oriented Computing: Research roadmap, Tilburg University, The Netherlands. (2006)
51. Papazoglou, M.P., Van Den Heuvel W.: Business process development life cycle methodology. Communications of the ACM, 50(10), 79-85. (2007)
52. Papazoglou, M.P., Van Den Heuvel, W.J.: Service oriented architectures: approaches, technologies and research issues, The VLDB Journal, 16(3), 389-415. (2007)
53. Riemenschneider, C., Hardgrave, B., Davis, F.: Explaining software developer acceptance of methodologies: a comparison of five theoretical models, IEEE transactions on Software Engineering, 28(12), 1135-1145. (2002)
54. Rodriguez da Silva, A.: Model-driven engineering: A survey supported by the unified conceptual model, Computer Languages, Systems & Structures 43 (2015) 139–155. (2015)
55. Rodriguez, L.C., Mora, M., Alvarez, F.J.: Process Models of SDLCs: Comparison and evolution, Handbook of Research on Modern Systems Analysis and Design Technologies and Applications 76-89, Minnesota State University, Mankato, USA, IGI Global. (2008)
56. Rodriguez-Martinez, L.C., Mora, M., Alvarez, F., Garza, L.A., Duran-Limon, H.A., Muñoz J.: Review of Relevant System Development Life Cycle (SDLC) in Service-oriented Software Engineering (SOSE), Journal of Applied Research and Technology, 10(2), 94-113. (2012)
57. Rodriguez-Martinez, L.C.: A Case Example for Illustrating SOCA Development Systems Engineering Methodology and Evaluation Questionnaires: Complementary Documents. (2018). http://hduran.cucea.udg.mx/publications/SOCA-paper-extra-resources.zip.
58. Sargent, R.: Validation and verification of simulation models, Proceedings of the 1999 Winter Simulation Conference, 39-48. (1999)
59. Sheng, Q.Z., Qiao, X., Vasilakos, A.V. Szabo, C. Bourne S., Xu, X.: Web services composition: A decade's overview, Elsevier, Information Sciences, 280(1), 218-238. (2014)
60. Sheskin, D.: Handbook of parametric and nonparametric statistical procedures, CRC Press. (2003)
61. Sigh-Walia, G., Carver, J.C.: A systematic literature review to identify and classify software requirement errors, Elsevier, Information and Software Technology. (2009)
62. Sommerville, I.: Software Engineering, 7th Edition, Pearson, Addison-Wesley. (2005)
63. Straub, D.: Validating instruments in MIS research. MIS quarterly, 147-169. (1989)
64. Van Lamsweerde, A.: Goal-Oriented Requirements Engineering: A Guided Tour, In Proc. of the Fifth IEEE International Symposium on Requirements Engineering, 249-262. (2001)
65. Vavpotic, D., Bajec, M.: An approach for concurrent evaluation of technical and social aspects of software development methodologies, Information and software Technology, 51(2), 528-545. (2009)
66. Venkatesh, V., Morris, M., Davis, G., Davis, F.: User Acceptance of Information Technology: Toward a Unified View, MIS Quarterly 27(3), 425-478. (2003)
67. Wohlin, C., Höst, M., Henningsson, K.: Empirical research methods in software engineering. In Empirical methods and studies in software engineering, Springer, 7-23. (2003)

68. Zikmund, W., Babin, B., Carr, J., Griffin, M.: Business research methods, Cengage Learning. (2013)
69. Zimmermann, O., Krogdahl, P., Gee, C.: Elements of Service Oriented Analysis and Design: an approach of interdisciplinary modeling for SOA projects, IBM Developer Works. (2004) [Online]. Available: https://www.ibm.com/developerworks/library/ws-soad1/index.html (Current 2017)

Laura C. Rodriguez-Martinez is a full Professor at the Systems and Computing Department, Institute of Technology of Aguascalientes, Mexico. She holds a PhD in Computer Science at Universidad Autonomous University of Aguascalientes, Mexico in 2009. Her research interests include Software Systems Development Processes, Service-Oriented Software Engineering and Graphical-User Interfaces Development Processes.

Hector A. Duran-Limon is currently a full Professor at the Information Systems Department, University of Guadalajara, Mexico. He completed a PhD at Lancaster University, England in 2002. His research interests include Cloud Computing and High Performance Computing (HPC). He is also interested in Software Architectures, Software Product Lines and Component-based Development.

Manuel Mora is a full-time Professor at the Autonomous University of Aguascalientes (UAA), Mexico. Dr. Mora holds an Eng.D. in Engineering (2003) from the National Autonomous University of Mexico (UNAM). He has published over 90 research papers in international conferences, research books, and journals listed in the JCR. Dr. Mora is a senior member of ACM (since 2008) and of the Mexican National Research System at Level I, and does research focused on Design Methodologies for IT Services Systems and Decision-Making Support Systems (DMSS).

Francisco Alvarez Rodríguez is a Professor of Software Engineering. He is PhD from the National Autonomous University of Mexico. He has published research papers in several international conferences in SwE and e-Learning process. His research interests are SwE lifecycles for small and medium sized enterprises and SwE process for e-learning. He is currently president of the National Council for Accreditation of programs and Computing, A.C. (CONAIC).

Received: July 3, 2017; Accepted: September 20, 2018

Computer Science and Information Systems 16(1):45–74 https://doi.org/10.2298/CSIS171005010M

Building a BPM Application in an SOA-Based Legacy Environment

Mladen Matejaš[1] and Krešimir Fertalj[2]

[1] Privredna banka Zagreb,
Radnička cesta 50, 10000 Zagreb, Croatia
mmatejas@gmail.com; mladen.matejas@pbz.hr
[2] Faculty of Electrical Engineering and Computing,
Unska 3, 10000 Zagreb, Croatia,
e-mail: kresimir.fertalj@fer.hr

Abstract. Modern organizations need to understand and constantly improve their business processes (BPs) in order to make successful business decisions. This paper describes an integration model for building a Business Process Management Application (BPMA) and connecting the BPMA with legacy systems based on Service-Oriented Architecture (SOA). A BPMA is an application developed to support a BP performed by legacy application/s. A combination of multiple BPMAs provides support for multiple BPs and forms a BPM solution. The presented model is characterized by a simple co-dependence of the BPMA and the existing systems, minimal changes to the legacy applications and a maximal utilization of the existing functionalities. It enables the existing applications to function independently from the BPMA and simplifies the business data used in the BPMA. An extensive evaluation of the model was undertaken by experts from the BPM area. Its feasibility is demonstrated on a real-life business use case scenario.

Keywords: BPM, SOA, business process, integration model, business layers.

1. Introduction

In the volatile modern economy, the relevance of information technology and business process modeling is constantly growing [1]. By closely observing the business environment, one can identify BPs as its core elements. A BP can be defined as a sequential flow of business activities (BA) over a certain period [2]. The performance of an organization depends upon BPs which provide the means for achieving the organization's fundamental objectives [3]. Furthermore, the process view of BPs provides a better understanding of the organization and facilitates the development of information systems that would successfully support these BPs.

Business Process Management (BPM) is a set of management disciplines whose goal is to continuously optimize the efficiency of processes and activities through automation [4]. BPM provides support for the efficient management of a business environment with the purpose of increasing its flexibility and productivity. BPM implementation is a process of building a BPM solution in an existing business environment. The main objective of such a project is to create an environment, with the help of BPM tools,

where the flow of processes is determined in real time by events and the results of the actual process execution. A set of BPM tools (sometimes called a BPM suite) is a software product designed to develop, maintain and optimize a composite process application [5].

Although a BPM solution has its benefits, several questions arise: How many resources need to be spent for its implementation? What changes need to be made in the existing legacy environment? Is it necessary to develop new applications or services to support the BPM solution? To properly answer these questions, it is necessary to analyze the amount of user activities the BPM solution needs to support. Functionalities that support user activities can be determined by reviewing the needs of the business owners as they have a clear understanding of how the entire business system should function and can therefore roughly define the business requirements. Business owners allocate resources, approve the funding, decide on the scope of the projects and that way create additional business value. However, extensive redesigns and upgrades of existing information systems are tasks that business owners do not approve lightly. While such improvements can provide stability to the existing information systems, from a business owner's perspective, they do not create a new product or service, nor do they attract additional customers.

Integrating BPM in legacy environments, where BPs are managed by existing applications, is a challenging task. Some BPM implementation options include rewriting legacy applications using a BPM suite and building a custom BPM solution within the existing environment. However, redesigning and totally rewriting the existing processes with BPM tools, or building new applications, would be an enormous effort and would require a large amount of resources and time. Also, the organization's ability to grow, respond to market demands and cope with ongoing challenges would be limited during the redesign period. Consequently, business owners hesitate in approving such a task.

In case of SOA-based legacy environments, i.e. environments comprising legacy applications based on an SOA architectural style, a solid service background is present. Solid service backgrounds enable the flexibility of legacy applications and constitute a good foundation for the endeavor of BPM integration.

The goal of this research is to develop an integration model which would enable BPM implementation by identifying business activities in existing SOA-based applications and reusing them in a newly created BPMA. The usage of existing business activities, supported by the functionalities of the legacy applications, removes the need for extensive redesigns of the existing environment. By applying minimal changes to the existing legacy environment, organizations can, to the satisfaction of the business owners, minimize the time and resources needed for BPM implementation while enjoying the benefits provided by BPM.

The rest of the paper is organized as follows. Section 2 is an overview of relevant research in the area concerning the subject of the article. The challenges, similarities, differences and co-dependence of SOA and BPM, along with the general idea on how to connect the two approaches are presented in Section 3. Additionally, Section 3 discusses the difficulties arising from a conventional approach to BPM-SOA integration, outlines the foundations for the new and improved integration model, and lists the needed characteristics of the new model. Section 4 defines the model in detail and explains its components and structure. A proof of concept implementation on a plausible business use case is presented in Section 5. The similarities and differences of

the integration model with Enterprise Service Bus (ESB) architectures, Enterprise Application Integration (EAI) principles and the role of SOA are discussed in Section 6. Further on, Section 7 describes the evaluation of the model while Section 8 presents some conclusions.

2. Relevant Research on the Topic

SOA, BPM, optimizations of BPs and the construction of BPM solutions are often discussed topics. Relevant research ranges from emphasizing the process-oriented view [1], [6], [7], [8] to providing methods for analyzing and building business information systems [1], [8], [7]. The literature includes theoretical models for identifying the success factors of a BPM project [7], and analyses of the factors that complicate BPM implementation [3]. Articles [2] and [9] demonstrate a connection between business optimization and service-oriented architecture, [10] explores how the flexibility of SOA affects the core BPs of an organization, while [11] presents a detailed meta-model of the interactions between the activities of a BP and the functionality provided by software services. The benefits that BPM and SOA architecture bring to a business environment are further discussed in [5], [6], [12] [13], whereas in [14] one can find an interesting comparison of the structure of companies based solely on BPM or solely on SOA principles.

The process of building, optimizing and using a BPM solution, based on the specific requirements of a use case scenario, is explored in [4], [8], [15] and [16]. Their research aims range from emphasizing the importance of monitoring key process indicators (KPIs) [8], [15] and honoring service level agreements (SLA) [15], to applying the basic elements that support agility in managing BPs by utilizing Capability Maturity Model (CMM) [16].

Prototypes of BPM solutions are also described. Research in [4] presents a solution, specific for Third Party Logistics or 3PL companies, that integrates the processes within the organization with those of customers and external partners, while the idea of a business process model based on a set of services offered by distributed service providers (Software-as-a-Service or SaaS) is explored in [12]. BPM solutions are also used to detect misalignments between a BP and the supporting software systems and the effect of the misalignments on the performance of the BP [17].

Several attempts to integrate BPM in existing systems were made, but the articles describing them often do not provide a sufficiently detailed discussion of the conceptual model presented, the principles it is based on, the role of its components or the technologies, tools, frameworks and guidelines for its implementation. Implementation on a business use case, or a real-life problem that inspired the creation of the model, along with some kind of evaluation are also lacking. The existing articles, however, outline some interesting ideas, as will be seen in the following review.

An example of incorporating existing SOA services into BPM by using the modular principles of object-oriented approach is presented in [13], but the idea presented is to create a BPM solution in an SOA environment from scratch and without using a BPM tool. Unlike in [13], the authors in [5] discuss Business Process Management Systems (BPMS) and emphasize the need to apply a BPM tool while constructing BPM solutions. However, they define the roles for development teams, business users and

project managers, which means that they mainly analyze the human aspect of the BPM integration effort.

Matei in [6] considers optimizing BPs by managing the existing services in an SOA architecture. He concludes that existing legacy applications should be re-engineered into a set of reusable modules or services. Still, the re-engineering of large and complex legacy applications can be difficult, if not even impossible.

The question of redesigning legacy systems is also discussed in [18] where the authors describe a three-layer SOA-BPM architecture which is based on a central service repository specially adapted for building an effective apparel quotation system. The authors also state the need for further research to determine the best way of integrating BPM into legacy systems. The dilemma is whether to build new services or to transform, adapt and reuse legacy systems.

The approach studied in [19] presents the integration of a BPM system in a healthcare environment. The focus lies on the dynamic allocation of long-lasting tasks to the currently available, or most suitable, practitioners. The authors define a middleware layer to connect the semantic layer (ontologies and rules used to define healthcare concepts and management principles of interdisciplinary healthcare teams or IHTs) and the execution layer (BPs defined in the BPM tool and existing legacy applications, or Hospital Information Systems, HIS). However, since healthcare is a branch where most work is done in practice, the role of the existing HISs is limited and not clearly defined, mostly used to start a BPM process on the arrival of the patient. With the help of the semantic layer, BPM is primarily used as a tool for assigning tasks and for further task and process management. The concept of reusing legacy application functionalities, or interfaces, is not studied as process actions are not executed in legacy applications and usually do not span through multiple HISs. The study presented in [20] addresses the need for business agility in the domain of insurance companies by creating a SOA architecture comprised of a Workflow Management System (WfMS), ESB, a Business Activity Monitoring (BAM) and a Business Rules Management (BRM) system. Although the architecture uses the ESB to invoke different services (via Remote Method Invocation or RMI), the option of replacing the exposed services with applications, interfaces and parts of applications' functionalities is not considered.

The usage of existing legacy applications as a part of the process monitoring system is presented in [14]. A service interlayer is constructed to connect systems across the organization and create an overview of the BPs. However, the additional service interlayer is built on top of legacy applications, and thereby uses the existing applications as a basic resource instead of incorporating the functionalities of these applications directly into the BPs.

A quite interesting idea is presented in [4] where the authors discuss the creation of software components that would interface legacy systems and use them as building blocks for the BPM solution. The reuse is limited to the fact that the existing legacy services based on Apache Tuscany implementation of a Service Component Architecture (SCA) are used to model a BP workflow in an OSWorkflow open source engine. Although the authors mention the reuse of legacy applications, the concept is not elaborated or studied further.

In [10] a complex and detailed view of the SOA-BPM integration is presented from the technical perspective. The model requires that parts of BPs (business rules and content specific routines) be extracted into separate components which would then be used to build a complex but effective BPM solution. The article does not clearly define

the role of existing applications, and specific BPs contained within those applications, in a final BPM solution. The focus of the presented model lies on redesigning legacy application architecture, rather than building upon and using existing structures as its foundation. Also, the usage of a BPM suite was not considered.

The need for further research arises from the fact that detailed models describing the integration of BPM into a service-oriented legacy environment are still lacking. Also, only a small number of studies analyze tools and implementation methodologies which would make such integration possible.

3. Building BPM Solutions, Overview, Case Study and the Characteristics of a New Model

An overview of SOA and BPM is given in the following subsections, along with the explanation of difficulties that organizations face while building BPM solutions in existing legacy environments. A live real-world example of a legacy business application is studied in detail. Based on the conclusions, a set of characteristics needed in the new integration model for building a BPMA in an existing SOA business application architecture is outlined.

3.1. Differences, Similarities and Connections of BPM and SOA

A general discussion about the challenges of SOA and BPM is given in this section. Similarities, differences and the importance of a proper co-dependence of the two approaches are illustrated.

SOA and BPM both have a common goal: to achieve success in a business environment by accomplishing the desired business goals, a set of stable states that are valid after the BP is executed [17]. Generally, while BPM deals with the execution of well-defined processes in a specific order, SOA aims to expose the functionality of a system using well-defined services [5]. BPM and SOA concentrate on the management and coordination of processes, or services, and organize them in different repositories where they can be monitored and controlled [10]. By their nature, both SOA and BPM are iterative approaches where a set of predefined phases, or steps, is used to build, test and analyze the desired functionality, service or process [14]. Although SOA and BPM complement each other, they exist on different levels.

Operating on the lower level, SOA represents a set of technical principles and gives a dose of flexibility to the architecture of information systems [9] [20]. It has the task of combining simple services, self-contained units of functionalities [8], into complex ones while taking into consideration a set of technological limitations [10, 14]. As such, SOA serves as a network for filling the gaps between the infrastructure, data, and a BPM solution [4]. It ensures the technological means that provide resources for the implementation of goals defined by BPM.

As a set of management principles, BPM is on a higher level. It is more about the decomposition of processes into sub processes and activities and it ensures that the processes are integrated, automated, optimized, monitored and documented, so they can effectively support the decision-making process [7], [20]. BPM groups, sorts and

presents the fetched data in a user friendly and logical way. The presented data is then subjected to different user input and user actions which, finally, create a quality service for the customer. Therefore, BPM refers to management of resources; it is a roadmap that integrates human-driven processes (processes in which human interaction takes place) with the use of technology to create the desired service or product.

Despite the availability of several BPM and SOA software solutions on the market, their full integration is still a tremendous practical challenge, mainly due to extensive organizational and IT changes that an organization needs to undertake [12]. However, there is evidence that a partnership between BPM and SOA produces an effective and indispensable solution that meets the challenges of the modern business environment [13].

3.2. Motivation, Difficulties and the Traditional Way of Building BPM Solutions

Let us explain the difficulties involved in the conventional approach to integrating BPM with an existing SOA-based legacy environment and analyze in detail the ensuing disadvantages using a real-world example of a legacy business application.

An inspection of the complex business environment in the banking sector reveals a multitude of different interconnected and intertwined business activities and processes. Changing a single process can affect multiple processes and cause unintended, negative consequences. The primary motivation for introducing BPM in that kind of environment is a systematic overview and control of this vast business architecture.

Satisfactory control can only be achieved if related business activities are first identified and grouped, as this step would lead to the use of individual applications for each set of related activities. That way, each separate business area, for example the entire business of corporate clients, would be controlled through only one complex module. A centralized management of a business area would allow a single point of entry, from where one could, using only electronic channels, create a business offer, arrange a meeting with a client or initiate processing and contracting of desired services. In addition, one would have total control over the situation and be able to see the status of issued requests at any point in time.

Most BPM solutions which are integrated into a SOA model are based on the fact that the presentation of business data and interaction with them are closely related to the BPM tool and, in most cases, presented by it [13]. That means that users can access, change or view the presented data using various predefined user interfaces and according to the given roles [10]. A business arranged in that manner is closely dependent on the BPM tool. Consequently, it is very difficult to sell a BPM-dependent software solution to a third party who do not use BPM-related software without extensive redesigns.

In addition, the implemented BPM tool must be able to handle large volumes of business data and massive objects, which complicates the implementation process. A comprehensive presentation of data hinges on the creation of complex forms with many data fields, menus, labels and presentation-related business logic, which is often not fully supported by the tool itself. Such a cumbersome architecture can produce quite a vulnerable application, which is then difficult to maintain or upgrade efficiently. The

differences in data presentation between an architecture that contains legacy applications and an environment with a BPM solution are outlined in **Fig 1**.

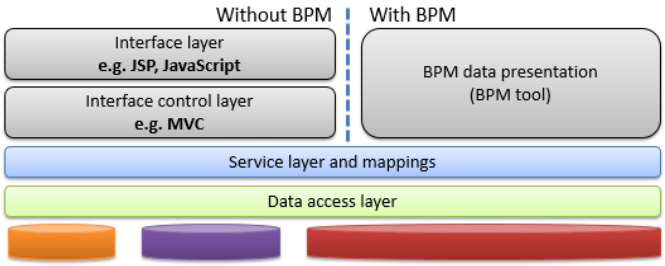


Fig 1. Architecture of a system with and without a BPM tool

Finally, a successful BP execution depends on the communication between different software applications supporting different parts of BPs. Much of the changes or reshaping of processes in the real world depends on how difficult it is to change the communication between the mentioned parties. If no standardized means of communication exist, communication is often ingrained in different, separate and monolithic software solutions. Processes built in that manner are strictly related to the technology used in the background and are, therefore, inert to changes which become very difficult and expensive to implement [18].

Legacy Business Application Case Study

To illustrate the problem accurately, let us look at a real-life candidate for a BPM implementation: a banking application managing the business of corporate clients. The application is based on the Java Spring framework and SOA architecture. It consists of four layers where each layer is implemented in a separate project. The data access layer is used for creating, updating, removing, and retrieving data from different databases. The service layer is like a wrapper around the data access layer that integrates the fetched data into reusable services using the basic business rules. The web presentation layer consists of interfaces and interface control components that manage the user's interaction with the data. Services from the service layer and additional business rules are also used to arrange, present and manipulate the acquired data. For example, in an application based on the Spring framework, the web presentation layer would consist of a Model–View–Controller or MVC pattern combined with JavaServer Pages (JSP) and JavaScript. The domain model, or layer, is common for object-oriented systems and describes the business area one is dealing with, for example banking, insurance, sales, etc. The domain model reflects the business model and consists of object representations of real-life artifacts specific to the domain in question. In the banking domain, one would have objects representing accounts, customers, orders and so on. Objects from the domain model are used in all other layers of the application.

Since its first operation 10 years ago, the application has constantly been developed and upgraded as the business environment grew and regulations changed. Over the 10 years, a team of 8 people on average continuously worked on the application. The application's core data is stored in two Oracle database instances, each containing a

database data model comprised of more than 400 data tables. The application is connected with 18 different internal systems and applications which provide a variety of banking services, e.g. credit rating management along with on demand credit score calculation, service and product catalog functionalities, client exposure calculation, depositing management, access to a variety of central records, liquidity management, document digital signature support, extensive reporting in conjunction with a data warehouse system, document management features, etc. Additionally, external service providers are consulted regarding matters like the Foreign Account Tax Compliance (FATCA) or the verification of black-listed clients.

To clearly see the challenges of redesigning and implementing the described banking application in a BPM suite, one must grasp the size of the application and the resulting amount of work invested in its development. Even more importantly, one must compare the sizes of the application layers. To achieve the above goal, let us analyze the application in terms of the number of lines of code (LOC), not counting reports, JavaScript, CSS and JSP pages which are all part of the web layer. An analysis using SonarQube, an open source platform for analysis and inspection of code quality, revealed that there are 396.560 lines of Java code in the entire application. The size of the application layers compared to the total size of the application is shown in **Fig. 2**.

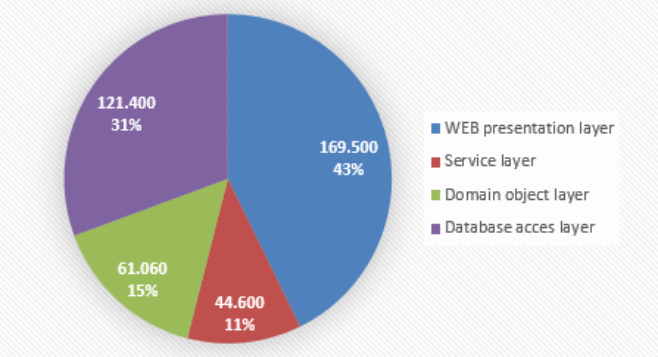


Fig. 2. Comparison of layer sizes in a banking business application based on the number of LOC

If we want to convert our banking application into a BPMA, a typical method would be to separate the service and database access layers and expose them as standalone services. Those services would then be used by the new BPMA. To create the BPMA, however, the entire Web presentation layer of the legacy application should be rewritten in a BPM tool. As shown in **Fig. 2**, the Web presentation layer makes almost a half of the original application (Java code + JSP + JavaScript + Jasper reports). Rewriting it from scratch would be a daunting and long-lasting task. To avoid that problem, an alternative BPM integration model is proposed in this article.

3.3. Characteristics Needed in a New BPMA Integration Model

To further clarify the contribution of the paper, let us look into the foundations of the new model for building a BPMA within an existing SOA business application

architecture. Based on the conclusions made in the previous section, the following characteristics of the new model were identified.

Redesigning and totally rewriting existing processes in custom BPM tools, new frameworks or building new applications must be avoided. According to the case study presented in the previous subsection, rewriting everything from scratch would require a large amount of resources, time, and would also limit the organization's ability to deal with the ongoing market challenges during the redesign period. Therefore, a new model should, before all, maximize the utilization of existing functionalities, existing applications and architecture while requiring as little adaptation and change as possible. Maximum utilization of the existing architecture reduces the cost of BPM implementation as the use of resources is kept to the minimum. Also, since less work is needed, the implementation time is reduced.

In order to produce a loose coupling between a BPMA and the existing architecture, emphasis should be on adaptability. The existing legacy applications need to be operational without a BPMA as the system may become unavailable due to failures or upgrades of the BPM tool. The organization can decide to switch BPM tools and use a different, better, tool from a different manufacturer. In extreme cases, the organization can decide to stop using its BPM tool in order to cut costs or if it aims to develop a custom solution that better fits their needs.

A BPMA should manipulate with only a reduced set of business data required to achieve the desired functionality.

A major requirement is to allow changes in the business process flow by modifying the business logic contained in the BPMA. This needs to be done with minimal changes to the background applications and systems, or ideally, without modifications.

The ideal solution should be independent of different hardware, operating systems or technologies used in the organization. The integration is more successful when communication is independent of the back-end systems.

Every business process flow must reflect the real-time status of the underlying data in the background system. Changes in BPs must be visible in the underlying applications automatically, and the state of the data in those applications needs to be promptly displayed in the process application.

To assure flexibility, the integration model must be able to operate with different BPM tools simply by adapting the implementation of its Application Programming Interfaces or APIs for the interaction with the new tool.

Finally, it is important that the integration model is scalable. Scalability would facilitate the integration and use of different legacy applications, developed in different technologies and from different parts of the organization in a BPM solution.

4. The Proposed Integration Model

In order to achieve the characteristics described in the previous section, create a fluid data flow, clearly define the business environment and enable the flexibility of the system, the following architecture is proposed (**Fig. 3**).

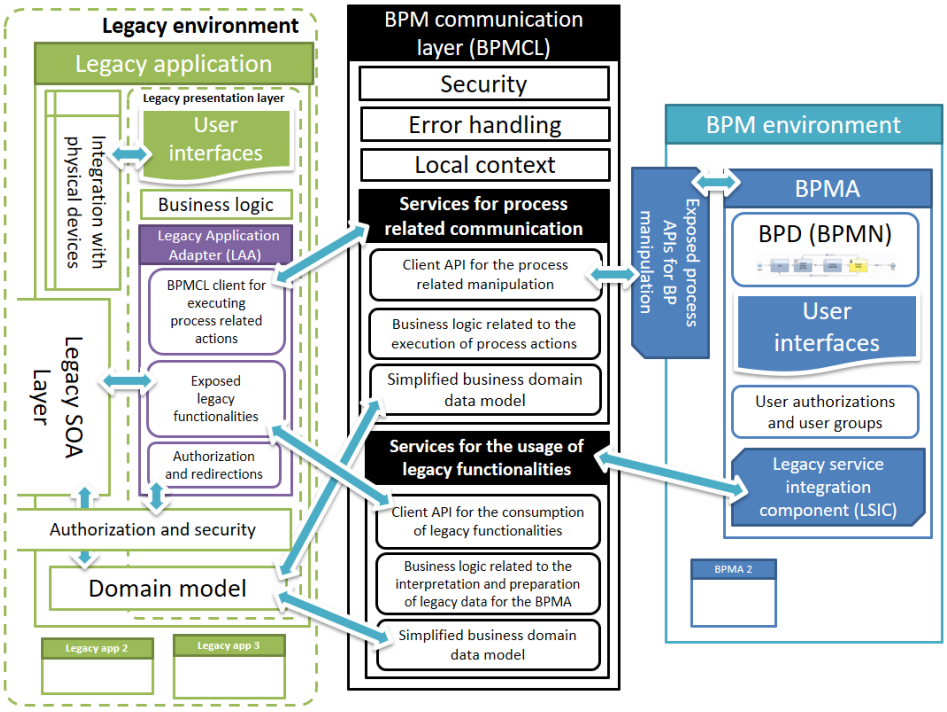


Fig. 3. Conceptual diagram of the proposed integration model

The integration model is based on three key layers or parts: the legacy application layer, BPM communication layer (BPMCL) and a BPMA built in the BPM tool. Legacy applications contain parts of the functionalities of existing business processes. BPMA describes the existing business process with a BPD modeled in congruence with the Process Model and Notation (BPMN) standard. BPMCL provides the methods for a universal way of communication between the legacy applications and the BPMA. The role, functionality and capabilities of each layer are shown and explained in detail in the following subsections.

The original contribution of this article resides in the definition of the BPMCL layer and the principles of integrating the BPMA with different legacy applications via the BPMCL in a way that exploits the existing functionalities and user interfaces while avoiding extensive redesigns of existing environments. Additional contributions lie in the conclusions of the presented case study, proof-of-concept implementation study and the extensive evaluation of the model by relevant experts from the BPM area.

4.1. Legacy SOA Layer

This layer represents the SOA implementation in the organization and enables the usage of existing functionalities. It contains the data access and service layer.

The data access layer consists of a variety of data servers, media for storage of all types of data, different databases on different machines and in different technologies.

All business data used in the system is initially retrieved through this layer. Some basic logic, required for different lookups, is also implemented in this layer.

The service layer consists of various services with specific purposes which can be divided into several groups or categories, based on the purpose. The division improves performance, enables better understanding and eases the maintenance of the overall system. Based on our analysis of the legacy environment (section 3.2) basic service categories were identified. If required, additional service groups can be added to the presented list. An overview of the proposed service categories (with examples) is shown in **Fig. 4**.

The first service group consists of services for a direct access to databases that manage the data required by the key functionalities of the applications. For example, a set of services that manages payment orders in an internet banking application. The basic characteristics of this category would be speed and high availability of the underlying databases.

The second group should consist of services accessing resources or databases that do not have a key role in the functioning of the system. A typical example is fetching data that are rarely used or accessing data archives that have no impact on the core functionalities. Another example of a non-essential resource would be a document management system (DMS). Such a system can be used to store a variety of contracts and confirmations that the clients use, but which are not crucial for performing core functions, for example execution of payment transactions.

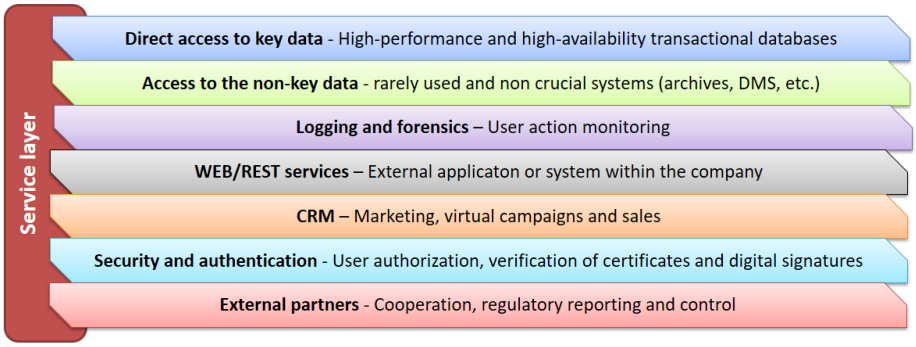


Fig. 4. Basic service layer groups

The next group encompasses monitoring and forensic services that enable tracking and logging of key actions that the user does while working with the application. A good practice here is to use NoSQL databases as they can store large quantities of data relatively fast.

Another part of the service layer would contain remote calls to the exposed web services of some external application, or system, to fetch additional data when needed. For example, in an online banking application, the data set to be retrieved from the external system can consist of detailed information about the clients from the central register, or verification of retail payment accounts. Services that would make such data exchange possible, broadly speaking, represent the communication between two applications over the Internet and can be implemented using Representational State Transfer (REST) or Simple Object Access Protocol (SOAP) based services.

An important part of the service layer is a connection with a specific Customer Relationship Management or CRM system which, as a marketing and sales tool, is used to inform the client about the current benefits, additional offers and new services.

The next addition to the service layer is the group of security and authorization verification services. A typical example is the use of a Validation Authority (VA) server to verify the certificates of users during the execution of key actions in the applications. Those actions include authorization during initial login, execution of orders, change of user authorizations, etc.

A final group of the service layer are services for communication with external companies, business partners or regulatory bodies. A classic example could be the delivery of mandatory reports to the tax administration offices or collaboration with a financial mediation organization.

4.2. Presentation Layer

The presentation layer integrates two types of user interfaces, parts of the presentation layers of legacy applications and simple BPMA user interfaces constructed in the BPM tool.

The presentation layer of the legacy application enables the interaction with data retrieved by the legacy SOA layer. It enables the usage of complete business functionalities and provides a comprehensive view of all business-relevant data. Interaction with the processes implemented in the legacy application is done entirely through this layer, along with the execution of all process-related actions.

The BPMA user interfaces deal with the initial user interaction as the user goes on to solve the tasks that were assigned to him/her. The primary role of the BPMA user interfaces is to show the crucial data regarding the process, and to redirect the user to the predefined forms in the presentation layer of the legacy application, where the rest of the work is done. To redirect the user, the BPMA utilizes the legacy service integration component (LSIC) which fetches the data needed for the redirection, and the authorization and redirection component of the legacy application adapter (LAA) which authorizes the user and executes the redirections trough the legacy application. The details of the redirection process are described in detail in the following subsection.

As an optional part of the presentation layer, one can show a BP diagram at each step of the process. This would allow users to clearly see the current and all remaining actions.

Legacy Application Adapter (LAA)

Legacy applications execute the actions of the existing business processes. To enable their communication with the BPMA, a Legacy Application Adapter (LAA) component is created. Integrated into the legacy application, the LAA has the task of sending instructions to the BPMA and responding to the requests of the BPMA with the help of the BPM communication layer, which will be described in the following section.

To help determine the intensity and the nature of the communication between the LAA and the BPMA, the term *process point* was defined. *Process points* mark locations where the LAA and the BPMA exchange information. Process points can be *standard*

and *background*. *Standard process points* are events in the process flow where some process-related data in the legacy application changes. At that point, the legacy application communicates the change to the BPMA by updating the state of the BP instance. Updates can consist of minor changes of the BP instance-related data, or they can indicate a major change in the state of the process; e.g. mark a completion of a task or a point where the process flow moves from one legacy application to another. *Background process points* are important for the BPMA and indicate a call of the BPMA to the LAA using LSIC in order to fetch business data, get direct links to the tasks in the legacy applications, or to perform simple background checks. *Process points* have proven to be important for the discovery of the processes in legacy applications and their integration with the BPMA.

As the user is redirected from the BPMA to the legacy application to complete his/her task, the first point of interaction with the legacy application is the authorization and redirection component of the LAA. The authorization and redirection component completes the user login process with the help of the mechanism that already exists within the legacy application. Next, it prepares the necessary data regarding the user task and redirects the user to the screen where he/she can perform the assigned task. Upon the completion of the task, the LAA updates the status of the process in the BPMA and returns control over the process to the BPM environment.

To further elaborate the user interaction with the legacy application interfaces, *interaction points* were defined. *Interaction points* identify parts of the process flow where the user begins/ends the interaction with the process in the legacy application. Interaction points usually include *standard and background process points*. *Interaction points* have proven to be important concepts during the design of the authorization and redirection component of the LAA.

Each LAA consists of two parts. Application specific components contain the information concerning the legacy application so that the adapter can interface the BP in the legacy application. According to **Fig. 3**, specific components of the LAA include the authorization and redirection component and the component for exposing legacy functionalities. Existing authentication mechanisms, available in the legacy environment, are used by the components of the legacy application, while the authorization and redirection component of the LAA uses those legacy components to perform the authorization of the BPMA users (BPM users access the legacy application via specific links, which is further explained in sections 5.1 and 5.2). The component for exposing legacy functionalities can expose the functionalities of the legacy SOA layer in a needed form (e.g. via Representational State Transfer or REST endpoints). More importantly, however, it can expose parts of the functionalities that exist only in the legacy application and are needed for the BP execution. The specific components of the LAA are different for each of the legacy applications due to the differences in technologies, frameworks or methodologies used in the legacy applications.

The second part of the adapter is a general component that knows how to use the BPMCL to communicate with the BPMA. The general component, BPMCL client, utilizes the services for process-related manipulation provided by the BPMCL and is basically the same for each of the LAA in different legacy applications. It is possible to apply a method for automatic generation of the general LAA component in different programing languages, as described in section 5.1.

To illustrate the scalability of the model, let us examine the process of adding an existing BP, or a part of a BP implemented in a legacy application, under the control of the BPM solution. The described task would require the following actions:

- Creating the BPMA by building a BPMN model of the process in a BPM tool.
- Integrating and adjusting the adapter in the legacy application.
- Modifying the specific part of the adapter and connecting it with the targeted process, or parts of the process, by identifying *process* and *interaction points*.
- Integrating the communication with the BPMA based on the process workflow of the created BPMN process model.

An illustration of integrating the adapter in a legacy application, to support a part of the BP, is shown in **Fig. 5**. The existing legacy application functionalities are reused to execute the activities of the BP.

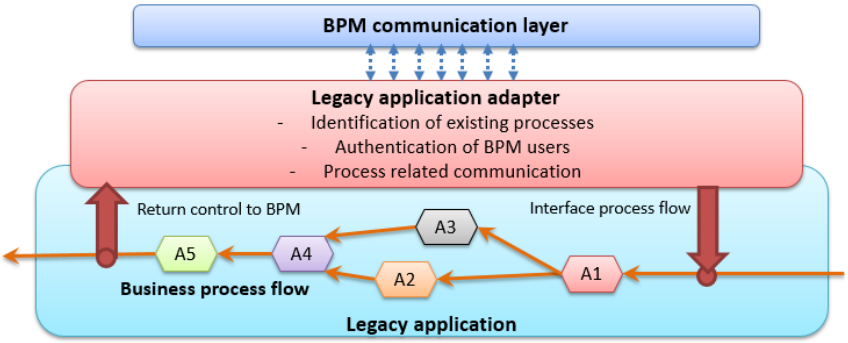


Fig. 5. Adapter interfacing a process in a legacy application

The adapter also has the task of separating the legacy application from the BPMA. If the need arises, it can be used to completely exclude the BPMA from the legacy application. The adapter allows the interaction for only specific user roles and in line with their authorizations. Legacy application users, with roles that do not require BPM access, can continue to execute their daily tasks in legacy applications.

4.3. BPM Communication Layer (BPMCL)

BPMCL is the component, a middleware layer, which enables the communications between the legacy systems that perform specific business actions and the BPMA. It contains the methods for manipulating process data by calling process APIs exposed in the BPM environment. BPMCL gives flexibility to the integration model by providing means for the orchestration of BPs built in a legacy environment using a generic BPM suite.

Communication, which takes place in the key parts of the business process flow, consists of precisely defined services performing specific actions. Actions include the creation of processes, change of status, retrieval of the allowed user tasks, etc. Additionally, methods for the retrieval and manipulation of data from SOA-based legacy systems are also available on the BPMCL. The BPMCL can fetch legacy data from multiple legacy applications, perform the needed operations over the data, and present it to the BPMA in a desired way.

The BPMCL has its own domain data model that mirrors the BP data structures of the BPMA and is used in the communication with the BPMA and the legacy applications. This greatly simplified model is created by reducing the domain model of legacy applications so that it contains only a basic set of data needed by the BPMA.

Special components within the BPMCL are responsible for security and error handling.

A component for managing the local context of service calls is also defined. The local context is a devised principle where a specific set of data is transferred with a service call to identify the caller and define the transaction specific parameters of the service call. The local context is valid for the duration of the service call and is mandatory for each of the service calls. The following exemplify local context-related data: an identifier of the legacy application initiating the call, process instance id, or data related to the user who is working in the legacy application.

All presented BPMCL components are further explained in section 5.

4.4. BPMA with its User Interfaces

A BPMA is built in a BPM environment by using BPM tools. The process of BPMA creation requires that flow charts of business processes and sub-processes be defined. A BPMA created in this manner contains a process model of an existing BP that is executed in a legacy environment with the help of several existing legacy applications.

Based on user input, business process manipulation (starting a BP, changing BP status, assigning user tasks, etc.) is performed using process APIs (programmable interfaces related to process-bound communication) which are exposed in the BPM environment. The exposed process APIs enable access to BPM artifacts, business processes, tasks and process-related data. Process APIs, which are part of the BPM tool used by the organization, depend on the specific implementation of the tool's manufacturer.

In order to retrieve business-relevant data needed in certain steps of BPs, the BPMA contains a legacy service integration component (LSIC). The LSIC, with the help of the BPMCL and the business logic implemented in the BPMCL, calls the legacy SOA services exposed by the adapters in legacy applications. This solution enables direct communication with the organization's SOA implementation while the business logic implemented in the BPMCL allows one to interpret, manage or group the returned data as needed.

User authorizations and user groups defined within the BPM environment are used in BPMAs when building BPDs.

A set of BPMA user interfaces, input and presentation forms, together with the legacy presentation layer, are used to handle user interaction. User interfaces constructed in the BPMA provide the central points for user interactions. They receive and redirect users to a legacy application where the functionalities of BP actions are performed.

Different complex data types used by legacy applications and further simplified by the BPMCL are mapped to simple data types that a BPMA can handle.

5. Utilization of the Integration Model, Proof of Concept

In this section the proposed integration model is further explained from a business point of view. In addition, the section discusses the technical implementation of the model in a real-life legacy environment. A simplified version of a business process from the banking environment, the Loan Utilization Request Approval Process (LURAP), is used as an illustrative proof-of-concept (PoC) scenario to test the implementation of the integration model and confirm the feasibility of the overall approach.

5.1. Technical Concepts and Implementation Principles

This section describes the implementation principles of the proposed integration model and presents the selection of tools, frameworks and methods used in a PoC implementation effort in a real-life business use case scenario. Due to the limited scope of this work, the used software solutions, tools and technology standards are not explained in detail.

A standardized method of communication between the components of the model needs to be defined and the issue of security must be addressed. In the PoC implementation effort, communication is achieved via REST API due to its simplicity (JavaScript Object Notation or JSON data format), ease of implementation, testing (e.g. Postman REST client) and compatibility with legacy systems written in different technologies. HTTP Basic authentication method, in combination with SSL (Secure Sockets Layer) protocol and additional custom HTTP headers, was used to secure the communication.

As part of the architecture proposed in this article, a general BPM tool for building the BPM process model is needed. Process modeling, the creation of BPDs and their elements in the BPM tool, needs to be consistent with the BPMN standard as it enables the construction of a BPMA with all the elements described in section 4.4. Interaction with the BPMA must be enabled using predefined APIs exposed on the BPM tool. APIs must provide control over process instances, BPDs, services, activities, users, user groups and user tasks. The BPM tool needs to have the ability to call external services so that the LSIC can be created as shown in **Fig. 3**. Additionally, the selected BPM tool needs to support the construction of basic user interfaces.

To ensure the validity and flexibility of the integration model as shown in **Fig. 3**, and to validate the listed requirements of the BPM tool, a study of the APIs and features of four BPM suites was conducted (Red Hat JBoss BPM Suite, Bonitasoft BPM, Alfresco Activiti BPM and IBM BPM).

For the purpose of this PoC case study, a BPM tool from the IBM organization (IBM Business Process Manager or IBPM), a leading commercial suite in the BPM space [19], was selected. All details about the IBPM tool can be found in [21]. No in-depth analysis of the tool's structure, functions, or the philosophy of an iterative BPM implementation process will be undertaken in this article. IBPM provides control over all process-related artifacts through a set of predefined REST APIs, while a combination of dashboard and task completion coaches is used to enable human interaction with the BPMA [21]. IBPM Integration Services are used in the LSIC, while General System Services manage the data within the BPMA. Decision Services are used to apply

business rules during the process execution with the help of BAL (Business Action Language) rules [21].

The role of BPMCL needs to be implemented in a standalone service application that is able to communicate with other legacy applications and the previously selected BPM tool. A variety of frameworks and tools are available for the task. For the PoC implementation Spring Boot framework along with Maven as a dependency management and build automation tool was used. Spring Boot is a favorable candidate for the construction of BPMCL as it features a minimal, non-XML based configuration, has an embedded servlet container, offers extensive support for nonfunctional requirements and enables a relatively simple installation in test and production environments. Detailed technical description of the basic BPMCL components used in PoC implementation is given in the following chapter along with a simplified Unified Modeling Language (UML) class diagram shown in the **Fig. 6**.

RestInterceptor implements the HTTP Basic authentication and has the task of processing *bpm-access-request-header* custom HTTP header. The data from *bpm-access-request-header*, along with *RequestLocalContext* and *RequestLocalContextHolder* classes, is used to implement the local context principle described in section 4.3. *CustomResponseEntityExceptionHandler* provides general exception handling while *ExceptionResponse* class represents a universal exception description. Services for process related communication are implemented via *LoanUtilizationBpmRestControler* REST endpoints, which contain the need business logic, use the simplified domain data model and execute the client APIs for the process related communication (*BpmRestManager*). One part of the *BpmRestManager* is a *BpmApiEndpoint* enum containing URI identifiers of a predefined format for the execution of the IBPM REST APIs. For example, to start a task `"/bpm/wle/v1/task/{taskId}?action=start&x-http-method-override=PUT"` URI is defined. Clearly, a verb tunneling principle [21] is used when the number of characters in a GET request exceeds the practical limitations of the HTTP protocol. Following the principles of services for the process-related communication, services for the usage of legacy functionalities are implemented via *LegacyRestController*, *LegacyManager* classes.

Springfox implementation of the Swagger 2 specification (*SwaggerConfig*) is used in BPMCL to document the functionalities of BPMCL based on an OpenAPI specification (OAS). The generated OAS document, in JSON or YAML (YAML Ain't Markup Language) format, enables computers to detect and people to understand the REST service endpoints without accessing the source code. This implementation exploits the primary advantage of the OAS: code generation tools can use it for the purpose of automatically creating REST clients, or REST endpoints, in different programming languages. To achieve the task, a Swagger Codegen tool is combined with AnthillPro continuous integration system to create BPMCL REST client artifacts and store the artifacts on a central repository of the organization (Sonatype Nexus).

The generated BPMCL REST client (general component of the LAA) needs to be adapted to the technologies and programming languages used in the legacy application. Therefore, to further ease the implementation, Swagger Codegen tool is used to generate REST clients in different programming languages.

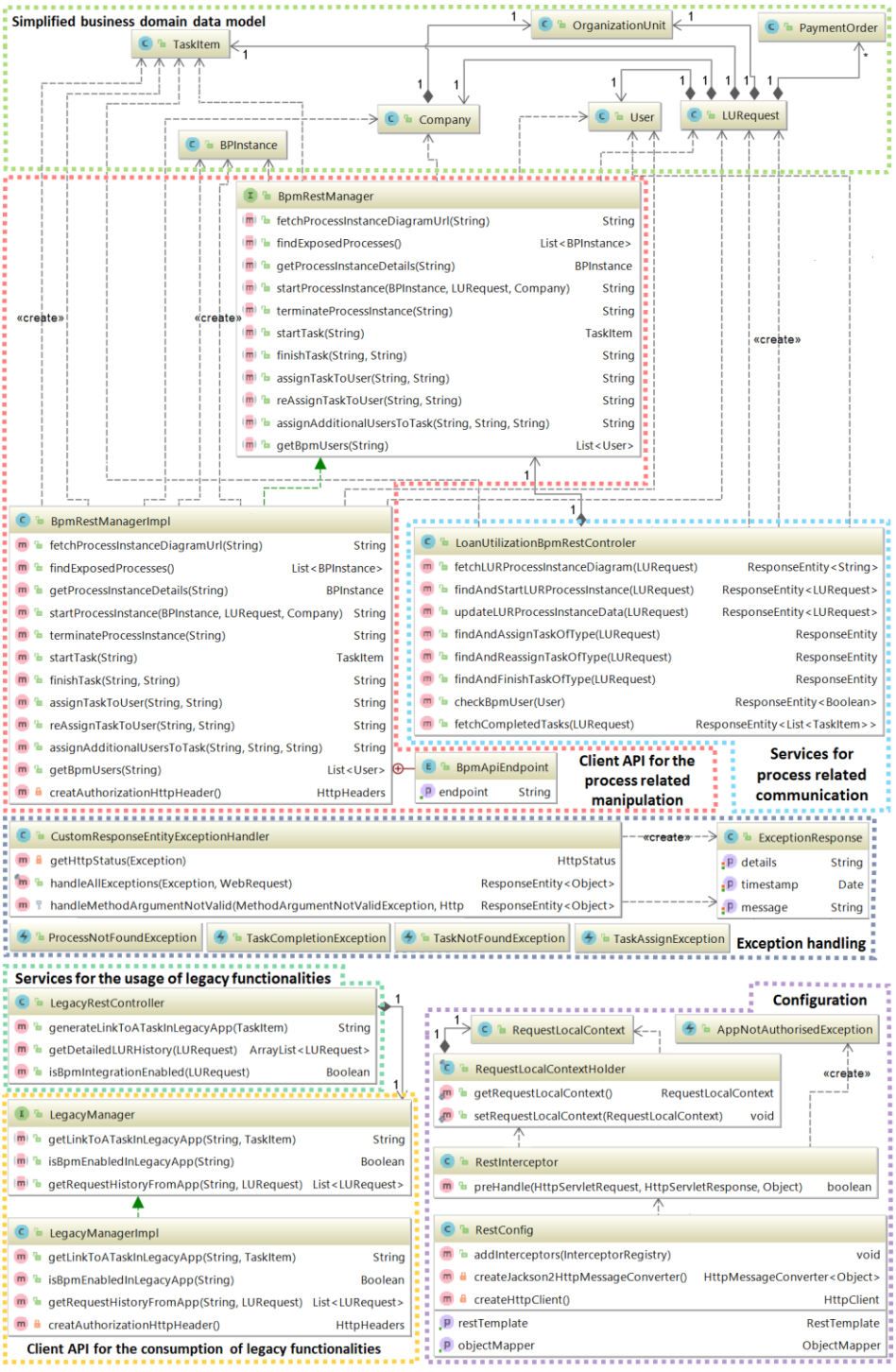


Fig. 6. Basic class diagram of BPMCL components

Once the generated BPMCL REST client artifacts are uploaded to the central repository of the organization, they can be included in a legacy application (LAA component) with the help of a dependency management system.

Additionally, based on OAS, IBPM can automatically discover implemented REST services. Integration services of the BPMA LSIC are created by importing the OAS of the BPMCL's services for the usage of legacy functionalities (*LegacyRestControler*). Due to the limited space in **Fig. 6**. the interdependence between the services for the usage of legacy functionalities and the simplified business domain data model of the BPMCL is not shown.

LAA components largely depend on the structure and the technologies used in the legacy applications. However, some technical implementation guidelines can be applied in general. In terms of the Spring MVC model, the role of an adapter can be implemented into a special set of controller and helper classes that will work as a central entry point to the application. If we have a single page application, an LAA can be created as a separate set of components with a dedicated routing module.

Let us discuss a Spring MVC legacy application with an LAA. A BPMCL Java REST client artifact, auto-generated by Swagger Codegen, is used by the legacy application to execute a process-related action in *standard* process points. Simple REST controllers expose the existing SOA functionalities for the usage by the BPMA in *background* process points.

Authorization part of the redirection component is implemented in a central controller that uses the existing login methods of legacy applications: certificate-based authentication via Public Key Infrastructure (PKI) and a simple username and password authentication. Certificate-based authentication requires integration with physical devices (smart card readers or USB based tokens). The code for integration was reused from the legacy application. The controller for the redirection part of the authorization and redirection component, based on the process and type of action, redirects the user to the *interaction point* of the legacy application. After the user executes the action, the legacy application is closed.

5.2. Business Use Case Scenario

The LURAP (**Fig. 7**) is chosen due to its complexity. This business process both requires multiple user authorizations on different levels and involves multiple legacy applications in the process flow. It is, therefore, suitable for the demonstration of the integration model. When a loan is approved, the funds need to be transferred to the desired accounts. To achieve the transfer, the client needs to issue a Loan Utilization Request (LUR).

The client creates the LUR in the internet banking application, fills in the necessary data, issues payment instructions and attaches the documentation if needed. Further on, the LUR is validated, signed and sent to the bank for a series of approvals. First *standard process point* is located at this step. The adapter integrated in the internet banking application contacts the BPMA, creates the LURAP instance (*findAndStartLURProcessInstance*) and starts the first task of the process (*findAndAssignTaskOfType*).

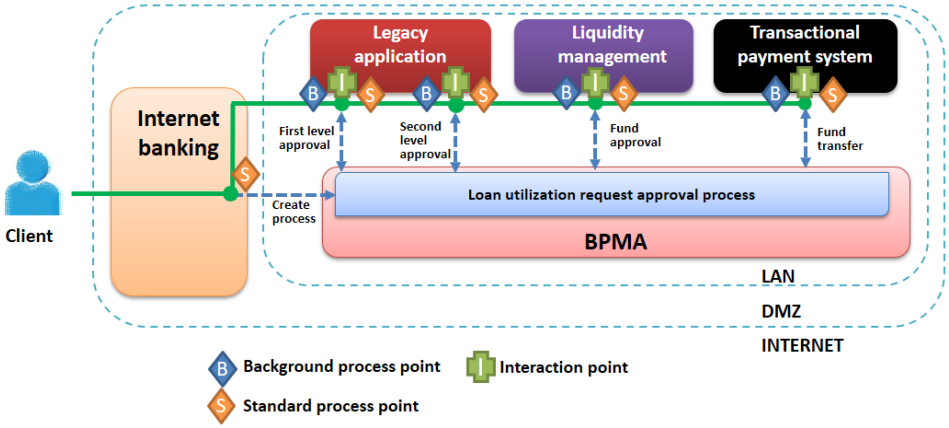


Fig. 7. LURAP BPM integration example

In the bank, working in a BPMA, a business information system user notices that a new first level approval task in the LURAP has just been assigned to him/her. Being a responsible user as s/he is, s/he immediately claims the task and goes on to solve it.

Based on the data available in the LURAP (status and type of the LUR), the BPMA asks the BPMCL to determine and create a link to the legacy application in the bank (to the LAA's redirection and authorization component) that is responsible for the necessary action (*generateLinkToATaskInLegacyApp*). The described action represents a *background process point*. Data contained within the generated link allows the redirection and authorization component to authorize the user (initial login data), prepare the data for the user task (LUR identifier, action type and business user information) and redirect the user to the exact form in the legacy application, where the required action can be executed (*interaction process point*).

Using this principle, the execution of actions directly from the BPMA is avoided. There is no need to implement complex forms and actions as a part of the BPMA because they already exist in the legacy applications. It is also not necessary to transfer and recreate the entire existing domain model (massive objects with many properties) in the BPMA, because the BPMA can function with a reduced set of data.

Let us look back to the business user and the approval process. Located on the approval form in the legacy application, the user selects the desired action from the list of the allowed actions. The application executes the selected action, verifies the LUR, and updates the status of the request in the back-end system or database. Here one can find the next *standard process point*, the legacy application notifies the BPMA about the changes (*findAndFinishTaskOfType* and *updateLURProcessInstanceData*) and the first level approval task is completed.

After the completion of the first level approval, the second level approval task is started and completed in the same manner, thereby fully approving the request. The following step in the process flow is to ensure the funds for the payments to the client. At this stage, the BPMA directs the user to the liquidity management application where s/he completes the task by reserving the necessary funds. Next, the BPMA initiates the task of creating payment orders with the help of the transactional payment system. Finally, after the payments are completed, the client is notified and the loan utilization request process ends successfully. Each of the described task executions include

background and *standard process points* along with the *interaction points* as shown in **Fig. 7**.

5.3. Remarks, Conclusions and Additional Comments

Besides the *process and interaction points* that are required for the execution of BP, one can also define optional *process and interaction points*. Optional *background process points* can be added by allowing the BPMA user to, on demand, fetch data from the legacy applications. For example, the user executes the action which fetches the history of the request from the legacy applications (*getDetailedLURHistory()*). With the introduction of additional *standard process points* users can fetch a BPD diagram of the process instance (*fetchLURProcessInstanceDiagram()*) to view the current state of the process execution while performing tasks in legacy applications. Optional *interaction points* allow the user to access different legacy applications in case some additional verification of process-related data is needed during the BP execution.

While considering *interaction* and *process points* the question of access control mechanisms arises. In order to ensure that the tasks of a business process are executed by authorized users, proper authorization mechanisms must be applied [22]. In the proposed model, user access is largely defined by the existing access control mechanisms of the legacy applications as they execute the actions of the process. Studied legacy applications are based on the standard RBAC (Role-based Access Control) model (explained in [22] and [23]), with a dynamic SoD (Separation of Duty) concept (described in [22]). User roles defined in the legacy applications contain the authorizations to execute tasks and access resources. They are defined at the process level and do not depend on the current task or process instance being executed, involved resources (documents or some sensitive data) or the elements outside the standard access control entities.

Therefore, the proposed integration model features a double-layer RBAC model. The first layer is implemented in the BPMA, where, during the construction of BPD, user groups were defined and assigned to the lanes containing tasks which the group is allowed to execute. The second layer includes the existing RBAC mechanisms of legacy applications.

Further improvements of the access control mechanisms in the proposed model could include a more advanced access control model. In that case, extensions of the RBAC, such as the TRBAC (Task-Role-Based Access Control) model where permissions are assigned to tasks and tasks are assigned to roles, could be used.

In case of a BP where context aware access control privileges (privileges depending on current context conditions) are needed, the model would require a Context-sensitive access control model for business processes (COBAC) [22]. The advantage of COBAC is that it extends the RBAC model with business processes, activities, context and resource categories [22,23]. It enables the definition of access control policies on process instance and process context level, assignment of activities to roles and definition of resource access permissions for the activities [22,23].

However, for a successful integration of COBAC concepts in the proposed integration model, further research and adaptations of the proposed model are required.

PoC implementation effort has shown that it is not always necessary to implement all LAA components in the legacy applications. For example, if a certain legacy application

has a single task of starting a process instance, only the auto-generated BPMCL client can be implemented. In a case where only certain legacy functionalities need to be used by the BPMA, only the component that exposes them can be created.

During the creation of a simplified BPMCL domain data model, several design options came into consideration. A BPMCL domain model can consist of parts of the domain models of legacy applications. In case of a shared code repository, during the AnthillPro automated BPMCL build process, classes of legacy applications are copied and packed with the BPMCL. The BPMCL can then use those data objects for the communication with the legacy applications and the BPMA.

If the domain classes of the legacy applications contain a large number of members not needed in the BPMA, one can use serialization to and deserialization from JSON to filter the not needed members. In the BPMCL, classes with a reduced number of identically named members needed for the BPMA are created. During REST calls to the legacy applications, JSON values of members that are not present in the BPMCL classes are automatically discarded in the deserialization process.

Another option is to create a separate domain data model in the BPMCL according to the needs of the BPMA. The created domain data model is included in the auto generated BPMCL client component of the LAA. During the execution of calls to the BPMCL, the data domain model from the legacy applications is remapped to the BPMCL domain data model.

The usage of non-SOA based legacy applications in the integration model, e.g. monolithic application architectures, was also considered. Due to many dependencies, intertwined functionalities and services, monolithic architectures are much harder to understand, which would cause problems in identifying and isolating parts of existing processes and difficulties in determining *process* and *interaction points*. Reusability is the main concept of the proposed integration model and successful reuse requires some sort of modularity in the existing systems. In a monolithic application architecture, the usage of separate, distinguishable components, along with parts of business functionalities is questionable. Services tend to get tightly coupled and entangled as the application evolves, which makes it difficult to isolate and use them for individual purposes. Additional modifications of the LAA component would be required, resulting in a tighter coupling with the monolithic application, which is the easiest way to complicate the code and increase interdependence.

Further detailed studies and adaptations of the integration model for use in non-SOA based environments are required, along with a proof of concept implementation effort.

6. EAI, ESB, SOA and the Proposed Integration Model

Let us discuss the similarities and differences of the proposed integration model with standard ESB architectures, EAI principles and the role of SOA.

ESB is a set of principles, an architecture, that allows multiple applications to communicate with one another via a bus-like infrastructure. Fundamentally, ESB architecture is a “communication agent”. It receives a message, translates it into the appropriate format and sends it to wherever the message is needed.

The proposed integration model serves to control the BPs and redistribute the work that certain applications must perform within a BP by utilizing existing functionalities.

Besides enabling the communication of different systems, the proposed model not only utilizes the existing user interfaces and the business logic but also coordinates human interaction in the execution of the BPs. ESB orchestrates services [20], while the proposed integration model orchestrates processes by identifying, extracting and integrating activities from different legacy applications with the goal of achieving a centralized point of interaction for the successful execution of BPs.

The integration of data across applications, with the goal of simplifying BPs extending through the connected applications, brings the proposed integration model closer to the EAI domain. To be precise, the undertaken presentation level EAI includes elements of the application level EAI and the user interface level EAI. The application level EAI is visible in the access to the BPs and to the data of legacy applications through legacy application interfaces. It allows the invocation of the existing business logic, is transparent to the integrated applications and also preserves the data integrity of the integrated applications. Elements of user interface EAI can be identified in the reuse of user interfaces of the legacy systems (along with any business logic embedded in those interfaces) which are combined with the user interfaces of the BPMA to create a central point for user interaction.

A BPMCL can be described as a custom extension to the ESB integration, specially adapted to the principles and needs of the proposed integration model. It is a middleware layer designed to provide the means for controlling BPs and legacy applications. While typical ESBs transport and translate data, a BPMCL also has the role of simplifying business data using its own data domain model (sections 4.3 and 5.3). Principles like the auto generation of BPMCL REST client artifacts (section 5.1) enable the compatibility of a BPMCL with legacy applications built in different technologies. The compatibility of the ESB implementation with the existing systems leads to a low-effort integration into the existing architecture [20]. A BPMCL not only enables the interaction between different systems but includes a business logic for isolating parts of existing applications as sets of reusable components, complete with functionalities and user interfaces, for the use by the BPMA. It also enables the orchestration of actions on BPs and helps in managing user interaction. Custom-defined exception handling, security and local context concepts further extend the ESB implementation.

In case that an existing ESB is successfully implemented in an organization, the principles of the BPMCL could be implemented with the ESB suite. Based on the possibilities and type of the implemented ESB suite, adaptations of the LAA, the BPMCL and the proposed principles of the integration could be made. However, further research into the topic, along with a concrete survey of the ESB suite used in the organization is required to provide an exact assessment the described scenario.

So, what is the role of SOA? ESB is one of the ways to implement SOA, a model of SOA implementation where one service is used for the management of communication between other services. Additionally, SOA can also provide the means for achieving EAI by facilitating the flow of information between applications. One can say that SOA makes the building blocks for ESB and EAI. It is the modularity of SOA, as discussed in the last two chapters of section 5.3, that enables the operation of the proposed integration model.

7. Evaluation of the Proposed Integration Model

The proposed integration model was presented to the group of 40 experts from five large commercial organizations. The group consisted of business information system users, project managers and business owners, and they were asked to give their opinions and identify possible use cases for the integration model. All interviewed experts were familiar with the BPM but differed in the amount of experience with different BPM initiatives.

The respondents with limited experience and limited practical knowledge in BPM were aware of the basic BPM concepts and principles. For the most part, they had studied examples of BPM initiatives, attended basic BPM software trainings or limited practical tests designed to demonstrate concepts and determine the abilities of BPM tools. The respondents with extensive practical experience had regularly designed and maintained business process management systems, were familiar with the role and operation of BPM tools and had insight into the practical advantages and disadvantages of BPM solutions.

The evaluation was conducted in two phases and was based on the principles of the Delphi method. The Delphi method is a structured communication technique used to achieve consensus forecasts from a group of experts in a structured and iterative manner. In the first phase, each of the experts individually attended a presentation on the integration model. They were then asked to state their opinion of the model and present a possible use case scenario, a conclusion about, or description of the situation where the proposed model would prove useful. Following the completion of the interviews, the obtained scenarios were systematized and summarized. In the second phase, the collected scenarios were verified. Each of the interviewed experts received a summary and was asked to rate each of the scenarios. In this way, based on the opinions of different experts, the most relevant use case scenarios related to the proposed integration model were obtained, which is the goal of the Delphi method. Detailed results of the evaluation are presented in the following subsections.

7.1. Research Results

Interviews with the experts produced the following use case scenarios in which the proposed integration model would be a valuable resource.

- ***Implementation of the BPM pilot project in the organization***
  People with less BPM experience viewed the proposed integration model as a possible way of doing a BPM pilot project. This was owing to the fact that only minimal adaptations of the existing systems are necessary, which makes it a faster and less costly implementation effort. Organizations could, utilizing the existing systems, build a BPMA to support one of the existing processes. In line with the needs of the organizations, the implementation effort would provide an insight into good and bad aspects of BPM and help them decide whether to proceed with a full BPM implementation. In case of a negative decision, BPM is simply decoupled from the existing systems.
- ***Practical assessment and comparison of different BPM tools***

The flexibility of the proposed integration model is suitable for examining the features and characteristics of different BPM tools. Since BPMCL separates legacy applications from the BPM tool and provides a universal way of communication with the BPM tool, the implementation of the same process can be relatively easy linked to BPMAs built into different BPM tools.

- ***Implementation in a stable, reliable, well-built legacy application environment***
  One of the experts with extensive BPM experience stated that an integration model which maximally utilizes the existing environment is great if the organization has a solid business system. A stable and solid business system implies well-built legacy applications which successfully support the business of the organization and are expected to meet the needs of the organization in the forthcoming future. Furthermore, the employees that already work on existing processes would not need an extensive BPM-related training as they would mostly use the existing applications.
- ***Further development and improvement of existing systems***
  The proposed integration model would enable further development of existing, or new, BPs in existing applications. If the organization had a complete picture of the process, i.e. if the performance data were gathered by a BPM tool, any problematic activities could be detected and the corresponding background application responsible for the activity could be modified accordingly. The experience of developers with existing technologies and their knowledge about existing applications would enable a faster implementation. Once a new BP is implemented in a legacy application and a BPMA is created, the two would be integrated.
- ***Potential way of building an iBPM system***
  Defined by Gartner, Inc. as an extension of a conventional BPM tool, iBPM can glue together the capabilities of pre-built BPs with business planning approaches by means of capabilities like systems cross linking, cloud computing, real-time decision-making, etc. iBPM sounds like a new concept, but it was already happening when enterprises were using BPMS systems with other tools to optimize their processes. The presented integration model could be used to achieve the described connection.
- ***Environments with limited resources***
  Organizations where limited resources prevent the re-engineering of existing applications, but there are initiatives for analysis and improvement of BPs.

In the second phase, the collected scenarios were verified. Experts needed to evaluate the quality of use case scenarios for the integration model by rating it on a scale from 1 to 5 (1 being *Very poor*, 5 being *Excellent*). The results of the evaluation are shown in **Fig. 8**.

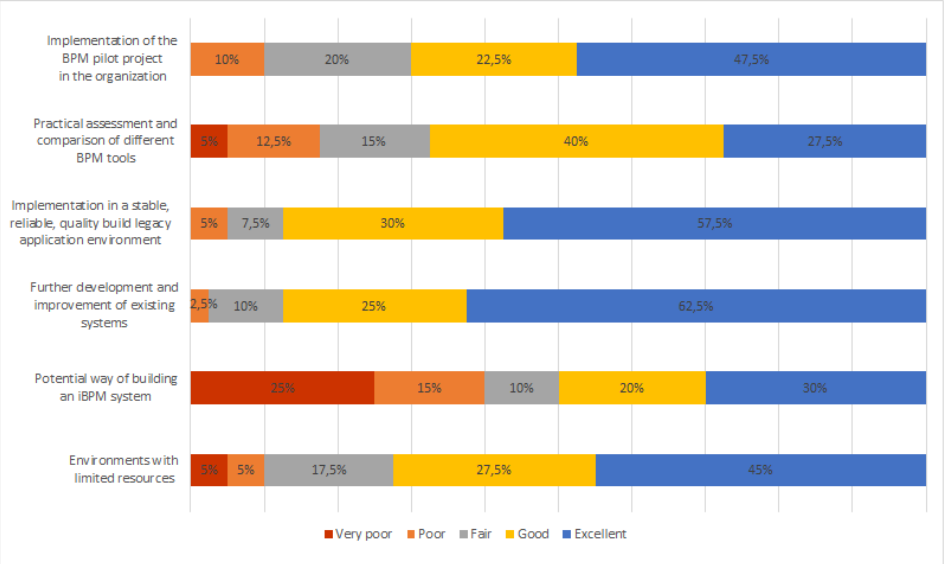


Fig. 8. Evaluation of the use case scenarios for the proposed integration model

To conclude the evaluation of the use case scenarios let us consider what one of the interviewed project managers said: ”Often, the best way to use a tool is to use only the best of its features.” A key part of any BPM tool is a feature for the modeling and managing of BPs. Moreover, a BPM tool incorporates predefined components that enable interaction with data flows in the modeled processes. These often include standard interfaces, predefined actions, or resources for solving some general problems. While BPM tools involve components for solving general problems, specific organizations that implement BPM have specific problems and specific needs that are already met in the existing systems and the information structure of the organizations. Consequently, one could get optimal results by combining the core management of business process flow from the BPM tool with the interfaces and specific solutions in the existing applications.

7.2. Additional Comments, Conclusions and Concerns

An interesting conclusion was made by one of the experts. He observed that the proposed integration model would not be the best choice for an environment with a lot of small, independent, legacy applications, each managing a limited number of processes. In such an environment one would need to create many separate adapters (one per application) and connect all the small legacy applications with the BPM solution. Since the process would increase the cost and the complexity of the BPM implementation, a better course of action would be to redesign the smaller applications into a set of services. Those services would then be used by the created BPMAs to support the required processes. Therefore, the proposed integration model is better suited for a business environment with a few large applications which control most of the BPs in the organization.

Next, businesses occasionally implement new processes that are not strictly related to the existing functionalities and existing applications. What should be done in such situations? The implementation of those processes into existing applications is not a good choice due to the disparity between the new processes and legacy applications. Also, it is not profitable to build separate applications to support the new processes, and then interface them with the BPM solution. The reasonable course of action in the described situation would be to develop a BPMA and the underlying services to support the created process.

Using existing applications as a part of the BPM solution is a bad idea if the implemented processes require redesigns. A better course of action would be to redesign and re-implement the processes using a BPM tool.

By processing the generated links which enable the users to access legacy applications, the authorization and redirection component of the LAA utilizes the existing authentication techniques. An organization's existing information security policies can thus be reused [16]. In case no valid security policies exist, a simple Single Sign On method (SSO) with the help of Lightweight Directory Access Protocol (LDAP) service, a central place to store authentication data, can be implemented. Different applications and services could connect to the LDAP server and validate users.

8. Conclusion

This research explores how important it is for organizations to control their BPs. Furthermore, it shows how control of BPs can be achieved by integrating BPM into existing SOA based legacy environments.

There are several examples of BPM integration in the literature. They, however, do not build upon the existing structures or utilize existing structures. Instead, they focus mostly on redesigning the legacy application architecture. As the process of redesigning legacy environments requires large amounts of time and resources, it is a luxury that many organizations cannot afford. Furthermore, a review of the literature showed that the usage of modern BPM tools which simplify the creation and management of complex processes is rarely considered when creating BPM solutions.

Both the case study presented in this article and the existing literature support the premise that a new integration model is needed to ease BPM implementation in the existing environments. This paper presents new considerations regarding the question on how to adapt the existing systems for BPM implementation. The undertaken research identifies the difficulties and problems relating to the integration of BPM in SOA legacy environments. Based on the results of the study, the requirements for a new integration model are defined.

A generic, scalable multi-layered integration model is proposed, the role of each layer in the overall solution is closely examined and the provided benefits are discussed. In order to save resources and time, which are usually spent on redesigns and refactoring during BPM implementation, modifications of existing legacy applications are kept to the minimum. The model enables companies to utilize business actions that already exist in legacy applications and connect them to form a BP. A loose interdependence between the BPM solution and the legacy environment is achieved. Consequently, legacy applications are operational without a BPMA. The model

provides a central place where business users can interact with a business environment instead of coping with multiple applications and different systems to perform daily tasks. A better understanding of business reality is provided and the model can be helpful when detecting flaws and suggesting improvements in an organization's business activities. The paper provides concrete illustrations of ideas and steps, and it also presents a specific business case of integration.

The evaluation section presents several possible applications of the model in various organizations and in different situations. The most suitable environments for the model's application are discussed in details. Technical aspects of the integration model are reviewed in a proof of concept implementation effort. Additional remarks and conclusions of the implementation effort are outlined. The paper also presents a partial answer to the eternal question whether and when one needs to use existing systems, and when some redesigns are necessary.

Several limitations of the proposed model are also outlined. The integration is reviewed in a single case study and evaluated by experts based on the principles of the Delphi method. Full scale application of the model in different organizational contexts would contribute to its consolidation, offer proof of its benefits, and suggest possible modifications or improvements. Additionally, the question of versioning and a more advanced exception handling will be covered in detail in future research. Next, the success of BPM implementation was discussed rather broadly and additional research into the criteria for measuring success is needed. This is particularly challenging since success is not something one can measure easily and since it can vary both in magnitude and over time.

The research presented in this article may provide organizations with the opportunity to use the proposed integration model as a platform for achieving greater business value and thereby help them to become more flexible and more focused on their core business objectives.

References

1. Elaine A. Carvalho, Tatiana Escovedo, Rubens N. Melo: Using Business Processes in System Requirements Definition, 33rd Annual IEEE Software Engineering Workshop, 125-130, 2009.
2. Matej Hertiš, Matjaž B. Jurič: Ideas on Improving the Business-IT Alignment in BPM Enabled by SOA, 2013 International Conference on Information and Communication Technology (ICoICT), 55-60, 20-22 March 2013.
3. Imanipour Narges, Talebi Kambiz, Rezazadeh Siavash: Obstacles in BPM implementation and adoption in SMEs; University of Tehran, 2012. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1990609 [17 February 2016]
4. Wen ZhenHua, Huang Yousen, Deng ZiYun, Zhang Wei: SOA – BPM Based Information System for Promoting Agility of Third Party Logistic, International Conference on Automation and Logistics, 248 – 252, Shenyang, China 5-7 Aug. 2009.
5. Fuhua Ge, Shaowen Yao, Architecture combining SOA and BPM, Institute of Electrical and Electronics Engineers, Jan 27, 2011.
6. Gheorghe Matei: SOA and BPM, a Partnership for Successful Organizations, Informatica Economică vol. 15, April 2011.
7. P. Trkman: The Critical Success Factors of Business Process Management, International Journal of Information Management, No. 30(2):125-134, April 2010;

8. Razvan Daniel Zota, Liviu Ciovica: Designing Software Solutions Using Business Processes, Procedia Economics and Finance Volume 20, 2015, Pages 695-699
9. Marinela Mircea: "Adapt Business Processes to Service Oriented Environment to Achieve Business Agility", Journal of Applied Quantitative Methods; Vol. 5 Issue 4, p679, 2010.
10. Nan Wang, Vincent Lee: An Integrated BPM-SOA Framework for Agile Enterprise, ACIIDS'11 Proceedings of the Third international conference on Intelligent information and database systems - Volume Part I, Pages 557-566, Springer-Verlag Berlin, Heidelberg 2011.
11. Bazán Patricia, Roxana Giandini, Gabriela Perez, Javier Diaz; Process-Service Interactions using a SOA-BPM-based Methodology, Chilean Computer Science Society (SCCC), 2011 30th International Conference, 9-11 Nov. 2011, Page(s): 100 – 107;
12. Alexandre Perin de Souza, Ricardo J. Rabelo: An Approach for a more Agile BPM-SOA Integration supported by Dynamic Services Discovery, 2010 14th IEEE International Enterprise Distributed Object Computing Conference Workshops, page 186-195,
13. Imran Sarwar Bajwa: SOA Embedded in BPM: A High Level View of Object Oriented Paradigm, WASET 2011 Spring International Conference, 304-308. Tokyo, Japan, (2011)
14. Imran Sarwar Bajwa, Rafaqut Kazmi, Shahzad Mumtaz, M. Abbas Choudhary, and M. Shahid Naweed: SOA and BPM Partnership: A paradigm for Dynamic and Flexible Process and I.T. Management, World Academy of Science, Engineering and Technology 45, 2008.
15. S. Al Aloussi: SLA Business Management Based on Key Performance Indicators, Proceedings of the World Congress on Engineering 2012 Vol III July 4-6, London, U.K.
16. Nivedita P Deshmukh: Leveraging BPM Discipline To Deliver Agile Business Processes In Emerging Markets, 2013 IEEE International Conference on Business Informatics, 338-345;
17. Lerina Aversano, Carmine Grasso, Maria Tortorella: Managing the alignment between business processes and software systems, Information and Software Technology, Volume 72, Pages 171–188, April 2016.
18. Li, Qian; Chen, Yu; and Zhang, Lanfang: An Apparel Trade Quotation Architecture Based on BPM and SOA; PACIS 2009 Proceedings. Paper 115.
19. Nihan Çatal, Daniel Amyot, Wojtek Michalowski, Mounira Kezadri-Hamiaz, Malak Baslyman, Szymon Wilk, Randy Giffen: Supporting process execution by interdisciplinary healthcare teams: Middleware design for IBM BPM; Procedia Computer Science, Volume 113, 2017, Pages 376-383, https://doi.org/10.1016/j.procs.2017.08.350, September 2017.
20. Hausotter, A., Koschel, A., Zuch, M., Busch, J., Hödicke, A., Pump, R., Seewald, J., Varonina, L.: Applied SOA with ESB, BPM, and BRM - Architecture Enhancement by Using a Decision Framework, in: Westphall, C.M., de Barros, M. (Hrsg.), Proc. SERVICE COMPUTATION 2016: The Eighth International Conferences on Advanced Service Computing, IARIA: Rome, Italy, S. 20-27. March 2016.
21. IBM Company: Overview of IBM Business Process Manager, ftp://ftp.software.ibm.com/software/integration/business-process-manager/library/ibpm_overview_pdf.pdf, April 2017.
22. Milosavljević, G., Sladić, G., Milosavljević, B., Zarić, M., Gostojić, S., Slivka, J.: Context-sensitive Constraints for Access Control of Business Processes. Computer Science and Information Systems, Vol. 15, No. 1., DOI: 10.2298/CSIS160628037M, 2018.
23. Stevan Gostojić, Goran Sladić, Branko Milosavljević & Zora Konjović: Context-Sensitive Access Control Model for Government Services, Journal of Organizational Computing and Electronic Commerce, 22:2, 184-213, DOI: 10.1080/10919392.2012.667717, 2012.

Mladen Matejaš was born on May 26, 1985 in Zabok, Croatia. After completing the mathematical high school in Krapina, in 2003 he enrolled at the Faculty of Electrical Engineering and Computing at the University of Zagreb. He received his B.Sc. degree (Dipl. Ing.) in computing in October 2008, by defending his graduate thesis entitled: "*Component Approach to Application Development in Heterogeneous Systems*". Since February 2009. he has been employed at Privredna Banka Zagreb in the Application Development Department, where he participated on different software development and system integration projects. With an interest in the Business Process Management research field he enrolled in doctoral studies of Computer Science at the Faculty of Electrical Engineering and Computing at University of Zagreb, Croatia. Since then, he participated in several domestic and international conferences, published an international reviewed article, participated in the EU funded project "HEUREKA-knowledge to success" and successfully completed several education courses and workshops in various fields. He uses English very well in language and written communication and possesses basic communication skills in German language.

Krešimir Fertalj is professor of computer science at the Faculty of Electrical Engineering and Computing at the University of Zagreb, Croatia, currently serving as Head of the Department of Applied Computing and Head of the Laboratory for Special Purpose Information Systems. He achieved his B.Sc. (Dipl. Ing.), M.Sc. and Ph.D. degrees in computing at the same university. Since graduation in 1988, he has been working at the Department of Applied Computing, where he currently lectures a couple of computing courses on undergraduate, graduate and doctoral studies. His professional and scientific interest is in automated software engineering, complex information systems, project management and in software security. He was the leader of several scientific, and research and development projects. He was the mentor of more than 200 graduate, 9 MS and 8 PhD theses. He is a senior member of IEEE and member of Croatian Academy of Engineering. He was one of the founders and member of management board of PMI chapter in Croatia.

Received: October 5, 2017; Accepted: August 10, 2018

Computer Science and Information Systems 16(1):75–104 https://doi.org/10.2298/CSIS180226038J

Reconstructing De Facto Software Development Methods

Marko Janković[1], Slavko Žitnik[1], and Marko Bajec[1]

[1] Faculty of Computer and Information Science, University of Ljubljana,
Večna pot 113, 1000 Ljubljana, Slovenia
{marko.jankovic, slavko.zitnik, marko.bajec}@fri.uni-lj.si

Abstract. Software development is a complex process that requires disciplined engineering approaches. Empirical studies show that companies still don't document their development practice, or if they do, these are not up-to-date and do not reflect how they really develop software. The main objective of this paper is to propose an approach that can help companies in documenting their real development practice. Comparing to existing approaches that require substantial effort on the side of project members, our approach extracts information on development practice directly from software repositories. Five companies have been studied to identify information that can be retrieved from software repositories. Based on this, an approach to reconstruct development practice has been developed. The approach has been evaluated on a real software repository shared by an additional company. The results confirm that software repository information suffice for the reconstruction of various aspects of development process, i.e. disciplines, activities, roles, and artifacts.

Keywords: Software development method; Software repository; Development practice; Development method; Development project.

1. Introduction

Software development requires a systematic and disciplined approach to assure the quality of the process and its results, i.e. the software that we develop 1. This has been recognized already in the early beginnings of the software development era and has led to the construction of many software development methods.[1] Over the years, it then turned out that there is no ideal development method that could fit to all kinds of projects, even in the context of a single organization. How suitable a particular development method is, actually depends on many factors, ranging from project and organization characteristics to the characteristics of the development team. These findings have been identified by many researchers, e.g. [2–9].

[1] In this paper, we use the term software development method to denote the work that we do to structure, plan, control and perform the development of an information system. With the term method we cover all important aspects of the development lifecycle, i.e. activities to be carried out, artefacts to be developed, techniques to be used, roles to be assigned etc.

One of the research fields that emerged as a result of the aforementioned problems, is method engineering. Researchers in this field devoted a lot of effort to find suitable solutions. One of them is the so called situational method engineering, which is a process of constructing development methods specifically attuned to the needs of projects [10]. Such development methods would either be composed of fragments of other development methods or created by tailoring the development method that is generally used and known to the organization. Unfortunately, there are several obstacles that hinder the application of situational method engineering in practice [11, 12]. One is for instance that we need somebody who is capable of applying the method engineering process (must be familiar with various development methods, method fragments etc.) and what is even more challenging, we need enough time before starting the project so that this person can do the job. In real settings, where projects are almost always run in very tight schedules, this is rarely the case [13–15].

Problem statement. Based on our experience from introducing the situational method engineering process in practice 16 and from related research findings (e.g. 17), companies see as beneficial if they are able to document and monitor actual work on development projects and compare it with their prescribed methods. In this way, they can detect deviations if they occur. Doing this manually is however, very time-consuming and perceived by developers as an unnecessary burden. An approach is thus needed that does not require more than just a minimal effort from developers.

Objective. The objective of our research is to solve the above problem by reconstructing information about the project performance from the data that is captured in software repositories. We assume that software repositories contain enough data to reconstruct at least the main method elements (i.e. disciplines, activities, artifacts). Moreover, the objective is to support post-development analysis to learn how the project was performed and to possibly identify its positive and negative aspects in relation to its outcome, and also during the project performance, so as to detect situations that might lead to project failures.

Contribution. The main contribution of the research presented in this paper is the approach that facilitates the reconstruction of the development method elements from software repositories. In contrast to existing approaches that only enable the reconstruction of disciplines (e.g. analysis, design, development) or focus on the development phase only, our approach enables the reconstruction on a more detailed level, including additional development method elements.

Outline. The paper is organized in ten sections. In Section 2, we explain how the research was performed (research design), giving also a brief information on the participating companies. In Section 3 and 4, we describe the suggested approach with Section 3 focusing on the analysis of the software repository content and Section 4 on the reconstruction of the development method elements. Section 5 covers the evaluation and its results, which are discussed in Section 6. Section 7 provides the information on threats to validity. In Section 8, we position our research within related work and, finally, Section 9 concludes the paper.

2. Research Method

This section gives a brief description of the research design. Our research was motivated by the following research question:
Does the information stored in software repositories suffice for the reconstruction of basic characteristics of the software development methods that were de facto used in the corresponding software development projects?

2.1. Data Collection Procedure

Data were first collected from five software companies whose business is software development (see Table 1 for their profiles). The companies shared their software repositories with us (limited to selected projects only) and provided their personnel (project managers) for qualitative analysis. For the evaluation step, an additional company joined the research. Its data (software repository) and personnel were used to validate the research findings.

Table 1. Profiles of participating companies.

Company	Company profile
Marand d.o.o.	Company with around 100 employees that develop innovative and easy to use healthcare IT products.
Comtrade d.o.o.	Large company with over 500 employees. They develop IT solutions for different industries, including government, financial institutions, healthcare, telecommunication providers.
Ekipa 2 d.o.o.	Company with over 200 employees. They are focusing on development of entertaining mobile apps and games.
Optilab d.o.o.	Small company of about 30 employees. They develop complex information systems for clients from the financial sector, utilities and healthcare.
Adacta d.o.o.	Company, with over 350 professionals that provides support to 400 regional and international clients. They are specialized in developing and implementing business IT solutions and business consulting.

2.2. Research Approach

To answer the research question, the following approach was used:

- **Step1**: **analysis of software repository content**: the purpose of this step was to find out what kind of supporting tools the participating companies are using within software development and, more importantly, what kind of attributes they capture in software repositories. For further research, we assumed that attributes which we found in all repositories (from the involved companies), are generic and could be

thus found also in any other software repository (i.e. from any other software development company).

- **Step 2**: **development of the algorithms for automatic reconstruction of development method elements from software repositories**: the purpose of this step was to develop algorithms (and tools support) that will allow us to reconstruct the development method elements from software repositories. The objective was to reconstruct the development method elements that represent valuable information for project managers and other project team members. Using semi-structured interviews with project managers from the participating companies, we identified the main development method elements of their interest. For each of these elements we then developed algorithms for their reconstruction.
- **Step 3**: **evaluation**: the findings from the first step and the algorithms developed in the second step were evaluated by involving another company in the research. At first, we checked whether our assumption about generic attributes holds in their case and then employed our algorithms to reconstruct the selected development method elements from the repository on the recently finished project. Finally, we discussed the accuracy and usefulness of the reconstructed software development elements with their project manager.

In the rest of the paper, each of these steps is described in more detail.

3. Analysis of Software Repository Content

As a first step we asked companies to provide us access to their development environments or just give us snapshots of their software repositories. This was not easy to get due to the privacy and security issues, but eventually we got enough data to get a good picture on what they use and collect. As we expected, the companies were quite similar in terms of the type of tasks for which they were using computerized support (e.g. revision control, issue/bug tracking, document management, etc.) but differed to a certain extent in the actual tools they were using. Among the tools that we found, the most common were:

- Jira, Bugzilla or DevTrack for issue/bug tracking (ITS)
- Subversion or Git for revision control (RCS)
- Sharepoint or LogicalDOC as a document management system (DMS)

Additionally, some companies were using tools for other tasks, such as for managing code reviews (e.g. Cruicable, Reitveld) or for managing team collaboration (e.g. Slack, Confluence, Skype). But since these tasks did not have computerized support in all companies and in some cases data cannot be obtained due to the privacy issues, we did not analyze them further.

For each company and tool, we then examined what kind of data they actually store in their databases or logs. In Fig. 1, you can see the set of attributes that we were able to find in software repositories of all five participating companies.

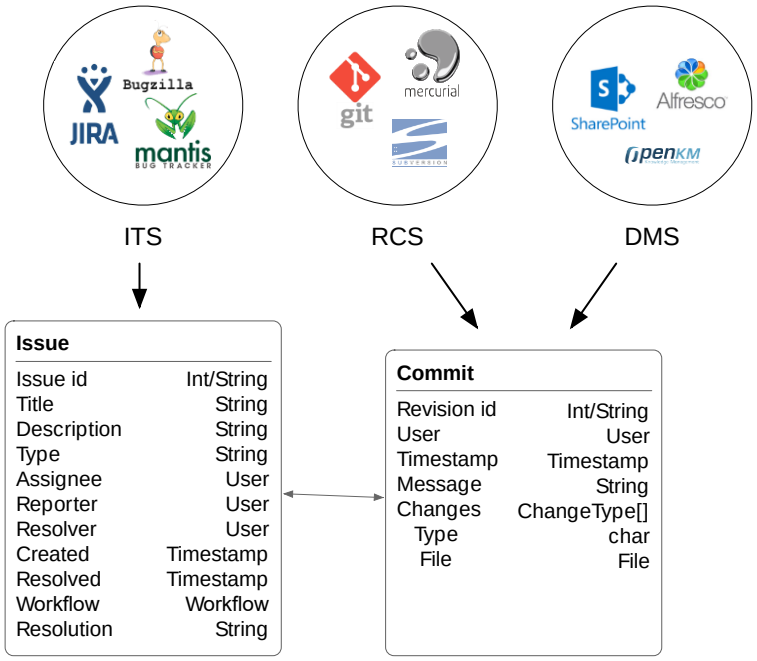


Fig. 1. Selected attributes from systems comprising the project's software repository.

The next step was to check how well the data from different tools comprising software repositories can be linked together. Remember that a software repository is not a standalone physical database but rather represents a logical view on several databases and logs from different systems. In our study, we found that important logical connections exist between issues and commits, which both carry important information and are kept track of in software development. While an issue represents a problem or associated tasks that need to be carried out in order to solve a specific problem, a commit refers to changes on specific files that are a result of solving the problem or task and are put back into the repository. Issues and commits are however managed in different systems and thus not necessarily linked. The link can be established if the commit message (Fig. 1) carries enough information so that we can identify which issue it is connected with. For unlinked commits and issues, techniques such as Frlink can be used 18. Let us also note that linking commits with issues is a good development practice that has been practiced in open source community for a long time 19 and should be enforced by the companies that want to raise the quality of their development processes.

Another challenge for establishing connections between data collected through various systems into a software repository, is to link user accounts created in these systems that refer to the same user. This is almost always the case, as software repositories usually comprise tools of different vendors and the single-sign-on option is not available. For this purpose, various existing entity resolution and identity merge algorithms can be used. In our case, we use the one published by Goeminne and Mens 20.

4. Algorithms for Automatic Reconstruction of Software Development Method Elements

As written in the introduction, our objective is to enable reconstruction of more detailed information about the development methods used on projects than existing approaches enable, and to do this without any substantial involvement of developers. The existing approaches [21-24] focus on the reconstruction of disciplines (i.e. they are able to tell how much effort was spent for analysis, design etc. or how these disciplines were following one another) or they go into details on the development phase only (i.e. by focusing on the issue lifecycle). In contrast, our goal is to focus on the whole project, and not just the development phase and to reconstruct more than just the disciplines. In this section, we first describe the meta model that was constructed in cooperation with the participating companies and then describe how this information can be reconstructed from the repositories. The steps for reconstructing the project specific software development method are shown in the Fig. 2. In the figure each step has a link to the section in which it is explained in more details.

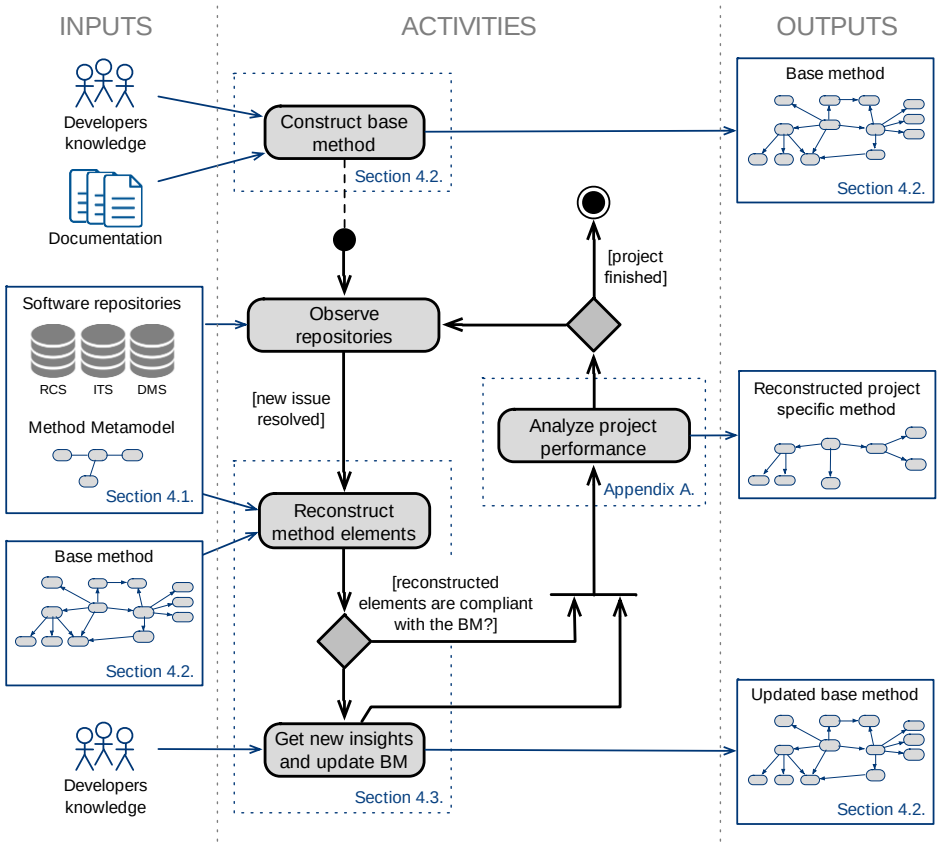


Fig. 2. Diagram showing the steps used during the reconstruction of the project specific software development method. BM is used as acronym for “base method”.

4.1. Metamodels of Software Development Methods

Development methods can be described by a number of different concepts, including phases, disciplines, deliverables, activities, techniques, examples, roles, experiences, life cycles, tools, etc. A number of different metamodels exist that underpin existing development methods 25. For the purpose of our research, we followed the SPEM metamodel 26. Its basics, including the separation of method and process concepts, were briefly described to the participating companies. Afterwards, the companies were asked to identify the main method meta elements that they would be interested in reconstructing from the software repositories. Linking data from software repositories to SPEM metaelements has also been done by others [27, 28]. The final selection, that was influenced also by the data that is actually captured in software repositories, included the following method meta elements:

Disciplines: a discipline presents a set of activities that are closely related in the sense that they all contribute to the same overall goal (e.g. Analysis, Design, Implementation).

Activities: an activity presents a general unit of work assignable to a specific performer (Develop a use case, Design GUI, etc.). As a result of an activity, different deliverables (artifacts) can be produced.

User roles: a user role is responsible for performing activities and producing the required artifacts (e.g. Developer, Analyst, Architect, etc.). Note that several project members can be assigned to a single user role.

Artifacts: an artifact is a result produced as part of performing a particular activity (e.g. Source code, Unit test).

The metamodel is represented in Fig. 3. Each *discipline* consists of one or many *activities* while an *activity* belongs to exactly one *discipline*. An *activity* can be connected with none, one or many *artifacts* and an *artifact* with one or many *activities*. For each *activity*, there is exactly one *user role* assigned, but for a particular *user role* there might be several *activities* that the *user role* is responsible for. The recursive relationship on the meta element *Activity* designates that *activities* are dependent on each other, which is due to the fact that they need to be performed in a certain order.

Fig. 3 also indicates the relationships of the metamodel elements with software repository concepts. The three most important concepts, originating from a software repository, are a *file*, an *issue*, and a *user*. The concept *file* designates a physical file that is stored in a revision control system or document management system, the concept *issue* represents an issue from an issue tracking system, and finally the concept *user* denotes user accounts from any of the software repository systems. The meaning of the relationships among metamodel elements and software repository concepts is as follows:

Artifacts are in relationship with *files* (stored in software repository) that represent the artifact. Each *artifact* can be related to none, one or several *files* while a file belongs to exactly one *artifact*.

Each *issue* requires a certain amount of work to be done. This work might be represented as an *activity*. A good development practice is that each *issue* is

connected to exactly one *activity* while *activities* might resolve several *issues* at once.

Users represent project team members with user accounts in a software repository. Several *users* can be assigned to a particular *user role* and vice versa, a particular *user* can play more than just a single *user role* in the observed project.

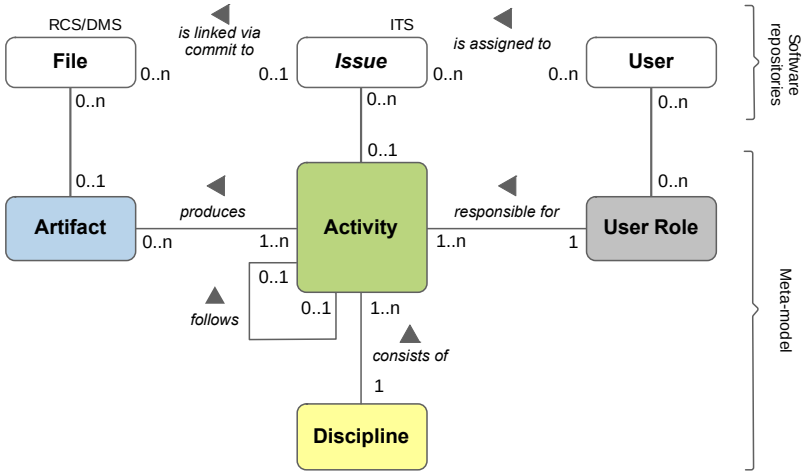


Fig. 3. Software development method metamodel and its connection with software repository concepts.

4.2. Construction of a Base Method

One of the concepts that plays an important role in our approach is the so called base method. With the base method, we denote the set of method elements that are typically used in a particular company when performing development projects. Taking into account the metamodel described in Section 4.1, the base method of a company includes its typical disciplines, activities, user roles, artifacts, and relations among them. The intensity of individual activities and the sequence of their performance is however not described with the base method as this depends on each particular project settings and its characteristics. In our approach, the base method represents the baseline to which we compare the reconstructed method elements and detect deviations.

The construction of a base method is a preliminary step before the reconstruction of the project specific development method. The simplest way to do this is by analyzing documentation of past projects or by acquiring this information from project managers. In some cases, companies even keep their software developments methods documented, in which case these documents can serve as a good starting point for the construction of the base method.

Fig. 4 represents an example of a base method constructed according to the existing documentation for one company that participated in the evaluation. Using the documentation, we were able to construct the company's base method consisting of 5

disciplines, 15 activities, 23 artifacts, 8 user roles, and the relations among them. During the construction we have also captured 14 rules, which define the presence of a relation between method elements based on the project characteristics. (e.g. Financial Calculation - If the project is small then Financial Calculation is not required; If the project is medium or large then Financial Calculation is required). More details about the construction are presented in the Section 5.3.

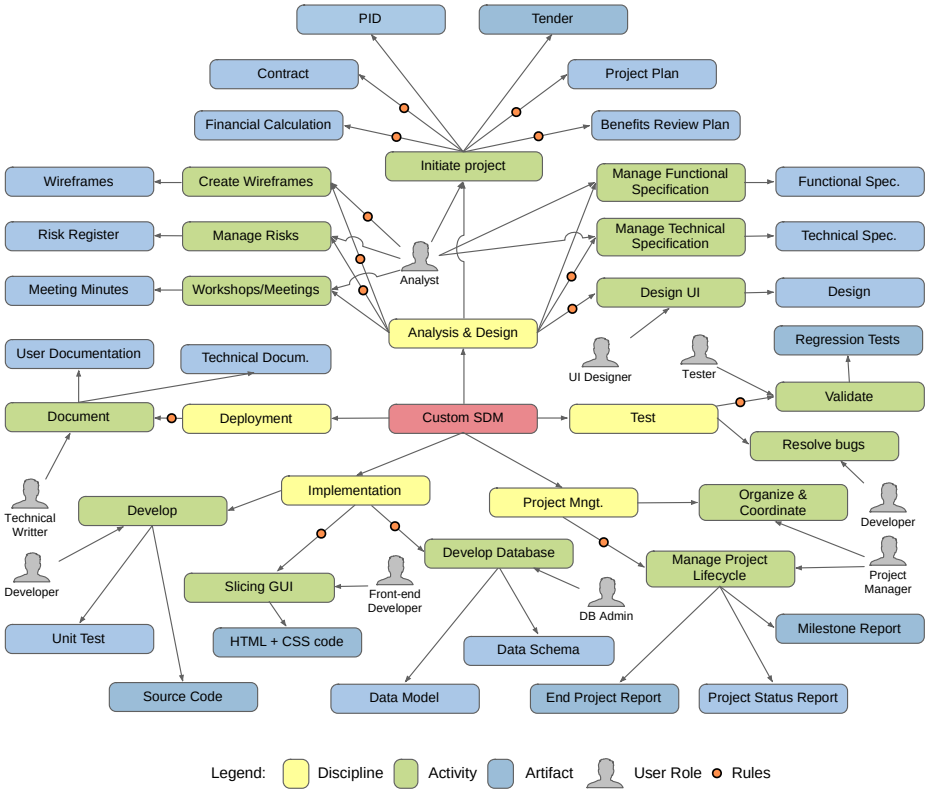


Fig. 4. The company's base method draft manually created out of the company's documentation. It comprises 5 disciplines, 15 activities, 23 artifacts, 8 user roles, and 14 rules. Examples of a rule: (a) Financial Calculation - If the project is small then Financial Calculation is not required; If the project is medium or large then Financial Calculation is required, or (b) If the project is small then Benefits Review Plan is not required; If the project is medium or large then Benefits Review Plan is optional.

4.3. Reconstruction of the Software Development Method Elements

Once the base method is defined, we use data from software repositories to reconstruct activities, artifacts, user roles, and disciplines. In this section we describe how this is

done. In the guidelines that follow, we take into account the findings on the data that software repositories include and how well this data is linked (see Section 3).

Reconstructing Artifacts

Artifacts are reconstructed from files that were committed to the repository. This is done immediately after an issue is being resolved. We check the repository and retrieve all files that were committed as a consequence of resolving this issue. Then we infer from the files (by checking their names, file types and if necessary also the file content) which artifacts from the base method they represent. Techniques that can be used for matching files to artifacts are many. Machine learning algorithms are useful when we have data to learn from, i.e. software repositories from past projects. In this case, we create a classifier which we then use for matching. If this is not available, we acquire additional information from the company employees so that matching can be done. The algorithm used in such cases is described at the end of this section.

Reconstructing Activities

After we reconstruct the artifacts as a result of resolving an issue, we go further and check which activities are connected to these particular artifacts. We do this by checking the base method where these relationships are defined. In most cases, activities are connected to one artifact only, thus the reconstruction of the correct activity is not a problem. The involvement of developers is only required in rare cases, a) when these relationships are not one to one (an activity produces several artifacts and vice versa, several activities might be responsible for the creation of one particular artifact) or b) when we have no artifact which we could use to infer the corresponding activity. This happens when we deal with an issue that did not commit any file to the repository. In such cases, we involve the developers to tell which activity is connected to that particular issue.

User Roles and Disciplines

Similarly, we also reconstruct user roles and disciplines. Since we already know the activity name, we simply check the base method to retrieve also the names of the associated user roles and disciplines. Furthermore, for each retrieved user role, we also retrieve the users (user accounts) that were assigned to this particular activity.

Algorithm for Reconstructing Artifacts and Activities

For matching files to artifacts and issues to activities we use the algorithm that is described below.

For the algorithm to work, the first step is to go through the base method and capture keywords that best describe each artifact. In addition, we capture file types that represent the format in which a specific artifact is usually created. Next, we capture keywords for the activities without artifacts.

Example: Keywords and file types for the artifact "Functional specification"

```
Keywords := {requirement, functional requirement,
non-functional requirement, usability, scalability…}
File types := {doc, docx, rtf, txt, pdf};
```

Such manual acquisition of this information is only required if we do not have any data to learn from. If this data is available, i.e. we have access to software repositories of finished projects, then techniques, such as Bag of Words, TF-IDF or similar can be used to automatically acquire this information.

The algorithm to match files to artifacts and issues to activities is as follows: for each resolved issue *I*, we find all connected commits *C*. For each such commit *C*, we classify each committed file *F* to an artifact *A* from the base method *BM*. We try to do that based on the file types. If several artifacts (artifact list *AL*) contain the same file type, we calculate individual artifacts weights. An artifact weight *w* for an artifact *A* tells what is the likelihood that the file *F* represents the artifact *A* from the artifact list *AL*. The higher the weight, the higher the likelihood. The weight *w* is calculated as a sum of TF-IDF values. The TF-IDF metric ((tfidf=tf(K,A)*idf(K,AL))) is calculated as a product between frequency of the keyword *K* in the file *F* ($tf(K,F)=f_{K,F}$) and the logarithm of the ratio between the number of all artifacts *A* in the artifact list *AL* and the number of these artifacts from the artifact list in which the keyword *K* appears ($idf(K,AL)=\log(|AL|/|K \in AL|)$). Once the committed files have been successfully classified to an artifact, we check in the base method which is the activity that is responsible for the delivery of this artifact.

If the issue under analysis cannot be connected to any commit, and thus we cannot identify the corresponding activity over connected artifacts, then we reconstruct the activity directly from the base method by employing a very similar approach (see lines 23-33 in Algorithm 1). Instead of searching for artifact keywords in committed files, we search for activity keywords in the issue title and description. These are two attributes that we can find in all issue tracking systems. For clarity reasons (to avoid duplicate lines), this part is not shown in the algorithm.

The algorithm is represented below. It uses three data structures:

Matrix Weight_KA: a two-dimensional matrix that tells for each keyword *K* and artifact *A* what is the likelihood that *K* represents *A*. The likelihood of *K* representing *A* is calculated using TF-IDF.

List Artifacts: a list of artifacts from the *BM* that contain a specific file type in their file types set.

List Activities: a list of activities from the *BM* that are in *BM* linked to the artifacts reconstructed during the classification of files linked to a specific issue.

Algorithm 1: Algorithm to reconstruct Artifacts and Activities

```
1  INPUT:
2  List ResolvedIssues; //ordered by resolved time ASC
3  BaseMethod BM;
4
```

```
5  OUTPUT:
6  // files are classified to artifacts
7  // issue are classified to activities
8  // relations between activities and artifacts
9
10 Matrix Weight_KA;
11 Set IssueArtifacts;
12 for each issue I in ResolvedIssues do
13   for each commit C in I.connectedCommits do
14     for each file F in C.files do
15       if F.type == IgnoreFileType then next file;
16       List Artifacts = artifactsByFileType(F.type, BM)
17         If Artifacts.size == 0 then
18           // new type -> ask project member, update BM
19         else if Artifacts.size == 1 then
20           Classify(F, Artifacts[0]);
21           IssueArtifacts.add(Artifacts[0]);
22         else
23           resetAndPopulateMatrix(Weight_KA, Artifacts);
24           for each artifact A in Artifacts do
25             for each keyword K in A.keywords do
26               Weight_KA[K,A]:= tf(K,F)*idf(K,Artifacts)
27
28           Artifact, Value = max(sumByA(Weight_KA))
29           if Value != 0 then
30             Classify(F, Artifact);
31             IssueArtifacts.add(Artifact);
32           else
33             // ask project member and update BM
34
35   if IssueArtifacts.size > 0 then
36     List Activities = findActByArtifacts(IssueArtifacts, BM);
37       if Activities.size == 1 then
38         Classify(I, Activities[0])
39       else
40         // ask proj. member (bad practice, new knowledge)
41   else
42         /*Issue without artifacts. Same as in lines 23-33,
43       but this time we use keywords from issue title and
44          description and as a list all Issues without
45       artifacts*/
```

As a part of the research described in this paper, we developed a computerized tool (iSPRToolset) that facilitates the application of the proposed approach in a company. The tool supports the following tasks:

a) The analysis and linkage of the data from tools that comprise software repositories
b) The creation of a base method
c) The automatic reconstruction of activities, artifacts, user roles, and disciplines using the algorithm described in Section 4.3.4
d) Various visualizations of the development method elements of the timeline

For more details about the tool check http://ispr.jmlabs.eu and the Appendix.

5. Evaluation

To make the evaluation unbiased, we invited an additional software company to join the project. In this way, we did not know what tools this company is using to facilitate development, neither what information it stores in its software repository. The evaluation comprised the following steps:

(1) Analysis of the company's software repository
(2) Construction of the company's base method draft
(3) Reconstruction of the development method elements
(4) Analysis of the project performance

The aim of the first step was to find out whether our assumptions about a) typical attributes that could be found in a software repository and b) linkage between repository data (issues, commits, user accounts) hold for this particular company.

In the second step, the goal was to create a draft of the company's base method.

The step three was dedicated to the evaluation to what extent specific development method meta elements can be reconstructed from the company's software.

Finally, the purpose of the fourth step was to evaluate how useful the reconstruction approach can be if used for controlling the project performance.

In the following, we report on the evaluation findings.

5.1. Profile of the Company

The company that we analyzed in the evaluation, develops e-business solutions for Health, Insurance and Telco industries and employs about 50 people. They develop software by following a combination of agile and traditional approaches. The decision to involve this particular company into our research was based on the following reasons:

- The company was willing to provide all the necessary information about the project that seemed appropriate for the evaluation.
- The project team members were allowed and willing to commit required time for the purpose of the evaluation.

- The company already had a prescribed and documented development method in place – i.e. guidelines for software development. This was useful as we could use it as starting point for the construction of the company's base method.

5.2. Analysis of the Software Repository Content

For the selected project, we imported data from three different tools that the company was using during the project. These were Jira (used as issue tracking system), SVN (used as revision control system), and LogicalDoc[2] (used as document management system). The first step was to retrieve issues, commits, and users (user accounts). For a summary report on the data collected see Table 2.

Table 2. Collected data.

Attribute	Value
# of issues retrieved (ITS)	186 – 13[a]
# of commits retrieved (RCS + DMS)	379 + 166
# of all files	3578
# of users (employees + stakeholders)	15 + 2

[a]In case of Jira, 13 issues were excluded as they were duplicates of other issues, could not be resolved, or were of the following type *meta task*.

On the next step, we tried to link commits from SVN and LogicalDoc to Jira Issues. At first, we did that by extracting issue IDs from commit messages using regular expressions, such as for example `"\b"+JiraProjectKey+"\b){1}[^\w[0-9]]+\d+"`. A similar approach was also used by others [29, 30]. In this way, we were able to link roughly 70% of all commits with corresponding issues and about 60% of issues with corresponding commits. These results alone were already promising, as we linked majority of the commits with issues and vice versa. To improve these results, we could have used the approach as suggested in 18, but we rather decided for a manual check via developers so that we also learned how consistent they are in using supporting tools.[3] Together with their help, we were able to link additional 139 commits and 54 issues. 34 (6.2%) commits and 18 (10.4%) issues were left unlinked.

By analyzing unlinked commits, we found 9 of them were made to restructure and move the repository to a new location. Additional 12 were related to specific changes, such as *upgrade library*, *add user as developer in pom.xml*, *fix typo*, *import files*, etc. The last 13 unlinked commits that were left were all found to be connected with project management activities and the creation/modification of various related documents.

Similarly, by analyzing unlinked issues we found that they mainly presented system administration activities, such as *increase RAM in test environment*, *update from java 6*

[2] In LogicalDoc, a commit is perceived as a new version of a file (check-in).

[3] It is important to note here that companies should require from their developers to link commits with issues otherwise important information is missing. The open-source community seems to be aware of that – in MongoDB and Hibernate, for example, over 90% of all commits from 2014 are linked to at least one issue from Jira 19.

to java 7, *install SSL certificate*, as well as activities that did not result in the creation of any artifact (e.g. *setup development environment*).

Regarding the resolution of users via user accounts, we had no problems, since the users were using the same usernames for all the systems. On this project, 17 different people participated, out of which 15 were employees and 2 stakeholders. For details on the results of linking the data see Table 3.

Table 3. Percentage of issues linked to commits and vice versa.

Attribute	Value
% of commits linked to issues (regex)	68.3
% of issues linked to commits (regex)	59.5
% of commits linked to issues (regex + manually)	93.8
% of issues linked to commits (regex + manually)	89.6

At the end of this step, we also checked how well the development method meta model corresponds to the company and its expectations from the development method reconstruction. The company's CIO was fine with the selected development method meta elements.

5.3. Construction of the Base Method Draft

When we asked the company to tell us how they usually develop software (i.e. do they have any predefined steps, deliverables, techniques, user roles etc. that project team members need to follow) they gave us a documentation in which they defined basics of their development method. This included the description of project disciplines and corresponding activities. For each activity, the documentation also provided a description of the activity goals and associated artifacts. Each activity was further linked with user roles responsible for its performance. The described development method also differentiated among different types of projects. Based on the project size, these were divided into three groups: small, medium and large. All this information was written in a series of word files and available to all employees. We constructed the base method by analyzing the provided documentation. We did that together with one of the company's project managers. The base method draft is depicted in Fig. 4. In the next sections, we describe how this base method served us to reconstruct the development method elements that were used on the observed project.

5.4. Reconstruction of the Project Software Development Method Elements

The most important part of the evaluation was to check how well can we reconstruct development method elements of an observed project by analyzing the data from the corresponding software repository.

To have a "golden rule", i.e. to be able to measure how accurate is the reconstruction, we asked the person that acted as the manager of the observed project, to help us manually reconstruct the development method elements that were used on the project.

For each issue, the manager identified connected commits and their files and based on that concluded what artifacts they represent.[4] In case an issue was found that had no connected commits (this happens when during the resolution of an issue no files are created), the project manager was asked to tell what activity this issue was about. Similarly, for the commits and related files that were not identified over issues, the project manager was asked to classify them into the artifacts they represent. From activities, we then inferred disciplines and user roles.

In the next step, we used this information as the baseline against which we compared the results obtained with our algorithm. To measure the quality of the reconstruction, we used the precision and recall measures, which are known from information retrieval and pattern recognition. The results are shown in tables below. To fairly judge the quality of the algorithm, we also compared manually and automatically classified file versions into artifacts and issues into activities. The reason for this is that some artifacts are created gradually, through many versions, and are thus connected to many issues. Consequently, it wouldn't be enough to limit the comparison on the artifacts only, as these might reconstruct well only for some of the commits.

Table 4. Precision and recall of the automatically classified file versions to artifacts and issues to activities.

	Manually classified	Automatically classified	Precision	Recall	F-measure
File version	5945	5908	0.997	0.991	0.994
Issues	173	155	0.98	0.88	0.93

Table 5. Precision and recall of the automatically reconstructed development method elements compared to the manually retrieved development method elements.

Method element	Manually retrieved	Automatically retrieved	Precision	Recall	F-measure
Artifact	15	12	1.00	0.8	0.89
Activity	12	10	1.00	0.83	0.91
Discipline	5	5	1.00	1.00	1.00
User role	8	7	1.00	0.88	0.94

To achieve good results with our reconstruction algorithm, it is crucial that artifacts are reconstructed with as high precision and recall as possible, as the reconstruction of other method elements depend on this.

As you can see from the results, the classification of file versions to artifacts yielded very good results (Table 4, row 1). The reason for this is that artifacts are quite different in terms of their names, content and formats in which they are created. Thus, we were able to differentiate among them with a high confidence. The results also show that file versions did not influence much on the classification accuracy. This is an important

[4] This was not that time consuming, as the company uses a special directory structure in revision control system and document management system to store files that belong to a certain result or artifact. This way the project manager could conclude already from the place in the directory structure what artifacts an observed commit's files most probably represents.

finding as it supports iterative reconstruction of development method elements, i.e. step by step through the project performance.

The files that were misclassified or were left unclassified, were not many and in most cases due to one of the following reasons: (a) the file was of an unknown type, i.e. not defined in the base method – in our case these were mainly fonts of file type woff, eot, tff…, (b) the file content couldn't be parsed – these were pdf files that contained images/scans and at the same time couldn't be classified based on keywords in their filenames and file paths, and (c) files that included keywords which are more typical for some other artifact – in our case this happened for files that represented meeting minutes (e.g. on one particular meeting they were discussing a lot on the functional specification, so this word occurred many times in the meeting minutes, and thus the file was classified as an artifact "functional specification" rather than "meeting minutes").

Good results in terms of precision and recall were obtained also for the classification of issues to activities (Table 4, row 2). In most cases, we were able to correctly identify activities that corresponded to issues by classifying files these issues created.

Table 5 shows results of reconstructing the development method elements. It compares manually retrieved development method elements with the automatic reconstruction. As you can see, all the automatic reconstructions were correct (100% precision) which is not surprising as the classification of files got such a high accuracy. The recall for activities and artifacts were however not that perfect on the first sight (0.8, 0.83, respectively). Several activities and artifacts were missing in the automatic reconstruction. The explanation that we got from the project members revealed that the missing elements were all newly introduced and thus couldn't be found in the base method. Let us emphasize however that these results are based on the fully automatic reconstruction, i.e. without any involvement of the development team. In other words, the results, presented in tables above, could be improved if the developers were asked for additional information in cases when classification or reconstruction couldn't be done.

5.5. Checking Project Performance

The algorithm for reconstructing development method elements can be used also during project performance. In this case, we reconstruct development method elements one by one, every time an issue is resolved, and check for their compliancy with the base method. There are several benefits of doing this during project performance. If the reconstructed development method elements are not compliant with the base method, the project manager is notified about that and can react by asking responsible project team members for clarification. In case no suitable argumentation is given, a bad practice was obviously detected and can also be prevented. However, if those responsible for the deviation can argument why they declined from what was expected, the base method can be supplemented by capturing new knowledge in terms of new development method artifacts or rules that bind project characteristics with some specific development method element. Additionally, if the project is being checked during its execution by reconstructing development method elements after each resolved issue, the information about activities and disciplines can be visualized on a timeline diagram which gives an interesting insight into the current state of the project – this is only possible if company

store information about the time planned and spent on the level of issues. It can even help to detect situations that might represent risk for the project. For example, if the majority of issues that are being resolved are still connected to activities and disciplines that should already be finished then it could be that we are at risk that the project will be late. Some of these analyses, produced with the iSPRToolset for this particular project are shown in the Appendix.

For the purpose of the evaluation, we simulated the project realization by passing the issues to our approach, as they were appearing chronologically (ordered ascending by the resolved date-time attribute) during the analyzed project. For each issue we have reconstructed the development method elements. The Fig. 5 shows the reconstructed development method. What is worth to mention is that we improved the base method by three new development method elements (two artifacts and one user role), which were detected during the reconstruction and present a new knowledge about development practice. This happened when the algorithm was not able to classify a file into any of existing artifacts or an issue into any of existing activities and we thus asked the project manager for an explanation. He explained that these development method elements are important but we obviously failed to capture them when we were creating the base method draft. Final assessment given by the project manager was that he would like to have our approach and iSPRToolset implemented and available for future projects he will be working on.

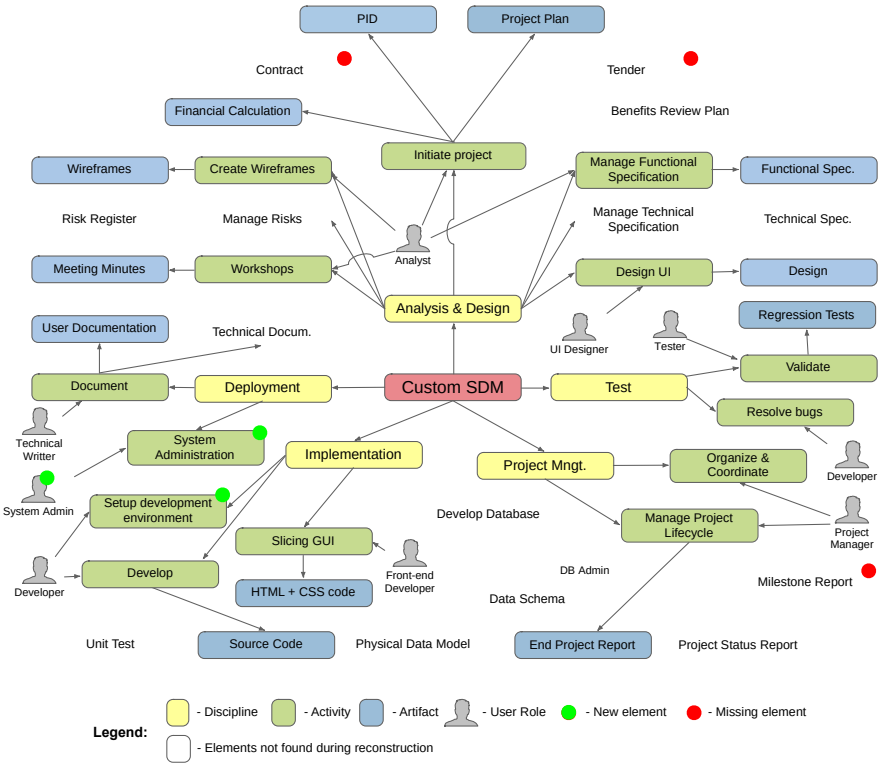


Fig. 5. Reconstructed software development method.

6. Discussion

The approach presented in this paper has some limitations that we need to acknowledge. First, it does not reconstruct all possible development method elements but only some of them, i.e. disciplines, activities, artifacts, and user roles, to be exact. It also does not reconstruct the workflow which would tell how exactly the activities followed one another. Furthermore, it depends on the quality of the data captured in software repositories. Finally, it requires some effort (although not substantial) from the project team members in order to work optimally.

On the other hand, to our knowledge, it achieves much more than existing approaches, which we are aware of. First, it helps the company in capturing and maintaining its base method, i.e. the method that really reflects how the company is developing software. Next, it helps the company to conduct development projects in a way that they are performed consistently and in line with what the company prescribes with the base method. Also, it helps to detect deviations from the base method as project is performed. Finally, it only requires small input from the development team.

To further validate the usability of the approach, semi-formal interviews were conducted with seven project managers of the participating companies. We asked them the following questions:

a) How do you perceive the suggested approach in terms of its complexity? Could it be introduced in your company? Do you think it would be accepted by your employees?
b) What do you expect the main benefits would be of using such approach in your organization?
c) What would you suggest to make the approach more useful?

The feedback that we received was generally positive. They all agreed the approach is simple enough to be adopted in their organizations. Since it doesn't require any substantial effort from developers or changes that developers would need to introduced in their everyday practice, the acceptance of the approach is also expected to be high.

As the main benefit, they emphasized the following possibilities offered by the approach: a) to do retrospective on finished projects, b) to observe project performance on the fly and identify steps that does not fit their regular practice (i.e. decline from their base method), c) to keep base method up-to-date, and d) to give more emphasis on methodological aspects of their development activities (as a side effect). Finally, as a suggestion for improvement they were all consistent that it would be very useful if we were able to reconstruct also workflow information, i.e. how exactly activities and their smaller counterparts (tasks) were performed during an observed project.

7. Threats to Validity

There are multiple threats to validity that face this research, we will address them in the context of construct validity, internal validity, external validity and reliability.

With respect to construct validity, we had to address the fact that we rely upon data that are created and annotated by project members and are stored in software repositories. To improve construct validity, we have validated the data and results with project members and constructed the reconstruction approach based on the insights that we got from five companies. The threat we face here is also that project members might not remember all the details from the project and there is no explicit evidence that the project has been performed as it was reconstructed. We presume that data in software repositories are unbiased and that project members possess enough information about how project has been conducted.

From an internal validity point of view, we do not face any threats, since our main goal was to show that using data from software repositories, we can reconstruct a development process and method followed on the project. In our evaluation, we analyzed an already completed project, hence the data should not be biased.

An external validity issue we face is that we evaluated our approach only on one case study, hence it is hard to justify how generalizable our results are. However, the approach to reconstruct method elements and perform different analyses is straightforward: if all the required data are available, it is reasonable to assume that reconstruction can also be done on other development projects. Among different organizations, the main difference, when using the proposed approach, is in the base method, which is specific to the organization and should be defined based on the company's development practice. Another threat we face is that the data are not of such a good quality as required. For example, commits and issues might not be linked to that extend as required. In our case study we were able to link 93.8% of all commits with an issue. However, this might not be achievable on other projects, since it is up to the development culture and rules inside a particular organization.

In terms of reliability the accuracy of the annotated data can be a concern as it can produce biased results. In case of the reconstruction this would give spurious results, but reconstruction would still be successful. So this threat is more related to the accuracy of the reconstructed method. To mitigate this threat, the reconstructed method was validated with project members.

8. Related Work

There are several works that can be considered related to our research. These can be grouped into four categories, according to the research fields they come from:

- Method engineering
- Software repository mining
- Software process discovery
- Software process mining
- Other related approaches

Note that there is some overlapping among these fields (specifically among the last three fields), in terms of approaches and techniques that they use. Different names that

they carry are more a result of the fact that they come from different research groups and times.

8.1. Method Engineering

Method engineering is an approach to create software development methods that are specifically attuned to organizations. In general, the idea lies in the conceptualization, and construction of new methods and tools (or in the adaptation of existing ones), so that they best fit requirements of a certain organization. The research on method engineering has a long tradition. A good introduction to the field can be found in 31. Based on the method engineering principles, a specific direction has emerged, called situational method engineering. As the name implies, situational method engineering deals with developing new methods or adapting existing ones on-the fly, i.e. to meet specific project situations. In the literature, a number of situational method engineering approaches were suggested [16, 32, 33]. For an excellent review see 10.

Method engineering and specifically situational method engineering works are related to our research in general, as they share the same motivation, i.e. to help software development companies develop software in more disciplined way. In both cases, development methods are the subject of research with a difference that situational method engineering approaches require much more human effort to properly work. As reported in 14 and 16, this is considered one of the main reasons why situational method engineering approaches hardly penetrate to practice.

8.2. Software Repository Mining

The analysis of software repository data is a research discipline that deals with the analysis of rich software repository information to get valuable insights about the development process and software itself. For a survey, see 34. The works that are directly related to our research, are for instance 35, 36, or 37. Here, the authors employ various statistical models, such as Latent Dirichlet Allocation and Latent Semantic Indexing to cluster unstructured and unlabeled textual data (commit log comments, source code, documentation, mailing lists, etc.) into topics. Although results show that this can be done efficiently, the topics do not convey much information on the underlying development method, except maybe the main activities, if they can be inferred from the topics.

In this group, we also include works that deal with software process recovery. Existing approaches mainly apply supervised and unsupervised techniques, such as bag-of-words, summary statistics, topic analysis, and Bayesian classifiers to recover the development process from a variety of artifacts that were created by developers and can be obtained from software repositories [21, 22]. These approaches allow for the recovery of the so called Unified Process Views, which illustrate how the relative emphasis on different disciplines changes over the course of the project. Detailed information on activities, user roles, and artifacts are out of scope of these works.

8.3. Software Process Discovery

In the field of software process discovery 38, their main objective is to automatically derive a formal model of a process from the data that was collected during the execution of a process. Several approaches have been suggested on how this can be done, for example [39–41]. What these works have in common, is that they only use information from revision control systems and not also from other systems that usually comprise a software repository.

8.4. Software Process Mining

In the field of software process mining, the authors linked data from different software repositories and apply process mining techniques to derive a process map and identify inefficiencies, imperfections, and enhance existing process capabilities [23, 42]. They have used data from software repositories to perform different control and organizational analyses. However, as part of their analyses, they only focus on the processes on the level of code review or bug life cycle. The same goes for other work in this field that employ process mining techniques to recover valuable information [24, 43, 44]. Here, the authors focus on the reconstruction of software processes on the level of disciplines or on the level of issues, but do not consider activities, user roles and artifacts, as we do in our research.

8.5. Other related approaches

There are also other research areas related to our research such as Organizational patterns, which also can be used to capture software development methods [45, 46, 47]. Mainly organizational patterns are still captured and documented manually, but some of the researchers are trying to use data from software repositories to detect bad practices (anti-patterns) [27, 28]. These approaches could benefit from our research since they could analyze the behavior and patterns on the higher level, level of activities.

9. Conclusion and Future Work

Software development is a complex and creative task, whose sophisticated results are increasingly influencing our daily lives in various ways. Due to the nature of this work, it is important that each company has a method in place to manage, control, and guide the work of software developers and project managers. Otherwise, confusion may ensue, leading to project failures, low quality of the developed software and higher maintenance costs.

To manage software projects, companies often use different supporting tools, such as issue tracking systems, revision control systems, document management systems, code review tools, and others. Their main goal is to support the work of developers. Each tool

per se contains a lot of information and valuable knowledge on how a project has been performed in practice. However, to obtain an even better overview of the development process as a whole, the information from these tools can be linked. Linked data can then be used to reconstruct what really happens behind those projects and eventually to learn, among others, why some projects go well and others do not.

In this paper, we described how the data from different software repositories (issue tracking system, revision control system, document management system) can be used to reconstruct valuable information on the project performance with only a little involvement of the developers. The aim of the paper was to demonstrate that using and linking the data from tools comprising software repositories, allows us to reconstruct the development method in more details than existing approaches do. Furthermore, the aim was to show that it is possible to capture the actual ways of working in an organization, in a form of a base method, which can be constantly kept up-to-date without any significant involvement of developers.

To identify the information that can be retrieved from software repositories we have cooperated with five companies, which shared their data with us. Based on the findings we have developed an approach to reconstruct development practice. We have evaluated the approach on a real software repository shared by an additional company. The results show that software repository information suffice for the reconstruction of various aspects of development process, i.e. disciplines, activities, roles, and artifacts.

As part of our future work we plan to gather data from other software repositories and include it into the process of reconstruction. With this we expect to rise the reconstruction accuracy and level of details reconstructed. We also plan to use the approach on other software projects to see how it performs in real-time manner (monitor, control, guide).

Acknowledgments. We thank Professor Martin Pinzger for his comments on the draft of this paper. We also wish to acknowledge financial support for this project by the Slovenian Research Agency ARRS within the research program P2-0359.

References

1. David P. and Parnas L.: Risks of Undisciplined Development. Communications of the ACM, Vol. 53, No. 10, 25-27. (2010) doi: 10.1145/1831407.1831419.
2. Hardy C. J., Thompson J. B. and Edwards H. M.: The use, limitations and customization of structured systems development methods in the United Kingdom. Information and Software Technology, Vol. 37, Issue. 9, 467-477. (1995) doi: 10.1016/0950-5849(95)97291-F.
3. Fitzgerald B.: An empirical investigation into the adoption of systems development methodologies. Information & Management, Vol. 34, 317–328. (1998) doi: 10.1016/S0378-7206(98)00072-X.
4. Huisman M. and Iivari J.: The individual deployment of systems development methodologies. Lecture Notes in Computer Science, 134–150. (2002).
5. Hansson C., Dittrich Y., Gustafsson B. and Zarnak S.: How agile are industrial software development practices? Journal of Systems and Software, Vol. 79, Issue 9, 1295-1311. (2006) doi: 10.1016/j.jss.2005.12.020.

6. Gonzalez-Perez C. and Henderson-Sellers B.: A work product pool approach to methodology specification and enactment. Journal of Systems and Software, Vol. 81, Issue 8, 1288-1305. (2008) doi: 10.1016/j.jss.2007.10.001.
7. Petersen K. and Wohlin C.: A comparison of issues and advantages in agile and incremental development between state of the art and an industrial case. Journal of Systems and Software, Vol. 82, Issue 9, 1479-1490. (2009) doi: 10.1016/j.jss.2009.03.036.
8. Clarke P. and O'Connor R. V.: The situational factors that affect the software development process: Towards a comprehensive reference framework. Information and Software Technology, Vol. 54, Issue 5, 433-447. (2012) doi: 10.1016/j.infsof.2011.12.003.
9. Ivarsson M., Gorschek T.: Practice selection framework. Int. J. Soft. Eng. Knowl. Eng., Vol. 22, No. 01, 17-58. (2012) doi: 10.1142/S0218194012500027
10. Henderson-Sellers B., Ralyté J., Ågerfalk P. J. and Rossi M.: Situational Method Engineering. Springer. (2014) doi: 10.1007/978-3-642-41467-1.
11. Kuhrmann M., Méndez Fernández D., and Tiessler M.: A mapping study on the feasibility of method engineering. J. Softw. Evol. and Proc., 1053–1073. (2014) doi: 10.1002/smr.1642
12. Ter Hofstede A. H. M., Verhoef T. F.: On the feasibility of situational method engineering. Information Systems, Vol. 22, Issue 6, 401-422. (1997) doi: 10.1016/S0306-4379(97)00024-0.
13. Fitzgerald B.: The use of systems development methodologies in practice: a field study. Information Systems journal, 201-212. (1997) doi: 0.1046/j.1365-2575.1997.d01-18.x.
14. Fitzgerald B., Russo N. L. and O'Kane T.: Software development method tailoring at Motorola. Communications of the ACM, Vol. 46, Issue 4, 65-70. (2003) doi: 10.1145/641205.641206.
15. Coleman G.: An Empirical Study of Software Process in Practice. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 315c-315c. (2005) doi: 10.1109/HICSS.2005.86.
16. Bajec M., Vavpotič D. and Krisper M.: Practice-driven approach for creating project-specific software development methods. Information and Software Technology, Vol. 49, Issue 4, 345-365. (2007) doi: 10.1016/j.infsof.2006.05.007
17. Gupta M., Sureka A., Padmanabhuni S. and Asadullah A. M.: Identifying Software Process Management Challenges: Survey of Practitioners in a Large Global IT Company. In 12th IEEE Working Conference on Mining Software Repositories, 346-356. (2015).
18. Yan Sun, Qing Wang and Ye Yang: FRLink: Improving the recovery of missing issue-commit links by revisiting file relevance. Information and Software Technology, Vol. 84, 33-47. (2017) doi: 10.1016/j.infsof.2016.11.010
19. Jankovic M. and Bajec M.: Comparison of software repositories for their usability in software process reconstruction. In proceedings of IEEE 9th International Conference on Research Challenges in Information Science, 298–308. (2015) doi: 10.1109/RCIS.2015.7128890
20. Goeminne M. and Mens T.: A Comparison of Identity Merge Algorithms for Software Repositories. Sci. Comput. Program, Vol. 78, 971-986. (2013) doi: 10.1016/j.scico.2011.11.004
21. Hindle A.: Software process recovery: Recovering process from artifacts. In 17th Working Conference on Reverse Engineering, 305-308. (2010) doi: 10.1109/WCRE.2010.46
22. Hindle A., Godfrey M. and Holt R.: Software process recovery using recovered unified process views. In IEEE International Conference on Software Maintenance, 1-10. (2010) doi: 10.1109/ICSM.2010.5609670
23. Gupta M. and Sureka A.: Mining Bug Report History for Discovering Process Maps, Inefficiencies and Inconsistencies. In Proceedings of the 7th India Software Engineering Conference, 1-10. (2014) doi: 10.1145/2590748.2590749

24. Poncin W., Serebrenik A. and Van den Brand M.: Process mining software repositories. In 15th European Conference on Software Maintenance and Reengineering, 5-14. (2011) doi: 10.1109/CSMR.2011.5
25. Hug C., Front A., Rieu D., and Henderson-Sellers B.: A method to build information systems engineering process metamodels. Journal of Systems and Software, Vol. 82, No. 10, 1730-1742. (2009)
26. OMG, Software & Systems Process Engineering Metamodel Specification (SPEM), Tech. rep. (Apr. 2008). URL http://www.omg.org/spec/SPEM/2.0
27. Pícha P., Brada P., Ramsauer R., and Mauerer W.: Towards Architect's Activity Detection Through a Common Model for Project Pattern Analysis. In Proceedings of 2017 IEEE International Conference on Software Architecture Workshops. (2017).
28. Pícha P. and Brada P.: ALM Tool Data Usage in Software Process Metamodeling. In Proceedings of 2016 42th Euromicro Conference on Software Engineering and Advanced Applications. (2016).
29. Fischer M., Pinzger M. and Gall H.: Populating a Release History Database from version control and bug tracking systems. In proceedings of the International Conference on Software Maintenance, 23-32. (2003) doi: 10.1109/ICSM.2003.1235403
30. Sliwerski J., Zimmermann T. and Zeller A.: When Do Changes Induce Fixes? In proceedings of the International Workshop on Mining Software Repositories, 1-5. (2005) doi: 10.1145/1082983.1083147
31. Brinkkemper S., Lyytinen K. and Welke R. J.: Method engineering: principles of method construction and tool support. In Selected Papers from the International Conference on ''Principles of Method Construction and Tool Support''. (1996) doi: 10.1007/978-0-387-35080-6
32. Ralyté J., Deneckère R. and Rolland C.: Towards a generic model for situational method engineering. In Proceedings of the 15th Conference on Advanced Information Systems Engineering, 95-110. (2003)
33. Karlsson F. and Ågerfalk P. J.: MC Sandbox: Devising a tool for method-user-centered method configuration. Information and Software Technology, Vol. 54, Issue 5, 501-516. (2012) doi: 10.1016/j.infsof.2011.12.009
34. Woosung J., Eunjoo L. and Chisu W.: A Survey on Mining Software Repositories. IEICE Transactions on Information and Systems, 1384-1406. (2012)
35. Savage T., Dit B., Gethers M. and Poshyvanyk D., TopicXP: exploring topics in source code using latent dirichlet allocation, In IEEE International Conference on Software Maintenance, 1–6, 2010. doi: 10.1109/ICSM.2010.5609654.
36. Chen T. H., Thomas S. W., Nagappan M. and Hassan A. E., Explaining software defects using topic models, In 9th IEEE Working Conference on Mining Software Repositories, 189–198, 2012. doi: 10.1109/MSR.2012.6224280.
37. Hindle A., N. Ernst A., Godfrey M. W. and Mylopoulos J.: Automated topic naming. Empirical Software Engineering, Vol. 18, Issue 6, 1125-1155. (2013) doi: 10.1007/s10664-012-9209-9
38. Cook J. E. and Wolf A.: Automating process discovery through event-data analysis. In 17th International Conference on Software Engineering, 73-73. (1995) doi: 10.1145/225014.225021
39. Kindler E., Rubin V. and Schäfer W.: Incremental workflow mining based on document versioning information. Unifying the Software Process Spectrum, no. 3840 in Lecture Notes in Computer Science, 287-301. (2006)
40. Akman B. and Demirors O.: Applicability of process discovery algorithms for software organizations. In 35th Euromicro Conference on Software Engineering and Advanced Applications, 195-202. (2009) doi: 10.1109/SEAA.2009.87

41. Duan B. and Shen B.: Software process discovery using link analysis. In IEEE 3rd International Conference on Communication Software and Networks, 60-63. (2011) doi: 10.1109/ICCSN.2011.6014218
42. Gupta M.: Process Mining Software Repositories to Identify Inefficiencies, Imperfections, and Enhance Existing Process Capabilities. In Companion Proceedings of the 36th International Conference on Software Engineering, 658-661. (2014) doi: 10.1145/2591062.2591080
43. Rubin V., Gunther C. W., Van Der Aalst W. M. P., Kindler E., Van Dongen B. F. and Schäfer W.: Process mining framework for software processes. In proceedings of the international conference on Software process, 169-181. (2007)
44. Lemos A., Sabino C., Lima R. and Oliveira C.: Using process mining in software development process management: A case study. In proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1181-1186. (2011) doi: 10.1109/ICSMC.2011.6083858
45. Coplien O. J. and Harrison B. N.: Organizational Patterns of Agile Software Development, Prentice Hall. (2004)
46. Sulaiman Khail W. and Vranić V.: Treating Pattern Sublanguages as Patterns with an Application to Organizational Patterns. In Proceedings of 22nd European Conference on Pattern Languages of Programs. (2017)
47. Frťala T. and Vranić V.: Animating Organizational Patterns. In Proceedings of 8th International Workshop on Cooperative and Human Aspects of Software Engineering. (2015)

Appendix A. Project Performance Analyses

In this appendix, we provide examples of visualizations that have been created during the evaluation phase based on the information created with the iSPRToolset (http://ispr.jmlabs.eu). The aim is to show that there are other insights on the project performance that we get by following the proposed approach and might be beneficial for project managers.

The reconstructed method elements in combination with other information available from the software repositories (e.g. worklogs – each user logs hours spent working on a specific issue) provide a basis for different project analyses. All figures in this section are created using the data provided by the participating company.

Using the information about the worklogs allows us to observe on a daily basis how much time (effort) was spent per discipline, activity, or user role. This tells us, for instance, for which part of the project the most time was spent. Furthermore, we also have a possibility to analyze the total time spent for a particular discipline, activity, or user role. The daily intensity of particular discipline/activity/user role is presented on the right-hand side of Figures 6, 7, and 8.

Before acquiring a project, companies typically prepare a tender for which they also estimate the time that will be needed for a particular activity on the project in order to estimate the total costs of the project. The comparison of the actual and estimated time allows us to detect for which disciplines/activities/user roles developers spent more time than expected. This information is very important to project managers and can help them to make better estimations on new projects. In our case, the observed company, in order to prepare a tender and to estimate the project costs, does an internal financial calculation as part of which they also estimate the time needed for a particular task on the project. All this information is documented in an Excel file, which is classified to the Financial Calculation artefact from the base method. We used this information to gather

information about the estimated time for each activity and consequently also for each discipline. The comparison of the estimated and actual time is presented on the left-hand side of Figures 6, 7, and 8. It shows what portion of time was planned for a particular discipline/activity/user role and what portion was actually spent. With this comparison, project manager can identify project activities (and roles) that required more time than was expected.

It is a rule in the observed company that at the beginning of each project they also prepare a project plan, which includes the timeline of a project - often presented with a gantt chart. From the gantt chart we gathered the information about the timespan of a particular task and when it was planned to be resolved. We used this information to visualize how the expected timeline deviated from the actual one. The information from the gantt chart is integrated into the actual timeline of a project and is presented with light blue rectangles in Figures 6 and 8.

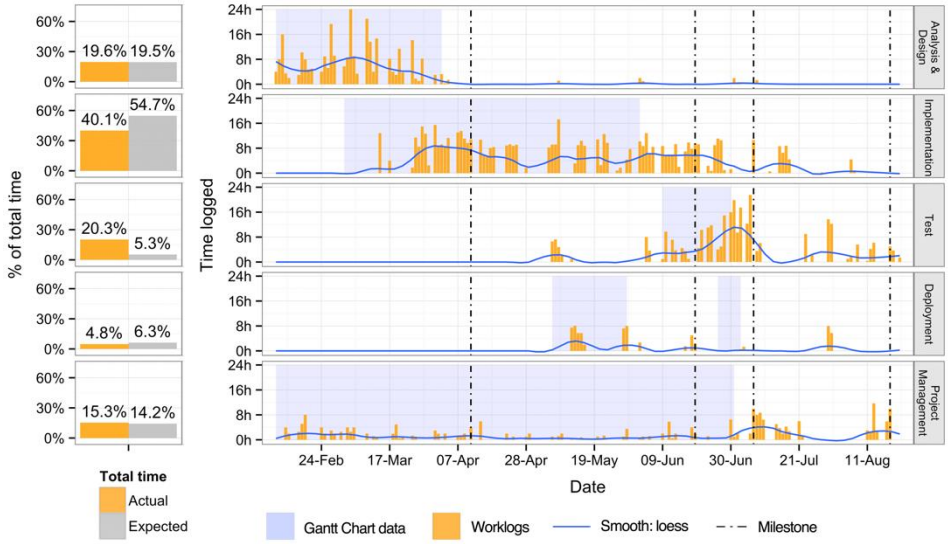


Fig. 6. Intensity of particular disciplines.

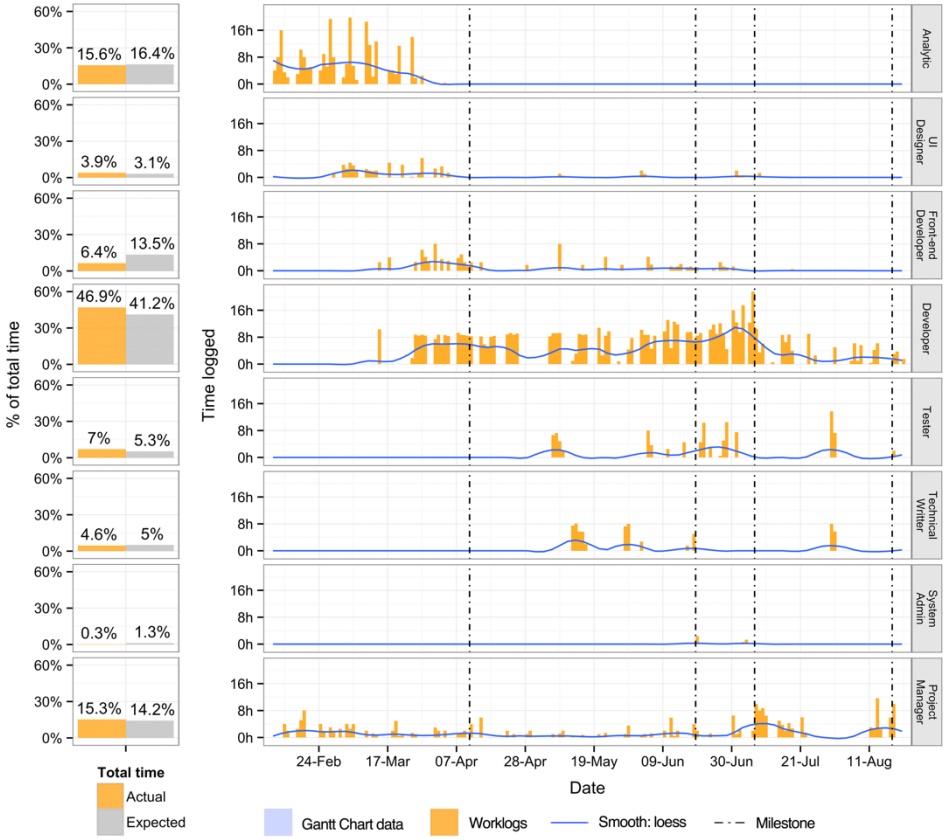


Fig. 7. Analysis on the level of user roles.

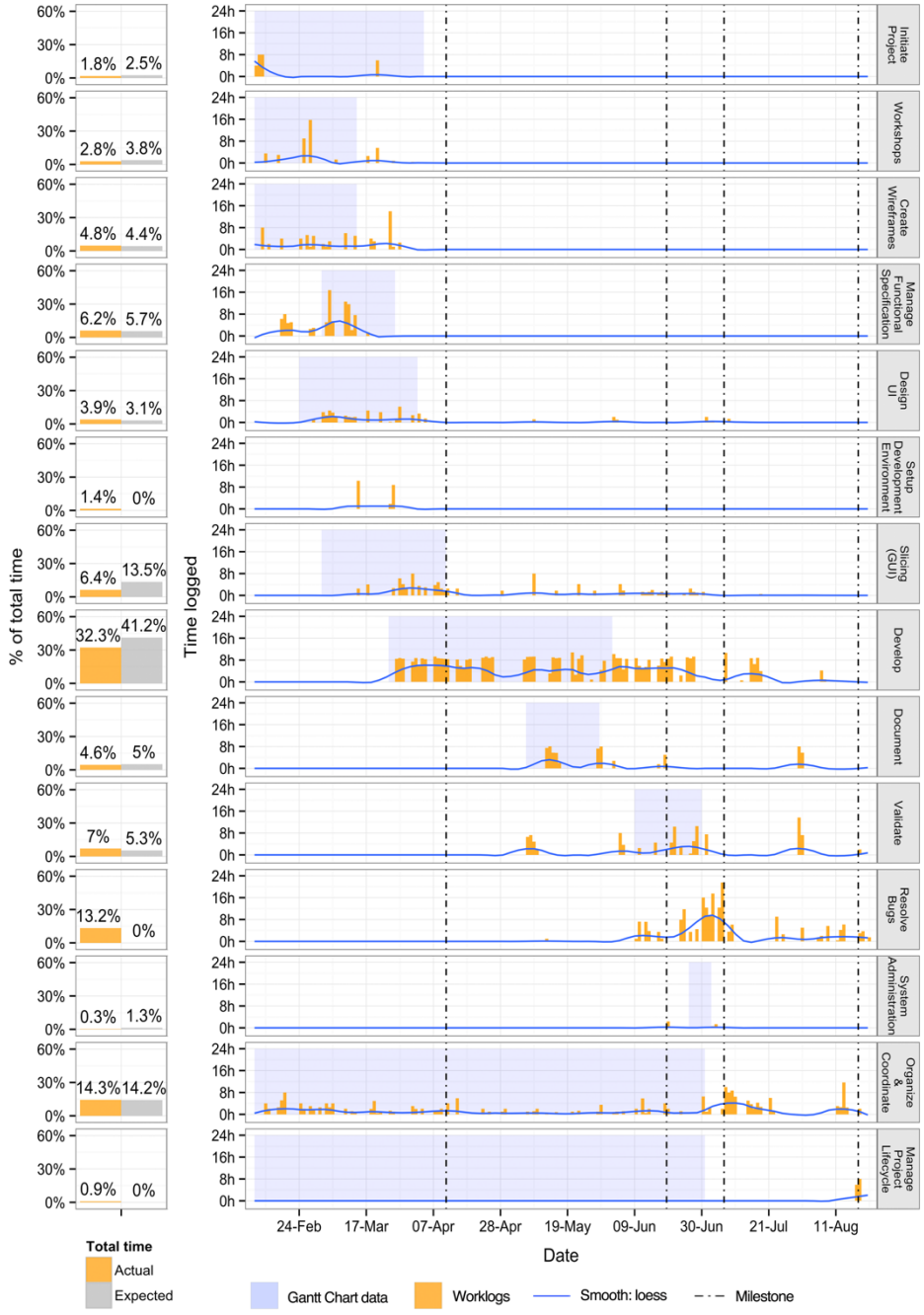


Fig. 8. Analysis on the level of activities

Marko Janković is a reseacher at the Faculty of Computer and Information Science, University of Ljubljana. He has been researching mainly in the field of mining software repositories and software development methods. His work has been published in journals and presented on several international conferences. Professionally he is also working on projects related to the Internet of things, big data, and data analysis.

Slavko Žitnik, PhD, is an assistant at the University of Ljubljana. His bibliography counts more than 20 items with a book chapter and a patent application. He has been researching mainly in the field of information retrieval and information extraction from textual sources. He designed an end-to-end information extraction system that uses an ontology and represents the extracted entities, relationships and coreferences against it. In the field of natural language processing he achieved first place at the BioNLP 2013 challenge in extracting relationships from text to build gene regulation network. Professionally he is also working on projects related to the Internet of things, big data, social networks and semantic data analysis.

Marko Bajec is a Full Professor and head of the Laboratory for Data Technologies at the Faculty of Computer & Information Science, University of Ljubljana. His research interests primarily focus on IT and data Governance include software development methods, IT/IS strategy planning, data management and manipulation. His work has been published in many high-ranked journals. In his career, he has led or coordinated over 30 applied and research projects. For his contribution in transferring knowledge to industry he received several awards and recognitions.

Received: February 26, 2018; Accepted: October 11, 2018

Computer Science and Information Systems 16(1):105–129 https://doi.org/10.2298/CSIS181010040G

CrocodileAgent 2018: Robust Agent-based Mechanisms for Power Trading in Competitive Environments

Demijan Grgic[1], Hrvoje Vdovic[2], Jurica Babic[3], and Vedran Podobnik[4]

[1] University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
demijan.grgic@fer.hr
[2] University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
hrvoje.vdovic@fer.hr
[3] University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
jurica.babic@fer.hr
[4] University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
vedran.podobnik@fer.hr

Abstract. Besides the smart grid, future sustainable energy systems will have to employ a smart market approach where consumers are able choose one of many different energy providers. The Power Trading Agent Competition (Power TAC) provides an open source, smart grid simulation platform where brokers compete in power brokerage. This paper presents CrocodileAgent, which competed in the Power TAC 2018 finals as a broker agent. The main focus in the design and development of CrocodileAgent 2018 was the creation of smart time-of-use tariffs to reduce peak-demand charges. CrocodileAgent 2018 was ranked third in Power TAC 2018 Finals, with a positive final profit and a positive result in each of three game types. In addition, CrocodileAgent 2018 had the highest percentage of "profitable games" (91%) from among all competing agents, the second highest level of "net profit per standard deviation" (0.48) and the third highest "net profit per subscriber" (79 monetary units).

Keywords: computer simulation, agent-based modelling, electricity trading, tariff design, Power TAC, CrocodileAgent

1. Introduction

Environmental challenges like pollution and CO2 emissions have recently become one of main drivers for innovations within the energy sector [9]. The common denominator for innovations is changes which make the world a better living place for current and future generations, and is a key concept behind sustainable development [28]. Given the importance that decision makers possess a holistic view when considering sustainability issues [27], the electricity market is a particularly relevant area of research as it directly deals with the triple bottom line, *i.e.*, social, environmental (or ecological) and financial aspects of the energy industry.

Evidently, the ever-increasing efforts to include renewable energy sources, such as wind turbines and solar systems, in the existing *energy layer* of a power system aim to

address the *environmental part* of sustainability. Large-scale producers (*e.g.*, coal-based power plants), as well as small-time producers, also known as *prosumers* (*e.g.*, individual households with solar panels), end-consumers and energy brokers (*i.e.*, entities which essentially act as mediators between large-scale producers and end-consumers), are examples of highly heterogeneous *social* groups that exhibit potentially conflicting objectives. For example, producers often seek to maximize profit which in turn negatively affects the consumer objective of minimizing electricity costs. In order for stakeholders to work in such progressively complex market environments, the *smart grid* is seen as an essential technical solution because it allows entities to connect and exchange information by implementing the *ICT layer* within the power system as well as enabling a two-way flow of electricity [12].

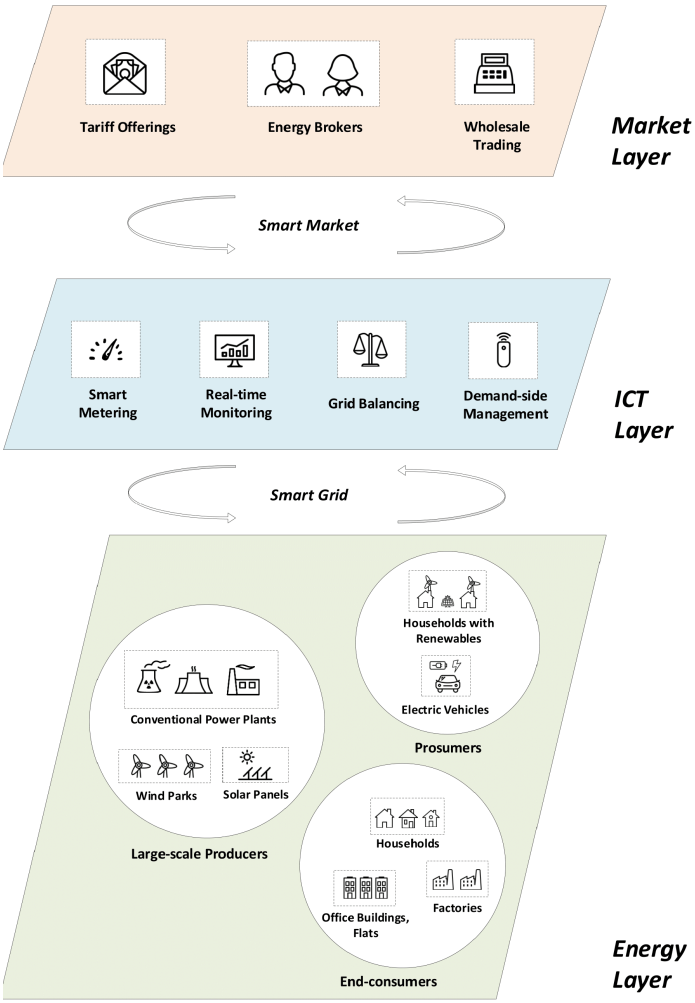


Fig. 1. A multi-layered concept of the next-generation power system (adapted from [2]).

However, as seen in Figure 1, it is evident that the smart grid is only a *technical foundation* for the next-generation power systems. In order to have a truly sustainable energy system of the future, one needs to *design*, *evaluate* and *implement* the *market layer* which ultimately provides added value for both market participants and end-consumers [2]. For example, until recently, end-consumers found themselves in highly-regulated environments and were tied to a single energy supplier, an approach that may not have been so market efficient. On contrary, with new functionalities now offered by smart grids (*e.g.* smart metering), end-consumers can receive and react to price signals from numerous energy suppliers, and thus have an active role in the newly-formed *smart market* [7].

Undoubtedly, designing this kind of market is not straightforward due to the need to overcome numerous inherent challenges. Perhaps the most intriguing challenge relates to system *complexity* which stems from complex interactions of heterogeneous entities. As opposed to other commodity markets (e.g., soft commodities like wheat, coffee and fruit), the electricity market is different due to certain intricacies related to electricity. That is, electricity must be used *instantly* when generated, meaning that *supply must match demand* exactly at any given time across the grid.

The scope of this work focuses on the trading mechanisms that a power broker usually utilizes on the electric power market in order to mediate electricity supply from producer to consumer. When it comes to brokering, a successful power broker is expected to maintain a profitable retail customer portfolio. To do so, a power broker is also required to have a compatible strategy on the wholesale market in order to meet customers' demand for energy. Since the electric power system is a critical resource for consumers, it is exceptionally important that the broker deploy mechanisms which have been *thoroughly tested* beforehand. Accordingly, researchers nowadays often resort to computer simulations that provide a highly-detailed computational model of the real-world system [1] [10]. Once the computational model is implemented, the entities within the model can be adjusted to evaluate and quantify the impact of certain interactions within the system in a risk-free environment. Hence, this paper follows the same principle by using an open source simulation platform called Power Trading Agent Competition (Power TAC) to evaluate robust trading mechanisms which are encoded within the agent-based [25] [6] power broker called CrocodileAgent. Within the scope of this work, *robustness* of trading mechanisms refers to the ability of a power broker to be competitive in the majority of scenarios regardless of the competition size. The result obtained from the annual Power TAC world competition 2018, *i.e.*, third place against 6 broker agents prepared by other research groups, demonstrates the applicability of the proposed trading mechanisms in competitive environments.

To summarize, the contribution of this paper comprises the following two parts: (i) design and implementation of robust agent-based mechanisms for power trading in next-generation power systems; and (ii) evaluation of the proposed mechanisms based on data from the Power TAC 2018 world competition.

The rest of the paper follows a specific structure. Section 2 presents the Power TAC simulation platform and its simulation scenario which resembles the real-world next-generation power system. Section 3 presents the proposed trading mechanisms by explaining the key design choices for the CrocodileAgent 2018 broker agent. Section 4 analyses and evaluates the CrocodileAgent 2018 using data generated from the Power TAC 2018 world competition. Section 5 concludes the paper and presents directions for future work.

2. Power Trading Agent Competition

2.1. Simulation Platform

The Power Trading Agent Competition (Power TAC) started as a collaboration between six European and North American universities. The goal was to create an open-source, competitive simulation that models a "liberalized" retail electric power market, where competing business entities offer energy services to customers through tariff contracts, and must then serve those customers by trading in a wholesale market [15] [13]. An attempt was made to introduce this kind of liberalized electric power market in California in 2000, eventually resulting in a major energy crisis [31], and leading to the conclusion that additional research and simulation was necessary in determining the regulatory framework for such a market where Power TAC, as a computational model of the electricity market, is an efficient way of achieving that in terms of time and costs.

Power TAC requires that competing teams establish trading agents or "brokers" that aggregate supply and demand in order to earn a profit. Brokers may buy and sell energy through contracts with retail customers and by trading in a wholesale market modelled after the European and North American wholesale energy markets. In the customer (tariff) market, brokers offer tariff contracts to a population of customers which may include fixed or variable prices for both consumption and production of energy, incentives for energy conservation, sign-up bonuses or early-withdrawal penalties. The wholesale market enables brokers to buy and sell energy for future delivery while the balancing market is responsible for real-time balancing of supply and demand on the distribution grid.

The initial version of the Power TAC platform was released in 2011. Power TAC is designed as an annual tournament with the first competition held in 2013. Usually, 6 to 11 participants from universities and research centres around the world compete in the tournament. In this tournament environment, simulations are run with different numbers and combinations of broker agents, and the most profitable agent over a range of scenarios is the winner [14]. After the tournament, participants typically present and discuss their work on relevant conferences and publish scientific papers [5]. Teams are encouraged to release their agent code, so that all teams can experiment with a range of different agent behaviours, thus improving their agent designs for the following year. Each year, the scenario is updated with new challenges, tuning market designs and adding a new level of realism. Changes in the 2018 scenario were focused on stability, simplifying interaction with the balancing market, and on encouraging more vigorous competition [15].

2.2. Brokers

In Power TAC, brokers are trading agents competing against each other to maximize their profits. In each time slot (one-hour simulated time period), a broker can offer or modify tariffs on the *tariff market*, submit a balancing order on the *balancing market* as well as submit asks and bids on the *wholesale market* to sell or procure energy for future time slots. The simulation server simulates *large-scale producers*, *prosumers* and *end-consumers* and their communication with corresponding markets, exposing only those markets to the energy brokers as shown in Figure 2. Brokers get weather forecast information, wholesale market clearing data, tariff transactions, portfolio supply and demand, including market and cash position from the simulation server each time slot. Apart from

trading in markets, broker should balance its overall supply and demand, as any surplus or deficit of energy will be covered by the distribution utility at a very unfavourable price. A broker balances its portfolio by acquiring producers and consumers that balance each other in real time, by acquiring storage capacity, controllable consumption and production capacity that is used as needed, and by trading future contracts on the wholesale market. Given that consumer consumption of electricity and production of energy from renewable sources is a variable factor, brokers need to have methods for forecasting energy consumption and production as precisely as possible in order to provide consumers with the required amount of energy and balance supply and demand. To provide further insight into a broker's main responsibilities and best strategies, this paper will next review literature on the most successful brokers that competed in Power TAC, i.e., TacTex, cwiBroker and AgentUDE.

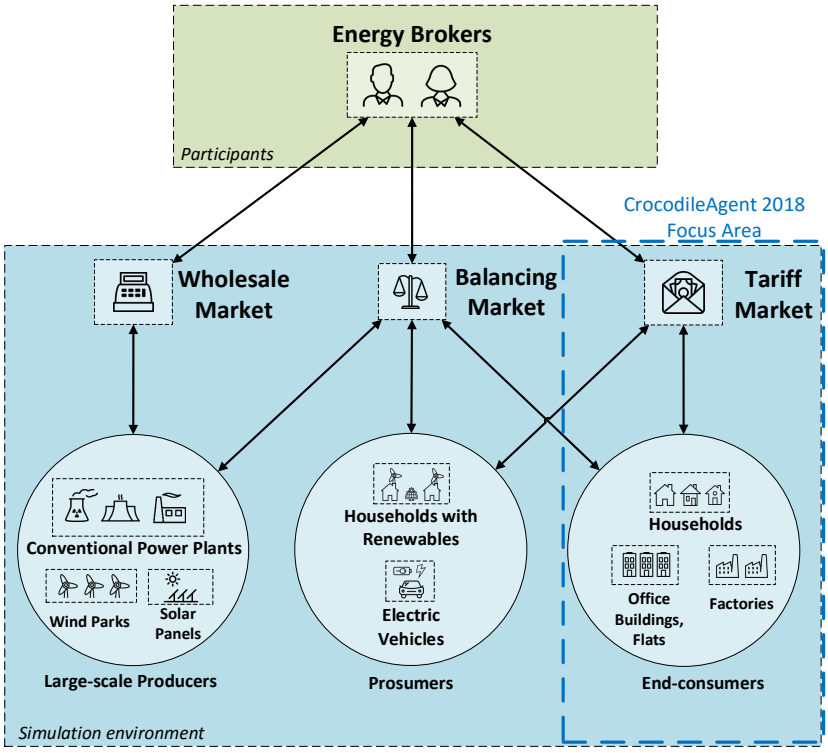


Fig. 2. Relations between Power TAC brokers and the simulation environment; the focus of CrocodileAgent 2018 was on designing and implementing successful tariff market mechanisms

TacTex was the champion in the 2013 inaugural competition [29]. It operates simultaneously in multiple markets and aims to increase the cash amount in its bank account which is treated as its utility measure. It maximizes its utility by simultaneously optimizing energy-selling prices, total energy demand from its customers and energy procurement

costs. The wholesale market bidding strategy of the TacTex 2013 broker is to collect market data and estimate market demand in order to minimize energy procurement costs. Its estimates improved during the game as more and more data was collected using an online reinforcement learning algorithm, optimization and adaptation of the bidding strategy to the specific market conditions in the game. Given the predicted energy demand and cost, TacTex in 2013 optimized future demands and selling-prices in the tariff market. Furthermore, the broker performed best in games with a small number of competitors, winning 6 out of 6 two player games and 15 out of 16 four player games.

The second-placed broker in the 2013 and 2014 Power TAC finals was cwiBroker [18]. The competitor cwiBroker endeavoured to balance its demand and supply by estimating the production and consumption of each type of customer. In the tariff market, its strategy varies depended on the type of game. If it was a duopoly game, the broker was initially competitive and then, after obtaining a reasonable number of customers, it copied its competitor's tariff price. In games with three or more players, the broker was more competitive by devising a set of tariffs, estimating the amount of profit achieved using a particular tariff and publishing the most profitable tariff. The wholesale bidding strategy used by cwiBroker was to buy energy at bargain prices in the first auction for each timeslot and then sell the surplus in later auctions at higher price. The competitor cwiBroker performed best in games with a high number of players, winning all seven player games.

AgentUDE was the winner of the 2014 [22] [20] and 2017 [23] [24] Power TAC finals and was ranked among the top 4 in 2015 and 2016. It stores historical data in a local repository which was subsequently evaluated using the wholesale and tariff module while creating future values for the retail and wholesale markets. In the wholesale market, AgentUDE predicts market trends regardless of weather conditions by tracking historical market data [21]. In the tariff market, it deploys an aggressive tariff strategy. The broker endeavors to offer the cheapest tariff, thus provoking other brokers to publish cheaper tariffs and triggering tariff penalties which translated into profit. The change in the 2015 Power TAC specification which added peak-demand charges for brokers whose customers spend the largest amount of energy in peak times required a change in broker behavior. To reduce peak-demand charges, AgentUDE used a time-of-use tariff price scheme with different rates depending on the time of day and day of week to encourage customers to avoid consumption during peak hours.

Previously mentioned brokers were designed with the aim to participate in the Power TAC world competition. However, Power TAC as a platform can also be used in a so-called *research mode* in which researchers are able to conduct *in vitro* experiments for exploring various research questions related to the next-generation power system. One such a research example is the work by Rubio el al. [26]. Concretely, the authors used fuzzy models to develop the so-called TugaTAC broker agent. The key mechanism is based around on updating tariffs using a conceptual model for agent's interest on selling or buying energy. Based on the input the broker receives from the environment, the fuzzy model helps the broker in improving tariffs with the aim to attract the best profile of clients. The proposed mechanism is benchmarked against the default broker and several brokers from the Power TAC 2014 competition. The same evaluation approach was used by Wang et al. for the sake of testing mechanisms implemented within the GongBroker agent [30]. Such a broker utilizes a hybrid-learning approach which includes a data-driven

method for predicting short-term customer demand, a Markov Decision Process (MDP) for activities on the wholesale market and reinforcement learning for activities on the retail market. The use of MDP in wholesale activities was also investigated by Kuate et al [17] within the AstonTAC broker agent. It is important to notice that our work is different than previously mentioned brokers in terms of the focus area, *i.e.*, improvements on the retail market activities based on the idea of avoiding peak-demand charges. Furthermore, one can conclude that our work has a stronger validation due to the fact the analysis was performed based on data from the latest Power TAC 2018 world competition.

Considering the approaches used by the above-mentioned broker agents, it is evident that Power TAC is the community-driven simulation platform evolving after each annual cycle [16]. Once the game specification is updated, broker agents need to make careful adjustment to stay competitive in new tournaments. Furthermore, given the complexity of the simulation platform, successful research teams often opt to specialize their broker for a certain trading task. For example, instead of developing an all-around broker, research teams might only focus on developing solutions aimed at retail consumption customers. Finally, judging from the results of the previous annual competitions, it becomes evident that different broker teams were successful in different years. This suggests that broker teams who fail to devise innovate designs for each annual competition are likely to be less competitive and underperform. For example, the CrocodileAgent broker [19] [3], after initially obtaining encouraging results, including fourth and third places in the 2013 and 2014 tournaments, experienced a gradual decline in performance before eventually dropping to last place in the 2017 tournament. The next section explains how the latest iteration of the CrocodileAgent broker, CrocodileAgent 2018, succeeded in significantly improving its competitiveness by focusing on innovations in tariff structure pricing with the specific aim of stabilizing unwanted risk exposure and negative fees.

3. Power Broker Design: CrocodileAgent

In general, broker agent actions in Power TAC can be summarized by three key interactions that occur in the following sequence: (i) publishing tariffs on the retail market to attract customers at an attractive consumer price point; (ii) purchasing energy at a competitive price for future time-slots to satisfy customer energy demand; and (iii) balancing energy supply-demand for each time slot to reduce over- or under-purchasing of electric energy. Given that the broker is a software product, organizers of the Power TAC provide dummy agent code in the Java programming language which is then used by research teams to code their brokers.

The key principle behind the CrocodileAgent broker is a modular design to facilitate integration of newly-developed mechanisms for power trading. That being said, broker functionalities are categorized into three sections: (i) wholesale market operations; (ii) retail market operations; and (iii) internal portfolio tracking that enables the agent to successfully participate in both the wholesale and retail market. Figure 3 showcases a modular approach used in the CrocodileAgent design and implementation. The *Wholesale Manager* module is responsible for activities in the wholesale market. The main goal of the module is to implement the efficient purchase of energy with respect to the price of power because the arrangement enables the broker to earn potentially significant profits. However, to make the appropriate trades on the wholesale market, the module requires

the information on demand for power from retail customers. To enable such an information, the *Portfolio Manager* module is used to actively track customer behaviour of all subscribed customers. In the retail market, the main objective is to have the most efficient resale of inventory, while taking into account prices from the wholesale market as well as demand for power from retail consumers. Operations in the wholesale and retail markets are highly interconnected. The behavior of CrocodileAgent 2018 regarding retail activities is encoded in the following modules: *Tariff Manager* is used to evaluate a tariff price, based on input from the wholesale market; *Smart Tariff Creator* is used to create tariffs based on multiple options (*e.g.,* tariff types and special conditions such as early withdrawal fees); and *Tariff Rate Creator* is used to create individual tariff rates which depend on the desired margin effects that each tariff is expected to exhibit.

3.1. Data-driven Strategic Planning

Given that agent trading within the Power TAC framework is characterized by competitive game interactions, the assumptions and initial approach can be modelled using *game theory* and *financial economics* based on participant interactions and profit opportunities.

Power TAC trading interactions simulate an environment using several players (brokers). The assumption is rational behavior by brokers endeavouring to maximize their individual profit. It is worth pointing out that the rules of Power TAC forbid any kind of *collusion* and that all competition information is publicly available to the agents. Hence, one can use the *Efficient Market Hypothesis (EMH)* to assume that the power brokerage

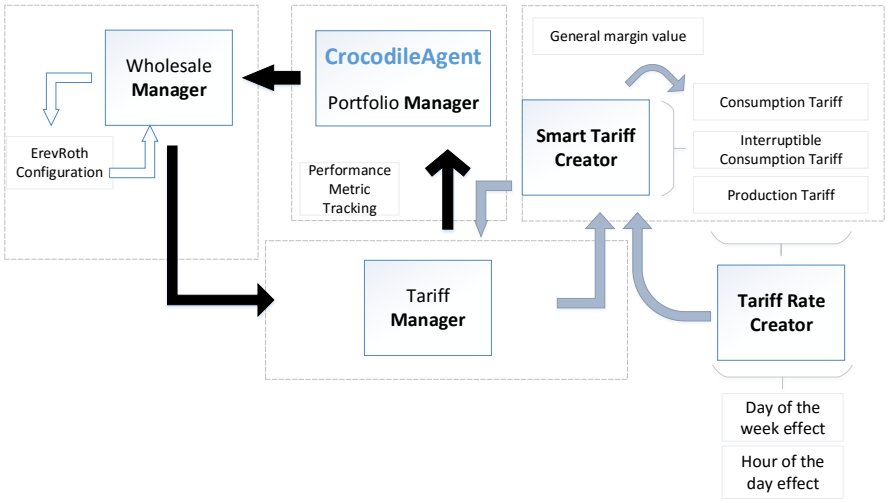


Fig. 3. Interactions among CrocodileAgent modules for wholesale and retail operations

market should be efficient or weakly efficient, creating a stable equilibrium where the profits of individual agents converges towards zero or into negative territory in a competitive environment. The EMH direct implies the impossibility of "earning profit" consistently on a risk-adjusted basis given that market prices in both the retail and wholesale sectors should incorporate all available information [11].

To establish novel trading mechanisms, an extensive post-mortem analysis of CrocodileAgent performance was conducted based on data from the previous world competition (*i.e.*, Power TAC 2017). In general, organizers of the Power TAC world competition record the following data which can be used for elaborate analysis:

- *aggregated per-tournament data* reports the financial balance of each broker and determines the winner of the tournament;
- *aggregated per-game data* reports the financial balance of each broker for a game; and
- *raw per-game data* allows for detailed inspection of a particular game.

For reference, Table 1 summarizes the results of Power TAC 2017. Performance from that year was used to initially target CrocodileAgent behavioral changes.

Table 1. Power TAC 2017 final results (in Mln "monetary units" and normalized)

	Profit by game size (in Mln)				*Z-score by game size*			
Broker	**7 brokers**	**4 brokers**	**2 brokers**	**Total**	**7 brokers**	**4 brokers**	**2 brokers**	**Total**
AgentUDE	-7.37	5.12	66.79	*64.53*	0.54	1.22	1.83	*3.58*
fimtac	-4.23	-7.61	0.07	*-11.77*	0.58	1.10	1.17	*2.85*
SPOT	-12.75	-11.21	0.06	*-23.90*	0.47	1.07	1.17	*2.70*
VidyutVanika	-9.33	-69.92	-36.93	*-116.18*	0.51	0.54	0.80	*1.85*
ewiBroker	-37.44	-32.98	-40.54	*-110.96*	0.14	0.87	0.76	*1.77*
COLDPower17	-25.30	-63.44	-55.28	*-144.02*	0.30	0.60	0.62	*1.52*
maxon17	-41.48	-158.20	-110.94	*-310.62*	0.08	-0.26	0.07	*-0.11*
CrocodileAgent	-241.96	-310.75	-176.95	*-729.65*	-2.61	-1.63	-0.58	*-4.82*

A highly competitive environment is evident due to small profits achieved in Power TAC 2017. The majority of brokers achieved a negative profit and loss (PnL) irrespective to the game size, implying that retail energy brokerage is effectively efficient (see Table 1). After taking a closer look at the reasons for sub-optimal performance in 2017 (*i.e.*, CrocodileAgent was ranked last), the conclusion is that CrocodileAgent 2017 was too aggressive on the retail market which ultimately resulted in high costs on the balancing market. The reasonable solution to this detrimental behaviour is to design a more *defensive approach* in an attempt to preserve capital and minimize risk exposure both on the wholesale as well as retail market. In practical terms, the key mechanism behind the risk aversion strategy for purchasing and selling energy lies in prudent markup pricing for both trading markets.

When it comes to modeling brokerage interaction, the primary goal for CrocodileAgent 2018 was to correctly *price in* the risk exposure from its actively participation in the wholesale and retail markets. Adverse risk exposure consists of direct costs (*e.g.*, electric energy purchasing price) and indirect costs (*e.g.*, inventory imbalance), hence the

main target actions on the market are twofold: (i) *tariff publishing* needs to implement markup for pricing exposure to potential customer actions; and (ii) *wholesale purchasing* needs to be performed at an adequate price point to minimize negative exposure on the balancing market.

Given general strategy of CrocodileAgent, the implemented margin change needs to be sufficient to cover the mentioned costs using the following equation for achieving a return in a specific time slot:

$$r_t = profit_R + profit_W - cost_t \tag{1}$$

where r_t is the net return in time slot t, $profit_R$ is the profit from the retail market, $profit_W$ is profit on the wholesale market and $cost_T$ includes total time slot fees and costs from brokerage operations.

If we assume that profit is achieved directly through brokerage operations, net return can be summarized using a more detailed formula:

$$r_t = quantity_t * price_W * margin_R - cost_t \tag{2}$$

where $margin_R$ is the mean margin across mean energy quantity $quantity_t$ bought at the wholesale price $price_W$ and reduced by the total time slot fees and cost from brokerage operations $cost_t$.

The main hidden cost in Power TAC trading that may impose an *unexpected cost* for the broker are stochastic events resulting from indirect or direct broker actions on the retail market. These costs may occur if a broker fails to purchase an adequate amount of energy. In that case, as a negative consequence, the distribution utility steps in at the specific time slot to provide the required energy. In doing so, the distribution utility also penalizes the broker for failing to maintain the energy balance. Also, the broker may face severe losses in the form of transmission capacity fees for energy transmission when exposed to the grid capacity utilization event. Since the Power TAC scenario considers a broker's contribution to the overall grid load peak by looking into the broker's peak exposure (*i.e.*, broker's contribution to peak demand) and average grid load, each broker is taxed for the maximal grid load proportional to the broker's contribution to peak demand events [15].

The initial analysis of historic Power TAC logs and experimentation with CrocodileAgent has shown that the transmission capacity cost was several orders of magnitude higher than the expected gain. Furthermore, data from experimental simulations and the Power TAC 2017 world competition confirmed two findings. First, it was discovered that the margin for attracting customers in Power TAC is generally low due to competition. Furthermore, as the number of brokers increases, the average tariff margin decreases, and it approaches zero or even a negative value of having a chance to attract customers. Second, simulations have confirmed that penalized costs imposed on Power TAC agents were primarily driven from capacity fees and distribution fees. This conclusion is valid for Power TAC 2018 as well, as is evident in Figure 4 and which presents the cost breakdown for the Power TAC 2018 competition.

The tendency of setting attractive rates for customers is heavily mitigated by the fact that the stochastic nature of the electric peak grid cost provides a heavy bias towards losses in a competitive environment. To mitigate this effect, our approach included in the CrocodileAgent 2018 design was to offer tariffs which amplified margins during hours where the probability of peak energy exposure was the highest.

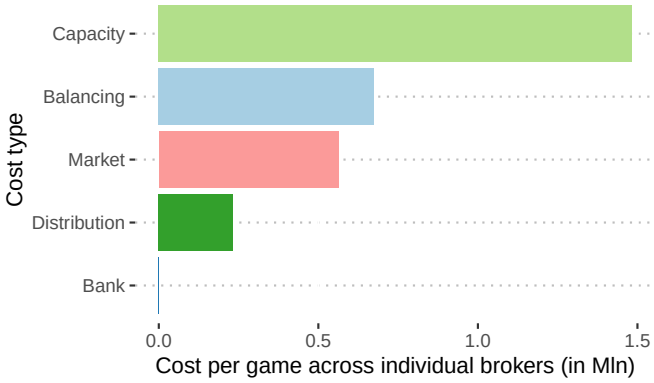


Fig. 4. Breakdown of average broker costs in a Power TAC 2018 game

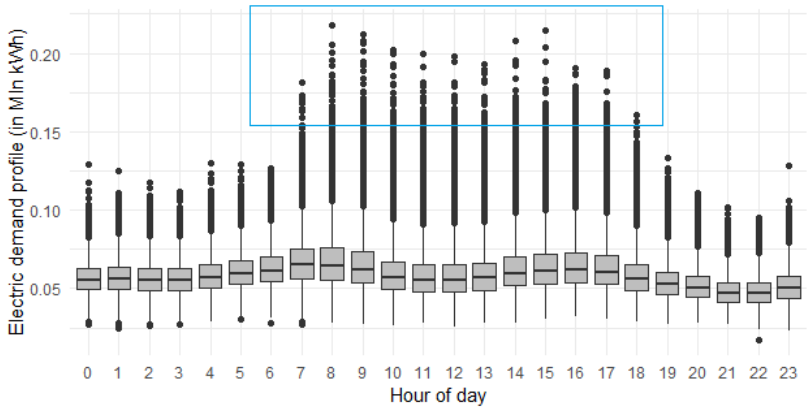


Fig. 5. Box-plot of Power TAC 2018 average in-game electric net demand per hour of day showing that the peak demand occurs usually between the hours 7 am and 6 pm

Experimentation confirmed two major areas that may affect a broker with the highest peak exposure (see Figure 5): (i) the early morning period when people are going to work; and (ii) the afternoon period when people are returning from work. Additionally, sporadic capacity events were identified for the midday peak period (*e.g.*, air-condition period during summer). In CrocodileAgent 2018, we divided the necessary margin hikes (*i.e.*, margin increases) during those periods into internal "medium" and "high" coefficients depending on the peak exposure. By design, this provides a twofold utility for the broker: (i) it hikes the tariff price necessary to cover the expected average exposure to potential losses; and (ii) it discourages consumers from using power during those periods in order to minimize penalization costs for the broker.

Peak exposure is penalized on a weekly basis by pricing the top energy peaks which were above peak threshold (*i.e.*, dynamically calculated value based on mean and standard deviation of the net demand from the previous time slots). Figure 6 shows a load curve

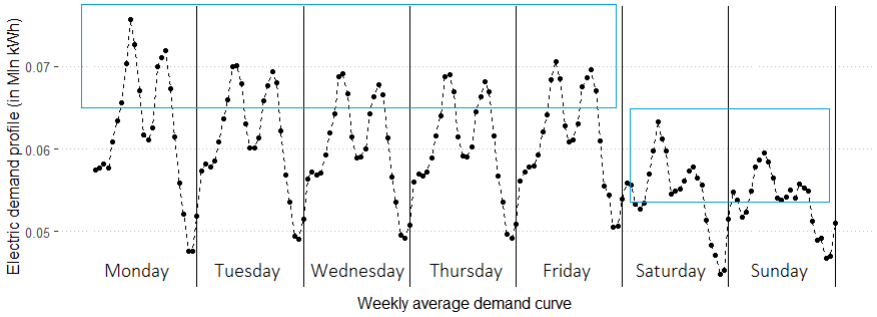


Fig. 6. Characteristic Power TAC 2018 load curve showing the lower peak average energy demand during the weekend (compared to working days)

for a typical week generated by consumers in Power TAC. Since working days usually consumer higher amounts of energy, peak exposure is primarily dictated by a probability distribution heavily favoring Monday to Friday. In contrast, exposure on the weekend follows a comparatively lower energy load and, consequently, the probability of a lower peak cost exposure.

3.2. Retail Mechanisms

Retail mechanisms at the disposal of the Power TAC agents revolve around tariff broadcasting and managing customer portfolios. Obviously, a profitable tariff design, a crucial prerequisite for an agent's competitiveness, is based on evaluation of a customer's tariff expense throughout the subscription lifetime. Power TAC models customers realistically whereby it evaluates all competing tariffs and generally subscribes to the most favorable tariff (one that has the highest profitability or lower cost for the customer). It is worth pointing out that customers are based on a logic choice model augmented with a stochastic (random) parameter for tariff assessment. Nevertheless, the parameter proved to be non-significant in the tariff assessment across the customer portfolio given that the random effect (*i.e.*, client irrationality) during the tariff assessment phase is evenly distributed across competing tariffs. Hence, generally speaking, no broker has an effective advantage over this process [15].

CrocodileAgent 2018 uses three types of tariffs for the retail market: *Consumption Tariff* is a general-purpose consumption tariff for broad retail customers; *Interruptible Consumption Tariff* is a consumption tariff that can curtail the electric load for a desired proportion of the client power usage and shift a clients energy load to other time slots; and *Production Tariff* is a tariff which enables the purchasing of energy on the retail market from producers such as solar homes.

Taking into account the above-mentioned tariffs, an agent's responsibility is to find a profitable solution to the brokerage problem of selling and buying energy. The following mechanisms were used to optimize the broker's performance on the retail market:

– *Time of use effect*: a rate which is directly tied to the hour of the day to which it applies. It provides an opportunity to have higher or lower prices during the day in order to effectively exercise tiered pricing on the consumer tariffs.

- *Weekday / weekend rate effect*: an effect that explicitly governs the intra-week behavior and adjusts the tariff price by tiers.
- *Signup payment value*: a positive payment value for the customer as a bonus for signing up to the tariff.
- *Minimal subscription duration value*: the minimum subscription length duration that governs the early withdrawal penalty trigger.
- *Early withdrawal penalty*: a penalty for withdrawal during the lockup period. This modifies a customer switching options in order to penalize customers from exploiting the sign-up payment opportunity and turning to a more favorable tariff.

In addition to the above-mentioned factors, two additional risk pricing values in tariff creation have been used as a floating component (corrective mechanism for tariff price modification) in defining the rate of the tariff:

- *Margin impact coefficient* (Ci): a coefficient used to update different rates with a desired margin hike to adjust for potential transmission load exposure and risk exposure.
- *Margin coefficient corrections* (Cc): a discounting factor used to attract customers by lowering impact coefficients.

CrocodileAgent 2018 corrects the tariff price according to the following equation:

$$TF = Price * (1 + margin_R) * Ci/Cc \tag{3}$$

where TF is the tariff price structure, $Price$ is an average wholesale price, $margin_R$ is a general margin set by the broker and Ci and Cc are the above-mentioned coefficients calculated separately for a different hour of the day or day of the week. The calculation of Ci and Cc is based on an extensive set of experimental simulations with publicly available brokers from the Power TAC 2017 world competition.

3.3. Wholesale Mechanisms

Even though the focus of CrocodileAgent 2018 is on the retail market, taking into account the wholesale market for procurement of required energy remains is important. In general, the agent estimates the subscription count and energy demand for future time slots and uses bidding strategies to place orders for estimated energy while endeavouring to minimize negative effects from the balancing market.

Implementation of bidding strategies relies on the Erev-Roth reinforced learning method for finding the optimal strategy for minimizing the above-mentioned cost function [4]. The main sub-module consists of (i) *bidding strategies module*, (ii) *reward module* and (iii) *weighted randomizer* module. The optimization cycle follows a pattern where bidding strategies are chosen based on the reward module, while their probability of selection is dictated by a weighted randomizer from the probability distribution.

Since brokers can trade up to 24 hours in advance (*i.e.*, power trading in the Power TAC simulation framework operates as a day-ahead market), the wholesale mechanism is modified to trade up to 24 times for the desired time-slot until the time-slot effectively starts. This enables CrocodileAgent to track progress of numerous wholesale trades as well as to bid for every time slot while continuously updating its reward function. It is worth mentioning that the initialization parameters of CrocodileAgent (*e.g.* learning

parameters for Erev-Roth and margin impact coefficient) are empirically derived and calibrated from an extensive set of experimental simulations publicly available to brokers at the Power TAC 2017 world competition.

4. Evidence from a World Competition: Power Trading Agent Competition 2018

4.1. Competition Setup

As already mentioned, Power TAC is a community-driven project that keeps evolving after each competition cycle on a yearly basis. That said, the organizers of the Power TAC 2018 world competition have introduced several changes in comparison to previous iterations. Apart from many minor updates, the key changes in Power TAC 2018 aim to increase software reliability, simplify interactions within the balancing market as and encourage a higher level of competition [15].

Competition rules specify that each research team needs to prepare a software agent based on the provided Java-based dummy agent, which then acts as a competitive broker within the Power TAC simulation scenario. Furthermore, during the competition, each research team is expected to run two instances of the same agent so that the organizers can schedule multiple independent games which can be played simultaneously in order to reduce time. The main idea behind multiple independent games is to run different scenarios according to the plentiful game initialization parameters, including game size (*i.e.*, number of competing brokers in a particular game), random seed number (*i.e.*, to influence stochastic elements in the game), initialization parameters for customers (*e.g.*, number of consumers), weather (*e.g.*, input data for weather reports), and the wholesale market (*e.g.*, input data for wholesale prices). The main objective of this arrangement is to expose the brokers to various simulations scenarios which in turn has two important implications:

- research teams are discouraged to fine-tune their brokers for a specific default game scenario, instead, they need to come up with *trading mechanisms which can perform in different contexts*; and
- the data produced during the world competition is particularly valuable for *evaluating trading mechanisms as it enhances generalizability of the research results*.

Obviously, one can conclude that Power TAC, when used in a world competition setting, is a reasonably complex distributed system which assumes the active involvement of many stakeholders, including *organizers* who supervise the competition, a *technical crew* for maintaining hardware infrastructure where Power TAC simulation servers are deployed, and *research teams* who need to monitor their brokers in order to secure a reliable connection with remote Power TAC simulation servers. To make sure competition is credible and complies with the rules, the organizers of Power TAC 2018 allowed research teams to test their brokers before the start of competition during two trials which took place in May 2018 and June 2018. After the trials, the qualifying round took place in June 2018. All broker agents that performed to the rules during the qualifying round were allowed to enter the finals which took place from 16 to 27 July 2018. All analyses presented in the next subsections are based on data from the Power TAC 2018 finals.

4.2. Competitors

The entry list in the 2018 competition included the following broker agents from seven international research teams:

- **AgentUDE**, Universitaet Duisburg-Essen / DAWIS (*Germany*);
- **Bunnie**, Nanyang Technological University (*Singapore*);
- **COLDPower18**, INAOE/CICESE-UT3 (*Mexico*);
- **CrocodileAgent**, University of Zagreb (*Croatia*);
- **EWIIS3**, University of Cologne (*Germany*);
- **SPOT**, University of Texas at El Paso (*Texas, USA*); and
- **VidyutVanika**, IIIT Hyderabad (Machine Learning Lab) and TCS (*India*).

The following three different game types were used in the Power TAC 2018 finals: (i) two-broker games where a broker is competing against another broker; (ii) four-broker games where a broker is competing against three other brokers; and (iii) seven-broker games where all of the above-mentioned brokers compete in the same game.

The CrocodileAgent performance in Power TAC 2018 finals was extensively assessed from a total of 324 games with the following breakdown:

- 84 instances of 2-broker games (24 games with CrocodileAgent participating)
- 140 instances of 4-broker games (80 games with CrocodileAgent participating)
- 100 instances of 7-broker games (all 100 with CrocodileAgent participating)

Table 2 shows the final results. It is evident that CrocodileAgent in Power TAC 2018 secured a respectable third place with a positive final PnL in the total result and a positive individual result in each of the three game types.

Table 2. Power TAC 2018 final results (in Mln and normalized)

	Profit by game size (in Mln)				*Z-score by game size*			
Broker	**7 brokers**	**4 brokers**	**2 brokers**	**Total**	**7 brokers**	**4 brokers**	**2 brokers**	**Total**
AgentUDE	49.96	62.14	134.91	*247.01*	1.09	0.63	1.57	*3.29*
VidyutVanika	48.20	101.94	47.54	*197.68*	1.06	1.06	0.34	*2.45*
CrocodileAgent	27.66	45.44	62.88	*135.98*	0.65	0.45	0.55	*1.65*
SPOT	-6.98	32.98	49.18	*75.19*	-0.04	0.32	0.36	*0.64*
COLDPower18	2.06	10.29	0.52	*12.88*	0.14	0.08	-0.33	*-0.11*
Bunnie	-67.98	-25.05	-19.60	*-112.63*	-1.25	-0.30	-0.61	*-2.16*
EWIIS3	-87.27	-206.96	-109.80	*-404.03*	-1.64	-2.25	-1.88	*-5.77*

4.3. Game Size Sensitivity Analysis

The most favorable game type for CrocodileAgent 2018 were the “2-broker games”, where the agent went head to head against other brokers. It is worth pointing out that other game sizes were also ranked favorably by CrocodileAgent 2018 and were generally within the similar competitive advantage area. This leads to the conclusion that the performance of CrocodileAgent 2018 was balanced across all game sizes.

Figure 7 shows that profit distribution in "2-broker games" generally follows a two-mode distribution. Most of the time there was a small positive advantage within games, whereas a larger advantage was found in a smaller number of games. Detailed inspection of Power TAC 2018 data suggests that the large advantage primarily came from games where CrocodileAgent 2018 has played against COLDPower18 and EWIIS3 brokers.

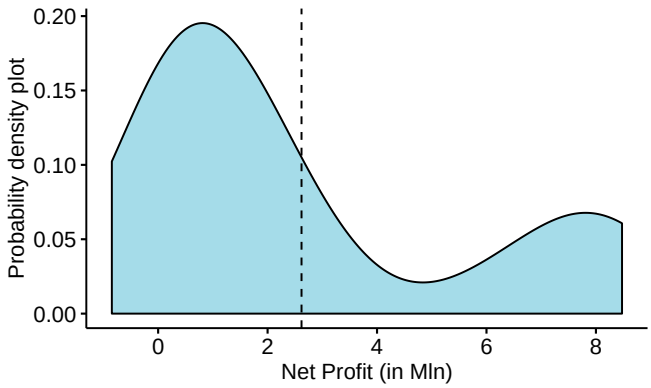


Fig. 7. Profit distribution for CrocodileAgent 2018 in "2-broker games" (with dotted mean of 2.62 Mln "monetary units")

Table 3 shows CrocodileAgent broker's performance relative to other brokers in Power TAC 2018. The relative performance was measured as a difference between final profit results for "2-broker games". Although both brokers could have finished positive, the one that had the highest score had the positive relative difference and vice versa. Evidence suggests that CrocodileAgent 2018 exhibited better performance relative to most of the competing brokers (*i.e.*, COLDPower18, EWIIS3, SPOT [8] and Bunnie).

Table 3. CrocodileAgent 2018 broker's relative performance in Power TAC 2018 (in Mln "monetary units")

Agent Name	*Total difference (aggregate for all games)*	*Average difference (per game)*
COLDPower18	29.62	7.40
EWIIS3	27.11	6.78
SPOT	18.96	4.74
Bunnie	4.17	1.04
VidyutVanika	-6.04	-1.51
AgentUDE	-8.18	-2.05

Figure 8 shows the profit distribution for "4-broker games". What is noticeable is that this distribution is heavily centered around the mean value indicating a constant advantage

over other games. This observation is also supported by the fact that the agent achieved a positive profit in 69 out of 80 four broker games (approx. 86 percent of games played).

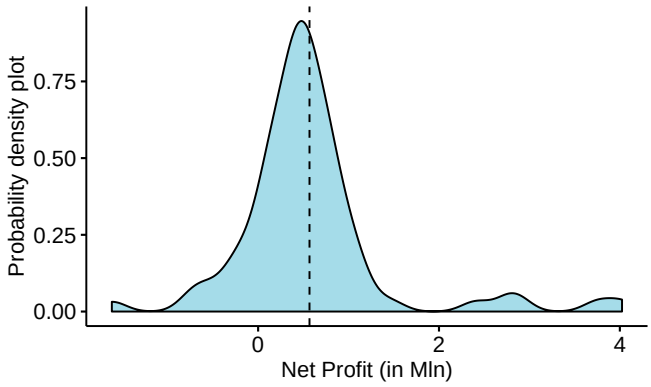


Fig. 8. Profit distribution for CrocodileAgent 2018 in "4-broker games" (with a dotted mean of 0.57 Mln "monetary units")

The "7-broker game" is perhaps the most relevant type of game as it reveals how an agent preforms in a highly competitive and heterogeneous environment with the most varied broker strategies. Figure 9 shows that the profit distribution is positive but significantly lower when compared to "2-player games" and "4-player games". Next, Figure 10 shows the probability for probability-probability (P-P) plot of profits as well as suggests that the profit distribution may follow a theoretical normal distribution. The mean profit as an adequate measure of broker's competitive advantage can be confirmed by testing the distribution to perform like a normal distribution around its average. Therefore, a one-sample Kolmogorov-Smirnov test was used (p-value: 0.3669; alternative hypothesis: two-sided) which confirmed the assumptions given that there was were not enough evidence to reject the null hypothesis.

Even though the broker exhibited small profits, it is worth mentioning that the profitability rate (*i.e.*, the ratio between profitable games and all games) was 95 percent. This observation supports the objective of designing a robust agent that protects against high losses and, at the same time, is able to achieve a positive profit in the majority of the games.

4.4. Benchmarking Against Competitors

In this section we present a benchmark analysis of all competing brokers using several key performance indicators: (i) *subscriber count* to assess the attractiveness of tariffs; (ii) *net profit per subscriber* to quantify the financial performance of tariffs; (iii) *capacity and balancing cost* to analyze the proposed mechanisms in terms of peak exposure; (iv) *total cost per subscriber* to get a sense of the customer portfolio quality; (v) *information ratio* to assess broker's profitability while taking into account profit variability across the games; (vi) *profitability rate* to give an indication as to whether the mechanisms are robust

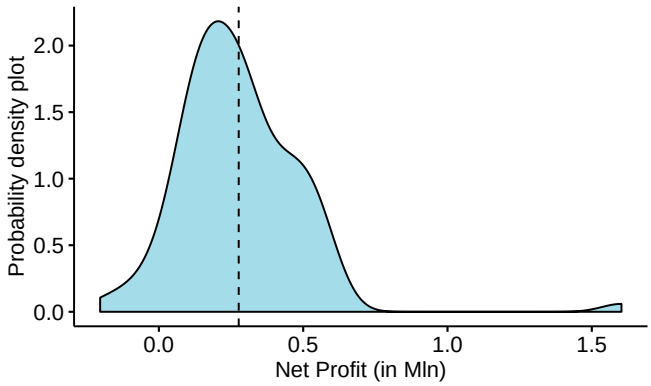


Fig. 9. Distribution of profit for CrocodileAgent 2018 in "7-broker games" (with dotted mean of 0.28 Mln "monetary units")

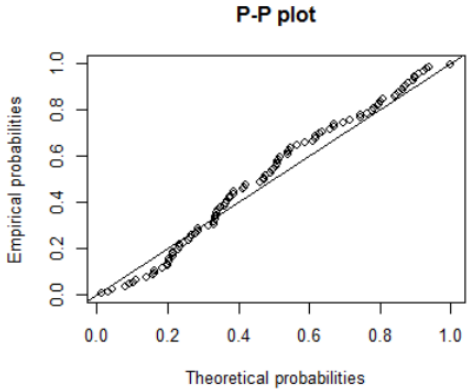


Fig. 10. Probability-probability (P-P) plot of CrocodileAgent performance in 7-broker games vs theoretical normal distribution

with respect to different scenarios; and (vii) *performance bubble chart* to link profitability rate with the final normalized scores.

Subscriber count Figure 11 reveals that CrocodileAgent 2018 had a generally low average subscription count when placed against multiple brokers and indicating it had a pricey tariff structure. The CrocodileAgent 2018 had a relatively higher average subscription amount in "2-broker games" but it was still below the top three brokers according to this KPI. These results validate the tariff mechanisms given that it supports the objective of having a defensive broker on the market.

Net profit per subscriber We calculate the net profit per subscriber as the ratio between the net profit and the average subscriber number per time slot for any given time. To reduce

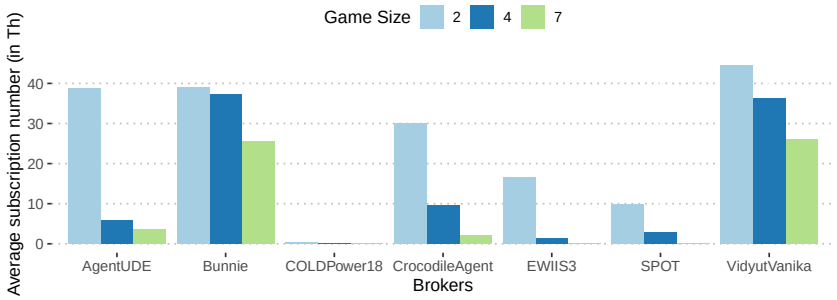


Fig. 11. Power TAC 2018 Finals: average subscription count for individual competing broker per specific game type

the impact of outliers and other situations which are not related to retail tariff subscription (*e.g.,* positive results from a wholesale price differential, huge monetary amounts due to random chance but where the broker had no additional subscribers etc.), we filter out games where the broker had less than 10 subscribers on average. Table 4 shows that CrocodileAgent 2018 had a medium level of profit per subscriber, indicating that it had the ability to extract a "sufficient" income for profitable brokerage operations (statistics for agent COLDPower18 is not shown in the table as it had less than 10 subscribers on average per time slot in each of the games in which it participated).

Table 4. Net profit per subscriber in the Power TAC 2018 Finals competition

Broker	*Net profit per subscriber (in "monetary units")*
SPOT	166
AgentUDE	140
CrocodileAgent	79
VidyutVanika	30
Bunnie	-18
EWIIS3	-767

Capacity and balancing cost Figure 12 shows that CrocodileAgent 2018 was able to reduce the peak exposure and minimize the unwanted negative cost to the level of a profitable brokerage. Balancing costs in comparison to other brokers were on the higher side, indicating a potential for improvement in future work.

Total costs per subscriber Figure 13 shows that only two brokers (*i.e.,* EWIIS3 and SPOT) were effectively heavily penalized when they had subscribers and the rest, including CrocodileAgent 2018, were able to keep cost per subscribers at a relatively lower level.

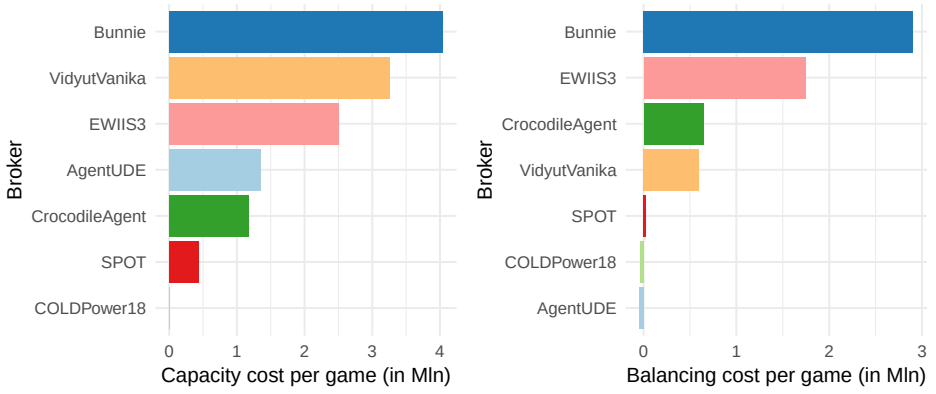


Fig. 12. Power TAC 2018 Finals: average capacity and balancing cost for individual competing brokers (in Mln "monetary units")

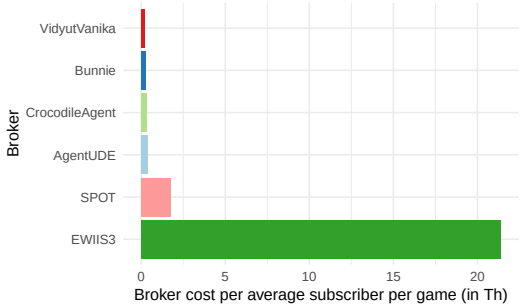


Fig. 13. Power TAC 2018 Finals: total costs per subscriber across all games by competing broker (in Th "monetary units")

Information ratio Broker information ratio is defined as the average net profit divided by the standard deviation of net profit. The significance of this KPI is that it creates a relationship between the net profit of individual brokers and the variability of that net profit. It shows us which brokers perform better per unit of risk (standard deviation). Figure 14 shows that the highest "risk adjusted" returns were realized by VidyutVanika on a 0.79 per game basis. CrocodileAgent 2018 was second with an information ratio value of 0.48 and AgentUDE was third with 0.45.

Profitability rate The broker profitability rate is defined as the ratio between the number of games a broker finished with a positive net amount and the total number of games in which the broker participated. Figure 15 shows that CrocodileAgent 2018 had the highest profitability rate of 91% across all games. This suggests that CrocodileAgent 2018 is a stable agent for power brokerage that would achieve a profit in the majority of market situations.

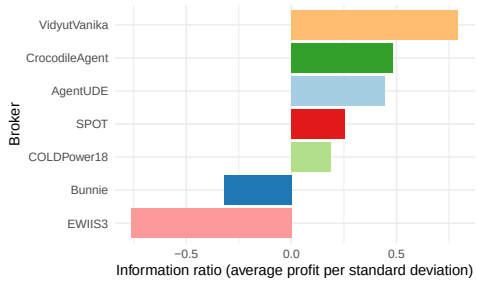


Fig. 14. Power TAC 2018 Finals: competing brokers' information ratio levels

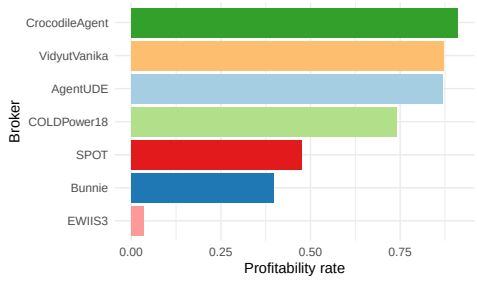


Fig. 15. Power TAC 2018 Finals: profitability rate in all simulated games by competing broker

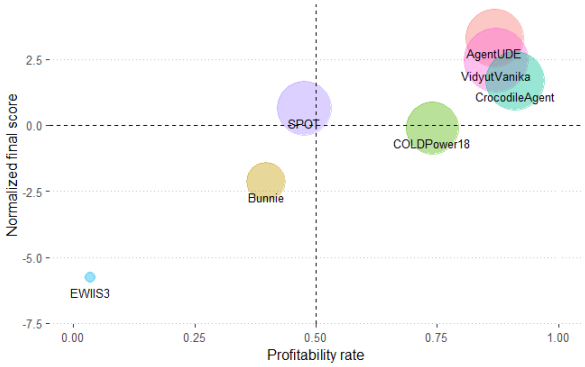


Fig. 16. Power TAC 2018: summarized performance bubble chart

Performance bubble chart The bubble chart in Figure 16 shows broker position based on the profitability rate (x-axis), a normalized final score (y-axis) and the information ratio (bubble size). The chart shows that three brokers: AgentUDE, VidyutVanika and CrocodileAgent performed similarly while EWIIS3 was the biggest outlier in terms of performance with a sizable negative performance. This also correspondents to the final rankings of agents, where AgentUDE, VidyutVanika and CrocodileAgent took podium places while EWIIS3 was ranked last in the finals.

5. Conclusion

Sustainable energy systems of the future will not only include advanced infrastructure but also innovative markets. The Power Trading Agent Competition is a competitive simulation platform which models energy markets. The scope of this paper required designing and implementing the proposed trading mechanisms using a software agent called CrocodileAgent 2018. To perform a proper analysis, the mechanisms were put to test in the Power TAC 2018 world competition against six research teams from Germany, Singapore, Mexico, Texas and India. In particular, a detailed inspection of data from the Power TAC 2018 finals showed that tiered margin weighting was adequate to create a stable and well-rounded robust solution for the broker agent problem of buying and selling electricity al in a competitive environment. Intra-day seasonal components were identified as the primary driver of risk exposure to potential negative costs stemming from peak charges and were consistent on a week-to-week basis. Analysis of customer demand exposure and power grid load were suitable for creating a "profitable structure" for the broker agent. Moreover, it effectively minimized the downside risk of extreme losses and maximized the positive upside in the retail market. The robustness and consistency of CrocodileAgent 2018 was evident through its profitable performance in almost all games throughout the competition. The third place in the Power TAC 2018 world competition, along with the highest profitability rate of 91%, suggests that the broker agent presented in this paper is a profitable framework to build-upon for future competitions.

In future work which will be aimed at designing and implementing CrocodileAgent 2019, the focus will be on improving the retail trading mechanisms, such as lowering average balancing cost, and designing new solutions to tackle complexities of the wholesale trading aspect in the Power TAC environment. Therefore, apart from using the well-established offline learning mechanisms, future work will seek to design and implement new mechanisms which are based on online learning mechanisms.

References

1. Babic, J., Carvalho, A., Ketter, W., Podobnik, V.: Evaluating Policies for Parking Lots Handling Electric Vehicles. IEEE Access 6(99), 944 – 961 (2018)
2. Babic, J., Podobnik, V.: A review of agent-based modelling of electricity markets in future energy eco-systems. In: 2016 International Multidisciplinary Conference on Computer and Energy Science (SpliTech). pp. 1–9. Split, Croatia (2016)
3. Babic, J., Matetic, S., Matijas, M., Buljevic, I., Pranjic, I., Mijic, M., Augustinovic, M.: The CrocodileAgent 2012: Research for Efficient Agent-based Electricity Trading Mechanisms. In: Spec. Sess. Trading Agent Compet. KES-AMSTA 2012. Dubrovnik, Croatia (2012)
4. Babic, J., Podobnik, V.: Adaptive Bidding for Electricity Wholesale Markets in a Smart Grid. In: Ceppi, S. ; David, E..R.V.S.O..V.I. (ed.) Proc. Work. Agent-Mediated Electron. Commer. Trading Agent Des. Anal. (AMEC/TADA 2014) @ AAMAS 2014. pp. 1–14. Paris, France (2014)
5. Babic, J., Podobnik, V.: An Analysis of Power Trading Agent Competition 2014. In: Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets, pp. 1–15. Springer, Cham (2014)
6. Badica, C., Budimac, Z., Burkhard, H.D., Ivanovic, M.: Software agents: Languages, tools, platforms. Computer Science and Information Systems 8(2), 255–298 (2011)

7. Bichler, M., Gupta, A., Ketter, W.: Designing Smart Markets. Information Systems Research 21(4), 688–699 (Dec 2010)
8. Chowdhury, M.M.P., Folk, R.Y., Fioretto, F., Kiekintveld, C., Yeoh, W.: Investigation of Learning Strategies for the SPOT Broker in Power TAC. In: Ceppi, S., David, E., Hajaj, C., Robu, V., Vetsikas, I.A. (eds.) Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets. pp. 96–111. Lecture Notes in Business Information Processing, Springer International Publishing (2017)
9. Chu, S., Majumdar, A.: Opportunities and challenges for a sustainable energy future. Nature 488(7411), 294–303 (Aug 2012)
10. Davis, J.P., Eisenhardt, K.M., Bingham, C.B.: Developing theory through simulation methods. Academy of Management Review 32(2), 480–499 (2007)
11. Fama, E.F.: Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25, 1–25 (Dec 1969)
12. Gungor, V.C., Sahin, D., Kocak, T., Ergut, S., Buccella, C., Cecati, C., Hancke, G.P.: Smart Grid Technologies: Communication Technologies and Standards. IEEE Transactions on Industrial Informatics 7(4), 529–539 (Nov 2011)
13. Ketter, W., Peters, M., Collins, J., Gupta, A.: A Mutual Competitive Gaming Platform to Address Societal Challenges. Management Information Systems Quarterly 40(2), 447–460 (Jun 2016)
14. Ketter, W., Collins, J., Reddy, P.: Power TAC: A competitive economic simulation of the smart grid. Energy Economics 39, 262–270 (2013)
15. Ketter, W., Collins, J., Weerdt, M.d.: The 2018 Power Trading Agent Competition. SSRN Scholarly Paper ID 3087096, Social Science Research Network, Rochester, NY (Dec 2017), https://papers.ssrn.com/abstract=3087096
16. Ketter, W., Peters, M., Collins, J., Gupta, A.: Competitive Benchmarking: An IS Research Approach to Address Wicked Problems with Big Data and Analytics. Management Information Systems Quarterly 40(4), 1057–1080 (Dec 2016)
17. Kuate, R.T., He, M., Chli, M., Wang, H.H.: An Intelligent Broker Agent for Energy Trading: An MDP Approach. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. pp. 234–240. IJCAI '13, AAAI Press, Beijing, China (2013)
18. Liefers, B., Hoogland, J., La Poutré, H.: A successful broker agent for power TAC. In: Agent-Mediated Electronic Commerce. Designing Trading Strategies and Mechanisms for Electronic Markets, pp. 99–113. Springer (2014)
19. Matetic, S., Babic, J., Matijas, M., Petric, A., Podobnik, V.: The CrocodileAgent 2012: Negotiating Agreements in a Smart Grid Tariff Market. In: Agreement Technologies (AT 2012), Proceedings of the 1st International Conference on (2012)
20. Özdemir, S., Unland, R.: Trading strategies of a champion agent in a multiagent smart grid simulation platform. In: Multiagent System Technologies - 13th German Conference, MATES 2015, Cottbus, Germany, September 28-30, 2015, Revised Selected Papers. pp. 218–232 (2015), https://doi.org/10.1007/978-3-319-27343-3_12
21. Özdemir, S., Unland, R.: Wholesale Bidding Approaches of an Autonomous Trading Agent in Electricity Markets. In: Derksen, C., Weber, C. (eds.) 3rd and 4th International Conference on Smart Energy Research (SmartER Europe 2016 and 2017). Smart Energy Research. At the Crossroads of Engineering, Economics, and Computer Science, vol. AICT-495, pp. 90–104. Springer International Publishing, Essen, Germany (Feb 2016), part 1: SmartER Europe 2016
22. Özdemir, S., Unland, R.: The strategy and architecture of a winner broker in a renowned agent-based smart grid competition. In: Web Intelligence. vol. 15, pp. 165–183. IOS Press (2017)
23. Özdemir, S., Unland, R.: AgentUDE17: A genetic algorithm to optimize the parameters of an electricity tariff in a smart grid environment. In: International Conference on Practical Applications of Agents and Multi-Agent Systems. pp. 224–236. Springer (2018)

24. Özdemir, S., Unland, R.: AgentUDE17: Imbalance Management of a Retailer Agent to Exploit Balancing Market Incentives in a Smart Grid Ecosystem. pp. 1911–1922. Luneburg, Germany (Mar 2018)
25. Perisic, M.M., Storga, M., Podobnik, V.: Agent-Based Modelling and Simulation of Product Development Teams. Tehnicki vjesnik - Technical Gazette 25(Supplement 2), 524–532 (Sep 2018), https://hrcak.srce.hr/205953
26. Rubio, T.R.P.M., Queiroz, J., Cardoso, H.L., Rocha, A.P., Oliveira, E.: TugaTAC Broker: A Fuzzy Logic Adaptive Reasoning Agent for Energy Trading. In: Rovatsos, M., Vouros, G., Julian, V. (eds.) Multi-Agent Systems and Agreement Technologies. pp. 188–202. Lecture Notes in Computer Science, Springer International Publishing (2016)
27. Tschopp, D.J.: The sustainability advantage: seven business case benefits of a triple bottom line by Bob Willard, 2002. New Society, 203 pp (pbk). ISBN 0-86571-451-7. Business Strategy and the Environment 12(6), 405–405 (Nov 2003)
28. United Nations World Commission on Environment and Development: Our Common Future. Oxford University Press, Oxford ; New York, 1 edition edn. (May 1987)
29. Urieli, D., Stone, P.: TacTex'13: a champion adaptive power trading agent. In: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. pp. 1447–1448. International Foundation for Autonomous Agents and Multiagent Systems (2014)
30. Wang, X., Zhang, M., Ren, F.: A hybrid-learning based broker model for strategic power trading in smart grid markets. Knowledge-Based Systems 119, 142–151 (Mar 2017)
31. Weare, C.: The California Electricity Crisis: Causes and Policy Options. Public Policy Institute of CA (2003)

Demijan Grgic is a research assistant at the University of Zagreb, Faculty of Electrical Engineering and Computing (FER). He received his M.Sc. degree in information and communication technology and M.Econ degree in Economics in 2011 and 2013, respectively. He is currently studying for a doctoral degree at the FER. His current research interests relate to data science applications within the energy domain. He is a member of the Social Networking and Computing Laboratory.

Hrvoje Vdovic is a research assistant at the University of Zagreb, Faculty of Electrical Engineering and Computing (FER). He received his M.Sc. degree in information and communication technology in 2017 and is currently studying for a doctoral degree. His current area of research are energy informatics, electric vehicles and automotive software. He is a member of the Social Networking and Computing Laboratory.

Jurica Babic received the M.Sc. degree in information and communication technology and the Ph.D. degree in computer science from the University of Zagreb, Faculty of Electrical Engineering and Computing (FER), in 2012 and 2018, respectively. He is currently a postdoctoral researcher within the same institution. His teaching and research activities are in a field of energy informatics, which aims to foster Information System (IS) skills and IS thinking to create sustainable systems of the future. He is the energy informatics team leader within the Social Networking and Computing Laboratory and a member of the Laboratory for Assistive Technologies and Alternative and Augmentative Communication.

Vedran Podobnik received the M.Eng. degree in electrical engineering and the Ph.D. degree in computer science from the Faculty of Electrical Engineering and Computing (FER), University of Zagreb, Zagreb, Croatia, in 2006 and 2010, respectively, and

the M.Phil. degree in technology policy from the Judge Business School, University of Cambridge, Cambridge, U.K., in 2013. Since 2006, he has been with the Department of Telecommunications, FER, where he has been an Associate Professor since 2016. He is the Founder and Director of the Social Networking and Computing Laboratory and Co-Coordinator of the FERs student startup incubator SPOCK. His teaching and research activities are in transdisciplinary fields of network and data science, social computing, and technology policy. He has co-authored over 80 papers, including publications in Information Sciences, Information Technology and People, AI Magazine and Cybernetics, and Systems journals.

Received: October 10, 2018; Accepted: November 11, 2018.

Computer Science and Information Systems 16(1):131–154 https://doi.org/10.2298/CSIS171212011S

Estimating point-of-interest rating based on visitors geospatial behaviour

Matej Senožetnik[1,2], Luka Bradeško[1,2], Tine Šubic[1], Zala Herga[1,2], Jasna Urbančič[1], Primož Škraba[1], Dunja Mladenić[1,2]

[1] Institut Jožef Stefan,
Jamova 39, 1000 Ljubljana
E-mail: name.surname@ijs.si
[2] Jožef Stefan International Postgraduate School,
Jamova cesta 39, 1000 Ljubljana

Abstract. Rating of different services, products and experiences plays an important role in our digitally assisted day-to-day life. It helps us make decisions when we are indecisive, uninformed or inexperienced. Traditionally, ratings depend on the willingness of existing customers to provide them. This often leads to biased (due to the insufficient number of votes) or nonexistent ratings. This was the motivation for our research, which aims to provide automatic star rating. The paper presents an approach to extracting points-of-interest from various sources and a novel approach to estimating point-of-interest ratings, based on geospatial data of their visitors. Our research is applied to campsite dataset where the community is still developing and more than thirty percent of camps are unrated. Our study use case addresses a real-word problem of motorhome users visiting campsites in European countries. The dataset includes GPS traces from 10 motorhomes that were collected over a period of 2 years. To estimate star ratings of points-of-interest we applied machine learning methods including support vector machine, linear regression, random forest and decision trees. Our experimental results show that the duration of visit, which is a crucial part of the proposed approach, is an indicative feature for predicting camp ratings.

Keywords: machine learning, geospatial analysis.

1. Introduction

Crowd-sourcing services and product ratings are a great way to allow the community to comment on their experiences with various services or products. As such, it is not surprising that integrated rating and review systems have become extremely popular, leading to the rise of sites (like GoodGuide and Yelp) dedicated to specific products and services. Recommending hotels and restaurants based on text reviews is a commonly addressed research topic [11,15,16,19]. In this research we focus on addressing a problem of estimating campsites rating based on their available facilities and data about visitors geospatial behavior. In particular, we focus on *motorhomes* which are visiting campsites and other point-of-interests. The data have been collected over more than 2 years from 10 motorhomes traveling across Europe.

Motorhome community is still in the process of development. Motorhome companies have been seeking ways for the motorhomes to be able to dynamically assist the user in

discovering new destinations and advise them by taking into account different aspects, such as availability of electricity, fresh water and other resources. With the popularity of motorhomes on the rise (registration of motorhomes has increased by 17.9% in 2016 in comparison to 2015 [9]). We applied our research to their users' geospatial patterns and analyzed their common accommodations - camps, service areas, etc.

The main contribution of this article is a new approach to automatically estimate (predict) point-of-interest ratings based on user behavior and camp characteristics. The next contribution is evaluation of the proposed approach in real-world data. For predicting camp rating we compared several machine learning algorithms including linear regression, k-NN, random forest and SVM. To be able to perform the analysis, we have collected two years of location data for motorhomes that traveled around Europe. The third contribution of our research is building the definitive database of points-of-interest for campers and along with the analysis providing it as a service to the motorhome community.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related work. In Section 3 we provide a detailed description of architecture and interaction between components. The data description is presented in Section 4. Section 5 presents the approach used to check the hypothesis. Finally, we conclude the paper and describe some of the future work directions in Section 6.

2. Related work

To the best of our knowledge, the problem of predicting camp rating based on location history was not addressed before. However, similar works on different datasets and using different methods exist and had been tried before. This section provides a brief overview of that related work. Understanding and predicting what users really want and what suits their needs has become an important and very popular research topic. A lot of work has already been done on predicting ratings for hotels, restaurants, etc. The most popular datasets for research and building personalization applications are provided by Yelp and HRS.com (major European platform for hotel reservations). If users rate places highly, then algorithm would recommend places similar to those that the user gave high ratings. This can go beyond overall ratings, by taking into account advantages of multi-criteria ratings and improving recommendation accuracy compared to single-rating recommendation approaches [10,20]. We could not perform such kind of analysis on our data, because we do not have information on specific users rating of the location. However, our system is designed to work with the data like this, once available.

Other useful information could be extracted from different reviews of the same place. The most common approach is to use sentiment analysis to evaluate hotel ratings [12,20]. There have also been studies dealing with the correlation between hotel star-rating and review rating [18].

Recommendation systems collect information based on the preference of users on different items, such as movies, jokes, books, hotels, restaurants, camps. Some recommendation systems went further and additionally take into account the location as an important component of the user context [22,25,26]. As opposed to our approach, the usual recommendation systems rely on the availability of large amounts of data.

Tiwari and Kaushik [27] propose a rating method for points-of-interest based on fuzzy logic which can take a location stay time, user categorization (native, regional and tourist)

and weighted frequency (number of times when the visitor enters a place and stays there for more than a defined time threshold) as inputs. They also propose to categorize users into five categories (global, national, regional, local and native) [28] reporting that user categorization has the impact on stay time. This method returns a rating between 0 and 5. The main issue, as explained by the authors, is that only 0.5% of places from their dataset of 26,807 places have any ratings. Secondly, it is hard to compare ratings between places. If places are internationally known, they have the significantly higher volume of ratings than small places which are more regionally/locally known. Another potential problem is that ratings could be biased by the deliberate promotion of a location as a marketing strategy. However, to measure any impact of a rating method on the users, we would have to include the users in the evaluation.

3. Proposed Approach and Framework Description

Figure 1 shows an overview of the proposed architecture that consists of four main components: data capture, Mobility patterns, analysis and API. The *Data capture* component primarily captures GPS locations from motorhomes, but it also takes care of building point-of-interests (POI) data (*Building POI*). *Building POI* prepares OpenStreetMaps (OSM) files for camp data extraction and collects information about camps from publicly available online resources. Main two sources of the data for our approach are real-time GPS coordinates stream in WGS84 format coming from connected motorhomes, and a POI database which is matched to the stream, to find visited POIs for each user (motorhome). This process consists of cleaning the data and transforming it into manageable form for further analysis. The *Mobility patterns* is a service for processing raw GPS data and performing clustering, global location detection and next place prediction. *Analysis* component aims to determine the correlation of ratings to time spent on location, predicts star rating and analyze number visits to camps.

The main function of the *API* service is to enable access to analyzed data to external parties; in case of camp data to share it with the motorhome community and Optimum mobile application. Optimum is an EU project which aims to improve transit, freight transportation and traffic connectivity throughout Europe [5]. Within the project one of case studies is based on motorhome and their users [6].

3.1. *Mobility patterns* component

Mobility patterns service is capable of accepting and parsing continuous GPS data (Definitions 1 and 2) from various sources, like mobile phones and other devices. It also collects vehicle GPS data from motorhomes. In addition to collecting the data it is an analytics service which classifies raw GPS coordinates into the appropriate paths and staypoints [29]. Staypoints (Definition 3) are classified into global locations (Definition 4) and paths into global routes. Based on the historical data, the service can also predict next user location and detect anomalies in paths. Here we only use the information about visits (staypoints) and resulting global locations classified as camps. We provide definitions related to staypoints in a similar way as in [24,30] .

Definition 1 (GPS point) *GPS point p is a pair of longitude (lng) and latitude (lat) values which can be formally defined as $p = (lat, lng)$.*

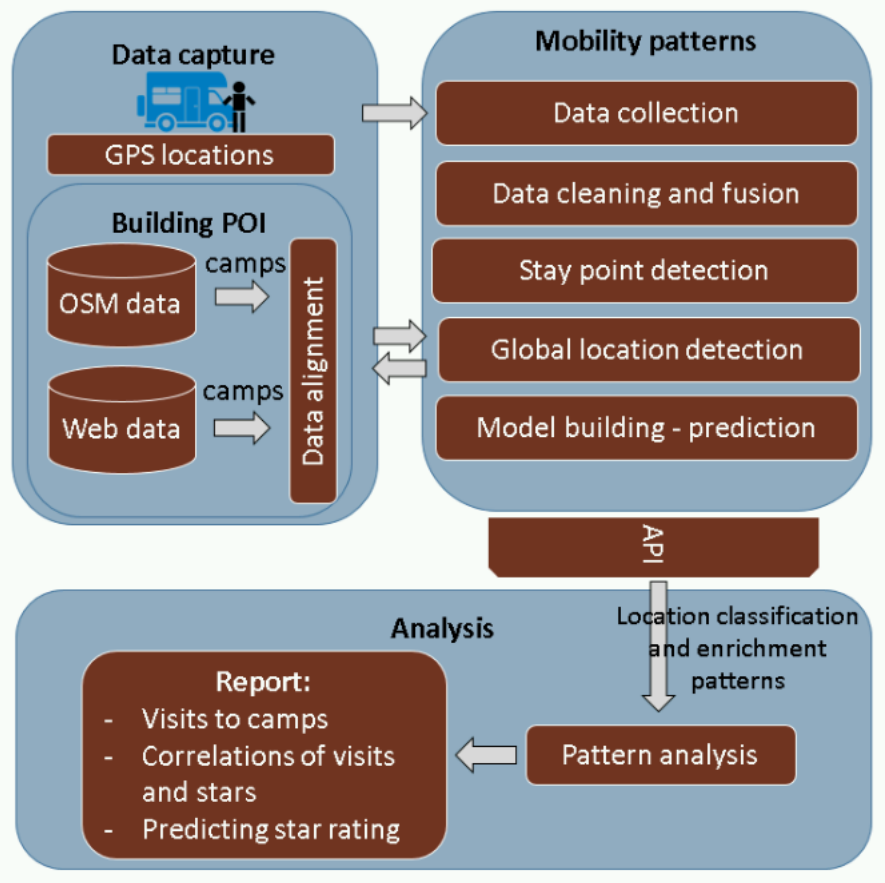


Fig. 1. The architecture of our approach to estimating (predict) points-of-interest rating based on visitors behavior.

Definition 2 (GPS trajectory) *GPS trajectory is a spatio-temporal sequence which can be formally defined as*

$$Traj = \{(p_1, t_1), (p_2, t_2), \ldots, (p_n, t_n), \}$$

where $p_i, i = 1, \ldots, n$ *are GPS points and* t_i, $i = 1, \ldots, n$ *monotonically increasing time values.*

Definition 3 (staypoint, path, activity) ***Staypoint*** *S is a limited geographic area which is shown in Figure 2 at which the trajectory stayed for a certain time. More formally, a staypoint S is a set of pairs* $S = \{(p_i, t_i)\}_{n \leq i \leq m}$, *where*

1. $t_m - t_n > T$, *for a threshold duration* T
2. $Distance(p_i, p_j) < D$, *for each* $i = n, \ldots, m$, $j = n, \ldots m$, $i \neq j$, *for a threshold distance* D.

Point	Coordinates	Time
p_1	(lat_1, lng_1)	t_1
p_2	(lat_2, lng_2)	t_2
	.....	
p_n	(lat_n, lng_n)	t_n

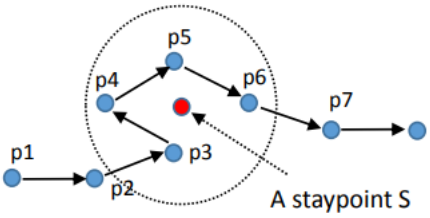


Fig. 2. GPS trajectory and staypoint

3. $S_i \neq S_j$ *for* $p_i = p_j$ *and* $t_i \neq t_j$ *(staypoint is time dependent).*

Staypoint is a state of a trajectory. A section between two consecutive staypoints defines a ***path****. Path is a sequence of pairs* $P = \{(p_i, t_i)\}_{n \leq i \leq m}$*, such that* $(p_{n-1}, t_{n-1}) \in S_j$ *and* $(p_{m+1}, t_{m+1}) \in S_{j+1}$*. A common name for paths and staypoints is* ***activity****.*

Several algorithms have been used for staypoint detection [31,32,23]. We use our own modification which was originally described in [14] and is uses a two-stage approach, first clustering GPS coordinates by employing a staypoint detection (SPD) algorithm grouping together GPS coordinates which are located inside a parameter given by threshold distance D and visited within the time period T. In our implementation we utilized thresholds of 50 meters for radius (distance) and 5 minutes for the duration. To handle potential errors in the GPS coordinate reading after performing the clustering we eliminate locations that are detected as outliers (e.g., a big jump in the distance back and forth over a short time) and merging staypoints if appropriate. The output of our 2-pass SPD algorithm is a list of cleaned activities either staypoints or paths.

Our users (motorhomes) $\{u_1, \ldots u_n\}$ collect GPS points and produce unique trajectories which are clustered into staypoints. Some users' staypoints overlap in space. We are interested in grouping those staypoints that correspond to the same geographical location together into a common *global location*.

Definition 4 ***Global location*** *is a circular geographic area defined by location center (point) and a radius. It is not time-dependent and it holds a semantic meaning.*

Each staypoint is assigned to exactly one global location. For each global location we store the center of the location (GPS coordinate pair) and based on included staypoints we calculate of a visit statistics, time and date of the first visit, the last visit and the average duration of visit.

3.2. Data capture component

OpenStreetmaps OpenStreetMap (OSM)[13] provides a free geographic database of the world. While this started with mapping streets, it has already gone far beyond that, and now includes footpaths, buildings, camps, parkings, and many other geographic properties. For downloading OSM data we are using a service called Geofabrik [3]. Geofabrik

provides pre-built shape files, maps and map tiles, allowing us to bypass some of the manual extraction and preprocessing work. The data is organized by regions that are regularly updated and saved in PBF format.

We focus on the camps in the European Union, however the approach can use any geographical region worldwide. The data is stored as a collection of SQL tables for points, polygons, roads and lines which enables easy querying and data processing (using Osm2pgsql [7] to convert OSM data to PostGIS-enabled PostgreSQL databases). We were mainly interested in point and polygon tables which contain camp data that can be retrieved by the following query:

```
SELECT * FROM table_name WHERE ((tags->'tourism')='camp_site'
OR (tags->'tourism') = 'caravan_site');
```

In OSM, camps are defined by the $camp\_site$ tag, while $caravan\_site$ is generally reserved for short-term parking of motorhomes. Besides spatial point or polygon data, OSM provides other semantic information such as location name, fee, opening hours, phone, website, address and others [4]. However,the presence of this data is very inconsistent and varies from record to record.

Additional data We used data provided by different online forums [1,2] and we composed two datasets. One consists of regional datasets (consisting of Slovenia and its neighbor countries) and the other is a general European dataset.

European dataset consists of over 20,000 records of motorhome stops, separated into following categories: motorhome parking, service areas and motorhome-friendly campsites. Service areas are places, where motorhomes can discharge wastewater and chemical toilet and refill their fresh water supply. Motorhome friendly campsites need to have the possibility to discharge wastewater and chemical toilet and to supply fresh water, as well as allowing overnight stays. Motorhome parkings are areas appropriate for motorhomes and are further divided into multiple subcategories such as mixed parking, private parking near a restaurant, near a marina, motorhomes only, etc...

The Regional dataset is focused on Slovenia, Croatia, Bosnia and Herzegovina, Montenegro and Serbia. Every camp record contains three categories of detailed information:

- Camp is suitable for: families with children, dogs, surfing, rowing, bicycling, etc.
- General info about camp such as open hours, good shaded pitches, location near the sea, outdoor swimming pool, camp at the spa, tents for rent, caravans for rent, mobile homes for rent, bungalows, children playground, animation for children, sanitary facilities for disabled people, baby showers, service station for campers, shop, restaurant, paid wireless internet access, washing machine, mooring for boats, etc.
- Sports and other activities available, such as tennis, volleyball, rent a bike, rent a boat, diving center, table tennis, basketball, football, mini golf, rowing, surfing, climbing wall, etc.

The resulting datasets are merged with OSM data which we consider the main data source. We tried to match coordinates with corresponding position or polygon in OSM. Merging function, shown in Algorithm 1 iterates through all of the additional camps and attempts to find the appropriate camp within a certain radius (delta) inside OSM. If it finds more than one suitable OSM camp, we update the records for all polygons in the vicinity.

Algorithm 1: Algorithm for merging OSM and web data.

```
1 function Merge (webCamps,Δ);
  Input : All camps which are obtained through from web forums. Δ is distance in meters
2 for camp in webCamps do
3 |   osmCamps ← findCampInTheDatabase
  |     (camp.latitude,camp.longitude,Δ);
4 |   if len (osmCamps) ≥ 1 then
5 |   |   updateAll (osmCamps);
6 |   else
7 |   |   insertNewCamps (osmCamps);
8 |   end
9 end
```

If, however, no camp is found in a range we conclude that OSM data does not contain it yet and we add a new camp record.

Cleaning data In this step we remove possible duplicates which occur due to combining OSM with other data sources. Some of the camps cover a very large area and may even cross national borders, which causes them to exist in more than one country file and are then present as multiple polygons when combining country data. Because of this, we need to merge these duplicates into single record using PostGIS. Due to that, we need to merge these polygons into a single record. Additionally, we detected that some camp records were not completely consistent, appearing as either points or polygons. Such conflicts must be resolved manually. For this we generate a rule file by hand, where we define which polygons or points will be deleted and which polygons or points are to be merged into one record.

Data alignment This step is necessary in order to improve our analysis. In some cases multiple *Mobility Patterns* global points were identified on one OSM camp polygon. We solved this issue by merging global points into one location. Hence, our aim is to find appropriate point matching between camps and *Mobility patterns* global location. Our assumption is that if a motorhome user is at (or close to) a global location then we treat this as a camping place. Our method goes through all global locations and in each iteration tries to find the containing polygon or point within 100 or 120 meters of distance.

3.3. Data analysis methods

In this section we describe methods used in our analysis. In data processing stage using *Mobility patterns*, we calculated a number of visits to each camp and histograms about arrival and departure times of motorhomes. With *regional* and *European* datasets, we applied Pearson correlation analysis in order to understand correlations between the features (including target value). Then we used machine learning methods such as linear regression, random forest, decision tree and SVM on *European* dataset in order to predict the star rating. We also fitted a model with *the regional* dataset, but the outcome was not

promising. Due to mismatching attributes in the datasets, we were unable to perform the merging into a single unified dataset, since that would cause either a certain amount of data loss or empty values in some records.

Visits to the campsite is a simple analysis which counts the number of visits to a specific campsite, provides time and date of first and last visit, and its duration. The practical example is shown in Section 5.2. This analysis is provided by *Mobility patterns*, where each campsite is represented by one global location.

Pearson correlation analysis In order to understand the relationship between attributes we used Pearson correlation coefficient, which is calculated as follows:

$$r = \frac{n \sum xy - (\sum x \sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \tag{1}$$

where n represents the number of samples. Pearson correlation depicts a linear relationship between two variables. We checked the correlation of all possible feature pairs (variables x and y), including target values. From these correlations, we built a heatmap which is shown in Figure 7. In general heatmap for correlations has values between -1 and 1. In our case the more intense is color the higher is correlation among attributes. The pros are if we have a small number of attributes it can be easily interpretable and we gain more insight into the data. When we have many attributes it is hard to show everything and we may have an issue of interpreting the heatmap.

Machine learning methods for prediction To predict star rating we used the following machine learning methods implemented in *Scikit learn* [8]:

- **Decision tree** is an easy to understand model used for classification and regression, with multiple available decisions and outcomes included in the tree. To avoid data over-fitting we tune the depth parameter. We used the algorithm for classification and regression.
- **k-Nearest Neighbors**(k-NN) algorithm for classification and regression is a simple method of making predictions by searching through the entire training set for the *k* most similar instances and summarizing the output variable for those *k* instances. We used the algorithm for classification and regression purposes.
- **SVM** using linear, polynomial or radial basis kernel. In our experiments we used radial basis kernel. In general it performs well with non-linear boundary depending on the kernel, and handles high dimensional data. It is used for classification and regression.
- **Random forest** is an alternative to decision trees and reduces variance by generating a whole collection of trees. However, the results are not easily interpretable. It is used for classification and regression.
- **Linear regression** is a standard method commonly used when dealing with numeric features. It is used for regression.
- **Logistic regression** is simple model for classification rather than regression.
- **Baseline** We calculated a mean of target values for regression and take majority class for classification.

Feature selection We used the method called stepwise forward selection, which finds the subset of features that produce the best results. The algorithm first processes every feature and evaluates the result to find the feature with the best performance. On every iteration it adds to the subset a feature that produces best results in combination with previously added features. Each step minimizes RMSE error. This process is repeated until we either run out of feature or the addition of new features does not improve the result anymore.

Performance evaluation For the evaluation we used 10-fold cross-validation. To asses the errors and performance, we used the following measures:

- **Accuracy** is the fraction of correct predictions over n calculated as

$$accuracy(y, \widehat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{1}(\hat{y}_i = y_i) \tag{2}$$

  where $\hat{y}_i$ is the predicted value of the i^{th} sample, y_i is the corresponding true value and n is the number of samples.
- **F1** is a weighted average of the precision and recall which is calculated as

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{3}$$

  Precision and recall are defined as $precision = tp/(tp + fp)$ and $recall = tp/(tp + fn)$, respectively, where tp are true positives, fp represents false positives, and fn false negatives. F1 score, precision and recall all take values between 0 and 1, where 0 is the worst and 1 is the best value.
- **Root mean square error** is calculated as

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2} \tag{4}$$

  which is the square root of the average of squared differences between prediction $\hat{y}_i$ and actual observation y_i. Values can range from 0 to infinity and it is a negatively-oriented score, which means lower values are better. Also, this metrics penalizes large errors more. This is an appreciated property for us since we want to avoid large error, whereas small errors are acceptable because the target value could be biased.
- **Coefficient determination** is calculated by equation

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1} (y_i - \widehat{y_i})^2}{\sum_{i=0}^{n-1} (y_i - \overline{y_i})^2} \tag{5}$$

  and reflects the proportion of the variance in the target variable that is predictable from the explanatory variables. Its values range between -1 and 1. The higher the value the better the result.

3.4. API

To provide the access to collected campsite data to outside users and easier utilization in our internal applications and algorithms, we use a REST API backed by PostgreSQL database, which contains records of campsites and allows querying and filtering by different properties - campsites near a certain point, campsites in a specific country, campsites with specific facilities, etc. Some of this data is already being used by an application in development, intended for multi-modal route-planning application (developed in the scope of Optimum project) that features smart notifications and proactive recommendations to encourage sustainable transport behavior.

4. Data description

This section presents the dataset, which we used to train the models and predict star ratings of the campsites.

4.1. Feature description

Campsite features We represented each campsite as a feature vector where features describe static properties of the campsite. Some features are the boolean type, where the position in the vector represents the property and its value indicates whether the place has this property or not. Category of location is a categorical attribute and continuous attributes are the location near water, distance from the city, camp categorization, star rating and number of votes. Features described below represent the union of features from both datasets.

The specific properties (features) are:

- **Terrain** (hardened, grass). Basic information about camp; whether it is covered with grass or hard material (e.g. concrete) or both.
- **Illumination.** Information whether camp has electric lighting or not.
- **Security.** Information whether camp has taken security measures, like camera coverage or security guards.
- **Facilities** (restroom, service station, dustbin, shower, playground, dogs not allowed, washing/dryer available, wheelchair accessible, camping-like behavior, internet services, wastewater discharge, chemical toilet discharge, fresh water, electricity). The facilities are important for every camp and more facilities camp has the better it is.
- **Possible nearby activities** (swimming, good fishing possibilities, bicycle tours, hiking tours, winter-sports). The best camps provide not just facilities but also activities which are important for longer stay.
- **Category of location.** Motorhome POIs are divided into three groups: service station, parking, campsite.
- **Location near water.** Distance in meters from the POI to the nearest water.
- **Distance from the city.** Distance in meters from campsite to the center of the nearest town.
- **Camp categorization.** Camp categorization is rating between 1 and 5 provided by organization.

– **Star rating, number of votes.** Ratings are provided as a number between 1 and 10, where higher number means better rating. *Regional* dataset additionally provided separate star ratings for campsite position, tidiness, tidiness of toilets, sport and other activities and value for price.

Temporal (*Mobility patterns*) features The following data was retrieved from *Mobility patterns* service. In the last two years the service detected more than $14,000$ staypoints, where the staypoints consisted of more than $7,361$ unique global locations. Global locations are not necessarily POIs as they include stops like border control, gas stations and other general service areas. This was general information about stays for motorhome, but we are showing in Table 3 just stays in campsites, parkings and service areas. Exact numbers of those places are described in Section 4.2 in Table 3. We collected the duration of each place visit; all other attributes are derived. The specific feature items include:

– **Average duration**: average time of visit at specific location (POI).
– **Min time**: minimum time spend on specific POI.
– **Max time**: maximum time spend on specific POI.
– **Time distribution**: histogram of POI visits durations. Each duration interval represents one feature vector. Duration intervals: (0h,1h], (1h,3h], (3h,6h], (6h,9h], (9h,12h], (12h,15h], (15h,18h], (18h,21h], (21h,24h], (24h,248h]).
– **Time distribution percent:** similar like above, but features represent percentage of specific visits, not real number.
– **Seasons** (winter, spring, summer, fall): Number of visits of specific POI in specific season.
– **Proportion of seasons** (winter, spring, summer, fall): Similar to above - proportion of visits in specific season.
– **Average number of days:** we calculate average number of days motorhomes stayed at the specific location. Our assumption is that the more interesting activities camp has to offer the longer the average stay.
– **Average number of nights:** number of nights spent at camping place.
– **Number of visits:** number of all campsite, parking or service area visits by whole motorhome group.

Complex features Time spent at the location and number of its visits could be dependent on the district that it's located in. For example, some places are much more touristic and near really interesting locations, while other places could be of really good quality, but away from the tourists points of interests. Derived features were intended to help us better understand differences between areas. After performing several tests, we decided on a radius of 10 kilometers. Complex features are derived from Mobility patterns and basic camp features [17,21]. We calculated:

– **Competitiveness** as the proportion of POIs in 10km radius whose categories are the same as the category of the target location.
– **Density** as number of POIs in the neighborhood.
– **Neighbor entropy** measures the heterogeneity of POI categories in the neighborhood.
– **The Popular area** as the total number of check-ins for POIs in the neighborhood.

4.2. Type of resting place distribution analysis

European dataset provides multiple different categories of resting places such as:

- **Campsite**: these are places where only motorhomes can park. Campsites are in general destinations for motorhomes users. Usually they are required to facilitate wastewater discharge, supply with water and electricity.
- **Motorhome parking**: reserved spots at a general parking lot, where motorhomes often park among cars.
- **Service area**: places in the parking lot of a restaurant, hotel or cafe. The facilities are not necessarily there and overnight stay is not allowed at these types of resting places.

Cleaned *European* dataset consists of $23,971$ service areas, motorhome parkings and campsites. Figure 3 shows that the most common type is motorhome parking with $15,859$ parkings. There are $7,056$ campsites and the dataset contains $1,056$ service areas. *Mobility patterns* detected 528 visits in different service areas, parkings and campsites. Out of these visits 249 were to parkings, 57 to service areas and 222 to campsites. The analysis (Table 3) presents the average time motorhomes spent at each type of the place. At the service areas campers in general spent 0.77 hours, at campsites 13.9 hours and at parkings 8.32 hours. This shows that there is a difference in the time spend for different types of POI. We can notice that the average time spent at campsites is 13.9 hours, which is lower than one would expect. However we understand that campsites could be used as the transit points to an end location or the user may leave the campsite to visit nearby locations during the day, which in our case is considered as a new visit to the campsite. This is not ideal in the context of our research as it decreases the time spent at a campsite and increases the number of visits. We plan to address such daily exits in the future work. Notice that the proposed approach does not distinguish between regular costumers and not regular costumers when estimating star ratings.

Type	**Average time (h)**	**# of visits**	**# of visits unique locations**
Service area	0.77	126	57
Motorhome parking	8.32	337	249
Campsite	13.9	570	222

Table 1. Type of point of interest and average time spent on location

4.3. Star rating distribution analysis

Place rating is a score between 1 and 10 which is shown in Figure 4. From the data about places, more than $6,000$ locations have no information about the rating. The most common score is 7 which is attributed to $5,166$ places, followed by score 8 attributed to $4,840$ places. Motorhomes that we tracked visited places with the ratings shown in Table 2. The distribution remains the same and the most common visited camps are still the ones with rating 7 and 8.

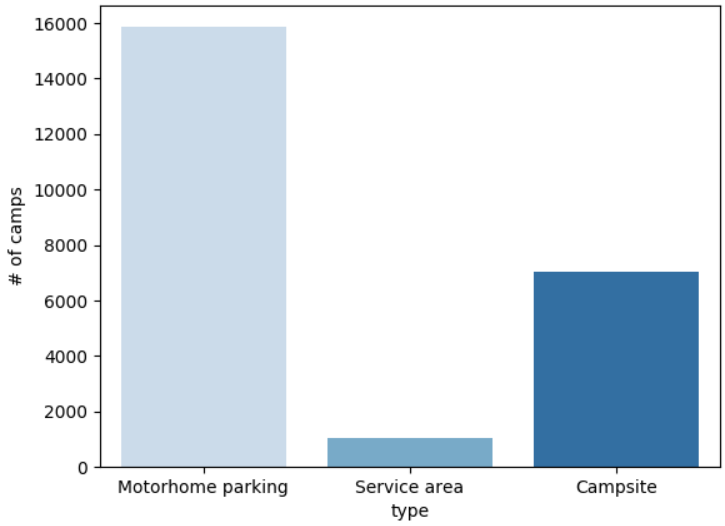


Fig. 3. Category distribution

Star rating	1	2	3	4	5	6	7	8	9	10	Unknown
# of visited camps, parking and service areas	4	7	8	18	44	92	137	110	26	5	77

Table 2. Number of visited camps, parkings and service areas by motorhomes.

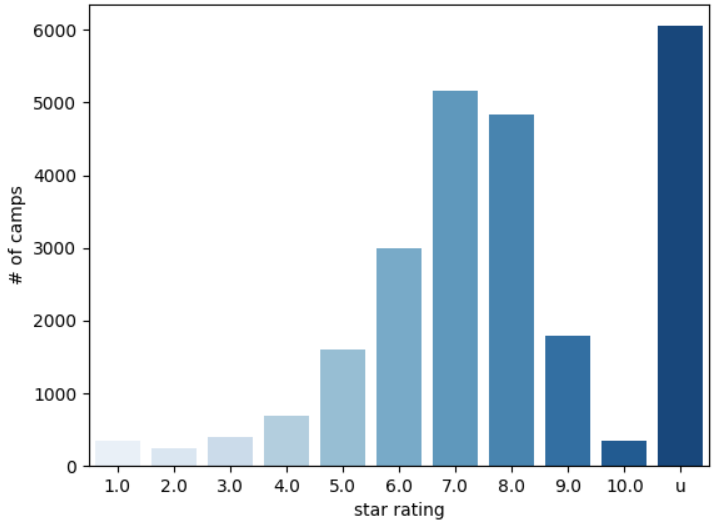


Fig. 4. Star rating distribution

The highest number of camps has France (5,958 camps) and then the number of camps is followed by Germany (4,841 camps), Italy (4,182), Netherlands (1,344 camps).

5. Evaluation

This section presents the methodology of experiments, analysis of the experimental results, and the key findings.

5.1. Experimental setting

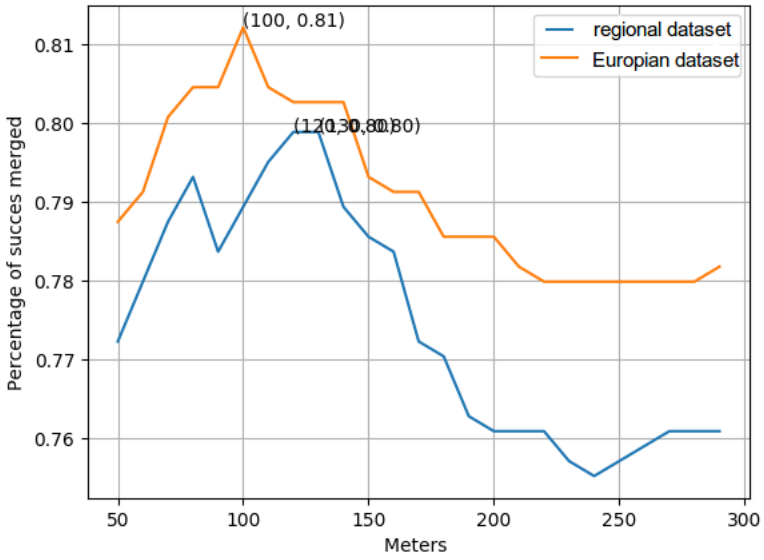


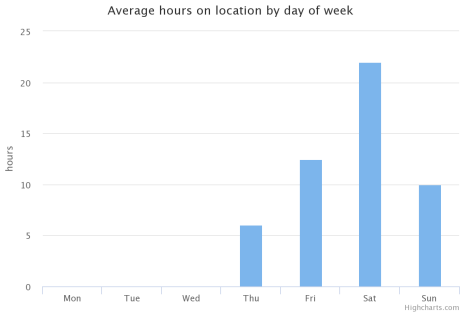
Fig. 5. Tuning the Δ parameter for matching campsites on regional in European dataset.

Parameter selection Multiple data sources can have slightly different coordinates for the same place. In order to overcome this issue we used Algorithm 1, mentioned in Section 3.2. After the algorithm utilization, we perform the evaluation of the merging results. Evaluation dataset was built manually by checking 630 camps from OSM, trying to find a suitable match in other data sources. Some examples were impossible to match even manually. We evaluate the algorithm by comparing its output values to manually annotated data. With this, it calculates the accuracy using following formula:

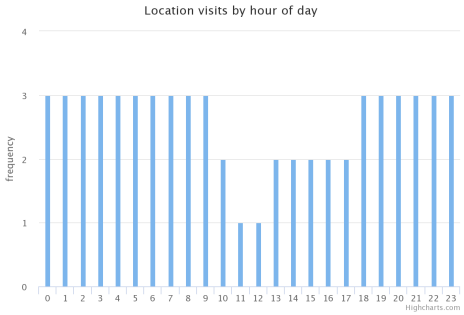
$$accuracy_{matching} = \frac{true_{matched}}{all_{matched}} \tag{6}$$

The result of accuracy is highly dependent on the appropriateness of the chosen Δ parameter. We wrote evaluation script which checks the accuracy of matching for Δ between 50 and 300 meters within a 10-meter interval (as shown in Figure 5). The higher the number the better the result and the best possible score is 1.

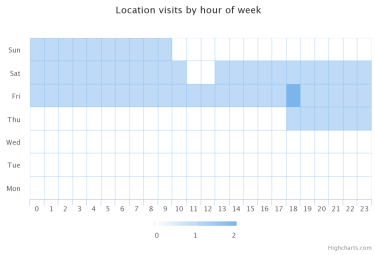
For camps from *European* dataset, the best results have been achieved with Δ of 100 meter (81% correctly merged locations). For camps from *regional* dataset, the best results are with Δ 120 – 130 m with 80% of correct merges. Both very high and very low Δ produced worse results.



(a). Number of hours spent on this location on particular day of week.



(b). Frequency of visits of a location that corresponds to visit of camp by hour of day.



(c). Frequency of visits of a location by multiple users that corresponds to visit of camp by hour of week.

Fig. 6. Mobility patterns.

5.2. Visits to camps

For each global location Mobility patterns service stores the center of the location (latitude, longitude). Based on the underlying visits of this location (staypoints) it calculates statistics like number of visits, time and date of the first visit, last visit and its duration, etc. This type of statistics can also be filtered by time window (take into consideration only the visits that occurred during a specific time range), user (taking into account only visits of a specific user) or group of users, etc. To understand better the semantics of a location it also creates histograms and heatmaps of location visits. A typical example from

the collected campsite data is a daily histogram of location visits by motorhomes. Figure 6a shows number of hours spent on this location on the particular day of week. We can observe that the most time in this campsite was spent through weekdays. Figure 6b shows frequency of visits of a location that corresponds to visits for camp an hour of day. We can observe a gap around midday, meaning that motorhomes users more frequently stay in the camp at night. Figure 6c shows heatmap of visits of a location by an hour of week for every week in our dataset. We can observe that the camp is most visited at the end of the week.

5.3. Correlation analysis

Correlation analysis was performed on *regional* and *European* dataset. *European* dataset results are described in the text, but the heatmap matrix is not shown.

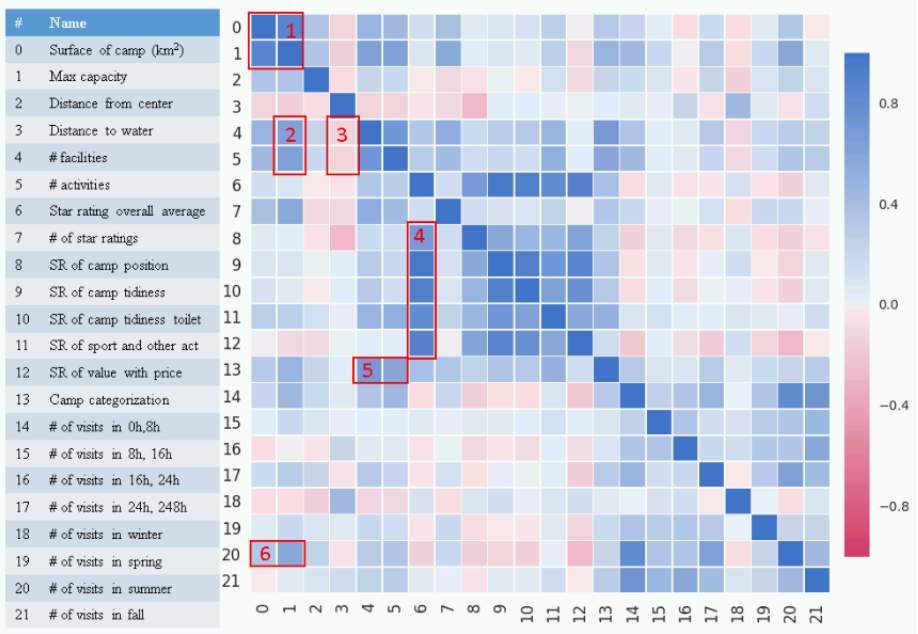


Fig. 7. Performing correlation analysis on *regional* dataset. The result in red boxes is explained in text from 1 to 6 in text.

Correlation analysis on *regional* dataset On the *regional* dataset we checked which attributes intersect and Figure 7 shows some of the more interesting correlations we found. The red squares in Figure 7 are numbered accordingly with their explanations below:

1. Surface of camp (km^2) is proportional to the maximal capacity of people and campers. This is easy to understand, since higher number of motorhomes and people takes up more space.

2. Number of activities and facilities is also correlated with maximum camp capacity. It is hard for smaller camps to provide so many diverse activities because they can take up a lot of space and can be expensive to maintain. We counted 12 distinct activities: inside swimming pool, tennis, volleyball, bicycle tours, diving, table tennis, basketball, boats, mini golf, rafting and rock climbing. We designated following items as facilities: shop, restaurant, restroom, restroom for wheelchair, chemical toilet discharge, bathroom, service station, dogs allowed, washing machine, yacht quay, a barrier for boats, playground, children bathroom, animation for children, dryer. While smaller camps again cannot provide all the facilities available in larger camping places, some common facilities do exist, such as bathrooms, chemical toilets, service station, dogs allowed, washing machines and dryers, etc.
3. Additionally, number of facilities is negatively correlated with location near water (Distance from water is measured in meters in our analysis). Our explanation is that for many facilities, the presence of lake, river or sea is important.
4. The average star rating of camp is strongly correlated with star rating of camp position, star rating of the tidiness of camp, star rating of the toilet, the star rating of sports activities in camp and star rating of price/value ratio. The most correlated star rating is for tidiness of camp and toilet, because these both express overall cleanliness of camp, making them strongly connected.
5. Camp categorization is a common approach to assessing camps and whether they provide service, activities, facilities from 1 to 5 and also number of activities facilities affect correlation.
6. In the summer people in general visit bigger camps.

Correlation analysis on *European* dataset We also performed correlation analysis on *European* dataset, but this dataset is less interpretable than *the regional* dataset. We can distinguish rules which are that time spent on location is correlated with the presence of facilities.

If people spent more than 24 hours in general area of a camp, they chose a camp with grass, restroom, shower, playground, washing/dryer, internet services and electricity. This is an important discovery, because this shows what is important for users when they are staying at a camp for multiple days. On the other hand, people who stay just up to 3 hours, do not care for the availability of such facilities and we found a strong negative correlation to confirm this.

5.4. Prediction of star rating

Regression Prediction of star rating was performed on *regional* and *European* dataset. In the *regional* dataset the result was close to baseline or even worse, mainly due to a lack of learning examples. We investigated and figured out that the number of examples is too low to predict point-of-interest rating well. Another possible issue could be that the features are not informative enough and maybe additional feature engineering could help. Additionally, there seemed to be a lack of meaningful correlation between camp features and temporal features. Due to mismatching attributes in the datasets, we were unable to perform the merging into a single unified dataset, since that would cause either a certain amount of data loss or empty values in some records.

For every camp we predicted star rating using machine learning methods (described in Section 3.3). In order to make better models the best feature set was chosen for every algorithm. All regression tests were performed on/using training set due to lack of available data. This was done by utilizing stepwise feature selection (described in Section 3.3).

We performed analysis on *European dataset*, and then also separately on camps which were visited in spring and summer, and again separately camps which were visited in fall and winter. Dataset for visited camps in spring and summer will be called *camps visited in warmer months*. The other datasets that include camps which were visited on fall and winter will be called *camps visited in cooler months*. Our assumption is that visitors have different requirements if they visit camp in fall and winter than spring and summer.

Each dataset has a distinct number of visited camps due to different frequency of visits in different seasons. We treat each camp as learning example when it has at least one visit from the motorhomes we are tracking. For evaluation we used 10-fold cross-validation in all three datasets. We tune the parameters for each algorithm separately on every dataset as follows:

- **k-NN**: We varied parameter k between values 1 and 10. Parameter k is number of neighbors.
- **Decision tree:** We varied parameter *max depth* from 1 to 10.
- **Random forest**: We varied parameter *n_estimators* from 1 to 10.
- **SVR**: We varied penalty parameter C of the error term, between 1 and 15. Also we varied ϵ which specifies the epsilon-tube within which no penalty is associated with the training loss function, from 1 to 15. When the SVM is used for regression in this paper is referred as SVR.

This dataset contains 528 camps, which were visited at all seasons. The best k-NN performance was observed with parameter setting k=7. For the random forests algorithm we set *n_estimators*=5. For the decision trees algorithm, we tuned *max depth* to 2. For the SVR algorithm we tuned C=7 and ϵ=1.

Algorithm	**Features**									
	Camps		**Temporal**		**Camps + Temporal**		**Camp+ Complex**		**All**	
	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**
k-NN	0.02	1.53	0	1.5	**0.06**	**1.5**	0.02	1.53	**0.06**	**1.5**
Random for.	0.06	1.5	0.01	1.46	0.11	1.46	0.06	1.5	**0.12**	**1.45**
Decision tree	0.06	1.49	0.02	1.53	0.08	**1.48**	0.06	1.49	**0.09**	**1.48**
Linear reg.	0.05	1.5	0.03	1.52	**0.08**	1.48	0.06	1.5	**0.08**	**1.47**
SVR	0.07	1.49	0.03	1.52	0.01	1.46	0.11	1.45	**0.19**	**1.39**
Baseline	-0.04	1.57	-0.04	1.57	-0.04	1.57	-0.04	1.57	-0.04	1.57

Table 3. Prediction of star rating with different features on dataset where all camps were visited. To evaluate the performance of our model we used measures to evaluate Root mean square error (RMSE) and coefficient of determination (R^2). The best results for every algorithm is in bold.

Table 3 shows results of predicting star rating. SVR outperforms other algorithms with observed RMSE of 1.39. Results indicate that temporal features from *Mobility patterns* improve predicting star rating. Also results improve if we add complex features. For every algorithm we bold the best results in Tables. This means that time spent on location is correlated with star rating. The best features for SVR are

- camp features (internet services, electricity, grass, dog are not allowed, shower)
- temporal features ((24h,248h],(3h,6h],(12h,15h],winter)
- complex features (competitiveness, entropy)

Every algorithm has its own set of features where it performs the best.

Algorithm	**Features**									
	Camps		**Temporal**		**Camps + Temporal**		**Camp+ Complex**		**All**	
	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**
k-NN	0.05	1.51	-0.01	1.49	0.06	1.49	0.05	1.51	0.06	1.49
Random for.	0.06	1.5	0.03	**1.44**	**0.14**	**1.44**	0.06	1.05	**0.14**	**1.44**
Decision tree	0.06	1.5	0.03	1.49	0.09	1.49	0.06	1.5	0.1	1.49
Linear reg.	0.05	1.51	0.02	1.49	0.06	1.49	0.06	1.5	0.08	1.49
SVR	0.06	1.5	0.02	1.46	0.11	1.46	**0.11**	1.46	**0.15**	**1.43**
Baseline	-0.03	1.57	-0.03	1.57	-0.03	1.57	-0.03	1.57	-0.03	1.57

Table 4. Predicting star rating with newly calculated temporal attributes taking to account only spring and summer. To evaluate the performance of our model we used measures to evaluate Root mean square error (RMSE) and coefficient of determination (R^2). The best result for every algorithm is in bold.

Due to different service demand in cooler and warmer months we analyzed camps separately for those seasons. For instance in cooler months we take into account only camps which were visited in fall and winter. We ignore visits in other months. Thus, the temporal feature values change. In this two datasets we again used 10-fold cross-validation method to evaluate results. The dataset for warmer months consists of 486 camps, and dataset for cooler months contains 75 camps.

For dataset in warmer months we set parameters of the algorithms as follows. For k-NN, the number of *k* is 7, for random forest *n_estimators* is 5. For the decision tree we tuned *max depth* to 4. For SVR we set *C*=3 and ϵ=0.9. The parameters for fall and winter are as follows: for k-NN we set *k*=5, for random forest the parameter *n_estimators* was set to 8. For the decision tree we set *maximum depth* to 2. For SVR we set $C = 5$ and $\epsilon = 0.4$. SVR features for spring and summer are the following:

- camp features (internet services, dogs are not allowed, electricity)
- temporal features (time intervals (3h,6h], (9h,12h], (12h,15h], (24h,248h], percent of visits)
- complex features (competitiveness, entropy)

The best features for fall and winter are

- camp features (presence of Internet services, camping-like behavior),
- temporal features((0h,1h], (6h,9h], (21h,24h], number of visits, average duration, spring, min and max stay)

We can see in Table 4 and Table 5 that SVR outperforms the other methods with All features. The best result was obtained on the dataset for cooler months and are presented in Table 5. The best score for R^2 is 0.56 achieved by SVR with Camps + Temporal features and All features. The best RMSE is 0.87 achieved by Linear regression with All the features. We started modeling with POI features only and then added temporal and complex features. Results show that adding temporal and complex features helps with star rating prediction.

Algorithm	**Features**									
	Camps		**Temporal**		**Camps + Temporal**		**Camp+ Complex**		**All**	
	$\boldsymbol{R^2}$	**RMSE**	$\boldsymbol{R^2}$	**RMSE**	$\boldsymbol{R^2}$	**RMSE**	$\boldsymbol{R^2}$	**RMSE**	$\boldsymbol{R^2}$	**RMSE**
k-NN	0.09	1.17	0.16	0.91	0.49	1.22	0.09	1.2	0.49	0.91
Random for.	0.16	1.17	0.15	0.96	0.49	1.17	0.14	1.17	0.49	0.96
Decision tree	0.1	1.17	0.16	1.11	0.24	1.21	0.1	1.21	0.24	1.11
Linear reg.	0.04	1.2	0.12	0.87	0.53	1.25	0.04	1.25	0.53	**0.87**
SVR	0.29	1.06	**0.23**	**0.9**	**0.56**	1.1	**0.29**	1.1	**0.56**	0.9
Baseline	0	1.2	0	1.2	0	1.2	0	1.2	0	1.2

Table 5. Predicting star rating with newly calculated temporal attributes taking to account only fall and winter. To evaluate the performance of our model we used measures to evaluate Root mean square error (RMSE) and coefficient of determination (R^2). The best result for every algorithm is in bold.

The only feature which was selected to every feature subset is internet services also important features are winter sports, when this visit occurs (summer, spring, fall, winter).

Classification Prediction of star rating was performed only for regional dataset. We split star ratings provided by users into three categories:

- (0,4] is range between 0 and 4 and we can mark this as bad camps.
- (4,7] is range between 4 and 7 and it is considered as decent camps.
- (7,10] is range between 7 and 10 and it is considered as very good camps.

We decided to discretize available data in this way because every user has different idea of what they imagine as rating 7 or 8. Due to different personalities and lifestyles, their opinion on what makes a good camp can be biased. We split dataset into test and training set in ratio 80:20. Using cross-validation on the training set we have tuned the parameters and choose the best features subset for each of the algorithms using the algorithm presented in Section 3.3:

- **k-NN:** We varied parameter *k* between values 1 and 10. Parameter *k* is number of neighbors.

- **Decision tree:** We varied parameter *max depth* from 1 to 10.
- **Random forest:** We varied parameter *n_estimators* from 1 to 10.
- **SVC** We varied parameter *C* of the error term between 1 and 15. When the SVM is used for classification in this paper is referred as SVC.

The best parameter values for which we reported the result are the following. For classification task k-NN, the number of nearest neighbors is 7, for random forest *n_estimator* is 3. For SVC we set *C*=1 and for decision tree will be expanded until all leaves are pure or until all leaves contain less than 2 samples. Table 6 shows results of classifying camps in three categories: bad, decent and very good. Among the tested algorithms, looking at F1 measure, decision tree performed the best with no difference in the performance results when using camps or temporal features. Combining all the features did not increase F1 score, though it did slightly improve accuracy for random forest (from 0.55 to 0.58). Accuracy of the majority class is 0.49, which means that all classifiers perform better or equal as the majority class.

Algorithm	**Features**									
	Camps		**Temporal**		**Camps+Temporal**		**Camps+Complex**		**All**	
	CA	**F1**	**CA**	**F1**	**CA**	**F1**	**CA**	**F1**	**CA**	**F1**
k-NN	0.54	0.56	0.49	0.51	0.52	0.69	0.4	0.39	0.46	0.48
Random for.	**0.55**	0.46	0.52	0.69	0.52	0.69	0.57	0.54	**0.58**	0.57
Decision tree	0.52	**0.69**	0.52	0.69	0.52	0.69	0.52	0.69	0.52	**0.69**
SVC	0.52	0.69	0.52	0.69	0.57	0.5	0.52	0.69	0.53	0.41
Logistic reg.	0.52	0.53	0.52	0.69	0.54	0.49	0.55	0.57	0.49	0.48
Majority class	0.49	0.0	0.49	0	0.49	0	0.49	0	0.49	0

Table 6. Predicting star rating with calculated temporal attributes. To evaluate performance of our model we used measures to evaluate Accuracy (CA), F-measure ($F1$). The best result for every algorithm are in bold.

6. Conclusions and Future work

In this paper, we proposed a framework to merge multiple different sources of point-of-interest data, including the real-time GPS coordinates from multiple vehicles, which can bring valuable additional information. We tested our system on motorhome camps, but it is easy to adopt the process for extracting other points with time component by simply configuring the properties file. We also conducted extensive experiments to investigate if temporal features help with predicting star rating and consequently assessing the place quality. We can say that it is possible to predict star rating via camp characteristic and temporal data of users. Our analysis showed that the temporal features obtained from geospatial information improve the regression prediction by 4-13 %. For classification we observed that classification accuracy improves for 3 % when using all the features set. From the experiments we can see that both classification and regression are better than the baselines. Another finding is that the temporal features improve accuracy of the models.

The regression model which outperforms other models most of the times is SVR. The best performing classification models are SVC and random forest.

In the future, this work could be extended in multiple dimensions. Firstly, by adding additional data about places like gas stations. Another option is to include more data in addition to the GPS coordinates. The motorhomes that we are tracking are already equipped with onboard sensors and are measuring other values, such as clean water level, wastewater, air quality within living space and car batteries state of charge. With this data we can build a recommendation system which suggests users where and when to discharge wastewater, sewage or refill the water.

Acknowledgments. This work was supported by the Slovenian Research Agency and the ICT program of the EC under project OPTIMUM (H2020-MG-636160). The authors would like to thank Anja Polajnar for proofreading of the manuscript.

References

1. Avtokampi description of data. http://www.avtokampi.si/, Accessed: 2017-05-06
2. Campercontact description of data. https://www.campercontact.com/en/, Accessed: 2017-05-06
3. Geofabrik. http://download.geofabrik.de/europe.html, Accessed: 2017-06-06
4. OpenStreetMap: Tag:tourism_camp site. http://wiki.openstreetmap.org/wiki/Tag:tourism%3Dcamp_site, Accessed: 2017-06-06
5. Optimum project. http://www.optimumproject.eu/, accessed: 2018-06-17
6. Optimum project. http://www.optimumproject.eu/about/pilot-cases/pilot-case-3/innovations-3.html, accessed: 2018-06-17
7. Osm2psql. http://wiki.openstreetmap.org/wiki/Osm2pgsql, Accessed: 2017-06-06
8. Scikit-learn regression models. http://scikitlearn.org/stable/supervised_learning.html#supervised-learning, Accessed: 2017-10-06
9. Europe: Registrations of new Motor Caravans per month 2016. Europian Caravan Federation (2017), http://www.e-c-f.com/fileadmin/templates/4825/images/statistics/europazul-9.pdf
10. Adomavicius, G., Kwon, Y.: New recommendation techniques for multicriteria rating systems. IEEE Intelligent Systems 22(3) (2007)
11. Asghar, N.: Yelp dataset challenge: Review rating prediction. CoRR abs/1605.05362 (2016)
12. Barbosa, R.R.L., SÃąnchez-Alonso, S., Sicilia-Urban, M.A.: Evaluating hotels rating prediction based on sentiment analysis services. Aslib Journal of Information Management 67(4), 392–407 (2015), https://doi.org/10.1108/AJIM-01-2015-0004
13. Bennett, J.: OpenStreetMap Be your own Cartographer. Packt Publishing Ltd. (2010)
14. Bradesko, L.: Knowledge acquisition through natural language conversation and crowdsourcing 150 (2018)
15. Carbon, K., Fujii, K., Veerina, P.: Applications of machine learning to predict Yelp ratings (2014)
16. Fan, M., Khademi, M.: Predicting a business star in Yelp from its reviews text alone. CoRR abs/1401.0864 (2014)
17. Hsieh, H.P., Li, C.T., Lin, S.D.: Estimating potential customers anywhere and anytime based on location-based social networks. In: Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II. pp. 576–592. ECMLPKDD'15, Springer, Switzerland (2015), https://doi.org/10.1007/978-3-319-23525-7_35

18. Hu, Y.H., Chen, K.: Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. International Journal of Information Management 36(6, Part A), 929 – 944 (2016), http://www.sciencedirect.com/science/article/pii/S0268401215301845
19. Huang, J., Rogers, S., Joo, E.: Improving restaurants by extracting subtopics from Yelp reviews (2014)
20. Jannach, D., Gedikli, F., Karakaya, Z., Juwig, O.: Recommending hotels based on multi-dimensional customer ratings. In: Information and communication technologies in tourism 2012, pp. 320–331. Springer, Vienna (2012)
21. Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., Mascolo, C.: Geo-spotting: Mining online location-based services for optimal retail store placement. CoRR abs/1306.1704 (2013), http://arxiv.org/abs/1306.1704
22. Levandoski, J.J., Sarwat, M., Eldawy, A., Mokbel, M.F.: Lars: A location-aware recommender system. In: Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. pp. 450–461. ICDE '12, IEEE Computer Society, Washington, DC, USA (2012), http://dx.doi.org/10.1109/ICDE.2012.54
23. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.Y.: Mining user similarity based on location history. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 34:1–34:10. GIS '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1463434.1463477
24. Lv, M., Chen, L., Xu, Z., Li, Y., Chen, G.: The discovery of personally semantic places based on trajectory data mining. Neurocomputing 173(Part 3), 1142 – 1153 (2016), http://www.sciencedirect.com/science/article/pii/S0925231215012916
25. Narayanan, M., Cherukuri, A.K.: A study and analysis of recommendation systems for location-based social network (LBSN) with big data. IIMB Management Review 28(1), 25 – 30 (2016), http://www.sciencedirect.com/science/article/pii/S0970389616000021
26. Ravi, L., Vairavasundaram, S.: A collaborative location-based travel recommendation system through enhanced rating prediction for the group of users. Intell. Neuroscience 2016, 7– (Mar 2016), http://dx.doi.org/10.1155/2016/1291358
27. Tiwari, S., Kaushik, S.: User category based estimation of location popularity using the road GPS trajectory databases. Geoinformatica (2014)
28. Tiwari, S., Kaushik, S.: Popularity estimation of interesting locations from visitor's trajectories using fuzzy inference system. Open Computer Science 6(1) (2016), http://www.degruyter.com/view/j/comp.2016.6.issue-1/comp-2016-0002/comp-2016-0002.xml
29. Vitorino, R., Herga, Z., Bradesko, L.: Integrated mobility decision support (318452). EU project Mobis (2015), http://www.europa.eu/whatever
30. Yan, Z.: Towards semantic trajectory data analysis: A conceptual and computational approach (2009)
31. Zheng, Y., Xie, X.: Learning location correlation from GPS trajectories. In: 2010 Eleventh International Conference on Mobile Data Management. pp. 27–32 (May 2010)
32. Zheng, Y.: Trajectory Data Mining: An Overview. ACM Trans. On Intelligent Systems and Technology 6(3), 1–41 (2015)

Matej Senožetnik received his B.Sc. in Computer science and engineering (2015) from University of Ljubljana. He is now pursuing M.Sc. in Information and Communication technologies at Jožef Stefan International Postgraduate School. He works on European and national projects as a student researcher at the Jožef Stefan Institute.

Luka Bradesko is part of the NLP Enrichment team at Bloomberg L.P. He obtained his PhD at Jožef Stefan International Postgraduate School, Ljubljana Slovenia with the topic

Knowledge Acquisition through Natural Language Conversation and Crowdsourcing. His research interests are in Natural Language Processing, Logical Inference and Knowledge Extraction. From 2008 to 2013 he worked as a principal software engineer for Cycorp Europe, which was at the time an EU branch of the American AI company Cyc Inc. During these years, he worked on an EU project developing distributed large scale inference engine (LarKC), and also on an natural language based AI assistant startup built on top of Cyc (Curious Cat). Some of the recent projects include a concept of an intelligent motorhome (reasoning engine with predictive analytics, interacting with sensors and actuators), geo-spatial based predictive analytics, and Named Entity Disambiguation algorithm which is already part of his work at Bloomberg L.P.

Tine Šubic has received B.Sc. in Computer Science and Informatics (2017) and is now pursuing M.Sc. in Electrical Engineering at University of Ljubljana. He works on European and national projects as a student researcher at the Jožef Stefan Institute.

M.Sc. Zala Herga is a researcher and a PhD student at Jožef Stefan Institute, Ljubljana. She finished her master's degree in Financial Mathematics at Faculty of Mathematics and Physics, University of Ljubljana. She has experience in AI, data mining, analytics and predictive statistics. She mainly applied her research to two areas: risk management and traffic domain.

Jasna Urbančič received her B.Sc. in Physics (2015) and is now pursuing M.Sc. in Computer and Information Science at University of Ljubljana. She works on European and national projects as a student researcher at the Jožef Stefan Institute.

Primoz Skraba is currently a senior lecturer of Mathematics at Queen Mary, University of London as well as a researcher in the Artificial Intelligence Laboratory at the Jozef Stefan Institute, Slovenia. He has also held positions assistant professor of computer science at the University of Koper and adjunct professor of mathematics at the University of Nova Gorica. He received his Ph.D. in Electrical Engineering from Stanford University in 2009. His main research interests are applications of topology to computer science including data analysis, machine learning, sensor networks, and visualization.

Prof. Dr. Dunja Mladenić `http://ailab.ijs.si/dunja_mladenic/` works as a researcher and a project leader at Jožef Stefan Institute, Slovenia, leading Artificial Intelligence Laboratory and teaching at Jožef Stefan International Postgraduate School, University of Ljubljana and University of Zagreb. She has extensive research experience in study and development of Machine Learning, Big Data/Text Mining, Sensor Data Analysis, Internet of Things, Data Science, Semantic Technology techniques and their application on real-world problems. She has published papers in refereed journals and conferences, co-edited several books, served on program committees of international conferences and organized international events. She serves as a project evaluator of project proposals for European Commission and USA National Science Foundation. From 2013-2017 she served on the Institute's Scientific Council `http://www.ijs.si/ijsw/Scientific_Council/`, as a vice president (2015-2017). She serves on Executive board of Slovenian Artificial Intelligence Society SLAIS (as a president of SLAIS (2010-2014)) and on Advisory board of ACM Slovenija.

Received: December 12, 2017; Accepted: August 7, 2018.

Computer Science and Information Systems 16(1):155–178 https://doi.org/10.2298/CSIS180110037A

An insight into the effects of class imbalance and sampling on classification accuracy in credit risk assessment

Kristina Andrić[1], Damir Kalpić[2], and
Zoran Bohaček[3]

[1] Privredna Banka Zagreb, Radnička 50,
10000 Zagreb, Croatia
kristina.andric@fer.hr
[2] University of Zagreb Faculty of electrical engineering and computing, Unska 3,
10000 Zagreb, Croatia
damir.kalpic@fer.hr
[3] Croatian Banking Association, Nova Ves 17,
10000 Zagreb, Croatia
zoran.bohacek@alumni.unizg.hr

Abstract. In this paper we investigate the role of sample size and class distribution in credit risk assessments, focusing on real life imbalanced data sets. Choosing the optimal sample is of utmost importance for the quality of predictive models and has become an increasingly important topic with the recent advances in automating lending decision processes and the ever growing richness in data collected by financial institutions. To address the observed research gap, a large-scale experimental evaluation of real-life data sets of different characteristics was performed, using several classification algorithms and performance measures. Results indicate that various factors play a role in determining the optimal class distribution, namely the performance measure, classification algorithm and data set characteristics. The study also provides valuable insight on how to design the training sample to maximize prediction performance and the suitability of using different classification algorithms by assessing their sensitivity to class imbalance and sample size.

Keywords: credit risk assessment, imbalanced data sets, class distribution, classification algorithms, sample size, undersampling.

1. Introduction

The aim of credit scoring is to differentiate borrowers and classify them into two groups: good clients (non defaulters) and bad clients (defaulters), using predictive models and historical information on clients' repayment behavior. It is widely used in financial institutions for lending decisions and asset quality monitoring and is one of the most widely used applications of statistical models and data mining methods in practice [1]. Therefore, making a correct assessment of the likelihood of the client defaulting is of utmost importance for financial institutions [2] and increasing accuracy in default prediction would be of great interest for financial institutions [3].

Class imbalance is a common occurrence in credit scoring, where the good clients greatly outnumber the bad clients. As one of the key challenges in data mining [4], it has received a lot of attention in recent years. Performance of standard classification algorithms tends to deteriorate when class imbalance is present as the cost of misclassifying a case in credit scoring is greatly asymmetrical. Frequent occurrence of class imbalance in credit risk assessment indicates the need for additional research efforts in handling imbalanced datasets. Several methods have been proposed to improve prediction accuracy in these settings, such as altering class distribution [5].

Despite having access to vast amounts of data related to customer behavior, standard practice is to develop predictive models using a sample of data, i.e. representative of the overall population. Even though a lot of research has been focused around prediction accuracy of different classification algorithms, topics of data sample design have been largely neglected regardless of the fact that data sample preparation is the cornerstone and the most time consuming step of the model development process [6]. Little research effort has been devoted to assessing the impact of class distribution on credit scoring, in particular when dealing with real life data samples, that can be very heterogeneous and differ in terms of size, imbalance ratio (IR, to be defined later), the number of features etc. Majority of studies use benchmark data whose characteristics are not representative of real-life data samples, which may introduce a bias, or use a very limited number of data samples that contain only a few independent variables [7], [8]. Sample size is especially important for low default portfolios, where, due to the lack of a systematic empirical evaluation, it is unclear what is the minimum number of instances that would justify model development. Nevertheless, this topic has not been evaluated systematically with empirical experiments, despite its apparent importance and applicability.

To address the observed research gap, a study of the interdependence of class distribution, dataset characteristics and prediction accuracy in a real life credit scoring environment was performed using logistic regression, neural network and gradient boosting algorithm on a wide array of real life data samples. Class imbalance was increased progressively to determine how classification accuracy was affected by different IRs. Performance of the proposed framework was evaluated with several evaluation measures. The second part of the study addresses the question of how sample size impacts classification accuracy and finding the minimal sample size (namely the number of defaulters) needed to develop a model with adequate prediction accuracy. The main contribution from this study can be summarized as: (1) finding the optimal class distribution that maximizes classification accuracy, (2) assessment of different classification algorithm performance when dealing with different imbalance ratios and samples of different size, (3) assessment of gradient boosting algorithm in credit risk domain, (4) finding the minimal data sample size that can be used for model development with adequate prediction accuracy and (5) analysis of the intrinsic characteristics of the data samples (IR, number of minority class cases, number of total instances).

The rest of this paper continues as follows: Section 2 provides a literature review. The classification algorithms are described in Section 3, and the experimental framework in Section 4. Results of the study are presented and discussed in Sections 5. and 6. Conclusions and possible future research direction are presented in Section 7.

2. Background work

A lot of research effort has been committed to evaluating classification algorithms in credit scoring, ranging from traditional statistical methods, such as logistic regression [1], to non-parametric algorithms, such as neural networks [9]. In the recent years there has been an increased interest for using hybrid and ensemble classifiers in credit risk, such as boosted regression trees, random forests, deep learning methods and other [10], [11], [12], [13] [14]. A number of benchmark studies have been performed, comparing classification accuracy of different classification algorithms [15], [8]. However, there is no consensus among researchers as to which algorithms yield the best performance. In fact, it seems that the choice of algorithm should take into account the domain, dataset and evaluation criterion [1], [16].

Class imbalance is inherent to credit scoring. In [17] the authors analyzed how class imbalance impacts several classifiers by steadily increasing the IR. They concluded that gradient boosting and random forest achieved good performance, in particular when the IR was very high. Logistic regression also achieved high accuracy, unlike other methods, such as SVM and C4.5. One of the methods proposed to alleviate the adverse impact of using imbalanced data sets includes using pre-processing techniques, which alter the original imbalanced data sample to produce a more balanced class distribution [18]. With oversampling, the number of defaulted clients is increased, whereas with undersampling the number of good clients is decreased. A few comparisons of different sampling techniques were performed, but with different conclusions reached on which technique yields highest accuracy. In fact, research suggests that intrinsic data set characteristics and the application domain have a greater impact on the performance than a particular pre-processing technique, indicating that further analysis is needed [19], [20], [21].

Adjusting class distribution can lead to disregarding potentially valuable information from the sample or overfitting. Therefore, finding the optimal class distribution and its empirical evaluation is of great significance. Although not related to credit scoring, in [22] the authors studied the C4.5 classifier, concluding that balanced class distribution in general achieves the best results. However, in credit scoring the number of defaulted clients is usually very small and a balanced class distribution is hardly ever encountered in real life. In [23] the authors analyzed the effect of different class distributions in credit risk assessment. Class distribution was altered by using random undersampling to produce various imbalance ratios. The results stressed the importance of assessment criterion, reaching a conclusion that performance deteriorated with all classification algorithms used when faced with larger class imbalance.

A common approach in credit scoring is to develop predictive models using a smaller sample of the overall population, that should resemble the target population as much as possible, due to various reasons. As first, these models are subjected to restrictive out of sample validation procedures, implying that all the data available cannot be used for model development. As second, using larger samples increases the cost, in terms of time and computational resources needed to develop a model. As third, when population drift occurs, where characteristics of the population change over time, the entire historical sample cannot be used for model development, since it is not representative of the target population. Population drift has become an especially important topic in recent years, since during the financial crisis of 2007 the overall macroeconomic environment changed and both customers and financial institutions changed their behavior. Therefore, sample size and construction are of utmost importance for the quality of predictive models.

Despite its relevance, studies on choosing the optimal sample size are limited. There is a common understanding among practitioners that a sample containing around 1500 cases of each class should suffice to build predictive models [6]. However, in [20] the authors performed an empirical study on sampling in credit scoring, using two datasets and four classification algorithms. The study indicated that using samples larger than those recommended in practice increases prediction accuracy.

The focus of research so far has been proposing new algorithms and comparing classification accuracy of different classification algorithms, whereas the issues of sample design were largely ignored. This study aims to address the observed gaps and provide insight into credit risk modelling with empirical evidence on how to design the training sample and which algorithm might be appropriate in given settings.

3. Classification algorithms

Classification algorithms used in the study include logistic regression, one of the most commonly used methods for credit scoring [20], neural networks, a classifier whose good performance was demonstrated in many studies [9], [15] and gradient boosting, a relatively new method, yet to be thoroughly evaluated in the credit risk domain.

3.1. Logistic regression

Logistic regression (LR) is the standard method used for credit scoring among practitioners and one of the most commonly used methods for credit scoring [20]. LR estimates $p(y = 1|\mathbf{x})$. y represents the dependent variable, equal to 1 if the client is in a default status. $\mathbf{x}$ is a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, or a collection of n independent variables. $p(y = 1|\mathbf{x})$ then represents the conditional probability that the outcome of the client defaulting is present. The relationship between the independent variables and the dependent variable is described by a logit function [24]:

$$\text{logit}\,(p(y = 1|\mathbf{x})) = \ln\left(\frac{p(y = 1|\mathbf{x})}{1-p(y = 1|\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \,. \quad \textbf{(1)}$$

β_0 is the intercept parameter and $\beta_1, \ldots, \beta_n$ are the model parameters. The model parameters are estimated through minimizing the log-likelihood function:

$$min\, l = -\sum_{i=1}^{n} y_i \, ln\big(p(y = 1|\mathbf{x})\big) + (1 - y_i)\, ln\big(1 - p(y = 1|\mathbf{x})\big)\,. \quad \textbf{(2)}$$

3.2. Neural networks

Neural networks (NN) are nonlinear classifiers modelled after the human brain [25]. The multilayer perceptron (MLP) is one of the most commonly used types of neural networks. It contains an input layer, a hidden layer, and an output layer. In each of the layers neurons process their inputs into output values, used by the neurons in the next layer. The independent variables are the input layer neurons. Initial weights are assigned to connections between neurons and adjusted during training. This iterative process can be

based on various methods, such as gradient descent, Quasi-Newton etc. Output of the hidden neuron h_i is determined through an activation function f:

$$h_i = f\ (b_i + \sum_{j=1}^{n} w_{ij} x_j)\ . \quad (3)$$

x represents a collection of n independent variables, **x** = ($x_1, x_2, \ldots, x_n$), w_{ij} is the weight connecting input j to neuron i and b_i is the bias term. Output of the output layer is determined using a similar function**:**

$$o = g\ (b_i + \sum_{j=1}^{m} v_i h_i)\ . \quad (4)$$

m is the number of neurons in the hidden layer and v_i is the weight connecting hidden neuron h_i to the output neuron. MLP use sigmoidal f and g functions. The model parameters w_{ij}, v_i, b_i are determined through minimizing the loss function using gradient-based algorithms.

3.3. Gradient boosting

Gradient boosting (GB), also referred to as stochastic gradient boosting, is an ensemble model that consists of a series of simple decision trees. Ensemble models aim to improve accuracy by combining predictions of multiple base models and minimizing the error term in an iterative manner. After the initial base model (tree) is set up, each subsequent base model is fitted to the residuals of the previous model to minimize the error term and avoid errors of the current ensemble [26]. Since it is a homogeneous ensemble classifier, additional randomness is introduced by bootstrap sampling [8]. The model is described as:

$$F(x) = G_0 + \sum_i \beta_i L_i\ (x)\ . \quad (5)$$

β_i are coefficients for the respective tree node L_i fitted to the residuals of the previous algorithm and G_0 the first value for the series.

4. Experimental design

4.1. Data samples

To conduct the experiments, ten different data sets were used. Their characteristics are outlined in Table 1. Japanese and German data set are public data sets, available in the UCI Machine Learning Repository [27] and include data related to credit card applications. The German dataset has 1000 cases, 300 of which are defaulted, with 20 features. The Japanese data set has 690 cases, 329 of which are defaulted, with 13 features. Data sets DS1 - DS8 are real-life samples obtained from a commercial bank. Even though

a limited number of real-life data sets were used, datasets used for credit scoring usually show a high level of comparability across different credit institutions [20]. This notion stems from the fact that types of consumer data gathered by institutions broadly represent same types of information for similar types of portfolios. For a number of countries (for example EU countries) using the types of data that were used in this study is also a legal requirement [28], [29]. Consequently, institutions are required to take into consideration a broad set of information, including borrower risk characteristics, (borrower type, demographics, financial information, etc.), transaction risk characteristics, (product types, collateral information, etc.) and behavior patterns (historical patterns on repayment behavior etc.).

Table 1. Data sets

Size	Data set	No of cases	No of features	No of good clients	No of bad clients	Imbalance ratio (IR)	
Small	German	1,000	20	700	300	2	Low
	Japanese	690	13	361	329	1	
	DS1	3,053	206	2,757	296	9	
Mid	DS2	25,226	218	25,016	210	119	Severe
	DS3	67,800	162	64,831	2,969	22	Moderate
	DS4	65,535	160	62,542	2,993	21	Moderate
	DS5	26,590	198	26,397	193	137	Severe
Large	DS6	205,720	160	198,772	6,948	29	Moderate
	DS7	197,462	160	189,479	7,983	24	Moderate
	DS8	232,275	160	223,569	8,706	26	Moderate

DS3, DS4 and DS6 - DS8 data sets include information on retail clients. Features include sociodemographic characteristics of the client (20 features) such as age, occupation etc.; income and other available funds information (29 features); loan characteristics (14 features) such as products type, collateral details or seasoning information; behavior patterns such as balances (39 features), repayment behavior (47 features) and external information (11 features). DS1-DS2 and DS5 data sets include information on corporate clients. Features reflect details on borrower characteristics such as type, industry, region and size (27 features), financial information such as balance sheet (40 features), or profit and loss details (33 features), cash flows analysis (41 features), debt details (8 features), various ratios (29 features) and trend indicators (14 features), as well as external information (11 features). Every data set includes a variable that indicates whether a default event was observed during the performance period of one year. Clients with outstanding debt for more than 3 months were marked as defaulted.

German and Japanese data sets are frequently used as benchmark datasets, especially for evaluating performance of different methods. Unlike the benchmark datasets, real-life data sets are characterized by a much greater number of features (over 150 features), as well as much greater size and class imbalance. The key characteristics analyzed in this study are sample size, class imbalance and the number of minority class cases. The extent of class imbalance is defined with imbalance ratio (IR), the ratio of the number of majority class cases and minority class cases. Data sets were divided by size, the number of bad clients and imbalance ratio. Those containing less than 10,000 observations were marked

as small sized, while the ones with more than 10,000 but less than 100,000 observations were marked as midsized data sets. Data sets with more than 100,000 observations were marked as large. Except for the Japanese data set, all the data sets are imbalanced, displaying different imbalance ratios. Data sets with IR smaller than 10 were categorized as having a low IR, those with IR between 10 and 50 as moderately imbalanced, while the data sets with IR higher than 50 were categorized as severely imbalanced.

4.2. Data transformation

A graphical illustration of the research process is presented in Figures 1. and 2. Fig 1. illustrates the first part of the study focusing on the impact of class imbalance on accuracy of different algorithms, whereas Fig. 2 illustrates the second one focusing on the impact of sample size. The analysis was performed in the context of a specific algorithm since different classification algorithms are expected to exhibit varying sensitivity on both class distribution and size.

Repeated stratified random sampling, or Monte Carlo cross-validation [30] was used, where the data sets were split randomly into a training and a validation data set four times, generating training samples T_1-T_4 and validation samples V_1-V_4. In each iteration, a case was assigned to either the training data set or the validation data set. Stratified sampling was used, with the cases being assigned either to the training or the validation data set randomly, but keeping the original imbalance ratio constant. The classifiers were built on the training data set, whereas the validation data set was used for performance evaluation. For both experiments the data sets described in Table 1 were split randomly into a training (T_i) and a validation (V_i) data set using the 70%/30% ratio, as shown under Step 2 in Fig. 1 and Fig. 2. The same approach was used throughout the study for all the algorithms.

In order to assess the impact of gradually changing the level of imbalance, the initial training data sets were gradually undersampled to reach target imbalance ratios. Target IR ranged from 90/10, with 10% of the defaulted cases, to 20/80, with 80% of the defaulted cases. Class distribution was altered by random undersampling, where either non-defaulted clients or both defaulted and non-defaulted clients were removed to reach the target imbalance ratio. In other words, the defaulted clients' ratio, for all the data sets was altered by a factor of 10% to create greater disparities in the level of imbalance. Eight new data sets T_{i1}- T_{i8} with different IR were therefore created for each original data set, resulting in overall 320 different train data samples. This stage is illustrated in Fig. 1 under Step 2. It is important to emphasize that only the training data sets were altered, leaving the validation data sets and their underlying class distribution unchanged. Classifiers C_1-C_8 were trained using train samples T_{i1}- T_{i8} (Fig. 1, Step 4.) and applied to the validation set V_i (Fig. 1, Step 5.). The performance metrics, outlined in Chapter 4.6., were calculated for each V_i (Fig. 1, Step 6.). However, with Monte Carlo cross-validation, the final performance metrics, reported in the remainder of the paper, were computed as an average over all validation sets V_i to ensure reliable estimates of the results (Fig. 1, Step 7.).

For the study on the impact of sample size a similar approach with repeated stratified random sampling was used. Data set DS8 was used, as a data set representative of real-life behavioral data samples. Under a specific IR, the size of the sample was progressively decreased, while keeping the IR constant (Fig. 2, Step 3.). For each IR, the sample size was decreased by a factor of 10% until reaching 20% of the original size, followed by a reduction of 5% until reaching 5% of the original size with an additional reduction to 3%

of the original size. As a result, for each subsample T_1- T_8 new train subsamples T_{1j}- T_{8j} of different sizes were created. Classifiers C_{1j}- C_{8j} trained using train samples T_{1j}- T_{8j} (Fig. 2, Step 4.) were applied to the validation set V_i (Fig. 2, Step 5.). The final performance metrics, reported in the remainder of the paper, were computed as an average over all validation sets V_i (Fig. 1, Step 7.).

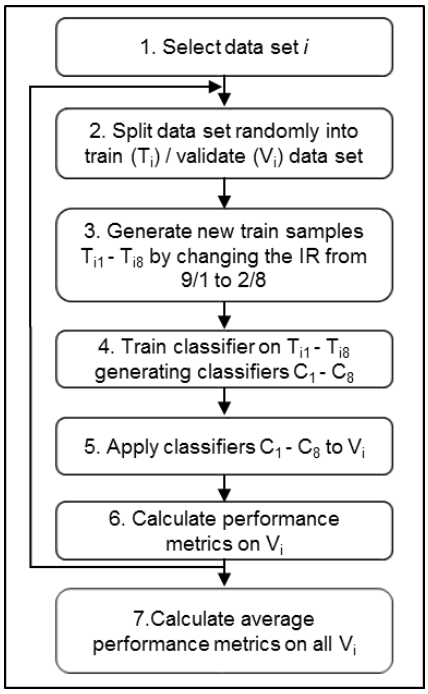


Fig. 1. Class imbalance impact evaluation

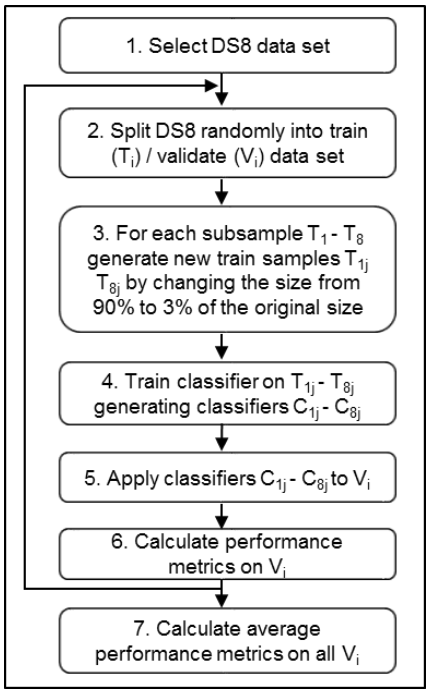


Fig. 2. Sample size impact evaluation

4.3. Classification algorithms and feature selection

For logistic regression we used the stepwise method to select independent variables. To determine the weights for neural networks we used the standard backpropagation algorithm. The hidden layer was set to include 3 units with no direct links between input and output nodes. For gradient boosting we used binary trees and the number of terms in the boosting series equal to 50. Features selection was performed by using Chi-square and R-square analysis. To prevent overfitting, the number of features was limited to 7.

4.4. Performance evaluation metrics

Choosing the right performance metrics when dealing with imbalanced data sets is crucial for appropriate evaluation of the classifier's accuracy [7]. Traditionally, the confusion

matrix was used to compare the predictive outcome of the algorithm with the true values [31].

Table 2. Confusion matrix

		Prediction		Accuracy rate (acc) = (TP+TN)/(TP+TN+FP+FN) True positive rate(TPR) = (TP)/(TP+FN) True negative rate(TNR)= (TN)/(FP+TN) False negative rate(FNR) =(FN)/(TP+FN) False positive rate(FPR) = (FP)/(FP+TN)
		Positive	Negative	
Actual	Positive	TP (true positive)	FN (false negative)	
	Negative	FP (false positive)	TN (true negative)	

From the confusion matrix, several metrics can be induced such as the *Accuracy rate (acc)*, that measures the ratio of correct predictions of the algorithm. When dealing with imbalanced data sets, performance measures insensitive to class distribution should be used. Classical measures, such as *acc*, exhibit an inherent bias toward the majority class, since for example the error of ignoring the minority class completely would be only 2% if the data set contained only 2% of minority class cases. Other, more appropriate metrics of performance can be used, monitoring accuracy for both classes independently. *True positive rate (TPR)* represents the ratio of all the defaulted clients that were correctly classified by the algorithm among all defaulted clients. It is also referred to as sensitivity or recall. *True negative rate (TNR)* is the ratio of all the non defaulted clients that were correctly classified by the algorithm among all non defaulted clients. It is also referred to as specificity. The goal of a classifier is to maximize the true positive and true negative rates. In credit risk assessment, default prediction is of greater interest since misclassifying a bad client attracts a much higher cost than misclassifying a good client.

In this study measures that consider prediction accuracy for both classes were used, the Receiver Operating Characteristic *(ROC)* and corresponding area under the ROC Curve *(AUC)*, as well as geometric mean of the true rates *(GM)*. The two measures represent different types of performance indicators and assess the predictive performance of the classifier from different angles. *ROC* is one of the most frequently used measures. *AUC* represents the probability that a randomly chosen positive case (a defaulted client) will be ranked higher than a randomly chosen negative case (a non defaulted client) [7]. It visualizes the trade-off between sensitivity and 1-specificity [32]. An algorithm that classifies all cases correctly would include point (0, 1) and a random algorithm point (0.5, 0.5). The geometric mean of the true rates measure *(GM)* on the other hand combines measures of correctness of the binary classification predictions, allowing for simultaneous maximization of the prediction accuracy for both classes. It is defined as follows [31]:

$$\text{GM} = \sqrt{TP_{rate} * TN_{rate}} \quad \textbf{(6)}$$

5. Empirical results - the impact of class distribution

5.1. Using geometric mean *(GM)* as performance metric

Tables 3, 4 and 5 summarize the results for all data sets at different IR, with logistic regression shown in Table 3, neural network in Table 4 and gradient boosting in Table 5. All are calculated using the validation data sets. The results show the rank of a given classifier for each imbalance ratio. The imbalance ratio where the highest result is achieved is marked by "1". Average result across all the data sets for a given class distribution is also shown, together with the coefficient of variation (*CV*), or the ratio of the standard deviation of *GM* to the mean *GM*. Fig. 3 presents how the *GM* is affected by the changes in the underlying class distribution for small, medium and large data sets.

Logistic regression - impact of sample size and imbalance ratio. The best results in terms of *GM* are achieved when the data sample is more or less balanced, or when the imbalance ratio ranges from 6/4 to 4/6. Similar results are observed for all data samples, regardless of their size. When analyzing the average *GM* across all data sets for a given imbalance ratio, the best results are achieved when the imbalance ratio is equal to 1. This can be observed also for small sized data sets, whereas for midsized data sets slightly better results on average are obtained when the imbalance ratio is greater than 1, 6/4 and for larger data sets when the imbalance ratio is smaller than 1. Choosing the optimal class distribution increases the *GM* by 55% on average. Overall, logistic regression did not show high sensitivity to IR.

If the absolute number of bad clients is taken into account, it can be observed that data samples with a smaller number of bad clients achieve the best results when the number of good clients is larger than the number of bad clients (6/4). Data samples with an adequate number of bad clients benefit most from a balanced class distribution and the ones with a higher number of bad clients when the imbalance ratio is smaller than 1, i.e. 4/6.

Neural networks - impact of sample size and imbalance ratio. Neural networks display similar performance, where the best result in terms of *GM* is achieved when the data sample is more or less balanced, more specifically when the imbalance ratio ranges from 6/4 to 4/6. However, neural networks show less sensitivity to changes in class imbalance than logistic regression, as can be seen by observing smaller disparities in the coefficient of variation (*CV*). When analyzing the average *GM* across all data sets we observed that a balanced distribution yields the highest accuracy. The same result can be observed for both small and large data sets, whereas for midsized data sets slightly better results on average are obtained when the imbalance ratio is greater than 1, i.e. 6/4. Choosing the optimal class distribution increases the *GM* by 54% on average.

If the absolute number of bad clients is taken into account, data samples with a smaller number of bad clients achieve the best results when the number of good clients is larger than the number of bad clients (6/4). Others benefit most from a balanced class distribution.

Table 3. *GM* rank for logistic regression

Data set	Imbalance ratio								
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8	CV
German	8	7	5	3	1	2	4	6	41%
Japanese	8	7	6	4	1	2	3	5	6%
DS1	8	7	4	2	1	3	5	6	41%
DS2	7	5	4	1	2	3	6	8	30%
DS3	8	7	5	3	1	2	4	6	27%
DS4	8	7	6	4	1	2	3	5	23%
DS5	6	3	2	1	4	5	7	8	40%
DS6	8	7	6	3	2	1	4	5	40%
DS7	7	6	5	4	2	1	3	8	50%
DS8	8	7	5	3	1	2	4	6	23%
AVG	7,6	6,3	4,8	2,8	1,6	2,3	4,3	6,3	16%

Table 4. *GM* rank for neural network

Data set	Imbalance ratio								
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8	CV
German	8	7	5	3	1	2	4	6	26%
Japanese	8	7	6	5	2	1	3	4	6%
DS1	8	7	4	2	1	3	5	6	32%
DS2	8	6	4	1	2	3	5	7	39%
DS3	8	6	3	1	2	4	5	7	23%
DS4	8	7	4	2	1	3	5	6	17%
DS5	7	4	2	1	3	5	6	8	40%
DS6	8	6	4	2	1	3	5	7	26%
DS7	8	7	5	3	1	2	4	6	40%
DS8	8	7	4	1	2	3	5	6	24%
AVG	7,9	6,4	4,1	2,1	1,6	2,9	4,7	6,3	12%

Table 5. *GM* rank for gradient boosting

Data set	Imbalance ratio								
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8	CV
German	8	7	5	4	2	1	3	6	40%
Japanese	8	7	6	1	2	3	3	3	8%
DS1	6	5	7	7	1	2	3	4	86%
DS2	5	5	5	5	1	2	3	4	120%
DS3	5	5	5	5	2	1	3	4	119%
DS4	5	5	5	5	3	1	2	4	119%
DS5	5	5	5	5	2	1	3	4	119%
DS6	5	5	5	5	2	1	3	4	119%
DS7	5	5	5	5	2	1	3	4	119%
DS8	5	5	5	5	2	1	3	4	119%
AVG	5,7	5,4	5,3	4,7	1,9	1,4	2,9	4,1	75%

Table 6. Highest average *GM*

	Imbalance ratio							
SAMPLE	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8
ALL	NN	NN	NN	NN	GB	GB	GB	GB
SMALL	NN	GB	GB	LR	LR	GB	GB	NN
MID	NN	NN	NN	NN	GB	GB	GB	GB
LARGE	NN	NN	NN	NN	GB	GB	GB	GB
REAL LIFE	NN	NN	NN	NN	GB	GB	GB	GB

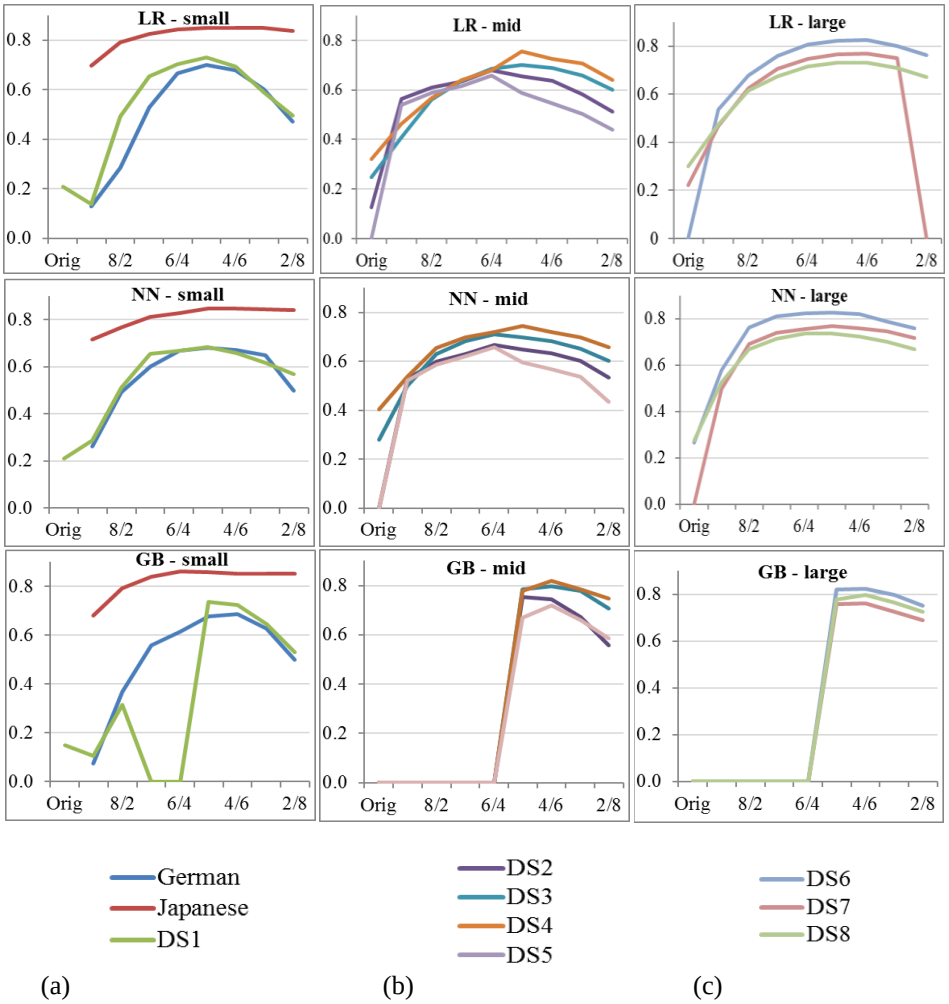


Fig. 3. *GM* for logistic regression (LR), neural network (NN) and gradient boosting (GB) for (a) small data sets, (b) mid data sets and (c) large data sets

Gradient boosting - impact of sample size and imbalance ratio. Table 5 clearly demonstrates that with gradient boosting the best results are also achieved when the data

sample is more or less balanced, with better results achieved when the imbalance ratio is smaller than one, 4/6. The same result can be observed for both mid and large data sets, whereas for small sized data sets slightly better results on average are obtained when the imbalance ratio is equal to 1. Gradient boosting displays much greater sensitivity to class imbalance than the previous two methods, as can be seen from the *CV*, but also from the fact that in vast majority of the cases when class imbalance is greater than 1, not a single bad client was identified by gradient boosting. As a result, *GM* is equal to 0%. This is especially relevant for real-life data sets (DS1-DS8). Choosing the optimal class distribution increases the *GM* by 75% on average. Similar results can be observed if the absolute number of bad clients is analyzed.

Comparison of class distribution across all classifiers. Table 6 shows the classification algorithm achieving the best results on average for a given imbalance ratio. The results are shown separately for all the data sets, depending on the size of the data set, and for real life data sets. This finding suggests that when imbalance ratio is greater than 1, neural networks in general achieve the best results, whereas for imbalance ratios smaller than 1 gradient boosting achieves the best results.

Kendall's coefficient of concordance (*W*) was used to compare the *GM*s with different class distributions. It is a non-parametric statistics based on the average ranked performance of the classification, and is calculated as follows:

$$\mathrm{W} = \frac{12S}{m^2(k^3-k)} \quad . \tag{7}$$

where *S* is the sum of squared deviations, equal to:

$$S = \sum_{i=1}^{k}(R_i - R)^2 \, . \tag{8}$$

R_i is the average rank across all classifiers, *m* the number of classifiers used to perform ranking (3) and *k* is the number of assessments subject to ranking (8). *W*= 0.81 has shown a level of agreement for all the algorithms, with *p-value* = 0.0174 < .05 = α, allowing rejection of the null hypothesis that there is no agreement among the different classification algorithms.

5.2. Using Area under the ROC curve (*AUC*) as performance metric

A similar analysis is displayed in Tables 7, 8 and 9, with *AUC* as the evaluation measure.

Logistic regression - impact of sample size and imbalance ratio. Logistic regression has shown to be insensitive to class distribution and a fairly robust technique. There is no single class distribution that yields the best performance for all the samples. However, larger disparities can be observed when the sample size and the actual number of defaulted cases is taken into account, especially with real-life data sets. For larger data samples the highest *AUC* is achieved when IR is smaller than 1 (more specifically 3/7).

For midsized data samples, the best results are achieved when the data sample is balanced, 5/5, whereas small data sets produce the best results when the IR is greater than 1 (6/4).

Table 7. *AUC* rank for logistic regression

Data set	Imbalance ratio								
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8	CV
German	8	6	5	1	4	3	1	7	1%
Japanese	5	1	4	3	6	2	7	8	0%
DS1	1	4	3	2	5	6	8	7	2%
DS2	1	2	5	4	7	3	6	8	3%
DS3	8	7	6	2	4	2	1	5	1%
DS4	8	7	6	4	1	2	4	3	0%
DS5	4	5	7	2	1	3	6	8	2%
DS6	8	7	5	5	4	3	1	1	0%
DS7	8	7	5	5	2	2	1	2	1%
DS8	7	6	5	8	3	1	3	2	1%
AVG	5,8	5,2	5,1	3,6	3,7	2,7	3,8	5,1	0%

Table 8. *AUC* rank for neural networks

Data set	Imbalance ratio								
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8	CV
German	7	1	2	3	4	5	6	8	3%
Japanese	8	3	1	6	2	4	5	7	1%
DS1	2	3	1	4	5	7	5	8	2%
DS2	2	1	3	4	7	4	6	8	3%
DS3	2	1	4	5	3	7	6	8	0%
DS4	2	1	5	6	4	3	8	7	0%
DS5	1	3	5	2	4	7	6	8	2%
DS6	2	2	1	2	5	5	7	8	0%
DS7	2	4	2	6	6	4	1	8	0%
DS8	1	1	6	3	4	7	4	8	0%
AVG	2,9	2,0	3,0	4,1	4,4	5,3	5,4	7,8	1%

Table 9. *AUC* rank for gradient boosting

Data set	Imbalance ratio								
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8	CV
German	8	2	3	4	6	1	5	7	3%
Japanese	2	1	5	4	3	7	8	6	1%
DS1	5	2	7	8	3	1	3	6	2%
DS2	7	6	7	5	1	1	3	4	6%
DS3	6	7	8	5	4	3	1	2	4%
DS4	8	6	7	5	4	2	1	3	4%
DS5	5	7	8	6	4	1	2	3	6%
DS6	8	6	7	5	4	1	2	3	15%
DS7	8	6	6	5	3	2	1	4	14%
DS8	8	7	6	5	4	2	1	3	15%
AVG	6,5	5,0	6,4	5,2	3,6	2,1	2,7	4,1	7%

If the actual number of defaulted cases is taken into account, it can be observed that data samples with a smaller number of bad clients achieve the best results when the number of good clients is much larger than the number of bad clients (8/2). Data samples with an adequate number of bad clients benefit most from a balanced class distribution and the ones with a higher number of bad clients benefit when the imbalance ratio is smaller than 1, 3/7.

Neural networks - impact of sample size and imbalance ratio. Neural network exhibits different behavior. With neural networks the highest *AUC* is achieved when the IR is significantly greater than 1, or more specifically when the ratio of good/bad clients ranges from 9/1 to 7/3. The only exception is the DS7 data set, where the distribution of good/bad clients of 3/7 yields the highest *AUC*. Similar to logistic regression, it seems that the sample size plays an important role when it comes to determining the optimal class distribution. For larger data samples the best result is obtained when the imbalance ratio is less than one, 3/7, whereas midsized data samples tend to produce the best results when the class distribution is equal to 8/2. Small data sets produce the best results when the imbalance ratio is 7/3. Nevertheless, neural networks did not show significant variation in performance depending on the imbalance ratio, as demonstrated by the coefficient of variation (*CV*), which is relatively small for all data sets.

If the actual number of minority class examples is considered, data samples with a smaller number of bad clients achieve the best results when the number of good clients is much larger than the number of bad clients (8/2).

Gradient boosting - impact of sample size and imbalance ratio. For gradient boosting class distribution ranging between 5/5 to 3/7 generally yields the best performance in terms of *AUC*. Gradient boosting is more sensitive to class imbalance than neural networks and logistic regression. When it comes to choosing the optimal class distribution, greater disparities are observed, depending on the sample size, especially with real-life data sets. Gradient boosting produces the best results on average when the imbalance ratio is greater than 1, or more specifically when the imbalance ratio equals 3/7 for larger data samples and 4/6 for midsized data samples. Class distribution of 8/2 yields best performance in terms of *AUC* for smaller data sets.

Data samples with a small number of bad clients achieve the best results when the number of good clients is smaller than the number of bad clients (4/6). Other data samples show the best results when the imbalance ratio is even lower (3/7).

Comparison of class distribution across all classifiers. We observed that regardless of the classification algorithm, changes in the underlying class distribution do not have a great impact on *AUC*. In fact, it seems that imbalance per se does not necessarily cause bad performance. As expected, when presented with large class imbalances, algorithms do not exhibit high classification accuracy, especially when a small number of minority class observations exist. In general, using a more balanced class distribution improves performance, which is especially evident with gradient boosting. Table 10 shows the classification algorithm achieving the best results on average for a given imbalance ratio. The results are shown separately for all the data sets, depending on the size of the data set, and for real-life data sets. This finding suggests that when imbalance ratio is greater than 1, neural networks in general achieve the best results, whereas for imbalance ratios smaller than 1 gradient boosting achieves the best results.

Table 10. Highest average *AUC*

	Imbalance ratio							
SAMPLE	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8
ALL	NN	NN	NN	LR	GB	GB	GB	GB
SMALL	LR	GB	GB	LR	GB	GB	GB	GB
MID	NN	NN	NN	GB	GB	GB	GB	GB
LARGE	NN	NN	NN	NN	NN	GB	GB	GB
REAL LIFE	NN	NN	NN	NN	GB	GB	GB	GB

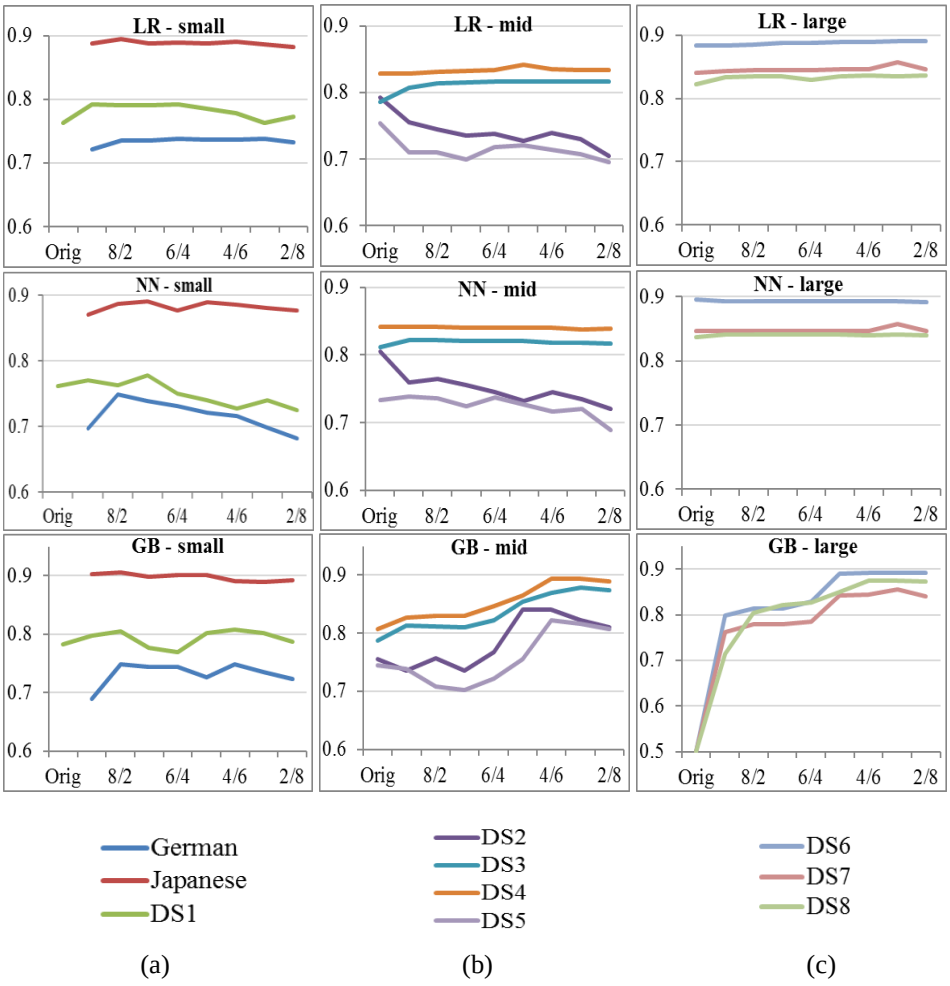


Fig. 4. *AUC* for logistic regression (LR), neural network (NN) and gradient boosting (GB) for (a) small data sets, (b) mid data sets and (c) large data sets

Kendall's coefficient of concordance (*W*) was used to compare the average *AUC* across three classification algorithms and eight class distributions. *W*= 0.24 has shown a low level of agreement for all the algorithms as regarding the ranking of different class

distributions, with *p-value* > .05 = α, not allowing rejection of the null hypothesis that there is no agreement among the different classification algorithms.

In any case, we observed that undersampling increases classification accuracy, with the largest impact observed with the gradient boosting algorithm. The impact of class distribution for samples of different sizes is displayed in Fig. 4.

5.3. Comparison of results across performance metric

The results presented in chapters 5.1. and 5.2. reveal that *GM* and *AUC* in some cases indicate a different imbalance ratio that provides the best performance for the same algorithm and data set. This finding can be explained by the fact that the two measures reflect different aspects of classifiers' performance. *AUC* reflects the ability of the classifier to rank the cases from negative to positive in the correct order. Although having appealing properties, such as averaging the performance over all possible thresholds and being objective (i.e. no subjective input from the user is required), there are some disadvantages to using *AUC*. For example, *AUC* can give misleading results if *ROC* curves cross. It can also provide a somewhat incoherent measure of performance, as *AUC* presumes different misclassification cost distributions for different classifiers [32].

On the other hand, *GM* measures whether a correct binary classification has been made, indicating the balance between classification performances on both classes. As a performance indicator, it is similar to accuracy rate. Even if positive cases are correctly classified, poor prediction accuracy of the negative cases will lower the *GM* value. Consequently, the *GM* measure can be used to avoid overfitting to the positive class, where the negative class becomes marginalized. However, *GM* does not distinguish the contribution of each class to the overall performance, as different combinations of *TPR* and *TNR* can produce the same *GM* value. Also, these types of threshold indicators disregard the absolute values of predictions. If the estimate is higher than the threshold, it is irrelevant what is the actual estimated probability is.

To get a clearer view on the alignment of the two measures, a correlation analysis of the rankings of the imbalance ratios suggested by both measures was performed, outlined in Table 11. The high correlation illustrates overall large agreement between the results for LR and GB on which imbalance ratio would yield the best performance. Somewhat lower agreement can be observed for small data sets. On the other hand, the results for NN are quite different and negative correlation between the measures indicates a disagreement between the measures. Therefore, for NN specifically, *GM* could offer an additional angle from which to prediction accuracy can be assessed.

In general, using different types of evaluation metrics when comparing different classifiers' performance is advisable, because a single metric cannot capture all related important aspects [33]. Nevertheless, when faced with conflicting results presented by *GM* and *AUC*, *AUC* could be considered as more appropriate since it enables measuring performance of a classifier and its discriminatory ability over its entire operating range. Credit institutions stand to benefit a great deal from having relative ranks of the clients determined accurately, since it enables a better differentiation of the degree of credit risk and a more accurate assessment of the portfolio and expected losses [34].

Table 11. Pearson's correlation coefficients among *GM* and *AUC* for imbalance ratios rankings

Data set	LR	NN	GB
German	0.725	0.725	0.405
Japanese	-0.190	0.238	-0.518
DS1	-0.119	-0.232	0.753
DS2	0.048	-0.179	0.946
DS3	0.763	-0.071	0.627
DS4	0.942	-0.286	0.848
DS5	0.429	0.381	0.811
DS6	0.572	0.067	0.811
DS7	0.535	-0.253	0.806
DS8	0.431	-0.313	0.738

6. Empirical results - the impact of sample size and the number of minority class cases

In addition to the impact of the IR, a detailed study of the impact of the sample size was conducted. For each IR, the sample size was decreased by a factor of 10% until reaching 20% of the original size, followed by a reduction of 5% until reaching 5% of the original size. An additional reduction to 3% was performed to illustrate inconsistencies that occur when using very small samples. Obviously, with smaller samples a greater deviation from the original population is introduced.

Based on the results outlined in the previous chapter related to class distribution, the impact of sample size was evaluated for IR ranging from 2.33 (70/30) to 0.43 (30/70), since the best performance for this data set was achieved at these IRs. Thus, for each IR additional 12 data sets of different sizes were created.

6.1. Using geometric mean (*GM*) as performance metric

Fig. 5 presents the impact of varying sample sizes on *GM* under imbalance ratios ranging from 2.33 (70/30) to 0.43 (30/70) by displaying the relative change of *GM* when compared to the sample of the maximum size for each of the algorithms.

The results document several important findings. A general trend of decreasing accuracy when the sample size is decreased can be observed, especially when the sample is more or less balanced. Surprisingly, the impact is much smaller than expected. Also, decreasing the sample size has a much greater impact on NN than LR. In addition to class imbalance, LR has demonstrated low sensitivity to sample size. Decrease in *GM*, compared to using the whole sample with all defaulters (100% of the size) did not exceed 1.5%. These results confirm suitability of using LR with small samples. NN exhibits greater sensitivity to the sample size and number of defaulters, with a much greater decrease in *GM* observed when faced with smaller samples. GB did not exhibit high sensitivity to sample size and seems to be more affected with the class imbalance than the sample size. NN and GB displayed unusual behavior when dealing with samples where

defaulters outnumber non defaulters. It appears that reducing the sample size unexpectedly leads to improvements with and a general trend of increased accuracy.

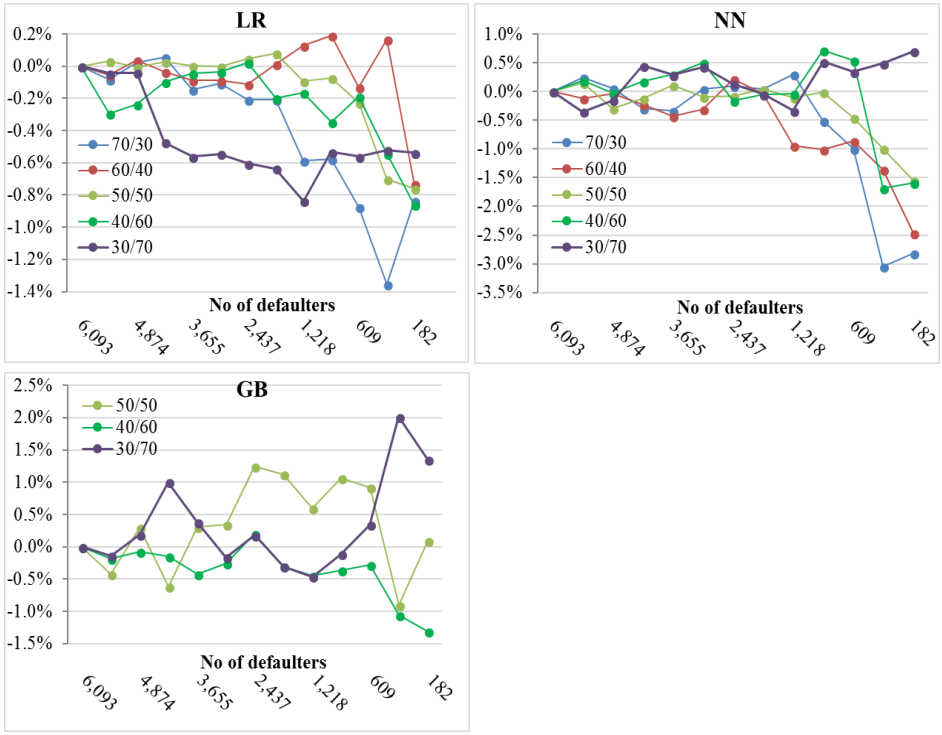


Fig. 5. *GM* % change for IR ranging from 70/30 to 30/70

When measuring performance with *GM*, NN outperforms LR when the good clients outnumber the bad clients, whereas LR outperforms NN when bad clients outnumber the good clients, regardless of the number of defaulters. However, when the sample is balanced, LR outperforms NN when the sample is small (less than 600 defaulters).

6.2. Using Area under the ROC curve (*AUC*) as performance metric

Table 12 presents the impact of reducing sample size to 20% or less of the original size on *AUC* under imbalance ratios ranging from 2.33 (70/30) to 0.43 (30/70). It enables comparison of different algorithms when using the same sample, as well as relative impact on performance when compared to the sample of the maximum size (size=100%) for each of the algorithms. Paired t-tests were conducted to evaluate if the performance when using the reduced sample differs significantly from using the entire sample.

Table 12. *AUC* - Impact of sample size with IR ranging from 70/30 to 30/70

LR											
SIZE	No Bad	IR = 2.33		IR = 1.50		IR = 1.00		IR = 0.67		IR = 0.43	
100%	6,093	0,84	***	0,84	***	0,84	***	0,84	***	0,84	***
20%	1,218	0.84	***	0.84	***	0.84	***	0.84	***	0.84	***
15%	913	0.84	***	0.84	***	0.83	***	0.84	***	0.83	***
10%	609	0,83	*	0,84	***	0,83	***	0,84	***	0,83	***
5%	304	0,83	*	0,83	***	0,83	***	0,83	***	0,83	*
3%	182	0,83		0,83	***	0,83	***	0,83	*	0,83	***
NN											
SIZE	No Bad	IR = 2.33		IR = 1.50		IR = 1.00		IR = 0.67		IR = 0.43	
100%	6,093	0,84	***	0,84	***	0,84	***	0,84	***	0,84	***
20%	1,218	0.84	***	0.84	*	0.84	*	0.84	*	0.84	*
15%	913	0.84	***	0.84	***	0.84	*	0.84	*	0.83	*
10%	609	0,84		0,84	***	0,83	***	0,83		0,83	*
5%	304	0,83		0,83		0,83	*	0,83		0,83	***
3%	182	0,82	*	0,83		0,82		0,81	*	0,82	*
GB											
SIZE	No Bad	IR = 2.33		IR = 1.50		IR = 1.00		IR = 0.67		IR = 0.43	
100%	6,093	0,82	***	0,82	***	0,84	***	0,87	***	0,88	***
20%	1,218	0.82	***	0.82	***	0.85	***	0.87	***	0.87	*
15%	913	0.82	***	0.82	***	0.86	***	0.87	***	0.87	*
10%	609	0,81	*	0,81	***	0,86	***	0,87	***	0,87	*
5%	304	0,82	***	0,82		0,85	***	0,87	***	0,87	*
3%	182	0,81		0,81	*	0,84	***	0,87	*	0,86	*

*** accuracy is not significantly different from the accuracy at size = 100% at α=95%
* accuracy is not significantly different from the accuracy at size = 100% at α=99%

Fig. 6 shows the relative change in *AUC* when compared to the entire sample with all defaulters included graphically.

The results document several important findings. With LR and NN a clear trend of decreasing accuracy when the sample size is decreased can be observed. Surprisingly, the impact is much smaller than expected and in almost all cases accuracy for small sample sizes of 600 or 300 defaulted clients, is not significantly different from the accuracy at size = 100%. This finding is surprising, given that it contradicts the widely recommended using of 1,500 – 2,000 cases. The finding that decreasing the size of the sample with this magnitude does not cause significant changes in performance, represents a novel and very useful discovery for credit scoring. In addition to reducing cost and providing more flexibility when choosing a training sample, it might also encourage practitioners to engage in model development for low default portfolios. For LR the decrease in *AUC*, compared to using the whole sample with all defaulters (100% of the size) never exceeds 1%. In almost all cases accuracy was not significantly different statistically from the accuracy achieved at size = 100%, even for very small samples. NN on the other hand is more affected by the sample size. However, *AUC* did not decrease by more than 4% even when faced with very small samples. When measuring performance with *AUC*, NN again outperforms other algorithms LR when the good clients outnumber the bad clients, whereas GB outperforms NN and LR when bad clients outnumber the good clients,

regardless of the sample size. However, when the sample is balanced, NN outperforms GB when the sample is larger (more than 4000 defaulters). GB did not exhibit high sensitivity to sample size and seems to be more affected with the class imbalance than the sample size. GB displayed unusual behavior when dealing with balanced sample, where reducing the sample size surprisingly lead to improvements in prediction accuracy.

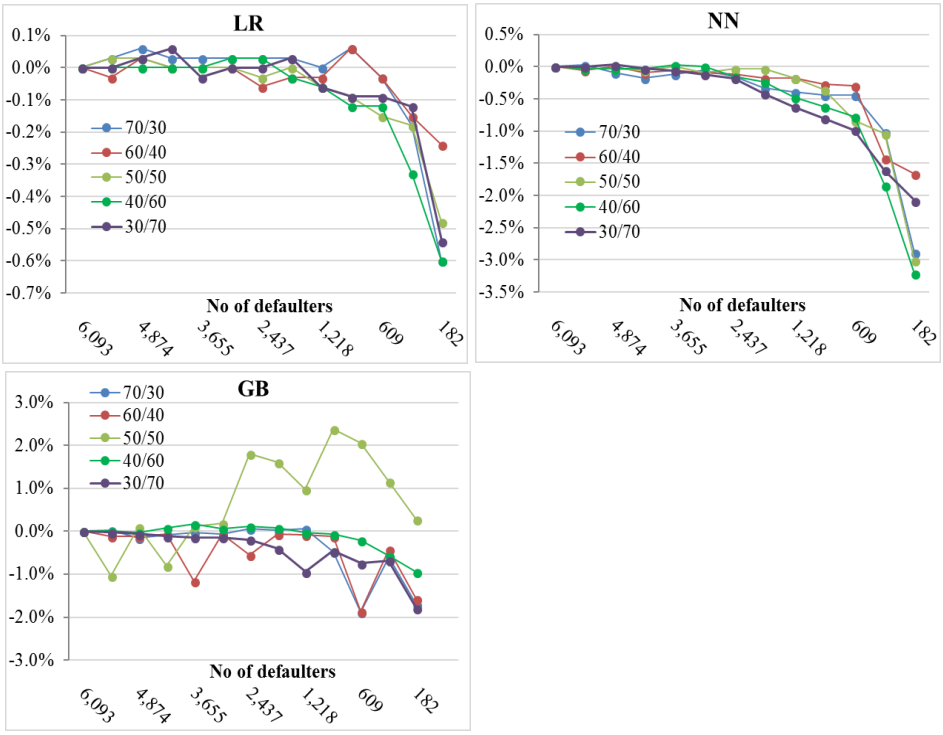


Fig. 6. *AUC* % change for IR ranging from 70/30 to 30/70

The results also confirm that GB is not suitable for dealing with imbalanced data sets, contradicting some of the previous research [17]. LR performs well when using small samples, indicating that samples smaller than the widely accepted benchmark of 1,500 defaulters can be used to develop models of good performance.

7. Conclusion

In this article, a comprehensive experimental study on the effect of sample size and class distribution in credit scoring was presented, using logistic regression, neural network and gradient boosting algorithm on a wide array of real life data samples.

All algorithms have shown a decrease in performance when challenged with higher imbalance ratios. The results suggest that gradient boosting might not be an appropriate classification algorithm when dealing with imbalanced data sets, while logistic regression

and neural networks are fairly insensitive to changes in the class distribution. Neural networks displayed the best performance on average when dealing with higher imbalance ratios. The results also indicate that class imbalance by itself does not necessarily cause a reduction in classification accuracy, and that sample size and classification algorithm play an important role when it comes to determining the optimal class distribution, especially the absolute number of minority class cases.

Classification algorithms have also shown different levels of sensitivity to sample size. With logistic regression and neural network a clear trend of decreasing accuracy was observed when the sample size was decreased. However, the impact was much smaller than expected, suggesting that decreasing sample size significantly, even down to as few as 300 defaulted cases does not cause a significant decline in performance, and represents an important discovery for credit scoring. Gradient boosting did not exhibit high sensitivity to sample size and seems to be more affected by the class imbalance than the sample size.

Given the recent advances in automating lending decision processes, the ever-growing richness in data collected, the big data trend and capital adequacy optimization, we foresee this to remain a very active research area. Leveraging on these findings, we plan to conduct experiments using other classification algorithms, such as random forests, and other, more sophisticated data pre-processing techniques.

References

1. Thomas, L. C., Edelman, D. B., and Crook, J. N., Credit Scoring and Its Application. SIAM, (2002).
2. Schmid, B., Credit Risk Pricing Models: Theory and Practice. Springer Finance, (2011).
3. Hand, D. J. and Henley, W. E., Statistical Classification Methods in Consumer Credit Scoring: a Review, J. R. Stat. Soc. Ser. A (Statistics Soc., vol. 160, no. 3, pp. 523–541, (1997).
4. Yang, Q., 10 challenging problems in data mining research, Int. J. Inf. Technol. Decis. Mak., vol. 5, no. 4, pp. 597–604, (2006).
5. Van Hulse, J. and Khoshgoftaar, T., Knowledge discovery from imbalanced and noisy data, Data Knowl. Eng., vol. 68, no. 12, pp. 1513–1542, (2009).
6. Siddiqi, N., Credit Risk Scorecards, John Wiley Sons, Inc., vol. 1, pp. 1–210, (2006).
7. García, V., Marqués, A. I., and Sánchez, J. S., An insight into the experimental design for credit risk and corporate bankruptcy prediction systems, J. Intell. Inf. Syst., vol. 44, no. 1, pp. 159–189, (2015).
8. Lessmann, S., Baesens, B., Seow, H. V., and Thomas, L. C., Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, Eur. J. Oper. Res., vol. 247, no. 1, pp. 124–136, (2015).
9. Malhotra, R. and Malhotra, D. K., Evaluating consumer loans using neural networks, Int. J. Manag. Sci., vol. 31, no. 2, pp. 83–96, (2003).
10. Ala'raj, M. and Abbod, M. F., A new hybrid ensemble credit scoring model based on classifiers consensus system approach, Expert Syst. Appl., vol. 64, no. July, pp. 36–55, (2016).
11. Feng, X., Xiao, Z., Zhong, B., Qiu, J., and Dong, Y., Dynamic ensemble classification for credit scoring using soft probability, Appl. Soft Comput., vol. 65, pp. 139–151, (2018).
12. Abellán, J. and Castellano, J. G., A comparative study on base classifiers in ensemble methods for credit scoring, Expert Syst. Appl., vol. 73, pp. 1–10, (2017).
13. Fitzpatrick, T. and Mues, C., An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market, Eur. J. Oper. Res., vol. 249, no. 2, pp. 427–439, (2016).

14. Kvamme, H., Sellereite, N., Aas, K., and Sjursen, S., Predicting Mortgage Default using Convolutional Neural Networks, Expert Syst. Appl., vol. 102, (2018).
15. Baesens, B., Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J., Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring, J. Oper. Res. Soc., vol. 54, no. 6, pp. 627–635, (2003).
16. Peng, Y., Wang, G., Kou, G., and Shi, Y., An empirical study of classification algorithm evaluation for financial risk prediction, Appl. Soft Comput., vol. 11, no. 2, pp. 2906–2915, (2011).
17. Brown, I. and Mues, C., An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Expert Syst. Appl., vol. 39, pp. 3446–3453, (2012).
18. López, V., Fernández, A., García, S., Palade, V., and Herrera, F., An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Inf. Sci. (Ny)., vol. 250, pp. 113–141, (2013).
19. Batista, G., Prati, R., and Monard, M., A study of the behavior of several methods for balancing machine learning training data, ACM Sigkdd Explor. Newsl., vol. 6, no. 1, pp. 20–29, (2004).
20. Crone, S. F. and Finlay, S., Instance sampling in credit scoring: An empirical study of sample size and balancing, Int. J. Forecast., vol. 28, no. 1, pp. 224–238, (2012).
21. García, V., Sánchez, J. S., and Mollineda, R. a., On the effectiveness of preprocessing methods when dealing with different levels of class imbalance, Knowledge-Based Syst., vol. 25, no. 1, pp. 13–21, (2012).
22. Weiss, G. M. and Provost, F., Learning when training data are costly: The effect of class distribution on tree induction, J. Artif. Intell. Res., vol. 19, pp. 315–354, (2003).
23. Andrić, K. and Kalpić, D., The effect of class distribution on classification algorithms in credit risk assessment, in 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2016, (2016), p. 7.
24. Hosmer, D. W. and Lemeshow, S., Applied Logistic Regression, *A Wiley-Interscience*. p. 375, 2004.
25. Bishop, C. M., Neural networks for pattern recognition, J. Am. Stat. Assoc., vol. 92, p. 482, (1995).
26. Friedman, J. H., Stochastic gradient boosting, *Computational Statistics and Data Analysis*, vol. 38, no. 4. pp. 367–378, 2002.
27. Dheeru, D. and Karra Taniskidou, E., {UCI} Machine Learning Repository. 2017.
28. European Parliament, Regulation (EU) No 575/2013, Off. J. Eur. Union, no. June 2013, pp. 338–436, (2013).
29. European Banking Agency, Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures, 2016.
30. Werner Dubitzky, Martin Granzow, D. P. B., Fundamentals of Data Mining in Genomics and Proteomics. Springer, (2007).
31. López, V., Fernández, A., García, S., Palade, V., and Herrera, F., An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Inf. Sci. (Ny)., vol. 250, pp. 113–141, (2013).
32. Hand, D. J., Measuring classifier performance: A coherent alternative to the area under the ROC curve, Mach. Learn., vol. 77, no. 1, pp. 103–123, (2009).
33. Japkowicz, N. and Shah, M., Evaluating Learning Algorithms: A Classification Perspective, Eval. Learn. Algorithms. A Classif. Perspect., (2011).
34. Chen, N., Ribeiro, B., and Chen, A., Financial credit risk assessment: a recent review, Artif. Intell. Rev., vol. 45, no. 1, pp. 1–23, (2016).

Kristina Andric received her MSc degree in computing from the Faculty of electrical engineering and computing, University of Zagreb, Croatia in 2006. She is currently pursuing her PhD in computer science at Faculty of electrical engineering and computing, University of Zagreb. Her primary research interests are data science, data mining and machine learning, with particular interest in algorithms used in credit risk management. She is currently head of the Credit risk models and Credit policy and methodology departments at a commercial bank.

Damir Kalpić graduated, received his PhD degree in computer science and worked as a professor at the University of Zagreb Faculty of electrical engineering and computing, Croatia. His main professional interest is in application of operational research in information systems, and in other applications with humans as direct users. He has published more than 100 internationally reviewed papers in journals and conference proceedings and lead about 40 R&D projects for industry, education, medicine, various services and administration. He has advised a few hundred students at student seminars, projects, graduation and PhD theses. He served as vice dean of the Faculty and was the first head of Department of applied computing after its establishment.

Zoran Bohaček graduated and received his PhD degree in computer science at the University of Zagreb Faculty of electrical engineering and computing, Croatia and his master' degree at Harvard University. He worked at McGill University, Bell Northern Research and served as Director of Operations at Fair Isaac International. He is now the Chief Advisor in Croatian Banking Association, where he also served as a Managing Director, and a member of the Management Board of the Croatian credit registry. He is also an Executive Committee member at the European Banking Federation. He has been teaching graduate course on quantitative methods in risk management at the University of Zagreb Faculty of electrical engineering and computing, Croatia.

Received: January 10, 2018; Accepted: October 10, 2018

Computer Science and Information Systems 16(1):179–203 https://doi.org/10.2298/CSIS171217012S

Towards Understandable Personalized Recommendations: Hybrid Explanations

Martin Svrcek, Michal Kompan, and Maria Bielikova

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovicova 2, 841 04 Bratislava, Slovakia
{name.surname}@stuba.sk

Abstract. Nowadays, personalized recommendations are widely used and popular. There are a lot of systems in various fields, which use recommendations for different purposes. One of the basic problems is the distrust of users of recommended systems. Users often consider the recommendations as an intrusion of their privacy. Therefore, it is important to make recommendations transparent and understandable to users. To address these problems, we propose a novel hybrid method of personalized explanation of recommendations. Our method is independent of recommendation technique and combines basic explanation styles to provide the appropriate type of personalized explanation to each user. We conducted several online experiments in the news domain. Obtained results clearly show that the proposed personalized hybrid explanation approach improves the users' attitude towards the recommender, moreover, we have observed the increase of recommendation precision.

Keywords: recommendations explanation, eye-tracking, collaborative filtering, personalized recommendation.

1. Introduction

Nowadays, information is an important part of our lives. The information we have available influences our decisions, opinions and ideas. Therefore, it is important that the available information is suitable and appropriate to our interests. Currently, there are lots of sources from which we can derive information, such as newspapers, books, television, friends or relatives. The availability of information is much higher than it was in the past, which is mainly due to the growth of the Internet[1].

The Web is integrated into everyday life and provides a large amount of information, which in most cases is freely available. People use the Internet to obtain information regardless of time or their physical location. However, the vast amount of data in the form of text, images, recordings or video brings many advantages but also some disadvantages (e.g., availability of information, amount of information).

In this context, there is a very important concept of information availability. Different search engines on the web allow us to get information quickly. Unlike books, we get specific information without a long search in the content or the registry.

The variety of information provided is another important aspect. The Internet provides a place where we can meet the different opinions, thoughts or different views on certain

[1] https://www.internetlivestats.com/total-number-of-websites/

issues. Internet is also a place of different qualities and styles of information such as professional, scientific or artistic as well.

The amount of information as another aspect has two main points of view. A lot of information on the Internet is on one side the positive feature, but on the other side it often results in information overload of users. Information overload relates to the problems of making certain decisions or understanding of certain issues due to the vast amount of information that is offered to the person concerned [36].

Information overload, unsuitability or poor quality of information causes a form of distrust among users. Often, the website displays information, products or services that are not interesting from the perspective of particular user or there are so much of these information that the user cannot find the important pieces. Problems like these can be solved by use of personalization.

Personalized Web minimizes the fundamental problem of classical websites and applications. This problem is that they provide the same information to all users without distinction of their knowledge or passions [7].

Personalized Web is based on the concept of adaptive hypermedia. This concept is related to creating, maintaining and using a user model as a base for adapting to the needs and preferences of particular users [7]. Just as there are several types of information that can be used for personalization such as search results, content or design, there exist a number of ways or opportunities of the personalization realization. One of these opportunities are recommendations as a form of an adaptation of the content to the needs of a particular user.

However, recommender systems in general often produce (recommend to the user) some type of information, product, article etc. without any explanation why they think that this particular item is suitable for particular user. Users usually do not know how these systems deduced presented recommendation and more importantly how recommender systems deduced information about the users, and how they used this information. We can speak about some kind of distrust of users in recommendations [41]. The problem of the trust in the human-computer interaction is a widely researched topic nowadays [38], as the need for such interaction is increasing day-by-day (often including new groups of users, e.g., elder people). Similarly, social media receive tremendous research attention in revealing, simulating and usage of the trust in such a complex environment [31,21].

One of the solutions for this type of problem is an explanation of recommendations, which has the power to clear the process of finding the recommended items and make the whole system more understandable. In this paper, we address the following research questions:

- RQ1: Can we increase the understandability and precision of recommendations with use of explanations?
- RQ2: Is there one explanation style that is preferred by users?

In order to find an answer to these questions, we have proposed a novel method of explaining recommendations. Several experiments were conducted with the effort to evaluate our method. Our contributions that we present in this paper are:

- Hybrid method of explanation of recommendations that combines different explanation styles.

– Improvement of precision for lower-positioned items by using proposed explanation method.

The paper is organized as follows. The brief overview of basic recommender approaches is in section 2. Section 3 is about explanations, reasons why we use them and also about different styles and attributes of explanations. In section 4 we analyze different approaches to recommendations and explanations related to them. Important part of this section is also detailed analysis of explanation approaches used in real web sites. Section 5 presents our novel method to explaining recommendation items. We evaluated our approach by user study which evaluation process consisting of experiment, its setup and results is described in Section 5. Summary and future works is in section 7.

2. Recommender Systems

Main purpose of recommender system is to provide (recommend) objects that would be helpful and suitable for the user [28]. Nowadays, such systems are relatively common and affect different domains as entertainment, content, e-commerce, etc.

In the context of the problem of information overload (which is related to various activities [12]), it is necessary to somehow reduce this overload. Our goal as users is to facilitate and simplify the process of selecting a particular item (ask friends, search for reviews). Recommender systems try to automate these activities (asking friends, etc.) and offer a recommendation, which should also be appropriate and interesting for the user [18].

There are many different types of recommended systems. Many of these systems are personalized to user needs. Personalized recommenders use personal preferences to generate recommendations. Source of such features and preferences is mainly implicit feedback (e.g., time spent, purchase, click) and explicit feedback (e.g., like, review).

There is a lot of systems that employ some of these methods to achieve better results [20]. Generally, the explicit feedback is considered more accurate then implicit feedback [3]. Thus, when choosing a method it is necessary to take into account the characteristics of the system and the opportunities that this system offers us according to specific domain.

Recommendations can be applied as part of systems in many different areas like news articles, movie databases, video portal or e-commerce. They are useful in each of these areas differently. Today, we can point to the different roles that recommender systems play and the different reasons why various service providers use them [28]: increase of items sold, diversity of sale, customer satisfaction or customer loyalty.

The most important part of the recommendation is the prediction of items that may be interesting to the user. Currently, there are several approaches that use different principles and methods to generate assumptions about what can help the user.

Collaborative filtering is considered the most popular and most used solution [8]. The basic idea of collaborative filtering recommendation is to recommend items that are interesting for users with similar interests [28]. An important part is therefore to identify these groups of users based on their preferences. In most cases, collaborative recommenders use for this purpose different types of ratings of specific items. Ratings are mostly binary, or express a preference for the wider range [14].

Content-based recommendation is not based on the similarity of users as collaborative filtering but on the similarity of items. Therefore, there are items recommended to the user, that they liked in the past [19,4]. For example, this is the case when the user rated or purchased the item. The similarity of items is determined based on the attributes assigned to it.

Demographic recommendations are based on demographic information about users, e.g. information related to personal attributes of the user. Classic examples are age, language or country, which create demographic profiles of users [28]. One of the first uses of this approach was the Grundy system [29], where books were recommended based on personal interview. Source of information for recommendations may also be part of various marketing research and customer segmentation [9].

Knowledge-based recommendations recommend items based on information about user needs and preferences [8]. This principle is, however, principle of many techniques but recommendation systems based on knowledge also have information on how the properties of items meet the users' needs [9]. In e-commerce, there are areas where users purchase certain items only once in a few years (electronics, bicycles, etc.). In these cases, it is appropriate to use knowledge-based recommendations.

Hybrid recommendations combine several techniques to achieve optimal recommendations to exploit their advantage and to eliminate the disadvantages [28,9,1]. They often combine collaborative recommendations with recommendations based on the content. Thus, we can remove problems with a small number of ratings in collaborative filtering.

The role of the recommendation technique is to generate suitable objects for a user. Therefore, the suitability of the recommendation depends on the quality of the recommendation technique. Unfortunately, there are plenty of different recommendation techniques (often based on latent features computation), which make white-box explanations impossible. In some domains, e.g. movies, users are not able to perceive the suitability of recommended item from some of its characteristics, e.g., name. Thus, our intent is to show the recommendations in the best way, provide reasons for the recommendations and in this way, convince the user to use the presented recommendations. Specific form of presentation and visualization of items of recommendations is the amount of information, structure of information, color, position or explanations of recommendations.

3. Explanation of Recommendations

Explanations have multiple definitions mainly because of the wide range of their application. We adopted following definition: recommendation explanation is an information that is designed to clarify why the item was recommended [2]. Explanations are also defined as a description that helps to determine the suitability of the item for a specific user [2]. However, generally speaking, they provide information about the recommendations and support the objectives defined by the creators of the system [32].

A very important point in the context of explaining the recommendations is to realize what the purpose of the explanation is. Recommender systems with explanations help make quick decisions and verify the suitability of buying a product [2]. But neither explanation can fix the problems with bad recommendations even though they can somehow compensate it [2].

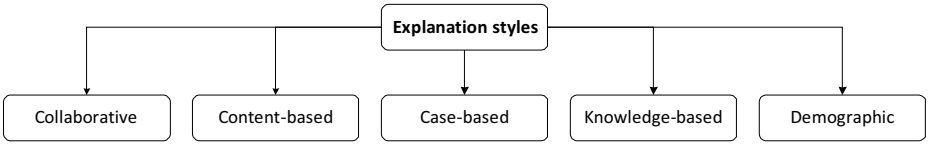


Fig. 1. Explanation styles categorization by [2].

Recommendations without explanations often act as a black box. This means that users have no clue how these systems get special knowledge about them and how they can know what is appropriate for them [26]. This is why the explanation itself is closely related to the recommendation technique. This means that the recommendation technique (eg, collaborative filtering, content-based) affects the explanation style of a specific recommended item. Therefore, the standard approach is when we use collaborative recommendations and the explanation will be based on the same technique.

Following the idea of the transparency, the explanations itself try to explain how the recommendation approach works. Unfortunately, the complexity and novel approaches to recommendation (e.g., based on some latent features [13]) cannot be explained to the average user. In such case, explanations aims as explaining the high level idea rather then the specific algorithm. Following the standard terminology, we will refer to it as the black-box explanation [37]. Pushing it forward, sometimes we have to abstract from the recommendation approach and generate explanations separately (e.g., we cannot explain the idea of neural network based recommender). On the contrary, if the explanations try to reveal the principle of the recommender, we will refer to it as the white-box explanations.

3.1. Explanations Styles

Since individual explanations are based on the recommendation technique which was used, they can also be categorized in this way (Figure 1). This approach relates to the fact that each technique can affect the style or form of explanation.

To make it clear, we provide short characteristic of different approaches to explanation styles referring to each style [32,2]:

Collaborative – Input for this technique are the ratings of items given by individual users. Such an approach is used e.g. in Amazon. An example of this style are explanations such as: *"Users like you positively rate this item."*

Content-based – In this case, an explanation tries to make clear to the user that the item was recommended based on the similarity with other items. An example is the explanation: *"This movie was recommended to you because it contains features of movies that you positively rated in the past."*
Another type of content-based explanation is case-based explanation, which clarifies that the item is similar to another item that has already been used as a recommendation and therefore the user liked it in the past. An example is the explanation: *"Recommendation was generated based on your most viewed items."*

Knowledge-based – This style of explanation is based on the description of user's needs or interests in the context of a recommendation. A classic example is the explanation: *"This destination has higher average temperature, which is better for sunbathing."*

Demographic – It clarifies the use of demographic data and its connection to the recommendations. One example is the explanation: *"This movie was recommended to you according to your age."*

3.2. Attributes of Explanations

The explanations can address certain goals, i.e., benefits we want to bring to the system [32]. The individual attributes cannot be achieved all at once. It is often a compromise. When we increase one attribute, the second one will be reduced. Important "beyond accuracy" attributes of recommender systems are:

Transparency or Justification – This is an explanation of how the system works, or explain how the recommendation item was generated [16].

Scrutability – This is particularly the possibility for users to tell the recommender system that something is wrong in the system or something is not working as it should.

Trust or Confidence – As the name implies this is about increasing user confidence in the system. Like the ability to respond to errors in the system (scrutability), confidence is closely linked with the transparency of the system.

Persuasiveness – This one is related to the effort to persuade the user to test or bought an item. However, it is necessary that the system did not push the user or force him to buy or choose any possibly unsuitable items.

Effectiveness or Education – This means helping the user to make good decisions. Explanation should help the user to evaluate the suitability of an item in terms of his preferences [16].

Efficiency – This is the combined effort to make a selection of items faster. In this case, explanations try to simplify the selection of items in order to reduce the time of decision making.

Satisfaction – User satisfaction with the recommendations may be necessary and useful also in the context of other attributes. Generally, this is an attempt to do the work with the system more enjoyable for user [2].

Explanations may vary in recommendation technique. However, they may also differ in the attributes or benefits that we can achieve by their use as part of a recommender system. Purpose of explanation then must make the recommendations appropriate to user goals or characteristics, because the particular user may prefer different presentation and explanation style.

4. Real-world Explanations Applications

There are a number of different approaches and studies that deal with recommendations and which bring new methods in the area of personalization models or explanations [16,24]. But especially explanations are still quite new areas of actual research. We describe several recent approaches to explanations that are included in real systems and applications.

Most of the techniques focus only on recommendations and their personalized models. Often they start using different technologies in order to make the personalization better. To evaluate impact of explanations, the eye-tracking seems to be a promising technology

which can show us if the user perceive them the way we expected. In [40] authors used gaze and incorporates it into the personalization models. Their results showed that eye-tracking data even from small number of users can significantly boost the accuracy of the recommendations [40].

One of the shortcoming of such studies is that they focus only on one part of the problem and that is accuracy of the recommendations. There are other important aspects (metrics), which describe the performance of the recommender system. In this context, the very important aspect is trust of users [41], their loyalty and increase of their satisfaction. Explanations can play an important role in this problem and make it easier for users to reach objects of their interest or even persuade them [17].

Some approaches also benefit from the way of creating the recommendations. System RecExp [17] is able to generate semantic recommendations by utilizing meta paths and thus to find various similar users. This type of recommender system makes possible to generate personalized explanations adapted to a specific user. In this case, explanations are presented as a fan chart, which shows weight of meta path representing user preferences [17]. Explanations are in this case presented in the form of three most similar users according to actual user.

Explanations itself are important to users for decision making. However, in case of recommendations, it is also important to present explanations in the right way so they attract user attention. One way how to present the explanations is by use of a natural language [11,25]. Powerful way is to use crowd-sourcing as an approach to generate natural language explanations [11]. However, in order to have good explanations, it is important to use domain experts as a crowd to generate these explanations. Therefore, this approach [11] builds on existing algorithms that generate personalized content-based explanations and combines them with texts from reviews. Thus, the authors were able to generate natural language explanations containing richer information than standard content-based explanations.

Framework ExpLOD [25] also generates natural language explanations, based on the Linked Open Data cloud as a source of information. The method is based on building a graph, which connect items liked by user and recommended items according to the Linked Open Data (LOD). ExpLOD framework thus outperforms baseline approaches in accuracy and also offers more interesting explanations.

There are also works, which specifically focus on explanations and comparing different explanation types [15]. This is also part of our research, however in different environment and domain. Another studies are trying to focus on trust of users towards the recommendation [39].

Nowadays, recommendations are used by large number of websites and applications. For related work here we present several different approaches. Specifically, we focus on popular websites such as Amazon, IMDB and Last.fm. Among these, we have analyzed the reasons for using recommendations and especially the presentation of recommendations.

4.1. Amazon

Amazon is globally known and popular e-commerce. It offers a huge variety of items and goods and therefore needs to direct its offer to the user. Amazon for this purpose use recommendations and their wide range of application. The main technique for generating

Fig. 2. Amazon: Explanation based on collaborative filtering.

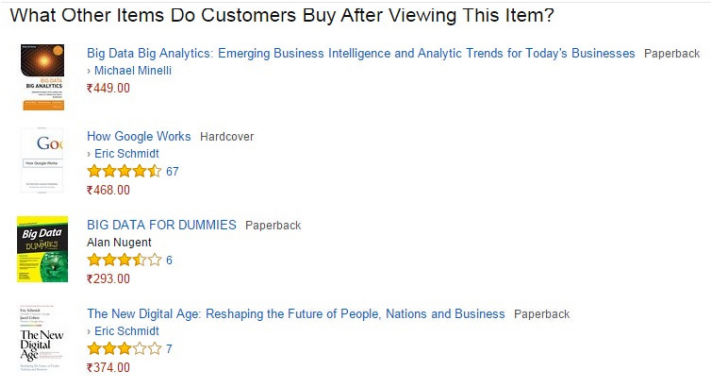


Fig. 3. Amazon: Explanation based on item-to-item recommendation.

recommendations is a content-based technique based on the similarity of items (item-to-item) [23].

The basic form of presentation of recommendations is a list of items along with an explanation of why they were recommended. Amazon uses especially the explanations based on the collaborative filtering (Figure 2) and on the recommendations based on the similarity of items (Figure 3).

4.2. Internet Movie Database - IMDB

IMDB is one of the largest film database and contains many items like Amazon. However, unlike the Amazon, IMDB is not try to sell these items. It is just a movie database that uses recommendations, specifically the item-to-item recommendation. Another difference is that the recommendations are not so promoted as in the case of Amazon.

The basic form of presentation of recommendations is a list of items along with an explanation of their recommendations (Figure 4). When a user hovers the cursor over one of the smaller images he will be able to see a fuller description on the right.

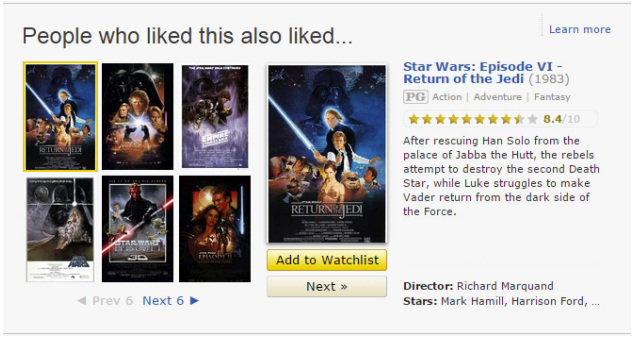


Fig. 4. IMDB: Explanation of recommendations.

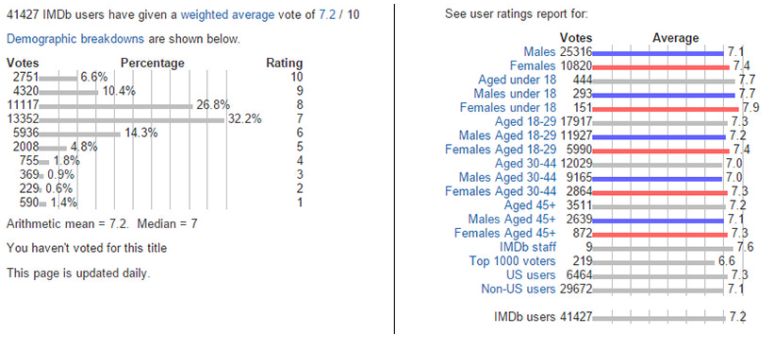


Fig. 5. IMDB: information about ratings.

Any rating of individual items (10 stars scale) is also available with more detailed information. These include a graph with the number of users who have rated the film and graph with demographic information about ratings (Figure 5).

IMDB does not use recommendations as much as Amazon but it still has some interesting solutions such as the list of recommended films (Figure 4). Explanations are of two types:

- Explanations based on collaborative filtering (Figure 4)
- Explanations of ratings (Figure 5)

The most interesting form are demographic explanations of ratings that should be very helpful for the users.

4.3. Last.fm

Last.fm is actually a music database that focuses on personalized recommendation based on the music the user is listening to. The most basic feature of this site is a social approach to the recommendations [10].

The main type of presentation is a list of similar artists. The basic form of this list is in the form of pictures with artist name. However, after clicking on the artist a user will

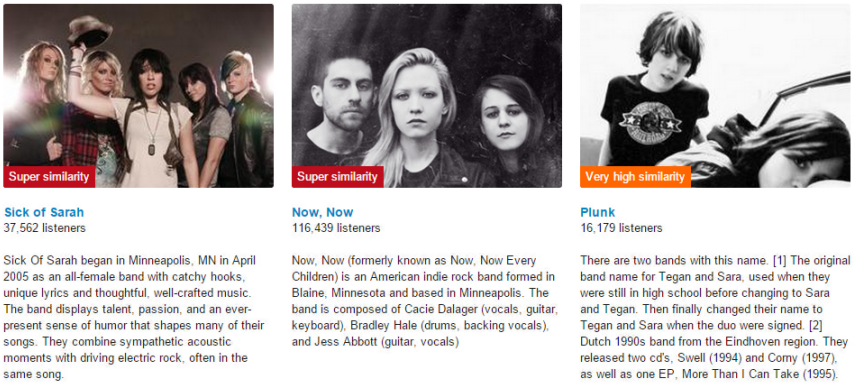


Fig. 6. Last.fm: List of similar artists.

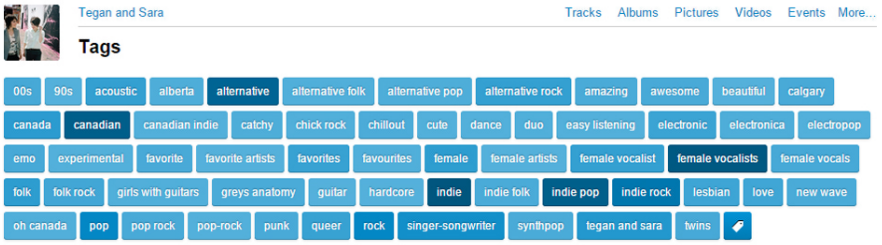


Fig. 7. Last.fm: Tags connected with artist.

be able to see a more specific list with a description and also with the level of similarity (Figure 6).

The strong feature of Last.fm is a labeling function (tag), which allows to add the artist or song to a certain group (Figure 7). These tags then form a new source of explanations for machine-generated and social-generated recommendations [10].

The advantage of Last.fm is fast and simple but very interesting recommender in the form of similar artists. Explanations are shown mainly by the level of the similarity (Figure 6). Tags are in this case also a form of explanations, which illustrates the similarity of the two artists.

4.4. Summary

Explanation of recommendations are widely used in many different ways. Amazon uses them as a part of their recommender system and also uses other features like star rating to make them more useful and usable for users. Main types of explanation are graphs with numbers of ratings, collaborative filtering based explanations and item-to-item based explanations. This approach is very easy to use and together with incorporating other features of the system (star rating, etc.) it is very interesting approach. However, the explanation is always the same and thus do not take into account the differences between users.

IMDB type of explanation is similar since they are using explanations together with ratings of items to make explanations more reliable for users. However, they are also providing a deeper analysis of these ratings with demographic breakdowns. This type of explanation can be very useful for users but there is harder to find it and thus the whole idea loses its meaning. In the same way, explanations itself are not personalized, which we see as a barrier.

Last.fm uses explanations to a lesser extent than other two systems before. However, their approach is one of the most interesting. They are showing the whole list of similar items together with the level of similarity and thus they make it easier to choose. However, personalization aspect is missing here too.

Most of the real world applications, web sites or recommender systems itself using the content-based recommendations. Subsequently, from this technique is derived the explanation style. This is the reason why the content-based explanations are that widely used. Nowadays, the most popular use of explanation is one general sentence, for example: *"People who liked this also liked ..."*. This approach is standard on recent systems and works quite well but it have some disadvantages. The biggest one is that the sentence is the same in every moment and for every user. However, every person is different and we need to somehow adjust even the explanation to these personal preferences of users. Moreover, these sentences are telling the truth on one hand, but just on high level of detail. For the users, it is hard to trust the system more, when they get this type of explanation on high level of detail without any personalization.

5. Hybrid Explanation Method

To answer our research questions in real-world scenario, we use a recommender service in the news domain. Our aim is to recommend news articles to users and present them on a website. For the purpose of presentation of the recommendations, our effort is to find optimal settings in the context of explanation of recommendations. According to our analysis, there are many recommended systems, which use explanations. However, in most cases, these explanations are fully based on recommendation technique (white-box). Such approach seems to us quite bonding, since it cannot be used for some current recommender techniques (e.g., latent features based). When focusing on specific domain, i.e. news, it is even more challenging to find explanation approach for the news recommender. In [6,5] authors proposed a news recommender systems, which uses a white-box explanations generated based on user previous feedback to similar articles. Therefore, we bring a novel, *black box* approach to explanations. For this purpose, we have proposed a method for hybrid explanation of recommendations – it generates content and collaborative explanations, independent of the used recommendation approach and tailored to user needs.

According to our method, *personalized explanation* is an approach which generates explanations with regard to user preferences. Thus, each user will be given an explanation adapted to what most impressed him (i.e., explanation style which he/she prefers). Another advantage is that proposed method is not dependent on recommendation technique that was used to recommend item. On the other hand, the recommendation techniques are used in the process of explanations generation. Based on the information about the

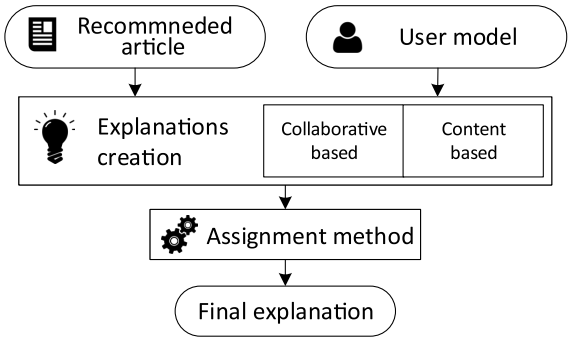


Fig. 8. Idea of proposed method for personalized back box explanation consisting of three steps: 1. Data acquisition; 2. Explanation creation; 3. Explanation assignment.

recommended items and about the user, we generate and present an explanation that is useful and interesting to the user.

Our method combines different approaches for explanation (explanation styles) to find suitable explanation for the actual user. Thus, we present an explanation that fits the user's preferences. For each recommended item, it is necessary to find explanation that is suitable in the context of characteristics of this item and which is also suitable for specific user. Proposed method is thus a hybrid and personalized type of explanation.

Our method follows three basic steps (Figure 8):

1. **Data acquisition**
   In the first step, there are two inputs: the first input is a list of recommended items together with their characteristics (id, title, author, content, etc.) and keywords (basic metadata). For the item representation we use only keywords extracted from the article content. Extraction of keywords from the content of the recommended items (articles) was accomplished by the following steps:
   - Removing special characters from text (html tags, punctuation, numbers, etc.)
   - Conversion to lowercase
   - Removing stop words
   - Getting individual words from the text - tokenization
   - Assessment of the number of occurrences of individual words

   The second input is the information about the users represented by their user model. The user model is determined by the keywords from the liked items (articles) read by the user previously. In other words, we use the bag of the words user model. With these two data sources, we can generate personalized explanation of recommended items (e.g., articles).
2. **Explanation creation**
   In the second step, the actual explanations are generated based on both methods of personalized explanations, which use these two approaches:
   - *Explanation based on similar users* – Explanation based on similar users uses the user's neighbors and items that they read. The basic procedure for generating explanations is as follows:
     (a) Selection of recommended items and their characteristics or keywords.

(b) Selection of the *k* - neighbors of the user and their preferences. Neighbors are other users who interact with (e.g., read, bought, clicked on) similar items as the actual user. Preferences consist of items that neighbors interact with in the past.
(c) Find a correlation between the recommended item and the items that have been read by neighbors of actual user. This comparison is based on the calculation of cosine similarity (based on users user models).
(d) If the correlation is above the selected threshold, then we will generate the explanation (in the news domain): *"This article was recommended to you because the similar user XY read this article too"*.

– *Explanation based on the content of items* - Explanation based on the content of items uses items that the user interact with in the past. The basic procedure for generation of the explanations consists of these steps:
(a) Selection of recommended items and their characteristics or keywords.
(b) Select items that user interact with in the past, together with their characteristics or keywords.
(c) Find a correlation between the recommended item and items that the user interact with in the past. This comparison is again based on the calculation of cosine similarity (based on items representation - keywords).
(d) If the correlation is above the selected threshold then we will generate explanation (in the news domain): *"This article was recommended to you because it is similar to article that you read before and his title was: XY"*.

3. **Explanation assignment**
The third step of proposed approach is the assignment of appropriate type of explanation for the specific user. Initial setting for a new user assigns to both explanation types the same weight. Results of these two explanations are presented in a standard way called the interleaving list. With this, we create a list of recommended items together with explanations in order to identify if users prefer a certain explanation style or approach. The learning phase is performed by continuous monitoring the user clicks on each item explained by different approaches and adjusting their weights based on user's preferences (Equation 1). This refers to the online learning, specifically test-then-train approach (idea similar as bandit algorithms). As a result, the users obtain more explanations generated by their preferred type (based on the item content or based on the similar users).

$$No_{cont} = Min\left(Round\left(\frac{\frac{Clk_{cont}+1}{Clk_{coll}+1}}{Clk_{cont}+Clk_{coll}+2} \times No_{to\ rec}\right), No_{to\ rec}-1\right), \quad (1)$$

where No_{cont} is the number of content explanations to generate, $No_{collab} = No_{to\ rec} - No_{cont}$, Clk_{cont} and Clk_{coll} is the number of content and collaborative explanations user clicks in his/her history and $No_{to\ rec}$ is the number of explanations to generate. The idea is to present to the user more explanations by the type hi/she likes. On the contrary, at least one of the explanations differ in the style and thus the users have a chance to adjust their preferences over the time (at last one explanation is generated by different method as the rest).

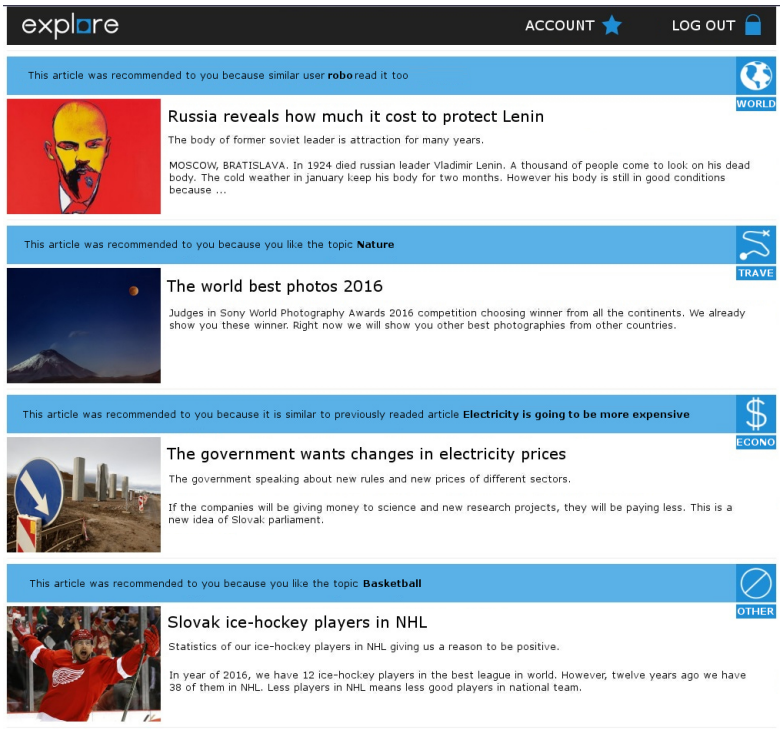


Fig. 9. Interleaved list for the new users (User interface of system ExplORe (setting ”with recommendation”).

6. Evaluation

As part of the evaluation of our method, we conducted several experiments where we used our method to explain recommendations to specific users in the news domain. In a live uncontrolled experiment with real users we recorded their activities while reading articles in our system. This system has been developed for the purpose of displaying the recommended news articles and related explanations.

6.1. ExplORe – a system for article recommendation

For the purpose of experimentation with our explanation method, we developed a system, which provides the place for displaying recommended articles and explanations generated by our method. System interface (Figure 9) contains news articles with explanation why they are recommended (setting ”with explanations”). Every user has displayed 10 most suitable articles and after he/she click on some of them, this article is no more part of the list of recommended articles. The system will replace it with the next most suitable item from the recommendation list.

ExplORe is the recommender system that uses recommendations to identify suitable articles for a particular user. We analyzed advantages and disadvantages of several options

of recommendation approaches. Consequently, we decided to use the library Apache Mahout, which is quite widespread [35,30] and reported to be well performing. Moreover, it provides a number of features for a recommendations and evaluation of the accuracy of the recommendations.

The system itself also implements our hybrid explanation method. It was implemented according to description in Section 5 and consist of two approaches. Both approaches work with the keywords associated with individual articles and compare them in order to find a suitable explanation. Specifically, the method works with these two basic models:

- Model of news articles – each article has assigned 10 keywords
- User model – each user is identified by keywords of articles that he/she read

These models help us to generate appropriate explanations for users. However, every user can perceive the explanations in different way even if they are equally suitable. Therefore, it is also important to know where and how to display explanations as a part of user interface. Thus as a part of design procedure, we conducted user eye-tracking study. The aim was to identify optimal position of explanations and amount of information which should be presented to users.

6.2. Eye-tracking user study

In this first experiment, we focused on obtaining basic information on the location and structure of information displayed in our system. We subsequently used this information to design the user interface of our system (Figure 10). We focused on:

- Presentation of recommendation: amount and structure of information
- Explanation of recommendation: positioning and visualization of explanation

Settings For each of the areas listed above, we created a few designs and showed them to participants of our eye-tracking study. This user study was conducted with 10 participants (8 men and 2 women), which is standard eye-tracking sample size and should eliminate major UX problems [33]. All participants were university students in age from 20 to 28. The participants should choose one article which was the most interesting for them. This showed us how difficult is for people to find this specific article in different conditions according to eye movements.

For the purpose of tracking the eye movements of participants, we used eye-tracker Tobii TX-300. This device has recorded both eyes at a rate of 300 Hz during our experiments. For evaluation of results from this study, we used basic gaze metrics: Fixation duration, Total fixation duration, Fixation count and Time to first fixation. According to I-VT fixation filter [27], eye-movement is considered as a fixation if its duration is above the minimum fixation duration (60 ms). Similarly, two fixations are merged when the duration between them is smaller than maximum time between fixations (75 ms).

After we finish eye tracking part of our experiment, we spoke with participants and ask them about their opinions on particular designs to find their subjective views.

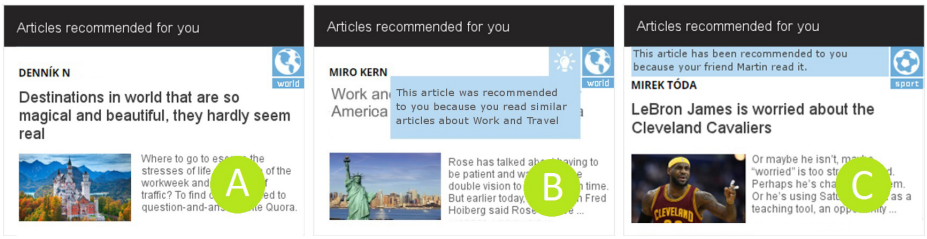


Fig. 10. Designs of explanation visualization presented to participants.

Results In case of presentation of recommendation, we found out that participants prefer articles described with short description, title, category and image. However, our main focus was to find out what is the proper way to visualize explanations of recommended articles. We created three designs of explanation visualization (Figure 10):

- Explanation next to image (Fig. 10 part A)
- Explanation in pop-up menu (Fig. 10 part B)
- Explanation above the title of article (Fig. 10 part C)

We analyzed fixation metrics as time to first fixation to determine how much time users took to find the explanation. Moreover we analyzed number and duration of fixations on explanations, along with the duration of viewing recommended items to uncover the difficulty of getting explanation information. According to these eye-tracking metrics and interview with participants, we decided to implement third option which displaying explanations above the title of articles. This options was easiest to find according to eye fixations (Figure 11). Participants also like the second option but we find out that some of them had a problem to find out that they have to point cursor on specific icon to open the pop-up menu with explanation.

6.3. Recommendations explanation method

In order to evaluate our hybrid method of explanation of recommendations, we conducted an experiment by use of our ExplORe system. Our goal was to find out what is the quality of explanation method in terms of impact on users while choosing articles and which approach will be more suitable in which specific conditions.

Settings To explore the features of proposed approach a long-term experiment with simulating the real media with news articles was conducted. Duration of the experiment was 18 days. Experiments were attended by a total of 17 people and 13 of them were university students. Up to 15 participants were aged 20-30 years. 17 participants were randomly divided into two groups without any specific criteria for the division: Group α (group consists of 8 participants), Group β (group consist of 9 participants). As the experiment toked 18 days, the number of participants is similar to other academia studies in recommender systems [34,22,40]. On the contrary, we are aware of limitations of the conclusions we address in this paper, which provide a chance for further exploration.

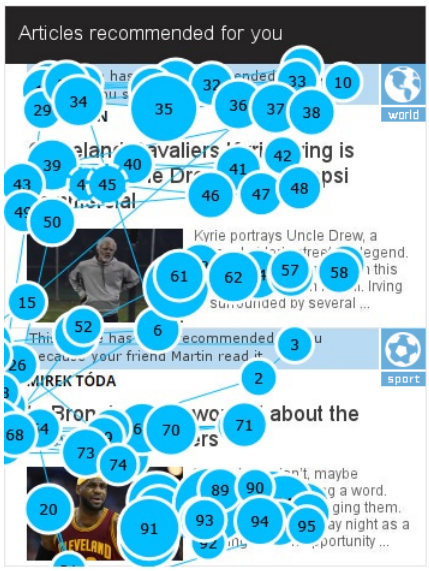


Fig. 11. Fixations on the explanation component of the ExplORe system.

	Group α	**Group** β
Phase 1	without explanations	with explanations
Phase 2	with explanations	without explanations

Table 1. Phases of the experiment and configuration settings.

Each group started with a different setups of the system. Group α started with newspaper articles without explanation. Group β started with newspaper articles along with explanations. In the middle of the experiment, the groups exchange their setups (Table 6.3). We also let participants complete two questionnaires. The first in the middle of the experiment, before the change of the setups. Second at the end of the experiment.

For the evaluation, we report the number of clicks on recommended articles and explanations. Article to which the participant clicked was always replaced by another article in the list. In further evaluation, we also report the Precision of the recommendation with and without explanation as a part of indirect evaluation.

Results We conducted the evaluation of the experiment in respect of all participants in the experiment but also in terms of individual groups so that we were able to point to individual differences in the data. The results can be divided into two main groups based on observed metrics:

- Impact of personalized explanations
- Comparison of different approaches

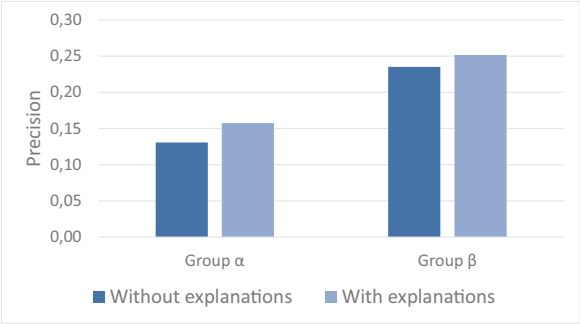


Fig. 12. Precision metric for both groups (settings).

RQ1: Can we increase the understandability and precision of recommendations with use of explanations?

Distrust of users in recommendations is relatively serious problem that the recommender system must face. During the experiments, we let participants to fill two questionnaires. The questionnaires were filled in by all participants and we find out that large part of them (14 out of 17 participants) reading news articles almost every day. The same number of participants also know the concept of recommendations.

In these questionnaires, we also asked participants about their experience with recommender systems. For many participants (12 out of 17) was unpleasant to know that recommender systems are watching them and that these systems know a lot of information about them. This confirms our assertion that users do not trust recommender systems too much.

However, we also asked them if it possible to reduce their distrust by using the explanations (as a tool for better understanding). A large majority of participants (15 out of 17) said that they will trust recommender systems more if they will use explanations. The results of our experiments and following questionnaires showed that our hybrid method of explanation of recommendations can reduce the distrust of user in recommender systems and thus to also increase transparency of these systems.

Basic statistics in the context of further evaluation is the frequency of clicks on articles without explanations and articles with explanations. We discovered that articles with explanations were clicked 1.24 times more (1 196 times) compared to articles without explanations (1 064 times).

Moreover, we looked at the same ratio in terms of precision between different groups (Figure 12). The graph clearly shows that in both cases there is an increase of the precision on articles with explanations. That means that users click more on articles with explanations (supported also by the questionnaire). Our assumption was that personalized explanations increase the number of clicks on articles. This was shown in both groups (Table 6.3) of participants. Thus, this result shows that explanations increase understandability and attractiveness of recommended articles.

Within the comparison of articles with explanations and without explanation, we looked on the ability to persuade the user to read an article, even if the article was not fully appropriate and recommender system ranked it on lower positions in the list of rec-

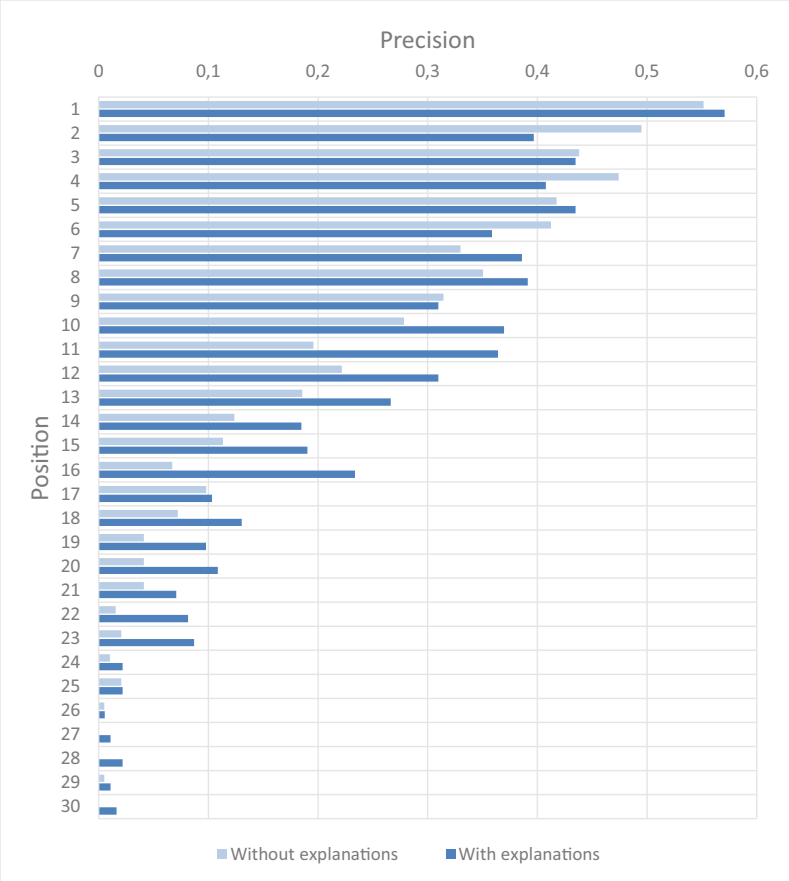


Fig. 13. Precision on positions in the list of recommended articles.

ommended articles. The following diagram (Figure 13) shows precision for each position in the list of recommended articles with and without explanations.

The figure shows that the articles in lower positions have higher precision when they are displayed with explanation. A paired t-test was conducted to compare precision over positions for two groups. On group had articles displayed with explanations and other without explanations. The difference is considered to be statistically significant ($t(29) = -3,363424$; $p = 0,002178$; $\alpha = 0,05$). This demonstrate that people click on articles which are less suitable for them more when they were explained by our explanation method. In other words, articles which are often ignored by users (presented on lower list positions), users found more interesting - while the explanations provide reasons for it. The figure thus clearly shows that explanation make articles more attractive and the recommendation more understandable.

Finally, we can conclude that participants clicked on the articles with explanations more than on the articles without explanations. This was supported in the context of precision, when we obtained higher values for individual users and also for whole groups.

Overall, we can evaluate the explanations in this case as a success in terms of impact on the understandably and precision of recommendations.

RQ2: Is there one explanation style that is preferred by users?

Firstly, we watched how often users clicked on various explanation approaches or styles. We discovered that the articles with explanations based on the content of articles had greater percentage of clicks (more than double) then articles with explanation based on similar users.

If we look at the precision in the context of each group we can see a similar result (Figure 14). In this graph, it is confirmed that the increase between the individual approaches are about twice as big. The first graph indicates that all users prefer one approach to explanation - content-based explanation.

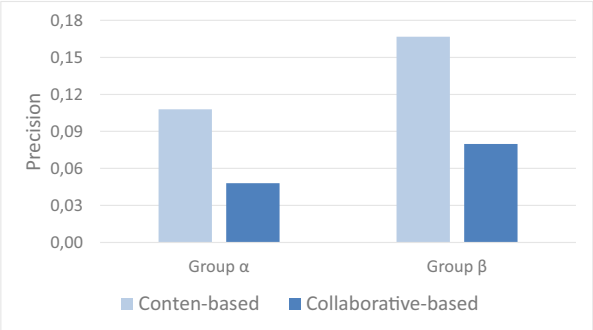


Fig. 14. Precision on articles with various explanations types for both groups (settings).

We hypothesize that the dominance of content based explanations was caused by two main reasons - Character of individual methods for explanation and Character of experiment. As our aim was to observe different preferences over various explanations types, the ratio between content and collaborative explanations was fixed to 0.5 (we ignored assignment mechanism from Equation1). First case was caused by nature of explanation based on similar users. In this approach, it was necessary to find someone else who reads the new article recommended for actual user. If there is nobody who reads this article, then we cannot use this type of explanation. This means that we explain a few more articles with content based approach then with approach based on similar users.

Second case was, in our opinion, caused by nature of experiment itself, which was attended by 17 participants. Problem was, that not all of these participants knew each other. Therefore, content based explanations could be more attractive for participants. If one does not know the person mentioned in the explanation based on similar users, he/she is less interested to click on the following article.

In this chapter we evaluate our approach with several experiments. Results of these experiments show that the idea of hybrid explanations make sense. We showed that different users prefer different type of explanations and that is first step towards the creation of the method of hybrid explanation.

7. Conclusions

In our work, we address the topic of recommendations presentation via web. Recommender systems are trying to solve the problems of information overload but often act as a black box. Users do not know how they recommend them each item or why these systems store a personal information about user preferences. In the context of these problems, we can speak about distrust of users.

To increase users' satisfaction and attitude to recommendations, the explanations are often used. Generally, the topic of the trust and human-computer interactions is a widely researched nowadays. In this paper, we proposed a method of presentation and explanation, which would be interesting for users while also address some of the problems of recommendation systems followed in our research questions.

To find out answers, we proposed an approach to explanation of the recommendations together with a suitable and transparent presentation of these explanations and recommendations themselves. This is the so-called hybrid method of personalized explanation of recommendations. The basic characteristics of this method are:

- The method is independent of the recommendation technique (black-box explanations)
- The method provides a personalized type of explanation that combines two approaches to explaining a) Content-based explanation; b) Collaborative-based explanations

We performed several experiments to evaluate proposed approach. We conducted a users study of system ExplORe using an eye-tracking tool (eye-tracker). We have mainly focused on obtaining basic information about the location and structure of presentation of explanations. Subsequently, we used this information to design the user interface of our system.

To answer our research questions, we conducted long term experiment in the domain of news. The results of the experiments have shown that users prefer articles with explanations compared to the article without explanations. We also found that users strongly preferred the content-based explanations compared to collaborative-based. The most interesting finding was that the explanations were able to convince participants to read less suitable articles (items that have been placed on the lower position in the list of recommendations (Fig. 15)). This is extremely useful from the business perspective, as we were able to increase the precision of recommendations, while the recommended method remains untouched. Last but not least, by using a questionnaire, we subsequently found that a large majority (15/17) of participants believe that the explanations can reduce their distrust in recommendation systems and thus actually increase the transparency of recommendations. This is a promising outcome, which can be used in further, large scale, online experiment.

As our aim was to explore the idea of personalized explanations, there are several other aspects, which should be addressed in following research. The idea of personalized explanations is to provide various explanations for various users. We proved, that users with similar demography prefers similar explanation approaches. However, this seems to be more important for users with different demographic (e.g., elderly users).

These results were obtained within the domain of news. However, we believe that the similar approach can be successful also in other domains, but this has to be proven

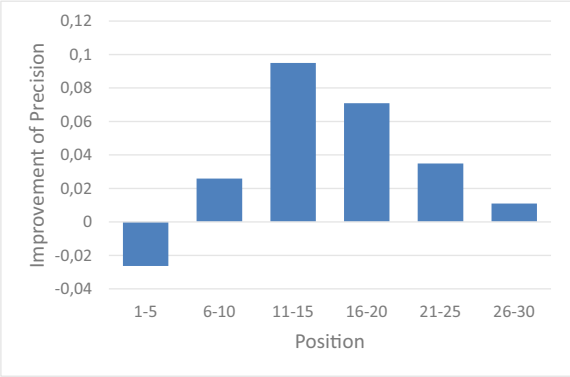


Fig. 15. The improvement of precision (recommendations "with explanations" compared to recommendations "without explanations") on recommendation list position.

by another experiment. Only big difference within these experiments will be the use of different characteristics of recommended items than with articles.

To sum up, we proposed a novel method of hybrid explanation of recommendations, which is independent on the recommender approach. Proposed method improved the precision of lower-positioned items in the recommendation list. Moreover, the participants express positive attitude to explanation as an approach for distrust reduce.

We see a great potential that our method brings into the area of recommender systems. This method has the potential of reducing the problems associated with the recommendations. However, in the future, the personalized explanation can be used in many other areas. The method is designed to include more explanation styles as explored in our experiments. This is useful for application in various domains. For instance the collaborative explanations can be considered as privacy issue in some domains (e.g., pharmacy e-shop). On the contrary, only users friends (after explicit agreement) should be used for such explanation style.

From our perspective, the most interesting area is to find which type of explanation is suitable for different type of articles. Thus, we can personalize explanation not only to user but also to items of recommendations (e.g., different type of articles). Second area is the context of user. Idea behind this concept is to show different types of explanation based on the context (e.g., season, weather, etc.) of the actual user.

Acknowledgments. This work is partially supported by the Slovak Research and Development Agency under the contract No. APVV-15-0508, the Scientific Grant Agency of the Slovak Republic, grants No. VG 1/0646/15 and VG 1/0667/18, and is the partial result of the Research & Development Operational Programme for the projects ITMS 26240220084 and ITMS 26240220039, co-funded by the European Regional Development Fund.

References

1. Aggarwal, C.C.: Recommender Systems: The Textbook. Springer (2016), https://doi.org/10.1007/978-3-319-29659-3

2. Al-Taie, M.: Explanations in recommender systems: overview and research approaches. In: Proceedings of the 14th Int. Arab Conf. on Information Technology, Khartoum, Sudan, ACIT. vol. 13 (2013)
3. Amatriain, X., Pujol, J.M., Oliver, N.: I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems, pp. 247–258. Springer, Berlin, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-02247-0_24
4. Bielikova, M., Kompan, M., Zelenik, D.: Effective hierarchical vector-based news representation for personalized recommendation. Computer Science and Information Systems (21), 303–322 (2012), http://eudml.org/doc/252774
5. Billsus, D., Pazzani, M.J.: A personal news agent that talks, learns and explains. In: Proceedings of the Third Annual Conference on Autonomous Agents. pp. 268–275. AGENTS '99, ACM, New York, NY, USA (1999), http://doi.acm.org/10.1145/301136.301208
6. Billsus, D., Pazzani, M.J.: User modeling for adaptive news access. User modeling and user-adapted interaction 10(2-3), 147–180 (2000)
7. Brusilovsky, P.: Adaptive Hypermedia: From Intelligent Tutoring Systems to Web-Based Education, pp. 1–7. Springer, Berlin, Heidelberg (2000), http://dx.doi.org/10.1007/3-540-45108-0_1
8. Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web, Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, vol. 4321. Springer (2007), https://doi.org/10.1007/978-3-540-72079-9_4
9. Burke, R.: Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction 12(4), 331–370 (Nov 2002), http://dx.doi.org/10.1023/A:1021240730564
10. Celma, O., Herrera, P.: A new approach to evaluating novel recommendations pp. 179–186 (2008), http://doi.acm.org/10.1145/1454008.1454038
11. Chang, S., Harper, F.M., Terveen, L.G.: Crowd-based personalized natural language explanations for recommendations. In: Proc. of the 10th ACM Conf. on Recommender Systems. pp. 175–182. RecSys '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2959100.2959153
12. Chen, Y.C., Shang, R.A., Kao, C.Y.: The effects of information overload on consumers' subjective state towards buying decision in the internet shopping environment. Electron. Commer. Rec. Appl. 8(1), 48–58 (Jan 2009), http://dx.doi.org/10.1016/j.elerap.2008.09.001
13. Costa, G., Ortale, R.: Model-based collaborative personalized recommendation on signed social rating networks. ACM Trans. Internet Technol. 16(3), 20:1–20:21 (Jul 2016), http://doi.acm.org/10.1145/2934681
14. Gedikli, F.: Recommender Systems and the Social Web: Leveraging Tagging Data for Recommender Systems. Springer Vieweg (2013)
15. Gedikli, F., Jannach, D., Ge, M.: How should i explain? a comparison of different explanation types for recommender systems. International Journal of Human-Computer Studies 72(4), 367–382 (2014)
16. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proc. of the 2000 ACM Conference on Computer Supported Cooperative Work. pp. 241–250. CSCW '00, ACM, New York, NY, USA (2000), http://doi.acm.org/10.1145/358916.358995
17. Hu, J., Zhang, Z., Liu, J., Shi, C., Yu, P.S., Wang, B.: Recexp: A semantic recommender system with explanation based on heterogeneous information network. In: Proc. of the 10th ACM Conf. on Recommender Systems. pp. 401–402. RecSys '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2959100.2959112
18. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction. Cambridge University Press, New York, NY, USA, 1st edn. (2010)

19. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems: An Introduction. Cambridge University Press, New York, NY, USA, 1st edn. (2010)
20. Jawaheer, G., Szomszor, M., Kostkova, P.: Comparison of implicit and explicit feedback from an online music recommendation service pp. 47–51 (2010), http://doi.acm.org/10.1145/1869446.1869453
21. Jiang, W., Wu, J., Wang, G.: On selecting recommenders for trust evaluation in online social networks. ACM Trans. Internet Technol. 15(4), 14:1–14:21 (Nov 2015), http://doi.acm.org/10.1145/2807697
22. Kowald, D., Seitlinger, P., Ley, T., Lex, E.: The impact of semantic context cues on the user acceptance of tag recommendations: An online study. arXiv preprint arXiv:1803.02179 (2018)
23. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing 7(1), 76–80 (Jan 2003), http://dx.doi.org/10.1109/MIC.2003.1167344
24. Muhammad, K., Lawlor, A., Rafter, R., Smyth, B.: Great Explanations: Opinionated Explanations for Recommendations, pp. 244–258. Springer, Switzerland, Cham (2015), http://dx.doi.org/10.1007/978-3-319-24586-7_17
25. Musto, C., Narducci, F., Lops, P., De Gemmis, M., Semeraro, G.: Explod: A framework for explaining recommendations based on the linked open data cloud. In: Proc. of the 10th ACM Conf. on Recommender Systems. pp. 151–154. RecSys '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2959100.2959173
26. O'Donovan, J., Smyth, B.: Trust in recommender systems. In: Proceedings of the 10th International Conference on Intelligent User Interfaces. pp. 167–174. IUI '05, ACM (2005), http://doi.acm.org/10.1145/1040830.1040870
27. Olsen, A.: The tobii i-vt fixation filter. Tobii Technology (2012)
28. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B.: Recommender Systems Handbook. Springer-Verlag New York, Inc., New York, NY, USA, 1st edn. (2010)
29. Rich, E.: User modeling via stereotypes. In: Cognitive Science: A Multidisc. Journal. vol. Vol. 3, No. 4, pp. 329–354 (1979)
30. Schelter, S., Owen, S.: Collaborative filtering with apache mahout. In: Proc. of ACM RecSys Challenge (2012)
31. Sutcliffe, A.G., Wang, D., Dunbar, R.I.M.: Modelling the role of trust in social relationships. ACM Trans. Internet Technol. 15(4), 16:1–16:24 (Nov 2015), http://doi.acm.org/10.1145/2815620
32. Tintarev, N., Masthoff, J.: Designing and Evaluating Explanations for Recommender Systems, pp. 479–510. Springer US, Boston, MA (2011), http://dx.doi.org/10.1007/978-0-387-85820-3_15
33. Turner, C.W., Lewis, J.R., Nielsen, J.: Determining usability test sample size. International encyclopedia of ergonomics and human factors 3(2), 3084–3088 (2006)
34. Verbert, K., Parra, D., Brusilovsky, P.: Agents vs. users: Visual recommendation of research talks with multiple dimension of relevance. ACM Transactions on Interactive Intelligent Systems (TiiS) 6(2), 11 (2016)
35. Walunj, S.G., Sadafale, K.: An online recommendation system for e-commerce based on apache mahout framework. In: Proc. of the 2013 Annual Conference on Computers and People Research. pp. 153–158. SIGMIS-CPR '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2487294.2487328
36. Yang, C.C., Chen, H., Hong, K.: Visualization of large category map for internet browsing. Decis. Support Syst. 35(1), 89–102 (Apr 2003), http://dx.doi.org/10.1016/S0167-9236(02)00101-X
37. Yoo, K.H., Gretzel, U., Zanker, M.: Persuasive Recommender Systems: Conceptual Background and Implications. Springer Briefs in Electrical and Computer Engineering, Springer, New York (2013)

38. Yuksel, B.F., Collisson, P., Czerwinski, M.: Brains or beauty: How to engender trust in user-agent interactions. ACM Trans. Internet Technol. 17(1), 2:1–2:20 (Jan 2017), http://doi.acm.org/10.1145/2998572
39. Zhang, J., Curley, S.P.: Exploring explanation effects on consumers trust in online recommender agents. International Journal of Human–Computer Interaction 34(5), 421–432 (2018)
40. Zhao, Q., Chang, S., Harper, F.M., Konstan, J.A.: Gaze prediction for recommender systems. In: Proc. of 10th ACM Conf. on Recommender Systems. pp. 131–138. ACM, USA (2016), http://doi.acm.org/10.1145/2959100.2959150
41. Zimmerman, J., Kurapati, K.: Exposing profiles to build trust in a recommender. In: CHI'02 Extended Abstracts on Human Factors in Computing Systems. pp. 608–609. ACM (2002)

Martin Svrcek received an M.S. degree in Information Systems from Slovak University of Technology in Bratislava, Slovakia in 2016. He is currently a PhD student at Slovak university of Technology in Bratislava. His research interests include web technologies, data science, eye tracking the user behaviour and user experience in general.

Michal Kompan received his PhD. degree in 2014, from the Slovak University of Technology in Bratislava. He is an associate professor, presently at the Institute of Informatics, Information Systems and Software Engineering, Slovak University of Technology. His research interests are in the areas of personalized recommendation for single and group of users, user modeling and machine learning. He is a member of IEEE Computer Society and ACM.

Maria Bielikova is a full Professor at the Institute of Informatics, Information Systems and Software Engineering, Slovak University of Technology in Bratislava. She is a head of User Experience and Interaction Research Centre at the same university. Her research interests include web-based systems, emphasizing user modelling and personalization, context awareness, and usability. She is a member of the Editorial Board of the Int. Journal of Web Engineering (JWE) and User Modeling and User-Adapted Interaction (UMUAI). She is a senior member of IEEE, a member of the IEEE Computer Society, and a senior member of ACM.

Received: December 17, 2017; Accepted: June 30, 2018.

Computer Science and Information Systems 16(1):205–225 https://doi.org/10.2298/CSIS180310015L

On the randomness that generates biased samples: The limited randomness approach

George Lagogiannis[1], Stavros Kontopoulos[2], and Christos Makris[3]

[1] Department of Agricultural Economics and Rural Development,
Agricultural University of Athens,
Iera odos 75, 11855, Athens, Greece
lagogian@aua.gr
[2] Department of Computer Engineering and Informatics, University of Patras,
26500 Patras, Greece
kontopou@ceid.upatras.gr
[3] Department of Computer Engineering and Informatics, University of Patras,
26500 Patras, Greece
makri@ceid.upatras.gr

Abstract. We introduce two new algorithms for creating an exponentially biased sample over a possibly infinite data stream. Such an algorithm exists in the literature and uses $O(\log n)$ random bits per stream element, where n is the number of elements in the sample. In this paper we present algorithms that use $O(1)$ random bits per stream element. In essence, what we achieve is to be able to choose an element at random, out of n elements, by sparing $O(1)$ random bits. Although in general this is not possible, the exact problem we are studying makes it possible. The needed randomness for this task is provided through a random walk. To prove the correctness of our algorithms we use a model also introduced in this paper, the *limited randomness* model. It is based on the fact that survival probabilities are assigned to the stream elements before they start to arrive.

Keywords: Biased reservoir sampling, Markov chain, random walk.

1. Introduction

Let us assume that we want to create a fixed size subset of a data stream of unknown, possibly infinite length. We call this subset a *sample* and let n be the given size of the sample, dictated by the available for this purpose memory. Creating such a sample falls into the general category of *synopsis maintenance* ([3], [6], [7], [10], [12], [14], [16], [17]). The problem of synopsis maintenance has been studied extensively for estimating queries ([4], [7], [17]) in data streams. A survey of stream synopsis construction algorithms can be found in [2].

In this paper we are going to focus on *reservoir sampling* [16], which is an important class of stream synopsis construction methods. Assuming that the length of the stream is known (let N be this length), we can easily create an algorithm to ensure that each stream element has n/N probability of being in the sample when the stream ends. However, the sample of a stream may need to be used before the stream ends, and it is even possible that the stream is endless. When the sample is unbiased and the length of the stream is unknown, the probability for each stream element to enter the sample reduces as the

stream progresses (see [16]). This means that as the stream progresses, fewer and fewer elements in the sample belong to the recent history of the stream. Such a sample may be proved useless for answering queries concerning the recent past. To address this problem one needs to create a biased sample. Such a sample is created in [1].

In [1], a bias function $f(r, t)$ is associated with the $r - th$ stream element at the time of arrival of the $t - th$ stream element ($r \leq t$). This function is associated with the probability $p(r, t)$ of the $r - th$ element belonging to the sample at the time of the $t - th$ element. In particular, the function $f(r, t)$ is monotonically decreasing with t (for fixed r) and monotonically increasing with r (for fixed t). For the specific function $e^{-\lambda(t-r)}$, an efficient algorithm is introduced in [1] for creating a sample consistent with that function. The parameter λ defines the bias rate and typically lies in the range [0, 1]. In general, λ is chosen in an application–specific way and it is the inverse of the number of stream elements after which the relative probability of inclusion in the sample reduces by a factor of $1/e$. Throughout this paper, we name this algorithm *Algorithm A*. Also, we assume that $f(r, t)$ is identical to $p(r, t)$.

The introduction of bias results in some upper bounds on the maximum necessary sample size, also called *maximum reservoir requirement*. It is proved in [1] that for the exponential bias function, the maximum reservoir requirement is bounded above by $1/\lambda$, assuming that λ is much smaller than 1, which happens to be the most interesting case. According to Algorithm A, each new stream element takes the place of another element of the sample which is chosen uniformly at random. Given that the size of the sample is n, Algorithm A needs $O(\log n)$ random bits per stream element, in order to choose an element of the sample at random.

In this paper we are going to present algorithms for creating a biased reservoir sample (over a very long, possibly infinite stream of data) consistent with the exponential bias function. The algorithms are based on the unit-cost RAM model, and use $O(1)$ random bits per stream element in the worst-case. Thus, our algorithms outperform Algorithm A in terms of *randomness complexity*.

The issue of reducing the randomness complexity of algorithms is not only of theoretical interest, but of practical as well. The analysis of randomized algorithms is always based on some assumptions concerning the theoretical machine-model used. The most frequently used machine-model is the unit-cost RAM. An interesting assumption concerning the unit-cost RAM, is that it needs $O(1)$ time to uniformly choose an element at random over n elements. This is because $O(\log n)$ random bits are needed for this task and the unit-cost RAM accesses these bits in O(1) time (we assume that w, the word size, is bigger than $\log n$). However this is not realistic, because the actual cost of generating these bits is swept under the carpet, through its assignment to some subroutine that has no cost at all. All the unit-cost RAM needs to do, is to spend $O(1)$ time in order to access $O(w)$ random bits, provided (free of charge) by this subroutine. The issue of the real cost of random bits is addressed in [8]. It is stated there that each random bit can be assumed to cost a constant amount of time "at best". This means that the time complexity of a randomized algorithm is at least equal to the number of random bits it uses. Minimizing the number of random bits is thus a straightforward direction for improving the time complexity of randomized algorithms, in cases where the randomness complexity of the algorithm dominates its overall time complexity. Algorithm A falls into this category as its time complexity is $O(1)$ in the worst-case in a unit-cost RAM, whereas its randomness

complexity is $O(\log n)$ (all the time complexities are "per stream element"). Assuming that each random bit costs $O(1)$ time (following the logic of [8]), the random bits in Algorithm A introduce a multiplicative factor of $O(\log n)$ per stream element in the time complexity. Our algorithms avoid this multiplicative factor and the intuition behind this result is given in the following paragraphs.

Let us assume that we have a coin, claimed to be fair. Intuitively, this means that if we toss the coin M times, "heads" or "tails" is supposed to occur at a fraction of M that approaches $1/2$, as M approaches infinity. However, to *evaluate* the coin (i.e. to decide whether the coin is fair or not) we do not need to toss the coin M times, for a big enough M. It suffices to inspect its shape (assume that we have special equipment for this). Clearly, the time of such an inspection is irrelevant (assuming that the shape of the coin remains unchanged over time). Let us now follow the same logic, replacing the coin by an algorithm that creates a sample over a data stream. The coin produces "heads" or "tails", whereas the algorithm produces samples. By setting the sample to be consistent with a bias function $p(r, t)$, we mean that if we execute the algorithm M times on a stream S, the $r - th$ stream element (i.e., $S(r)$) will survive in the sample at time t $(t > r)$ at a fraction of the M produced samples that approaches $p(r, t)$ as M approaches infinity. In order to evaluate such an algorithm (i.e. to decide whether or not the algorithm creates a sample consistent with $p(r, t)$), we need to check its instructions and using these instructions to calculate survival probabilities for the sample elements. If these survival probabilities are consistent with the given bias function, we conclude that the algorithm works as claimed. As with the coin, the time of the evaluation is irrelevant (clearly, the algorithm remains unchanged over time). Although the evaluation can be performed any time, the evaluation at time $t = 0$ (i.e. before the stream elements start to arrive) has a clear advantage, to be explained in the following paragraphs. Evaluating the algorithm at time $t = 0$, means that we have no sample to inspect but we do not actually need an existing sample. It is the "shape of the coin" that matters.

In order to successfully evaluate our algorithms, we must conclude that each time a new stream element arrives, each sample element is equally likely to be deleted from the sample (this is the basic property of the sampling algorithm in [1]). This means that each time a new stream element arrives, we need to perform uniform sampling. For this purpose, our algorithms use a Markov chain whose stationary distribution happens to be the uniform distribution. The random walk corresponding to this Markov chain is composed of *steps*. Each time, one of the elements of the sample is the *position* of the walk. Each step results in a new position of the walk and needs $O(1)$ random bits in the worst-case. Each accessed random bit is used for executing a step of the walk and we choose to delete the element of the sample implied by the position of the walk.

Using a Markov chain for performing uniform sampling is not new. Typically, one needs a Markov chain having the uniform distribution as its stationary distribution. In order to extract a random position, the Markov chain must be allowed to "run" for as many steps as necessary in order to reach its stationary distribution. However, when the random position is extracted, the position of the walk is not random any more and in order to retrieve one more random position, one has to let the Markov chain "run" again for as many steps as necessary in order to reach its stationary distribution. In this paper, we simply cancel this fact and instead we extract one random position per step.

The above claim may sound strange. In particular, one may argue that when an element is chosen by our algorithm at any time $t \neq 0$, the position of the walk is deterministically decided. Knowing the position of the walk at time t, how is it possible to choose the next element uniformly at random, by accessing only $O(1)$ random bits? The answer is the following: Accessing the sample at time $t \neq 0$ and calculating survival probabilities for the sample elements, we simply calculate conditional probabilities in regard to time $t = 0$, i.e. probabilities under the condition that the exact accessed sample of time t will occur. In this sense, calculating survival probabilities at any time $t \neq 0$, may be misleading.

To achieve uniform sampling while calculating survival probabilities at time $t = 0$, it suffices to make sure that by the time the stream elements start to arrive, the Markov chain will have reached its stationary distribution (i.e., the uniform distribution). As a result, it will remain in this distribution, which means that in regard to time $t = 0$, all future elements are equally likely to be chosen by our algorithm. More insight on this claim will be given in Section 3, where a more detailed intuitive example is given. We will also use the example of Section 3 for intuitively introducing a new model for evaluating our algorithms, which we call *limited randomness* model.

The structure of the paper is the following: In Section 2, we briefly present Algorithm A. In Section 3 we present the intuition behind the definitions that we present in Section 4. In Section 5, we present the random walk. This way we use this random walk as a black box in Section 6, where we present our new algorithms. We believe that this "black box" usage of the random walk simplifies the description of the algorithms, and helps the reader focus on the rest of the ideas that make our algorithms work.

2. Algorithm A: The exponential bias function

In this section, we will briefly present Algorithm A (i.e., algorithm 2.1 of [1]).

Algorithm A is based on the idea that each new stream element is deterministically inserted into the sample, and replaces one of the old elements of the sample. The old element to be replaced is chosen uniformly at random. Therefore, assuming that the reservoir contains n elements, each element "survives" in the sample with probability $(1 - 1/n)$ every time a new stream element arrives. Then, the element that arrives at time r will exist in the sample at time t with probability $(1 - 1/n)^{t-r} = ((1 - 1/n)^n)^{(t-r)/n}$. For large values of n (and it is reasonable to assume that n is large), the inner $(1 - 1/n)^n$ term is approximately equal to $1/e$, and from substitution, the exponential bias function follows.

It must be noted that Algorithm A is slightly more complicated, because the reservoir is filled gradually, to ensure that the sample is consistent with the exponential bias function from the beginning. We have skipped this part, firstly in order to keep things simple, and secondly because we are going to use the exact methodology, to achieve the same thing.

3. Intuition on the term "limited randomness"

We can use a roulette as an intuitive example. Let us assume that we have a roulette with n numbers (where each number is associated with a different slot), such that in the long run, each number is equally likely to host the ball. However, our roulette has a problem. Assuming that the ball is hosted in a given slot, the nearby slots are more likely than

distant ones, to host the ball after the next "throw". This weakness of our roulette comes from the fact that the ball is moved by means of a random walk algorithm (it is an e-roulette). The random walk uses less randomness than what is actually needed in order to guarantee that each time, the ball ends up in a slot that can be assumed to be fully random (it is therefore a *limited randomness roulette*). Although in the long run, one will observe that each number is equally likely to occur, the casino owners have a problem. If the players know the algorithm under which the roulette operates, they will bet their money on slots nearby the most recent winning slot. Even if the players do not know the algorithm under which the roulette operates, they can observe and notice that each time, the ball tends to end-up in a slot that is nearby the previous one. If the casino owners use this roulette in the "traditional" way, their casino is doomed.

To address the issue, the casino owners apply a curtain above the roulette, so that the players cannot see where the ball is before the throw. If the casino owners manage to hide all the information concerning the current position of the ball, so that the players conclude that below the curtain, the ball is equally likely to be in any slot, the problem is solved. This curtain is the basic notion of our mechanism and it will be modeled in the next section. The players now have *limited access* on what is going on.

One might wonder how is it possible to use this roulette instead of the traditional one. In particular, it is clear after a throw, the winning number must be revealed to the players and then the players will bet their money on nearby slots. Even if they cannot see the slots (i.e. they do not know which number corresponds to each slot), they will write down the winning numbers. Each time a winning number x appears, they will check their notes and see what is the most frequent winning number next of x. Then, they will bet their money on this number. It seems that the notion of the curtain did not make this roulette usable after all. Clearly, assuming that the players have no memory of the past, the problem is solved. But is such an assumption reasonable? For casino players, it certainly isn't.

However keep in mind that the central idea behind the algorithms presented in this paper, is that all evaluations (i.e., all probability calculations) are performed at time $t = 0$ (according to the intuition given in Section 1). Assuming that the players have no memory of the past, we simply incorporate into our model the fact that at time $t = 0$, when the evaluation occurs, there is nothing to remember. Thus, our limited randomness roulette can in fact be used, but not with a curtain (the curtain only exists in our evaluation model). The exact use is presented in the next paragraph.

The casino announces that the players are allowed to bet before the game starts, and the game involves N throws. No betting is allowed during the game. Each bet contains an amount of money, a slot where the money is put and the number of throw on which the bet is active. Therefore, each player is allowed to bet on any one of the N throws, before the first throw starts. After all the bets are placed, the roulette is initialized in such a way that the ball is placed in slot chosen uniformly at random among the n slots. Then the throws start. At any time $t \neq 0$, each player may find himself/herself to have an advantage or a disadvantage, based on the position of the ball (i.e. depending on how far from the slot he/she placed a bet, the slot that hosts the ball lies). However, this temporal feeling of stealing the casino, or being stolen by it, is just an illusion (and gambling involves quite a few illusions). The truth is that when the bets are placed, i.e. at time $t = 0$, (it can be derived that) each bet has $1/n$ probability to win and as a result, the game is fair.

4. Definitions

In general, we distinguish in this paper two approaches for a biased reservoir sampling algorithm: According to the *strict* approach, the reservoir is filled gradually and even when it is not full, the sample is consistent with the bias function. According to the *greedy* approach the reservoir is first filled and then the algorithm starts. Therefore it follows that initially, our sample does not comply with the bias function. But as time passes, fewer and fewer of the initial elements continue to exist in the sample, thus the fraction of the sample that was not built according to our function, continuously decreases. By the above definition, Algorithm A is strict.

We distinguish the memory positions used by an algorithm that creates a biased sample, into two parts. The *sample-space*, contains all the memory positions where the elements of the sample are stored. The *surrounding-space* contains all the memory positions used by the algorithm in order to obtain the sample, except the sample-space. The sample-space and the surrounding-space together, correspond to the *entire used-space* of the algorithm. Let us now introduce the concept of the *sample-user*, which informally, is a person able to calculate probabilities and comes in two flavors.

Definition 1. The *sample-user of full access* is an *evaluator* able to access at any time t i) the entire used-space, ii) the algorithm that created the sample and iii) the bias function associated with the algorithm. Based on what he/she can access at time t, the sample-user of full access is able to calculate for each existing or future sample-element x, the probability of x being in the sample for all future (in regard to x) time-points.

Definition 2. The *sample-user of limited access* is an *evaluator* able to access at any time t i) the sample-space, ii) the algorithm that created the sample and iii) the bias function associated with the algorithm. Based on what he/she can access at time t, the sample-user of limited access is able to calculate for each existing or future sample-element x, the probability of x being in the sample for all future (in regard to x) time-points.

It must be noted that in both definitions, we consider the sample-user to be *memoryless*, in the following sense: At any given time t, the sample-user of full (limited) access is able to calculate probabilities based on what he/she is able to access at time t only. In other words, the sample-user can not keep track of the changes that have occurred to the sample, in order to use the information provided by these changes when calculating probabilities.

The notion of the sample-users is introduced in order to capture the intuition described in the previous section. The different flavors of a sample-user are needed to capture the presence or absence of a curtain that covers all memory positions except the sample-space. According to the above two definitions, the only time-point where a sample-user of full access and a sample-user of limited access have the same "power" is the time before the stream elements start to arrive (i.e., when $t = 0$). This is because at time $t = 0$, the entire used space is empty. We have set the sample-users to be memoryless in order to reflect to the sample-users the fact that all calculations are done at time $t = 0$ because there is nothing to remember at time $t = 0$.

Definition 3. A biased sample *of full randomness* consistent with a function f is a sample where at any time t, a sample-user of full access is not able to find an existing or future element x in the sample, for which at least one time-point decreases the probability of x being in the sample in a way different than the one dictated by f.

Definition 4. A biased sample *of limited randomness* consistent with a function f is a sample where at any time t, a sample-user of limited access is not able to find an existing or future element x in the sample, for which at least one time-point decreases the probability of x being in the sample in a way different than the one dictated by f.

Definition 3 applies to the biased sample created in [1]. In particular, at any time t, the sample-user of full access will conclude by accessing Algorithm A, that each of the n elements in the sample has $1/n$ probability of being deleted from the sample at the next time-point, i.e. each element has $(1 - 1/n)$ probability to survive the next insertion, and this is consistent with the exponential bias function (allowing the substitutions explained in Section 2). Said otherwise, Algorithm A *convinces* a sample-user of full access.

In this paper we present algorithms for creating a biased sample of limited randomness, i.e. algorithms that convince a sample-user of limited access. According to the intuition given so far, convincing a sample-user of limited access suffices for proving that our algorithms produce samples consistent with a given bias function. This is because time $t = 0$ is as good as any for evaluating our algorithms and it turns out that at time $t = 0$, a sample user of limited access and a sample user of full access are identical (they both access the same thing, i.e. the instructions of the algorithm). We believe that this limited access model can be used in order to reduce the randomness complexity of algorithms, in cases where the randomness complexity dominates the overall time complexity.

5. The random walk

5.1. Description of the random walk

In this subsection we present an algorithm for performing a random walk on a list. The list is traversed in left-to-right order, with the leftmost node being the head of the list. Thus, the next node of a node u is the node on the right of u, and we set the next node of the rightmost node of the list to be the head of the list (i.e., we actually have a circle). Similarly, the previous node of u is the node on the left side of u, and the previous node of the head of the list is the rightmost node of the list. We also maintain a pointer called *cursor*, that points to a node of the list which we call *position* of the walk. The walk consists of steps, with each step having an initial, and a final node. The initial node is the position of the walk before the step starts, and the final node is the position of the walk after the step ends. The previous of the final node is called *hit-node*, and thus each step "produces" a hit-node. Initially, before the first step, the position of the walk is the head of the list. A step is briefly described by algorithm Step that follows.

Algorithm Step. We toss a fair coin on the position of the walk. If "tails" occur, the next node of the list becomes the position of the walk, and we repeat the same (i.e. we toss a fair coin again). Otherwise, the position of the walk becomes the hit-node of Step(), we set the cursor to point to the next node (which now becomes the position of the walk) and Step() ends. The maximum number of coins we are allowed to toss, is K. If we toss K coins and no "heads" occur, then we declare the node pointed by the cursor as hit-node, we set the cursor to point to the next node of the list and Step() ends.

A reader unfamiliar to Markov chains should hopefully find Appendix A (where the intuition behind Algorithm Step is given) helpful. An issue that must be discussed is the

value of K, the maximum number of coin-tosses we are allowed to have in each step. Obviously, the worst-case time complexity of Step is equal to K. However, as long as $K \geq 2$, observe that since the coin is fair, 2 coin tosses are performed on the average in each step, because this is the average number of tosses until "heads" occur. Thus, the expected number of random bits per step is 2 and Step costs $O(1)$ expected time. By allowing c coin tosses at most, where c is a constant greater than one, we achieve constant worst-case time per step. By allowing only one coin toss, no randomness is inserted into the scheme, and the position of the cursor becomes deterministic for all the future time points.

5.2. Analysis of the random walk

Let n be the length of the list. We will analyze the walk for K such that $2 \leq K \leq n$. Let $m_i = 1/2^i$ and

$$p_i = \begin{cases} m_i, & \text{if } i < K \\ m_{i-1}, & \text{if } i = K \\ 0, & \text{if } i > K \end{cases}$$

We can easily see that the walk can be described by a Markov chain with the following $n * n$ transition matrix.

$$P = \begin{bmatrix} p_n & p_1 & p_2 & p_3 & \cdots & p_{n-1} \\ p_{n-1} & p_n & p_1 & p_2 & \cdots & p_{n-2} \\ \vdots & \vdots & & \vdots & \ldots & \vdots \\ p_1 & p_2 & p_3 & p_4 & \cdots & p_n \end{bmatrix}$$

For example, by setting $K = n$ the matrix becomes:

$$P = \begin{bmatrix} 1/2^{n-1} & 1/2 & 1/4 & 1/8 & \ldots & 1/2^{n-1} \\ 1/2^{n-1} & 1/2^{n-1} & 1/2 & 1/4 & \ldots & 1/2^{n-2} \\ \vdots & \vdots & \ldots & \vdots & & \\ 1/2 & 1/4 & 1/8 & 1/16 & \ldots & 1/2^{n-1} \end{bmatrix}$$

For more details and an extended example see Appendix A. The key observations are that the factors are the same for each row and their sum is equal to 1. The position probability vector at time j is:

$$Pos_j = \left[P_{1,j} P_{2,j} \ldots P_{n,j}\right]$$

and can be derived by the initial probability vector at time 0, $\theta = \left[P_{1,0} P_{2,0} \ldots P_{n,0}\right]$ as follows: $Pos_j = \theta * P^{(j-1)}$.

Theorem 1 *The random walk described in this subsection simulates a uniform sampling process as time approaches infinity.*

Proof. It is easy to conclude that the Markov chain is irreducible and aperiodic. The states are irreducible because there is a path from any state to all the others. The states are aperiodic because they form one connected component in the graph and anytime the

probability of returning to a state is greater than 0, since we can reach any state from any other state with probability greater than 0. The above statements hold for any K. Thus the Markov chain has one limiting distribution π such that $\pi * P = \pi$ and $\sum_{i=1}^{n} \pi_i = 1$, where $\pi_i > 0$. It is straightforward to see that $\pi = [1/n, 1/n, \cdots, 1/n]$. □

5.3. Mixing time of the random walk

Theorem 1 establishes the fact that after "a while", the probability distribution of the Markov chain becomes very close to the uniform distribution. An important issue is how long it is going to take, i.e. what is the *mixing time* of the chain, defined as follows: Let V be the state space of our Markov chain. Let us assume that the chain will ultimately converge to the stationary distribution π which means that for every $i, j \in V$, $lim_{t\to\infty}(P^t)_{ij} = \pi_j$. The L^2 mixing time $r(\epsilon) = \max_{x\in V} min\{t : ||k_t^x - 1||_{2,\pi} \leq \epsilon\}$ is the worst-case time for standard deviation of the density $k_t^x(y) = \frac{(P^t)_{xy}}{\pi_y}$ to drop to ϵ [15]. For reasons of completeness we have calculated this mixing time in Appendix B and it is $O(n^2 \cdot \log n)$ where n is the number of list-nodes.

The mixing time is independent of K. Thus, by increasing K, one expects that the chain will reach the stationary distribution faster, but not asymptotically faster. While asymptotically there is no gain in setting K to be high, there is a gain in setting K to be a constant: The number of random bits per stream element becomes constant in the worst case. This is why in the following of the paper, we set $K = 2$.

6. New algorithms

Let $S(r)$ be the r-th stream element. Each arriving stream element defines the current time, i.e. the r-th stream element arrives at time-point r. Recall that in [1], it is proved that the reservoir size is bounded above by $1/\lambda$. As a result, initially, we have a list of $1/\lambda$ nodes, which we call *sample-list*. The sample-list is traversed in left-to-right order. The leftmost node is node 1, and the numbers of the nodes increment as we walk rightwards, however the numbers are only implied by the relative order of the nodes in the list and they are not stored in the nodes. The sample-list is the sample-space defined in Section 4.

We are going to present two new algorithms for creating a sample consistent with the exponential bias function. More specifically, we are going to show that according to both algorithms, the probability of $S(r)$ being in the sample at time t (for any $t \geq r$) is $e^{-\lambda(t-r)}$. In order to better state the problem we need to solve, we are going to first introduce a simple but problematic algorithm which we call *Algorithm 0*. In order to make things even simpler, let us assume that each sample element (stored in a node of the sample-list) is accompanied by its arrival time. That is, by accessing a node of the sample-list, we access an element of the sample, and the arrival time of this element.

Algorithm 0. When a new stream element arrives (after the reservoir is filled), we execute algorithm Step(), and replace the element stored in the hit-node by the new stream element.

The problem with Algorithm 0 is that it introduces a (disturbing) certainty for a sample-user of limited access. Let us assume that the user accesses the sample at time t.

Knowing the instructions of our algorithm and accessing the sample-list, the user is able to deterministically decide on the position of the walk: In particular, the cursor points to the next node of the node that contains the most recent element of the sample. This knowledge allows the user to determine for all the elements in the sample, a survival probability for the next time-point different than the one dictated by the exponential bias function. For example, the user will conclude that the element pointed by the cursor has 1/2 probability of being the next to be deleted. If the user starts calculating future probabilities, he/she will conclude that for any time-point more distant into the future than $r(\epsilon)$ time-points, each element can be assumed to be chosen uniformly at random and this is consistent with the given bias function. But for any future time-point between t and $t + r(\epsilon)$, the user will conclude that the instructions of the algorithm do not comply with the given bias function.

An issue that must be discussed, is whether Algorithm 0 manages to convince a sample-user of limited access, if we do not accompany each sample element by its arrival time. Then, the user cannot find the most recent element, therefore cannot decide on the position on the walk. However, this holds only for the elements that already exist in the sample, because the sample-user can obviously connect the future positions of the walk with the positions of the future elements. In particular, after the insertion of $S(t)$ for a future time-point t, the cursor will point to the node next to the one that contains $S(t)$. Therefore, assuming $K = 2$, the probability that $S(t)$ will be deleted at time $t + 1$ is 0 (we assume that the sample-list contains more than two nodes). This probability is not consistent with the exponential bias function, and thus a sample user of limited access is not convinced.

Since our goal is to convince a sample-user of limited access, Algorithm 0 does not solve the problem. In the following section, we are going to present two algorithms that convince a sample-user of limited access. *Algorithm 1* is a *mixing time dependent algorithm* meaning that its space-complexity is proportional to the mixing time of the Markov chain. *Algorithm 2* is a *mixing time independent algorithm* meaning that the mixing time of the Markov chain does not appear in the space-complexity. In both algorithms we do not accompany each sample element by its arrival time, as this is not mentioned in [1]. By adding the arrival times, Algorithm 2 is not affected whereas Algorithm 1 needs some more technical details which we omit for now.

6.1. Algorithm 1: A mixing time dependent solution

Algorithm 1 is based on the walk of Section 5, defined on the sample-list. It follows the greedy approach, i.e. initially we fill up all the nodes of the sample-list with the first $(1/\lambda)$ elements of the stream. We introduce a new list called *cursor-past* of length $r(\epsilon)$ (i.e. the length is equal to the mixing time of the Markov chain, defined in Subsection 5.3). A simple way to perceive the cursor-past list, is the following: A person watches a live program using streaming. There is a time-gap between what is really happening live, and what this person sees. As a result, he/she can only guess what is happening at present time, based on what he/she sees and knowing that what he/she sees happening now has actually happened in the past. In our algorithm, the cursor-past list will create such a time-delay effect on the position of the cursor, revealing to the sample-user the position of the cursor $r(\epsilon)$ time units ago. Based on this position, the user must decide on the current position of the cursor.

Before the stream elements start to arrive, we have the *preprocessing phase*, which includes two tasks. First, we need to set the cursor of the walk to point to one of the nodes of the sample-list, chosen uniformly at random. For this we use $O(\log(1/\lambda))$ random bits and then we need to access the corresponding (by the random bits) node of the sample-list in order to update the cursor, which means that we need to traverse the sample-list. Second, we have to traverse the cursor-past list from left to right. For each node u of the cursor-past list, we execute one step of the walk on the sample-list and store in u a pointer towards the hit-node of the step (keep in mind that the hit-node belongs to the sample-list).

The entire preprocessing phase needs $O(\log(1/\lambda) + r(\epsilon))$ random bits. Assuming that each random bit costs $O(1)$ time, the preprocessing phase needs $O(1/\lambda + r(\epsilon))$ time in the worst-case. One may assume that before the stream elements start to arrive, we have plenty of time to prepare our structures, i.e. the preprocessing time is not a problem.

After the preprocessing phase we have the *fill-up phase*, during which the sample-list is filled with the first $1/\lambda$ elements of the stream. When the sample-list becomes full, Algorithm 1 starts.

Algorithm 1. Let t be the current time and $S(t)$ the newly arrived stream element. We execute Step() and let y be the hit-node. We insert a new rightmost node into the cursor-past list, and store the address of y into this node. Let z be the sample-list node pointed by the leftmost node of the cursor-past list. We exchange the elements between nodes z and y. Then, we store $S(t)$ into z. We delete the leftmost node of the cursor-past list.

Algorithm 1 is graphically depicted in Figure 1 where the sample-list consists of 6 nodes. At time t, $S(t)$ is 60. According to Algorithm 1 we insert the red (rightmost) node in the cursor-past list. The two highlighted nodes of the sample-list will take place in the swap operation. The lower part of the figure visualizes our structures ready for the next insertion, at time $t+1$. Keep in mind that all the sample-user of limited access is allowed to "see", is the sample-list. Therefore, no node of the sample-list appears highlighted to the sample-user of limited access.

Theorem 2 *Associating Algorithm 1 with the exponential bias function, Algorithm 1 convinces a sample-user of limited access.*

Proof. Let us assume that the user accesses the sample just after $S(t)$ has arrived and let n be the size of the sample-list. In order to calculate the survival probabilities for the elements of the sample-list, the sample-user must decide on the position of the walk. Since the Markov chain has reached its stationary distribution from the preprocessing phase (remember that we have set the cursor to point to each element with equal probability), it remains in the stationary distribution from that point and on, thus the user concludes that any of the nodes of the sample-list can be the position of the walk with equal probability. This means that each element of the sample has $1/n$ probability to be the next we delete, therefore each sample-element will survive the next insertion with probability $1 - 1/n$ and this is consistent with the exponential bias function (see Section 2). In addition, the user cannot decide on the position of the walk for all the future time-points. For example, let t_1 be a future time-point. Let u_j be the node of the sample-list where $S(t_1)$ is going to be placed. In order to calculate the survival probabilities for all the future (in regard to t_1) time-points, the user must decide on the node pointed by the rightmost node of the

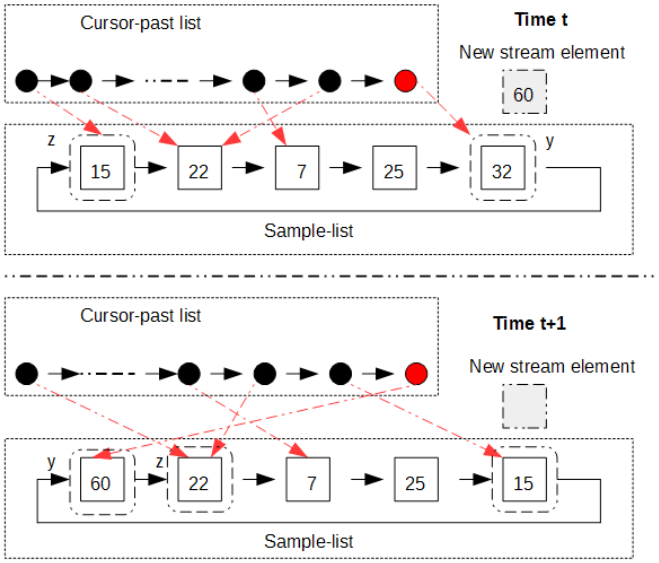


Fig. 1. Graphical representation of Algorithm 1.

cursor-past list. Being certain on the position of the walk implied by the leftmost node of the cursor-past list (it is the node next of u_j) does not help the sample-user. The length of the cursor-past list ensures that given a node pointed by the leftmost node of the cursor-past list, the node pointed by the rightmost node of the cursor-past list is fully random, i.e. the rightmost node may point to each node of the sample-list with probability $1/n$. Therefore each sample-element has $1/n$ probability not to survive at $t_1 + 1$, i.e. each element will survive the next insertion with probability $1 - 1/n$. □

Algorithm 1 uses O(1) random bits per stream element in the worst-case (remember that we have set $K = 2$ in Step). The rest of the actions need $O(1)$ time in the worst-case. The space-complexity is $O(r(\epsilon) + 1/\lambda)$, dominated by the size of the two lists (sample-list and cursor-past list). The space-complexity is critical. As long as it is dominated by the sample-list, Algorithm 1 is space-efficient. Although in our case this is not true, we believe that Algorithm 1 is interesting from a theoretical point of view, and as a "tool" that can be used elsewhere. To get rid of the mixing time in the space-complexity, we need to find another way (i.e. other than the cursor-past list) for disconnecting the position of the cursor from the most recent element of the sample. Such an algorithm is presented in the next subsection.

6.2. Algorithm 2: A mixing time independent solution

Algorithm 2 is presented in both the strict and the greedy approach and let us start from the strict one. We name *left(i)*, the part of the sample-list from node 1 to node i. It follows

that left($1/\lambda$) is the entire sample-list. For each i ($1 \leq i \leq 1/\lambda$), we associate a random walk on left(i) (we call it i-*walk*) based on algorithm Step. Thus we have $1/\lambda$ walks. We allocate another list of $1/\lambda$ memory positions called *cursor-list*. As with the sample-list, a number (between 1 and $1/\lambda$) is implied for each node of the cursor-list, according to the order (from left to right) of the nodes in the list. $\forall\ i$ ($1 \leq i \leq 1/\lambda$), the i-th node of the cursor-list is going to host the cursor of the i-walk.

If i elements of the stream have been inserted into the sample-list, they occupy $left(i)$ and the i-walk is the *active-walk*. Initially, since no stream element exists in the sample, all walks are *inactive* and clearly, if the i-walk is active, then all the j-walks $(1 \leq j < i)$ have been active in the past. When the i-walk becomes inactive, the $(i+1)$-walk will become active. Keep in mind that although we have defined $1/\lambda$ walks, we need to keep track of only two walks: the active-walk and the $(1/\lambda)$-walk.

When a walk becomes inactive, we free the node of the cursor-list corresponding to the cursor of this walk. As a result, the nodes of the cursor-list are set free in left to right order and at all times, the leftmost node of the cursor-list contains the cursor of the active walk (we call this pointer *active-cursor*). This means that we are able to access the active-cursor in constant time. It is also clear that the rightmost node of the cursor-list contains the cursor of the $(1/\lambda)$-walk (we call this cursor *final-cursor*). Thus, by maintaining a pointer towards this node, we are able to access the final-cursor in constant time. Finally, we maintain a pointer called *active-limit*, towards the rightmost node of the active walk (i.e. if the i-walk is the active walk, we maintain a pointer towards the i-th node of the sample-list).

Before the elements of the stream start to arrive we have the *preprocessing phase* where we traverse the cursor-list from left to right and store in the i-th node, a pointer to a position of the sample-list, chosen uniformly at random among all sample-list positions between 1 and i. After this task is completed, $\forall\ i$ ($1 \leq i \leq 1/\lambda$), the i-th node of the cursor-list contains the cursor of the i-walk. For the preprocessing phase we use $O(1/\lambda \cdot \log(1/\lambda))$ random bits. Assuming that each random bit costs $O(1)$ time, the worst-case time complexity for the preprocessing phase is $O((1/\lambda)^2 \cdot \log(1/\lambda))$ because for each i we need to traverse the sample-list from node 1 to node i in order to access the node chosen uniformly at random among all the nodes from node 1 to node i.

The above description is depicted graphically in Figure 2. According to the figure, the active walk is the 5-walk, the filled nodes of the sample-list are represented through rectangles and the empty ones through circles. Above the sample-list, we illustrate the cursor-list. The shaded nodes on the left part of the cursor-list have been deleted in the past, and the leftmost (black) node contains the cursor of the active-walk. The highlighted node of the sample-list is pointed by the active-limit pointer. The sample-user of limited access can only see the sample-list and since the active-limit pointer is "out of sight", no node of the sample-list appears highlighted to the sample-user of limited access.

Assume that at the time of (just after) the arrival of $S(t)$, the fraction of the reservoir filled is $F(t) = i \cdot \lambda$, which means that i elements have been inserted into the reservoir and they occupy left(i). It follows that the i-walk is the active-walk. Let w be the node pointed by the active-limit pointer, and z be the node next to w. We are now ready to proceed to the description of Algorithm 2.

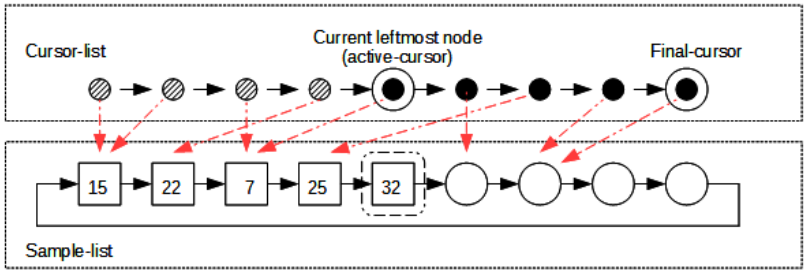


Fig. 2. The background for introducing Algorithm 2.

Algorithm 2. When $S(t+1)$ arrives, we deterministically add it to the sample. **If** the final-cursor points to a non-empty position of the sample-list, the new stream element is going to replace an existing element of the sample i.e. the number of elements in the sample will remain the same. We execute a step for the i-walk and appropriately update the active-cursor. Let g be the node pointed by the active-cursor. The previous node of g is the hit-node of the step, and let x be that node. We delete x from the sample-list by connecting the previous node of x to g. Then, we insert x between w and z. We store $S(t+1)$ in x. We set the *active-limit* pointer to point to x. **Else** the number of elements in the sample will increment and the $(i+1)$-walk will become the active-walk. We insert $S(t+1)$ in z. We delete and free the leftmost node of the cursor-list, making the $(i+1)$-th walk to be the active walk. We update the active-limit pointer to point to z. **End-If**. We execute a step for the $(1/\lambda)$-walk.

A first observation is that the elements are stored in increasing arrival order, from left to right. Clearly, the space-complexity is $O(1/\lambda)$ dominated by the length of the two lists. In practical terms, the entire used-space is initially twice the space of the sample-list. However, each time the number of elements inside the sample increases, we free the leftmost node of the cursor-list and as a result, the entire used-space decreases over time. When the entire sample-list is filled with elements, only one node from the cursor-list remains. It is the node that contains the cursor of the $(1/\lambda)$-walk. This means that in regard to Algorithm A, the only memory overhead that remains is the one due to the pointers that support the sample-list. Concerning the time complexity, Algorithm 2 clearly uses $O(1)$ random bits per stream element and beyond that, all the other actions cost $O(1)$ time. Assuming that each random bit costs $O(1)$ time, our algorithm needs constant time per stream element in the worst-case.

Theorem 3 *Associating Algorithm 2 with the exponential bias function, Algorithm 2 convinces a sample-user of limited access.*

Proof. Let us assume that the active-walk is the i-walk. The user concludes that the i-walk is the active-walk because the sample has i elements. To calculate future probabilities, the user must decide on the position of the active-walk. When the i-walk became the

active-walk, the active-cursor was set to point to a node of the i-walk chosen uniformly at random. Not being able to access the active-cursor, the user concludes that it may point to any node of the i-walk with probability $1/i$, because the Markov chain associated with this walk has reached its stationary distribution and obviously, each step from that point and on cannot change the distribution. Let us now assume that the current time is t, the length of the sample-list is n and the fraction of the sample-list occupied by elements is $F(t)$ (note that at this point, the proof of Theorem 2.2. in [1] starts). Then, according to the algorithm, when $S(t+1)$ arrives, an element of the sample is going to be deleted with probability $F(t)$ and if such a deletion is decided, each of the $n \cdot F(t)$ elements of the sample have equal probability $1/(n \cdot F(t))$ of being deleted (because, as stated above, the cursor may point to each position with equal chances). As a result, the probability of an element being deleted is $F(t)(1/(n \cdot F(t))) = 1/n$. It follows that the probability of a stream element to survive at $t+1$ is $1 - 1/n$. Thus, the element that arrives at time r will survive in the sample at time t with probability $(1 - 1/n)^{t-r} = ((1 - 1/n)^n)^{(t-r)/n}$. For large values of n, the inner $(1 - 1/n)^n$ term is approximately equal to $1/e$, and from substitution, the exponential bias function follows (note that at this point, the proof of Theorem 2.2. in [1] ends). In advance, the user cannot decide on the position of the walk for any future time-point and as a result, whatever holds for the existing sample-elements also holds for the future sample-elements. □

One issue of interest to be discussed is the following: The walk of Section 5 is defined on a static list. According to Algorithm 2, we are able to delete an internal node of the list and re-insert it as the new rightmost node of the active walk. This results in the position of the active-walk moving one node to the left. The same holds for the position of the $(1/\lambda)$-walk. This is not the walk we presented in Section 4. In particular it now follows that in each step the probability of doing nothing i.e. going one position forward and then one position backwards (thus ending up at the starting point) is 1/2, the probability to move one node to the right is 1/4 and so on. This is a lazy version of our walk and it is trivial to see that in this lazy version the stationary distribution remains the same. Thus, Algorithm 2 is not affected.

Going to the greedy approach is trivial. It turns out that we do not need the cursor-list, but instead only the cursor for the $(1/\lambda)$-walk. Therefore, the space remains the same and equal to the final space consumed by the strict approach. The time complexity remains identical to that of the strict approach.

7. Discussion

Let us now summarize the main issue introduced in this paper, which is derived by the comparison between the algorithm (and methodology) in [1] and the algorithms (and methodology) we present in this paper. In both cases, a bias function $f(r, t)$ is associated with the $r - th$ stream element at the time of arrival of the $t - th$ stream element $(r \leq t)$. This function is associated with the probability $p(r, t)$ of the $r - th$ element belonging to the reservoir at the time of the $t - th$ element. However, the issue here is the time of the association. In [1], the association can be done at any time t. In our paper, the function is associated at time $t = 0$, for all the future elements of the sample. This is a logical choice since the probability is implied by the algorithm, in the sense that we have manufactured

our algorithms to create samples consistent with the exponential bias function. In other words, survival probabilities are decided when the algorithm is created. It then follows that the randomness used in [1] is more than the randomness we really need in order to obtain a biased sample. It is a sample-user of limited access that we actually have to convince, and not a sample-user of full access. This observation has led to optimal randomness complexity for the problem we are studying and we believe that the technique we use for introducing randomness to our algorithms (i.e. the notion of limited access) may be useful in other problems.

References

1. Aggarwal C. C.: On biased reservoir sampling in the presence of stream evolution. In proceedings of the 32nd international conference on Very large data bases , pp. 607–618. Seoul, Korea (2006).
2. Aggarwal C. C. and Yu P. S.: A Survey of Synopsis Construction in Data Streams. Data Streams - Models and Algorithms, pp. 169–207 (2007)
3. Babcock B., Datar M. and Motwani R.: Sampling from a Moving Window over Streaming Data. In proceedings of ACM-SIAM Symposium of Discrete Algorithms, pp. 633–634, San Francisco, USA (2002).
4. Babu S. and Widom J.: Continuous queries over data streams. ACM SIGMOD Record Archives, Volume 30, Issue 3, pp 109–120 (2001).
5. Chung F. R. K. , Graham R. L. and Yau S.-T.: On Sampling with Markov Chains. In proceedings of the seventh international conference on Random structures and algorithms, pp. 55–77, Atlanta, USA (1996).
6. Gibbons P. and Mattias Y.: New Sampling-Based Summary Statistics for Improving Approximate Query Answers. In proceedings of ACM SIGMOD, pp. 331–342, Seattle, USA (1998).
7. Gibbons P.: Distinct sampling for highly accurate answers to distinct value queries and event reports. In Proceedings of 27th International Conference on Very Large Data Bases, pp. 541–550, Rome, Italy (2001).
8. Goldreich O.: Another Motivation for Reducing the Randomness Complexity of Algorithms. Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation, pp. 555–560, (2011).
9. Jerrum M. and Sinclair A.: Conductance and the rapid mixing property for Markov chains: the approximation of permanent resolved. In proceedings of the twentieth annual ACM Symposium on Theory of Computing, pp. 235–244. Chicago, USA (1988).
10. Hellerstein J., Haas P. and Wang H.: Online Aggregation. In proceedings of ACM SIGMOD, pp. 171–182, Tucson, USA (1997).
11. Lawler G. F. and Sokal A. D.: Bounds on the L^2 Spectrum for Markov Chains and Markov Processes: A Generalization of Cheeger's Inequality. Transactions of the American Mathematical Society Vol. 309, No. 2, pp. 557–580 (1988).
12. Manku G. and Motwani R.: Approximate Frequency Counts over Data Streams. In proceedings of the 28th International Conference on Very Large Data Bases, pp. 346–357, Hong Kong, China (2002).
13. Manku G., Rajagopalan S. and Lindsay B.: Random Sampling for Space Efficient Computation of order statistics in large datasets. In proceedings of ACM SIGMOD, pp. 251–262, Philadelphia, USA (1999).
14. Matias Y., Vitter J. S. and Wang M.: Wavelet-Based Histograms for Selectivity Estimation. In proceedings of ACM SIGMOD, pp. 448–459, Seattle, USA (1998).

15. Montenegro R.: Intersection Conductance and Canonical Alternating Paths: Methods for General Finite Markov Chains. Combinatorics, Probability and Computing, Volume 23, Issue 4, pp. 585–606 (2014).
16. Vitter J. S.: Random sampling with a Reservoir. ACM Transactions on Mathematical Software, Volume 11 Issue 1, pp. 37–57 (1985).
17. Wu Y., Agrawal D. and Abbadi A.: Applying the Golden Rule of Sampling for Query Estimation. In proceedings of ACM SIGMOD, pp. 449–460, Santa Barbara, USA (2001).

A. The intuition on Algorithm Step

In order to give the intuition behind this algorithm, let us name the list L. Let us set $length(L) = 5$ and $K = length(L)$ i.e., the maximum number of coin-tosses per step is 5. We assume that time increments (starting from 0) with each step, and suppose that someone is trying to guess the node pointed by the cursor, without accessing the cursor. At time 0, the node pointed by the cursor can be deterministically determined because the cursor points to the head of the list. Now, let us "travel" into the future and make our first stop at time-point 1. The probabilities for the cursor to point to each node at time 1 are shown in the following table (where "T" stands for "tails" and "H" stands for "heads") and let us be a bit more detailed on how these probabilities are derived: Let us focus on the probability that the cursor points to Node 1 at time-point 1. It follows that the execution of Step() that moved the cursor (starting from Node 1) back to Node 1 contains as many coin-tosses as necessary for the cursor to complete the circle. This means that the first coin-toss cannot cannot result to "heads" because then, the cursor will stop at Node 2 (keep in mind that a step ends when the coin-toss results to "heads" or when 5 coin-tosses have occurred). Given that the first coin-toss results to "tails" the second coin-toss cannot result to "heads" because then the cursor will stop at Node 3 (therefore the event "TH" leads to Node 3). Following this logic we conclude that in order for the cursor to complete a full circle, either the event "TTTTT" or the event "TTTTH" must occur during the first execution of Step(). The probability of each of these two events to occur is $1/32$. We conclude that the probability for the cursor to return to Node 1 at time-point 1 is $1/16$.

Node 1	Node 2	Node 3	Node 4	Node 5
$P_{1,1} = 1/16$ (event TTTTT OR TTTTH)	$P_{2,1} = 1/2$ (event H)	$P_{3,1} = 1/4$ (event TH)	$P_{4,1} = 1/8$ (event TTH)	$P_{5,1} = 1/16$ (event TTTH)

As time passes i.e. as more steps are performed, coin tosses gradually insert randomness into the scheme, introducing an increasing uncertainty on the position of the cursor. The probabilities at time-point 2 are shown below.

$P_{1,2} = P_{1,1} * 1/16 + P_{2,1} * 1/16 + P_{3,1} * 1/8 + P_{4,1} * 1/4 + P_{5,1} * 1/2$
$P_{2,2} = P_{2,1} * 1/16 + P_{3,1} * 1/16 + P_{4,1} * 1/8 + P_{5,1} * 1/4 + P_{1,1} * 1/2$
$P_{3,2} = P_{3,1} * 1/16 + P_{4,1} * 1/16 + P_{5,1} * 1/8 + P_{1,1} * 1/4 + P_{2,1} * 1/2$
$P_{4,2} = P_{4,1} * 1/16 + P_{5,1} * 1/16 + P_{1,1} * 1/8 + P_{2,1} * 1/4 + P_{3,1} * 1/2$
$P_{5,2} = P_{5,1} * 1/16 + P_{1,1} * 1/16 + P_{2,1} * 1/8 + P_{3,1} * 1/4 + P_{4,1} * 1/2$

In general, for time-point i, we have:

$P_{1,i} = P_{1,i-1} * 1/16 + P_{2,i-1} * 1/16 + P_{3,i-1} * 1/8 + P_{4,i-1} * 1/4 + P_{5,i-1} * 1/2$
$P_{2,i} = P_{2,i-1} * 1/16 + P_{3,i-1} * 1/16 + P_{4,i-1} * 1/8 + P_{5,i-1} * 1/4 + P_{1,i-1} * 1/2$

$P_{3,i} = P_{3,i-1} * 1/16 + P_{4,i-1} * 1/16 + P_{5,i-1} * 1/8 + P_{1,i-1} * 1/4 + P_{2,i-1} * 1/2$
$P_{4,i} = P_{4,i-1} * 1/16 + P_{5,i-1} * 1/16 + P_{1,i-1} * 1/8 + P_{2,i-1} * 1/4 + P_{3,i-1} * 1/2$
$P_{5,i} = P_{5,i-1} * 1/16 + P_{1,i-1} * 1/16 + P_{2,i-1} * 1/8 + P_{3,i-1} * 1/4 + P_{4,i-1} * 1/2$

It is not difficult to write a simple program and see that these probabilities converge to 1/5, as time progresses. Thus, after "a while", enough randomness has been inserted, for the cursor to be in each position with equal chances.

B. Bounding the mixing time

Bounding the mixing time of a Markov chain can be difficult. Here, we will use the conductance method ([9], [11]), according to which the ergodic flow between two subsets $S, T \subset V$ is defined to be

$$Q(S,T) = \sum_{i \in S, j \in T} \pi_i P_{ij}$$

The method could bound the mixing time for lazy reversible Markov chains, but it was extended in [15] in order to apply to general finite ergodic Markov chains. According to [15], the *intersection conductance* $\hat{\Phi}(A)$ of $A \subset V$ (and $\bar{A}$, the complement of A) is given by

$$\hat{\Phi}(A) = \hat{\Phi}(A, \bar{A}) = \frac{\sum_{u \in V} min\{Q(A,u), Q(\bar{A},u)\}}{2\pi(A)\pi(\bar{A})}$$

Then, the intersection conductance is $\hat{\Phi} = \min_{\emptyset \subsetneq A \subsetneq V} \hat{\Phi}(A)$. In [15] it is proved that the mixing time of a finite ergodic Markov chain is

$$r(\epsilon) \leq \frac{12}{\hat{\Phi}^2} log \frac{1}{\epsilon \sqrt{\pi_0}}$$

where $\pi_0 = \min_{u \in V} \pi_u$.
We are first going to bound the mixing time for $K = 2$, i.e. if at most 2 coin tosses are allowed in algorithm Step. For every $i, j \in V$, if $p_{ij} = 1/2$, we say that state j is the *next* of i. The left part of Figure B.1 visualizes the walk on 7 nodes. The black arrows correspond to probabilities of 1/2 and the red ones to probabilities of 1/4. The "contribution" of a node u in the numerator of $\hat{\Phi}(A)$, depends on the status of the three "preceding" nodes (according to the order defined by the black arrows). We distinguish 8 cases, depending on whether each of the 3 nodes belongs to A, or to $\bar{A}$. We name each case through the corresponding 3-bit number (each node corresponds to a 1, if it belongs to A, or to a 0, otherwise).

Let us introduce the following terminology: Set A *outperforms* set B, if $\hat{\Phi}(A) \leq \hat{\Phi}(B)$. Let $A \subset V$. Then, if all the states of A are connected through probabilities of 1/2, we call the set, *tight*. We are now ready to bound the mixing time of the Markov chain assuming that $K = 2$.

Theorem B 1 *Setting $K = 2$, $r(\epsilon) = O(n^2 \log n)$, where n is the number of states and $n \geq 8$.*

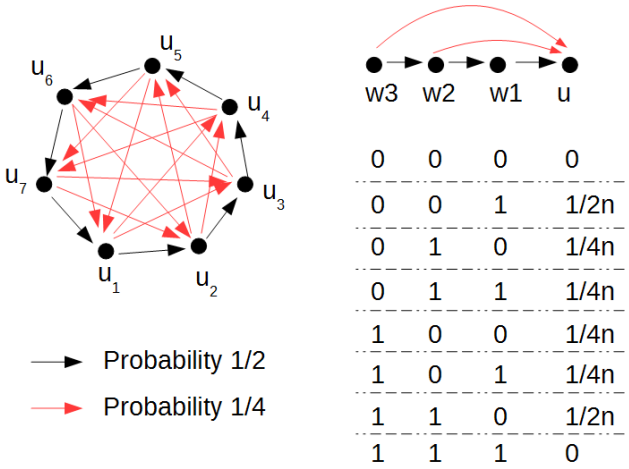


Fig. B.1. The random walk for 2 coin tosses. On the right, we show all the possible additives in the nominator of $\hat{\Phi}(A)$, because of node u.

Proof. We need to find the set A that outperforms all other sets. Let us first consider all the sets that contain d states, for a given d such that $1 < d \leq n/2$ (by allowing A to contain more nodes, $\bar{A}$ contains less than $n/2$ nodes and by renaming A to $\bar{A}$ and $\bar{A}$ to A, we fall into the same case). Thus, let A be such a set of cardinality d. Then, the denominator in $\hat{\Phi}(A)$ is $2 \cdot \frac{d}{n} \cdot \frac{n-d}{n}$. Since all the sets of cardinality d produce the same denominator, the one that results to the smallest numerator outperforms all the others. Observe, that in Figure B.1, the contribution of a node to the numerator of $\hat{\Phi}(A)$ is 0 if the three preceding states either all belong to A, or all belong to $\bar{A}$. As a result, among all the sets of cardinality d, the ones that outperform all the others are the tight ones. This is because in this case, the number of states for which the three preceding states either belong to A or not, maximizes and thus the numerator is minimized. For example, in Figure B.2 the states represented by squares belong to A, and the states represented by circles belong to $\bar{A}$. It follows that only u_2, u_3, w_2 and w_3 will contribute to the numerator. In particular, for u_2 we have the case 001 (see Figure B.1), for u_3 we have the case 011, for w_2 we have case 110 and for w_3 we have case 100. Observe that if A, B are tight sets (of states) of the same cardinality, then $\hat{\Phi}(A) = \hat{\Phi}(B)$.

We have now concluded that among all the sets of states having the same cardinality, anyone that is tight outperforms all the others. As a result, instead of trying to find the set A that minimizes $\hat{\Phi}(A)$ among all possible sets, we need to examine only the tight ones. Observe now that all the tight sets of more than two states produce the same numerator. As a result, among all the tight sets with more than two states, we must find the ones that result to the maximum denominator in order to minimize the fraction. The denominator is maximized for any set having $n/2$ states. We have now concluded that if A is tight and contains $n/2$ states and B is any other set of states that contains more than two states, $\hat{\Phi}(A) \leq \hat{\Phi}(B)$. For any tight set A of $n/2$ states we have the following:

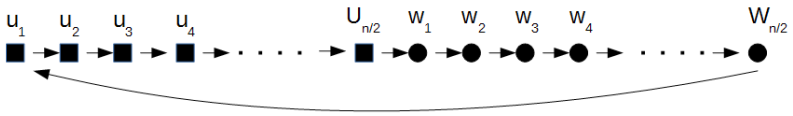


Fig. B.2. The rectangles correspond to the states of A. The arrows corresponding to probabilities with value less than 1/2 are omitted.

$$\hat{\Phi}(A) = \frac{\frac{1}{4n}+\frac{1}{2n}+\frac{1}{2n}+\frac{1}{4n}}{2\frac{1}{2}\frac{1}{2}} = \frac{3}{n}.$$

It remains to see if there is a tight set of cardinality 1 or 2 that outperforms the tight sets of cardinality $n/2$. For any tight set B of cardinality 2, only four nodes will contribute to the numerator of $\hat{\Phi}(B)$. In particular, the second (according to the order of the black arrows) node of B will fall into case 001, and the next three nodes (that belong to $\bar{B}$) will fall into cases 011, 110 and 100, respectively. It follows that for any tight set B of cardinality 2:

$$\hat{\Phi}(B) = \frac{3n}{8(n-2)}.$$

Similarly, for any single state set C, only the three states after the one that belongs to C are going to contribute to the numerator and they are going to fall into cases 001, 010 and 100 (according to the order defined by the black arrows). Thus, for any single state set C we have:

$$\hat{\Phi}(C) = \frac{\frac{1}{2n}+\frac{1}{4n}+\frac{1}{4n}}{2\frac{1}{n}\frac{n-1}{n}} = \frac{n}{2(n-1)}.$$

It follows that any tight set A of cardinality $n/2$ outperforms all tight sets of cardinality 1 and 2 for all $n \geq 8$, and as a result, it outperforms any other set of states. We have now concluded that by allowing at most two coin tosses, $\hat{\Phi} = \frac{3}{n}$. This means that setting $\epsilon = 1/n^c$ for some constant c, we get $r(\epsilon) \in O(n^2 \log n)$. □

It remains to find the mixing time for $K > 2$. Clearly, by increasing K we increase the area that each step can cover, and thus it is a fact that the mixing time is going to improve. However, asymptotically things remain the same, as the next theorem states.

Theorem B 2 *Setting $K = n$, $r(\epsilon) \in O(n^2 \log n)$, where n is the number of states and $n \geq 8$.*

Proof. Since we know that the chain with $K > 2$ reaches the stationary distribution faster than the one with $K = 2$, it suffices to find a set A such that $\hat{\Phi}(A)$ is small enough to imply $O(n^2 \log n)$ mixing time. Then, even if there exists another set B that outperforms A, it follows $\hat{\Phi}(B)$ can only imply $O(n^2 \log n)$ mixing time. We choose A to be a tight set of $n/2$ states, and we will prove that by setting $\epsilon = 1/n^c$ for some constant

c, $\frac{12}{(\hat{\Phi}(A))^2}\log\frac{1}{\epsilon\sqrt{\pi_0}} = O(n^2\log n)$. For the description we are going to use Figure B.2, where the states of A appear as rectangles, and the states of $\bar{A}$ appear as circles. Keep in mind that the arrows corresponding to probabilities with value less than 1/2 are omitted, however since $K = n$, each state is actually connected to all the other states. For states from w_2 until u_1 (following the arrows of probability 1/2), it is the set A that contributes to the numerator of $\hat{\Phi}(A)$. For states from u_2 until w_1 it is the set $\bar{A}$ that contributes to the numerator of $\hat{\Phi}(A)$. In particular:

Node w_2 contributes $\frac{1}{n}(\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^{\frac{n}{2}+1}}) < \frac{1}{2n}$

Node w_3 contributes $\frac{1}{n}(\frac{1}{8}+\frac{1}{16}+\cdots+\frac{1}{2^{\frac{n}{2}+2}}) < \frac{1}{4n}$

. . . . . .

Node u_1 contributes $\frac{1}{n}(\frac{1}{\frac{n}{2}+1}+\frac{1}{\frac{n}{2}+1}+\cdots+\frac{1}{2^n}) < \frac{1}{n2^{\frac{n}{2}}}$

It follows that the total contribution of all the states from w_2 to u_1 is smaller than $\frac{1}{n}$. Examining the remaining states (i.e. from u_2 to w_1), the results are identical. It follows that when set A is a tight set of $n/2$ edges we have the following:

$\hat{\Phi}(A) < \frac{\frac{2}{n}}{2^{\frac{1}{2}\frac{1}{2}}} = \frac{4}{n}$. Observe that the following also holds: $\hat{\Phi}(A) = O(\frac{4}{n})$

As a result, by setting $\epsilon = 1/n^c$ for some constant c, we get $\frac{12}{(\hat{\Phi}(A))^2}\log\frac{1}{\epsilon\sqrt{\pi_0}} = O(n^2\log n)$. □

George Lagogiannis was born in Xylokastro, a small town located in the north-east coast of Peloponnisos, Greece. He is an assistant professor at the department of Agricultural Economics and Rural Development in the Agricultural University of Athens. His research is focused on the field of data structures and algorithms.

Stavros Kontopoulos is a Phd student in the area of distributed indexing structures. His interests among others are: distributed system design, data streaming technologies, and NoSql databases. He is also a software engineer with more than 10 years of experience in various industries. His current focus is on machine learning and data streaming.

Christos Makris is an Associate Professor in the University of Patras, Department of Computer Engineering and Informatics, from September 2017. Since 2004 he served as an Assistant Professor in CEID, UoP, tenured inthat position from 2008. His research interests include Data Structures,Information Retrieval, Data Mining, String Processing Algorithms, Computational Geometry, Internet Technologies, Bioinformatics, and Multimedia Databases. He has published over 100 papers in refereed scientific journals and conferences and has more than 600 citations excluding self-citations (h-index:16).

Received: March 10, 2018; Accepted: July 25, 2018.

Computer Science and Information Systems 16(1):227–245 https://doi.org/10.2298/CSIS180220014B

Retinal Blood Vessel Segmentation Based on Heuristic Image Analysis

Maja Braović[1], Darko Stipaničev[1] and Ljiljana Šerić[1]

Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
Ruđera Boškovića 32, 21000 Split, Croatia
maja.braovic@fesb.hr

Abstract. Automatic analysis of retinal fundus images is becoming increasingly present today, and diseases such as diabetic retinopathy and age-related macular degeneration are getting a higher chance of being discovered in the early stages of their development. In order to focus on discovering those diseases, researchers commonly preprocess retinal fundus images in order to detect the retinal landmarks - blood vessels, fovea and the optic disk. A large number of methods for the automatic detection of retinal blood vessels from retinal fundus images already exists, but many of them are using unnecessarily complicated approaches. In this paper we demonstrate that a reliable retinal blood vessel segmentation can be achieved with a cascade of very simple image processing methods. The proposed method puts higher emphasis on high specificity (i.e. high probability that the segmented pixels actually belong to retinal blood vessels and are not false positive detections) rather than on high sensitivity. The proposed method is based on heuristically determined parametric edge detection and shape analysis, and is evaluated on the publicly available DRIVE and STARE datasets on which it achieved the average accuracy of 96.33% and 96.10%, respectively.

Keywords: Retinal blood vessels, fundus images, heuristic analysis, image segmentation.

1. Introduction

According to the World Health Organization (WHO) [27], some of the leading causes of visual impairment are glaucoma, diabetic retinopathy and age-related macular degeneration. These diseases can be identified and/or monitored by using a non-intrusive technique called retinal fundus imaging. Retinal fundus imaging is usually performed in hospitals or clinics with a specialized digital mydriatic or non-mydriatic fundus camera the output of which is an image of the retina of the eye. This image shows various retinal landmarks such as blood vessels, optic disc, fovea and macula, but it can also show abnormalities that could potentially represent a sign of disease. Traditionally, an analysis of these images and their subsequent diagnosis was performed by ophthalmologists, specialized medical doctors trained to be experts in the field of ophthalmology. Unfortunately, this analysis was often both costly and slow. Today, many automated methods for the detection of abnormalities in retinal images exist (e.g. [3], [7]), and they often include the segmentation of retinal blood vessels.

In this paper we propose a heuristically inspired automated method for the segmentation of retinal blood vessels from retinal fundus images. This method is based on a

modified SUSAN edge detector [37] [38] and shape analysis. For shape analysis we used commonly employed measures such as compactness, rectangularity and eccentricity, and gave detailed analysis of thresholds and constraints we used for their values. The proposed method was evaluated on two publicly available retinal image datasets, DRIVE (Digital Retinal Images for Vessel Extraction) [25] [41] and STARE (STructured Analysis of the Retina) [14] [15], on which it achieved the average accuracy of 96.33% and 96.10%, respectively. When compared with a number of existing methods on the same datasets, it achieved the overall highest accuracy and specificity, and the lowest false positive rate (*FPR*).

The proposed method can serve as one of the first steps in automatic systems whose goal is to detect various retinal abnormalities from retinal fundus images. The application of the proposed method is not limited to medical use only, as there are other applications that use retinal blood vessel segmentation as well (e.g. biometrics [4], [8], [11]).

This paper is structured as follows. In Section 2 we give an overview of the related work. In Section 3 we give an overview of the proposed method for the automatic segmentation of retinal blood vessels. In Section 4 we present an evaluation of the proposed method and compare it to various state-of-the-art methods on two publicly available datasets. In Section 5 we give a conclusion and discuss future work.

2. Related work

Methods for automatic retinal blood vessel segmentation can broadly be categorized as either being based on classifiers, matched filters or rules.

Classifier-based methods are fairly recent and they use classifiers to either segment the blood vessels or to improve the segmentation results obtained in some other way. Different classifiers are used for this purpose, such as SVM [42], Bayes [39] or neural networks [19] [21], and they usually perform binary classification, i.e. classify pixels into *vessel* or *non-vessel* categories.

Matched filters for blood vessel detection in retinal images were introduced by Chaudhuri *et al.* in [6], where they were constructed on the basis of several properties of the blood vessels, one of them being the property that the gray-level profile of the blood vessel's cross-section can be approximated by a Gaussian curve. Chaudhuri *et al.* constructed 12 kernels (for 12 different orientations with angular resolution of 15°) to detect piecewise linear segments that belong to blood vessels, and for each pixel of the image only kept the maximum of responses to those kernels. Matched filters or their improvements and modifications are commonly used today in retinal blood vessel detection, for example in [2], [5], [18], [40], [43] and [44].

Rule-based methods incorporate methods for retinal blood vessel segmentation that are based on mathematical morphology, line detection, various transforms, etc. Examples of these kinds of blood vessel segmentation methods are [1], [12], [20], [22], [23], [24], [29], [34] and [35]. Nguyen *et al.* [24] used multi-scale line detection for retinal blood vessel segmentation in order to avoid the drawbacks of the basic line detector. Santhi *et al.* [35] used Shearlet transform in combination with mathematical morphology, connected component analysis and length filtering in order to segment retinal blood vessels. Mendonça et al. [22] used differential filters and morphological processing for retinal blood vessel segmentation. Their method included the detection of vessel centerlines

that were later expanded by a region growing process. Miri and Mahloojifar [23] used a method based on a curvelet transform in order to enhance the retinal image and prepare it for blood vessel segmentation. They performed edge detection by using multistructure elements morphology, and removed false blood vessel edges by using morphological operators by reconstruction, connected component analysis and length filtering. Salazar-Gonzalez *et al.* [34] used a graph cut technique to segment the retinal blood vessels. Akram and Khan [1] proposed a retinal blood vessel segmentation method that involved 2-D Gabor wavelet and multilayered thresholding technique. Martinez-Perez *et al.* [20] proposed a method for segmentation of retinal blood vessels in both red-free and fluorescein images. It was based on multiscale analysis and able to detect blood vessels with different properites (e.g. width or length). Panchal *et al.* [29] presented an algorithm for retinal feature extraction and some of the methods that they used were edge detection, thinning operation, and scanning window (of size 3x3 pixels) analysis that searched for ridges and bifurcations. Hatanaka *et al.* [12] proposed a blood vessel detection method that used a square double ring filter. The width of the inner square was set to 3 pixels, while the width of the outer square was set to 39 pixels. The output image obtained when this filter was applied to the retinal fundus image represented the differences in average pixel values in the square and the rim regions. Hatanaka *et al.* [12] also presented a method for the detection of blood vessels crossing sections that used a ring filter the radius of which was experimentally set to 40 pixels. The edge detection method that we used in this paper is somewhat similar to the blood vessel detection method presented in [12].

Algorithms for the automatic segmentation of retinal blood vessels are commonly evaluated on DRIVE (Digital Retinal Images for Vessel Extraction) [25] [41] and STARE (STructured Analysis of the Retina) [14] [15] image datasets. DRIVE dataset consists of 40 retinal fundus images that are divided into training and testing sets, each of which contains 20 images. The images in the testing set have two ground truth segmentations of blood vessels each. The first of these two segmentations is commonly used as a gold standard. STARE dataset [16] contains 20 images that have two ground truth segmentations of blood vessels each. The first ground truth segmentations are provided by Adam Hoover [16] and are commonly used as a gold standard. In this paper we evaluated the proposed method on DRIVE and STARE datasets, as they are most commonly used in the evaluation of retinal blood vessel segmentation methods.

Additional information on the various blood vessel segmentation methods and techniques in digital retinal images acquired from fundus cameras can be found in [9].

3. Proposed method

The proposed method for retinal blood vessel segmentation encompasses iterative edge detection and shape analysis procedures. An iterative approach in this case assumes the modification of heuristically determined parameters used for edge detection and/or shape analysis at each iteration of the algorithm, thus allowing these methods to detect retinal blood vessels in different lighting conditions. 20 images from the training set of images from the DRIVE dataset were used in the heuristic determination of all of the parameters for edge detection and shape analysis procedures that we used in our method.

A brief overview of the proposed retinal blood vessel segmentation method is given in Figure 1. It can se seen that the proposed methods consists of 3 parts: preprocess-

ing step, heuristic image analysis step, and postprocessing step. Preprocessing step includes image sharpening and smoothing. Heuristic image analysis step is the main part of the proposed method and it includes the derivation of heuristic rules that we used in the proposed method. Postprocessing step encompasses morphological closing operation. Proposed method is schematically shown in Figure 2. Thresholds shown in Figure 2 are explained in detail in section 3.2.

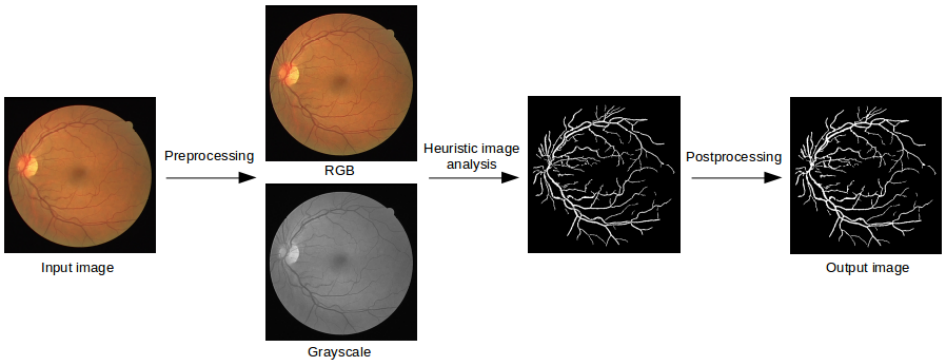


Fig. 1. Overview of the proposed method. The input image was taken from the DRIVE dataset

3.1. Preprocessing

Color fundus images often have low contrast between blood vessels and image background, so some image enhancement technique is usually necessary to improve later segmentation results. For this purpose we employ a 3x3 sharpening filter to the original input image, and a 3x3 Gaussian smoothing filter to the obtained resultant image. Gaussian smoothing filter is used to lower the effects of noise in the subsequent fundus image analysis.

The proposed method uses RGB values of image pixels in the process of extraction of the circular retinal fundus image area from the darker background, but uses grayscale images for every other step of the algorithm.

3.2. Heuristic image analysis

Heuristic image analysis is the main part of the proposed method for retinal blood vessel segmentation. As shown in Figure 2, it consists of 3 iterations that are detailed in the text below.

1. iteration

Edge detection is the first intuitive step in the segmentation of retinal blood vessels from fundus images. The edge detection method that we decided to use is similar to the SUSAN (*Smallest Univalue Segment Assimilating Nucleus*) edge detector [37], [38]. SUSAN edge detector examines a circular neighbourhood (whose radius is usually 3.4

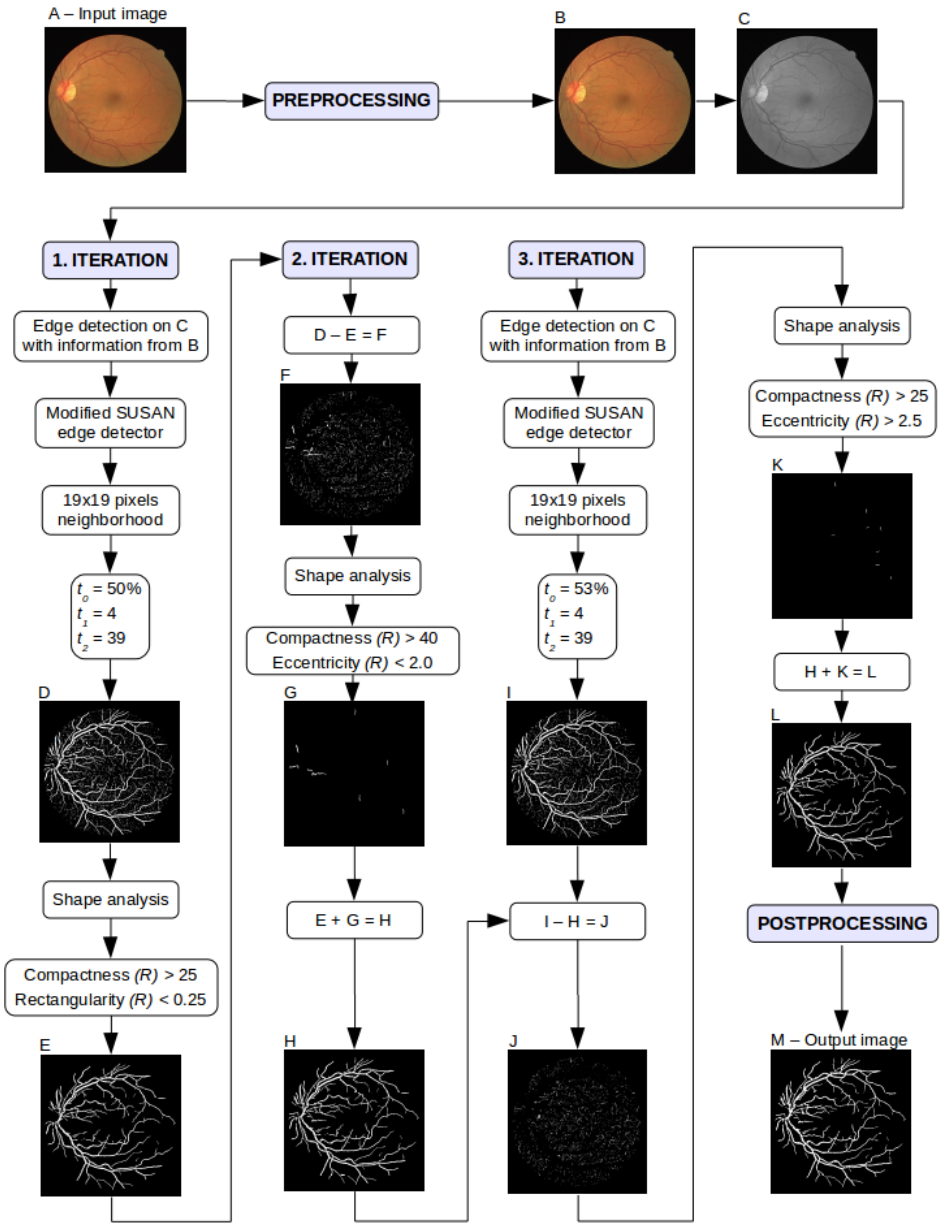


Fig. 2. Proposed method. The input image was taken from the DRIVE dataset

pixels) around each pixel and compares the brightness of the central pixel to those in its neighbourhood. In SUSAN edge detector, neighbouring pixels that have the same or similar brightness as the central pixel are called USAN (*Univalue Segment Assimilating Nucleus*), and from the USAN's size, centroids and second moments, features and edges

on the image can be detected. Since blood vessels in retinal images appear darker than the background [6], an edge detector like the SUSAN edge detector can provide good edge detection results. Instead of using original SUSAN edge detector, in this paper we used a modified version. This modified version uses larger size of the local pixel neighbourhood than it is usually used in the original SUSAN edge detector, and rectangular pixel neighbourhood instead of a circular one.

The edge detection method that we used in the segmentation of retinal blood vessels can be summarized as follows. For each pixel of the grayscale retinal image we examine a rectangular local neighbourhood and compare the grayscale intensity of each pixel in the neighbourhood to the one in its center. If the intensity of the central pixel is lower than the intensities (temporarily increased by a small threshold $t_1 = 4$) of more than $t_0 = 50\%$ of neighbouring pixels, and if its intensities in the RGB color space are all above a certain threshold $t_2 = 39$, then we assume that the center pixel potentially belongs to a blood vessel. Threshold t_0 was chosen with the purpose of locating pixels that are likely to belong to either the edges of blood vessels or to their inner pixels. It allows us to dismiss pixels that are part of a bright uniform area that does not have any sort of edge in its vicinity and therefore probably does not belong to a blood vessel at all. Threshold t_1 is used to lower the effects of noise on the final blood vessel segmentation and was chosen on the basis of a graph shown in Figure 3. This graph shows how the accuracy, sensitivity, specificity and false positive rate (*FPR*) change for this step of the proposed method when the threshold t_1 changes. The final value of $t_1 = 4$ was selected as a trade-off between the accuracy, sensitivity and specificity (which we tried to keep high), and *FPR* (which we tried to keep low) of the edge detection method that we used. Even though the accuracy was higher when the threshold t_1 assumed values higher than 4, the negative impact that this increase would have on the sensitivity of the edge detection method that we used would be greater that the positive impact this increase would have on the accuracy, specificity or *FPR*, i.e. sensitivity would decrease more than the accuracy or specificity would increase or *FPR* decrease, as can be seen from Figure 3.

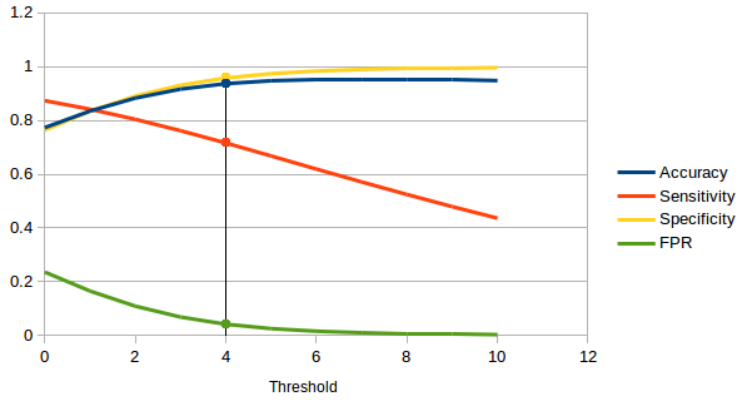


Fig. 3. Accuracy, sensitivity, specificity and *FPR* of edge detection algorithm given the local neighbourhood size of 19x19 pixels and different values of threshold t_1

Threshold t_2 is used to ignore the dark pixels that are outside of the circular fundus image area and was estimated heuristically. Other methods for blood vessel segmentation (e.g. [23]) often use Otsu's method [28] for this purpose, but since thresholding the channels of an RGB image with threshold t_2 usually accomplishes the separation of a circular fundus image from the darker background very well, we used that instead as it was simple to implement.

The size of the local pixel neighbourhood used in the edge detection part of the retinal blood vessel segmentation in this paper was determined heuristically and fixed at 19x19 pixels. If this size was smaller, the algorithm would have had problems with the detection of inner blood vessel pixels because the local neighbourhood around one of those pixels would be too uniform to be picked up by the modified SUSAN edge detector. If this size was larger, it would introduce too much noise, especially in the macula region of the retina. To illustrate our point, Figure 4 shows examples of blood vessel segmentation using modified SUSAN edge detector for local neighbourhoods of different sizes. Local neighbourhood size of 19x19 pixels was chosen as a trade-off between the amount of noise and the amount of accurately segmented blood vessel pixels.

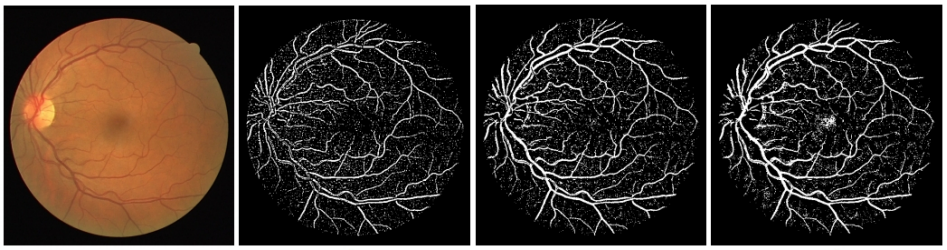


Fig. 4. Examples of edge detection for local neighbourhoods of different sizes. (a) Image from the DRIVE dataset, (b) edge detection for 9x9 neighbourhood, (c) edge detection for 19x19 neighbourhood, (d) edge detection for 35x35 neighbourhood

In order to remove the noise from the image obtained after edge detection, we first extract the connected components that represent possible blood vessel pixels (regions) from the image, and then perform shape analysis of each of those regions. For each of those regions we calculate 4 shape descriptors: their area, perimeter, compactness and rectangularity. The area is defined as the total number of pixels in the region. The perimeter is defined as the number of pixels on the border of a region, and in the process of its calculation we used 8-connectivity. 8-connectivity is commonly used in image processing and more information about it can be found in [31]. Compactness is defined as the ratio of the squared perimeter of the region and the region's area [17]. Rectangularity is defined as the ratio of the region's area and the area of the minimum bounding rectangle of that region [32].

The region R is likely to belong to a blood vessel if the constraints 1 and 2 are true. Image regions that satisfy these constraints are the output of the first iteration of the proposed method.

$$compactness(R) > C_{C_1} \tag{1}$$

$$rectangularity(R) < C_{R_1} \tag{2}$$

The parameters C_{C_1} and C_{R_1} were set to 25 and 0.25, respectively, and were chosen on the basis of graphs shown in Figure 5. These graphs show how the accuracy, sensitivity, specificity and *FPR* change for different values of parameters C_{C_1} and C_{R_1} after the first iteration of the proposed method. It can be seen from Figure 5 that the two best options for threshold C_{R_1} are 0.2 and 0.25, as the accuracies associated with them are two of the highest ones in the graph (the accuracy associated with $C_{R_1} = 0.2$ is 0.959701, and the accuracy associated with $C_{R_1} = 0.25$ is 0.959691). Additionally, specificity and *FPR* are similar for these two options - specificity is a little bit higher for $C_{R_1} = 0.2$ than it is for $C_{R_1} = 0.25$, and *FPR* is a little bit lower, but these are small differences, as can be seen from Figure 5. However, sensitivity for $C_{R_1} = 0.25$ is significantly higher than the one for $C_{R_1} = 0.2$, and this was the reason why we used $C_{R_1} = 0.25$ instead of $C_{R_1} = 0.2$ in the first iteration of the proposed method. When it comes to parameter C_{C_1}, we decided to use $C_{C_1} = 25$ because that was the point in the graphs in Figure 5 where the sensitivity was kept relatively high, *FPR* was slightly decreasing and specificity was slightly increasing. It can be seen from Figure 5 that $C_{C_1} = 20$ was also a good option, but we chose not to use it as it had slightly higher *FPR* and slightly lower specificity in comparison to $C_{C_1} = 25$, and it also decreased the accuracy of the overall proposed method.

2. iteration

Even though the majority of blood vessel regions satisfied the constraints 1 and 2, some of them did not. Thinner blood vessels and blood vessels the color of which is very similar to the color of the background might not satisfy these constraints, as illustrated in Figure 6. Figure 6 shows the results of edge detection for local neighbourhood of 19x19 pixels, and it also shows the regions that satisfy and do not satisfy constraints 1 and 2. It can be seen that the regions labeled as noise contain a small number of blood vessels that the edge detection method failed to extract in their entirety. These blood vessels are mostly represented only by their unconnected fragments, and each of those fragments is treated as a different region. Since the observed areas are very small, they do not demonstrate the same characteristics in terms of compactness and rectangularity as blood vessels that were not fragmented as much.

Since some of the blood vessel regions were segmented as noise because of their fragmentation, we performed additional shape analysis on all of the regions that were detected as possible blood vessel regions by an edge detection algorithm, but dismissed after the initial shape analysis procedure. In this additional shape analysis we used compactness and eccentricity (the ratio of the length and width of a minimal bounding rectangle of a region, where length is the longer side of the rectangle [10]) as shape descriptors in a way shown in constraints 3 and 4. Image regions that satisfy the constraints 3 and 4, alongside the output of the first iteration of the proposed method, represent the output of the second iteration of the proposed method.

$$compactness(R) > C_{C_2} \tag{3}$$

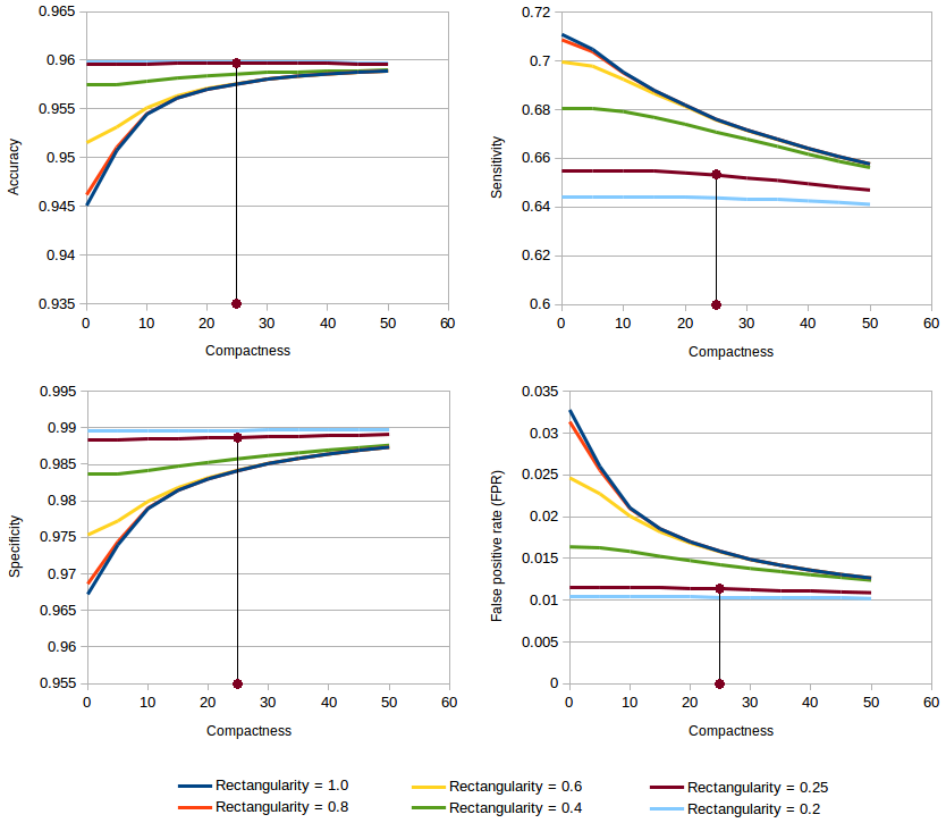


Fig. 5. Graphs used in the process of determination of parameters C_{C_1} and C_{R_1}

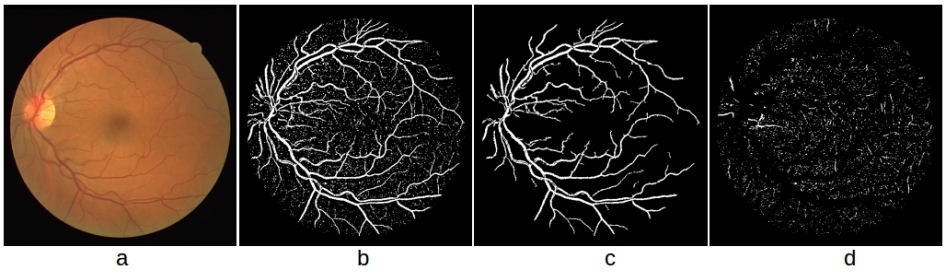


Fig. 6. (a) Image from the DRIVE dataset, (b) edge detection for 19x19 local neighbourhood, (c) regions labeled as blood vessels after shape analysis, (d) regions labeled as noise after shape analysis

$$eccentricity(R) >= C_{E_1} \tag{4}$$

The parameters C_{C_2} and C_{E_1} were set to 40 and 2.0, respectively, and were chosen on the basis of graphs shown in Figure 7. These graphs show how the accuracy, sensitivity,

specificity and *FPR* change for different values of parameters C_{C_2} and C_{E_1} after the second iteration of the proposed method. It can be seen from Figure 7 that the two best options for parameter C_{E_1} are 2.0 and 3.0, because the accuracy and specificity associated with them are high, and *FPR* is low. Since the sensitivity is visibly higher for $C_{E_1} = 2.0$ than for $C_{E_1} = 3.0$, we chose 2.0 as the value for parameter C_{E_1}. Furthermore, we used 40 as the value for parameter C_{C_2} because the accuracy associated with it was one of the highest in the first graph shown in Figure 7. The accuracy and specificity associated with $C_{C_2} = 45$ and $C_{C_2} = 50$ were slightly higher than the one associated with $C_{C_2} = 40$, and their *FPR* was also slightly lower, but $C_{C_2} = 45$ and $C_{C_2} = 50$ had lower sensitivity in comparison to $C_{C_2} = 40$. We decided to use $C_{C_2} = 40$ instead of $C_{C_2} = 45$ or $C_{C_2} = 50$ because, out of all the tests we performed, the use of this value gave the highest accuracy results for the overall proposed method. We should mention that the Sinthanayothin *et al.* [36] used a rule similar to constraint 3 in their blood vessel segmentation procedure. In the post-processing stages of their blood vessel detection method, one of the rules that they used was that the compactness of an isolated region should be above 39 in order to possibly belong to a blood vessel.

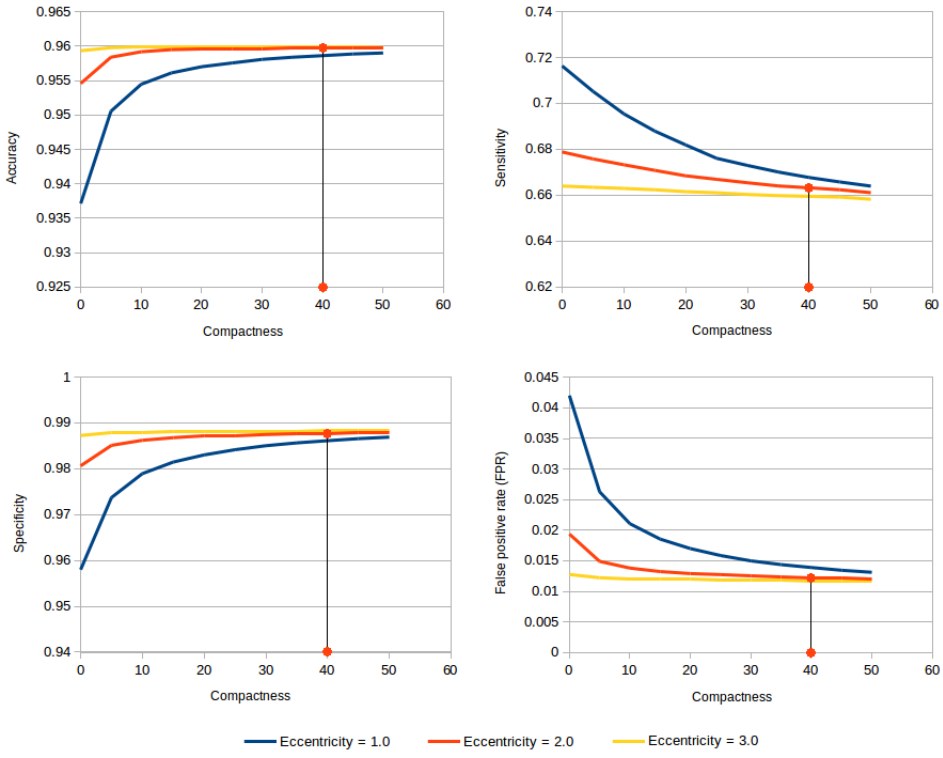


Fig. 7. Graphs used in the process of determination of parameters C_{C_2} and C_{E_1}

3. iteration

In order to locate possible blood vessel pixels that were left undetected by the previous two iterations, we perform two additional steps. In the first step we perform edge detection on the original fundus image with parameters $t_0 = 53\%$ and $t_1 = 4$, and neighbourhood size of 19x19 pixels. The reason why we set threshold t_0 to 53% instead of 50% as we used previously is because we wanted to make edge detection a little bit stricter in this iteration of the proposed method because most of the blood vessels pixels that exist in the input fundus image have been found until now. This stricter rule (which we derived heuristically) for edge detection allowed us to work with fewer pixels in this iteration of the proposed method, and therefore speed it up. The resultant image we obtain after the edge detection process is a binary image where white pixels correspond to detected edges, and black pixels to the background. Any white pixel in this image that was segmented as a blood vessel pixel in the first two iterations of the proposed method is transformed into a black pixel to eliminate it from further analysis. In the second step we perform shape analysis of the connected components of this new image, and only treat regions that satisfy constraints 5 and 6 as possible blood vessels. These regions, alongside the blood vessel pixels detected in the first two iterations of the proposed method, represent the output of the third iteration of the proposed method.

$$compactness(R) > C_{C_3} \tag{5}$$

$$eccentricity(R) > C_{E_2} \tag{6}$$

The parameters C_{C_3} and C_{E_2} were set to 20 and 2.5, respectively, were chosen on the basis of graphs shown in Figure 8. These graphs show how the accuracy, sensitivity, specificity and *FPR* change for different values of parameters C_{C_3} and C_{E_2} after the third iteration of the proposed method. It can be seen from Figure 8 that the two best options for parameter C_{E_2} are 2.5 and 3.0, because the accuracies associated with them are the highest. We decided to use $C_{E_2} = 2.5$ instead of $C_{E_2} = 3.0$ because the sensitivity associated with $C_{E_2} = 2.5$ was mostly higher than the one associated with $C_{E_2} = 3.0$. When it comes to parameter C_{C_3}, the value we decided to use for it was 20 because we found that to be a good trade-off between high accuracy, sensitivity and specificity, and low *FPR*.

3.3. Postprocessing

Morphological closing operation (dilation followed by erosion) [30] [33] is applied to the image obtained by heuristic image analysis. The purpose of this postprocessing step is to fill in any small openings that the image might contain. The resultant image represents the final retinal blood vessel segmentation, i.e. the output image of the proposed method.

Figure 9 shows an example of retinal blood vessel segmentation using the proposed method. Evaluation of the proposed method on DRIVE and STARE datasets is given in section 4.

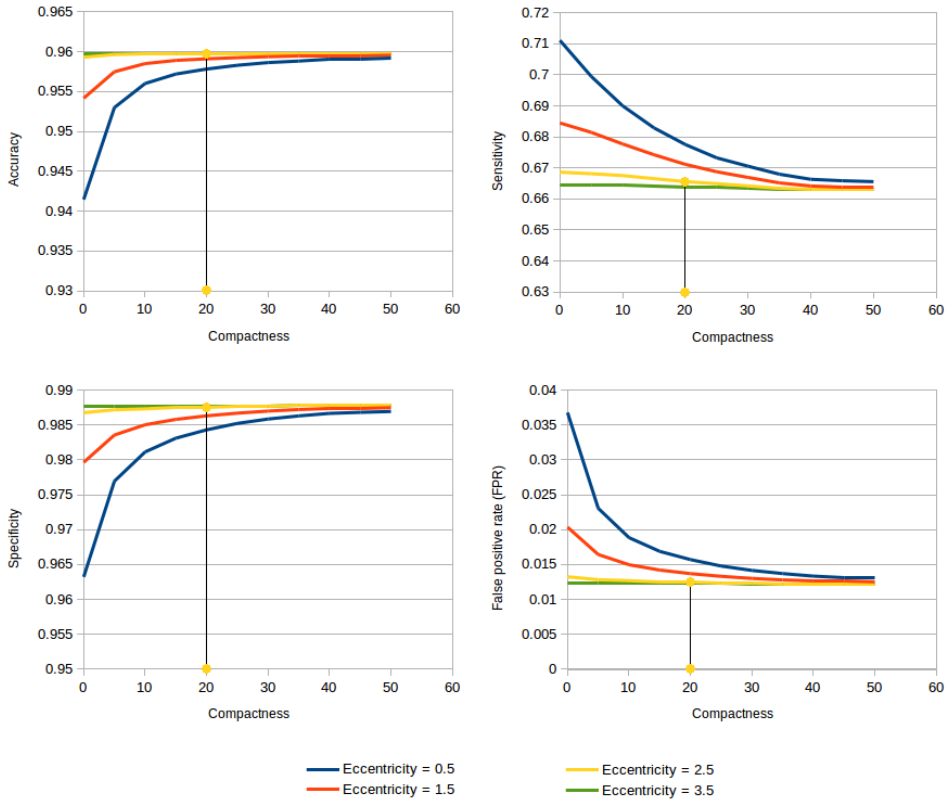


Fig. 8. The process of determination of parameters C_{C_3} and C_{E_2}

4. Results and discussion

The proposed blood vessel segmentation method was implemented in C++. For various image processing related functions we used the OpenCV library [26]. In the evaluation of the proposed method we used commonly employed statistical measures: sensitivity (true positive rate, *TPR*), specificity (true negative rate, *TNR*), false positive rate (*FPR*), and accuracy. These measures are defined in terms of true positives (*TP*), true negatives (*TN*), false positives (*FP*) and false negatives (*FN*). In our case, *TP* represents pixels accurately identified as belonging to the blood vessles, *TN* representes pixels accurately identified as belonging to the image background, *FP* represents pixels inaccurately identified as belonging to the blood vessles, and *FN* representes pixels inaccurately identified as belonging to the image background. Sensitivity, specificity, accuracy and *FPR* are defined by equations 7, 8, 9 and 10.

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

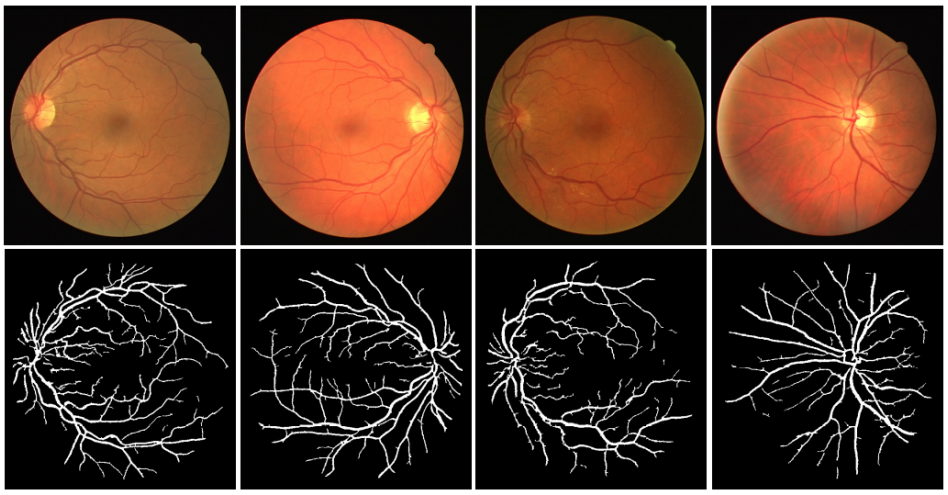


Fig. 9. Retinal blood vessel segmentation using the proposed method. First row shows images from the DRIVE dataset, while the second row shows corresponding segmentations obtained by the proposed method

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$FPR = \frac{FP}{FP + TN} \tag{10}$$

The proposed method was evaluated on DRIVE and STARE image datasets. For the evaluation of our method we used 20 images from the testing set of images from the DRIVE dataset, and for their ground truth segmentations of blood vessels we used the segmentations provided by the first observer as that is the usual practice amongst researchers. When it comes to the STARE image dataset, 20 images were accompanied by ground truth segmentations of blood vessels so we used them in the evaluation of our method. For their ground truth segmentations of blood vessels we used the segmentations provided by the first observer as that is the usual practice amongst researchers.

The values for *TP*, *TN*, *FP* and *FN* that we obtained for the DRIVE database were the following: 405834, 5950875, 70380 and 172111, respectively. The values for *TP*, *TN*, *FP* and *FN* that we obtained for the STARE database were the following: 447894, 7692080, 133867 and 196159, respectively. We used these values in the calculation of sensitivity, specificity, FPR and accuracy of the proposed method for blood vessel segmentation. Sensitivity, specificity, FPR and accuracy of the proposed method was rounded to 4 decimals, as can be seen from Tables 1 and 2.

Tables 1 and 2 show the comparison of the proposed method against various other methods on the images from the DRIVE and STARE datasets, respectively. Retinal blood vessel segmentation results of the methods that we compared against our own method were taken from the original papers where those methods were published, as shown and

cited in Tables 1 and 2. The results we obtained for the proposed method are shown in Tables 1 and 2 in terms of average sensitivity, specificity, *FPR* and accuracy.

Method	**Sensitivity (TPR)**	**Specificity (TNR)**	**FPR**	**Accuracy**
Mendonça and Campilho [22] (Input: gray-intensity images)	0.7315	/	0.0219	0.9463
Mendonça and Campilho [22] (Input: green channel images)	0.7344	/	0.0236	0.9452
Kande et al. [18]	/	/	/	0.8911
Xu and Luo [42]	0.7760	/	/	0.9328
Zhang et al. [44]	0.7120	/	0.0276	0.9382
Marín et al. [19]	0.7067	0.9801	/	0.9452
Miri and Mahloojifar [23]	0.7352	/	0.0205	0.9458
Yin et al. [43]	0.6522	0.9710	/	0.9267
Salazar-Gonzalez et al. [34]	0.7512	/	0.0316	0.9412
Melinscak et al. [21]	0.7276	/	0.0215	0.9466
Sreejini and Govindan [40]	0.7132	0.9866	/	0.9633
Zhao et al. [45]	0.742	0.982	/	0.954
Proposed method	**0.7022**	**0.9883**	**0.0117**	**0.9633**

Table 1. Performance evaluation of different retinal blood vessel segmentation methods on the DRIVE dataset

It can be seen from Table 1 that the proposed method has higher accuracy on the DRIVE dataset than any of the listed methods except for the one proposed in [40], the accuracy of which is the same as ours. On the other hand, it can be seen from Table 2 that our method has higher accuracy on the STARE dataset than any other listed method, and that includes the one presented in [40]. The results that we obtained on the STARE dataset are even more significant given the fact that we did not train our method on that dataset, but rather only on the training part of the DRIVE dataset.

It can be seen from Tables 1 and 2 that the sensitivity of our method is not as high as it is in most methods the results of which can be seen in those tables. We found that the most probable reason for this is the fact that our method has higher specificity than any other method listed in Tables 1 and 2. From the experience we gained while working on the proposed blood vessel segmentation method, the higher the sensitivity was of our method, the lower the specificity was. When faced with choosing between the higher sensitivity or higher specificity of the proposed method, we chose higher specificity. In this case, higher specificity meant that the proposed method would locate fewer of the blood vessels on the DRIVE and STARE datasets than it would have if it had higher sensitivity, but that the

Method	Sensitivity (TPR)	Specificity (TNR)	FPR	Accuracy
Kande et al. [18]	/	/	/	0.8976
Zhang et al. [44]	0.7177	/	0.0247	0.9484
Marín et al. [19]	0.6944	0.9819	/	0.9526
Salazar-Gonzalez et al. [34]	0.7887	/	0.0367	0.9441
Sreejini and Govindan [40]	0.7172	0.9687	/	0.9500
Zhao et al. [45]	0.780	0.978	/	0.956
Yin et al. [43]	0.7248	0.9666	/	0.9412
Proposed method	**0.6954**	**0.9829**	**0.0171**	**0.9610**

Table 2. Performance evaluation of different retinal blood vessel segmentation methods on the STARE dataset

blood vessels it would locate would have a higher probability of actually belonging to the blood vessels and not the image background or other retinal features or signs of various retinal diseases. Furthermore, Tables 1 and 2 also contain information about the *FPR*. It can be seen from Tables 1 and 2 that the *FPR* of our method is the lowest out of all the other compared methods listed in Tables 1 and 2 whose *FPR* is known, which means that the number of pixels that the proposed method inaccurately identified as belonging to retinal blood vessels (i.e. *FP*) is lower than in those other methods.

It can be seen from the text above that we decided to put higher emphasis on the specificity rather than on sensitivity of the proposed method. This decision was made in the wake of our future research where we plan to use the proposed method in the segmentation and classification of retinal diseases. In that research we will use the proposed method to remove pixels belonging to retinal blood vessels from the retinal fundus images in order to have a more reliable retinal disease segmentation and classification later on. If the sensitivity of the proposed method was higher, its specificity would decrease and this would cause problems because the proposed method would probably detect some signs of the disease as retinal blood vessels, which would exclude these pixels from the automated retinal disease segmentation and classification. This would, of course, perhaps cause the automated method to miss out on signs of a disease, which in our opinion is a larger problem than having false signs of a disease that can easily be disregarded by an experienced ophthalmologist once the automated system for retinal disease classification and segmentation raises an alarm.

Additionaly, we tested the proposed method on 20 images from the STARE dataset that we divided into two sets: images of healthy retinas that do not show any signs of the disease (9 images), and images of abnormal retinas that do show signs of the disease (11 images). Information about which image from the STARE dataset correspondes to which of these sets was obtained from [13]. Results that we obtained are given in the form of average sensitivity, specificity, *FPR*, and accuracy, and shown in Table 3. It can be seen from the table that the accuracy and specificity of the proposed method were high and the *FPR* was low for both of the image sets used. Sensitivity was significantly

lower on the images that showed retinas with signs of the disease, but this was expected given that a large number of retinal diseases can appear similar in color to retinal blood vessels. The proposed method is designed in such a way that it rathers fails to detect a pixel belonging to a retinal blood vessel than incorrectly detecting other pixels (usually connected to retinal abnormalities) as retinal blood vessels. As discussed previously, this decision was made in the wake of our future research where we plan to use the proposed method in the segmentation and classification of retinal diseases.

	Healthy retina	**Abnormal retina**
Sensitivity (TPR)	0.7647	0.6264
Specificity (TNR)	0.9802	0.9850
FPR	0.0198	0.0150
Accuracy	0.9620	0.9602

Table 3. Performance evaluation of the proposed method on the healthy and abnormal retinal images from the STARE dataset

Based on the results shown in Tables 1 and 2 and the discussion regarding them, we can conclude that the proposed method has the overall highest accuracy and specificity, and the lowest *FPR*, which would make it effective to use in retinal blood vessel segmentation. However, the reader should keep in mind that the proposed method was tested on low-resolution images from the DRIVE and STARE datasets. If it is to be used on high-resolution retinal fundus image datasets the heuristically determined parameters should probably be adjusted accordingly.

5. Conclusion

In this paper we presented a novel method for iterative retinal blood vessel segmentation based on heuristic image analysis. The presented analysis included the derivation of heuristic rules and thresholds that we used in edge detection and shape analysis of retinal fundus images with the ultimate aim of achieving automatic and high accuracy retinal vessel segmentation. The proposed method can be used in biometrics or in computer-aided diagnosis in the medical field of ophthalmology. It was evaluated on the publicly available DRIVE and STARE datasets on which it achieved the average accuracy of 96.33% and 96.10%, respectively. It achieved the overall highest accuracy and specificity and the lowest false positive rate out of all the methods it was compared against, and the obtained results indicate that it is effective in retinal blood vessel segmentation. In our future work we intend to expand the research proposed in this paper to the automatic detection of retinal abnormalities such as diabetic retinopathy, age-related macular degeneration and glaucoma.

Acknowledgments. This work was partly supported by the Ministry of Science, Education and Sport of the Republic of Croatia under grant "ViO - Vision Based Intelligent Observers" (in Croatian: "ViO - Vidom temeljeni inteligentni opserveri"). This work is part of activities of ACROSS

- Centre of Research Excellence for Advanced Cooperative Systems (http://across.fer.hr). The authors would like to thank Dr. Ljubo Znaor from the Clinical Hospital Center in Split for all the help he provided during their research on this project. The authors would also like to thank Dr. Stephen M. Smith from the University of Oxford for letting them know that the SUSAN edge detector was always free to use for non-commercial purposes.

References

1. Akram, M.U., Khan, S.A.: Multilayered thresholding-based blood vessel segmentation for screening of diabetic retinopathy. Engineering with Computers 29(2), 165–173 (2013)
2. Al-Rawi, M., Qutaishat, M., Arrar, M.: An improved matched filter for blood vessel detection of digital retinal images. Computers in Biology and Medicine 37(2), 262–267 (2007)
3. Arenas-Cavalli, J.T., Rios, S.A., Pola, M., Donoso, R.: A web-based platform for automated diabetic retinopathy screening. 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems. Procedia Computer Science 60, 557–563 (2015)
4. Barkhoda, W., Akhlaqian, F., Amiri, M.D., Nouroozzadeh, M.S.: Retina identification based on the pattern of blood vessels uzing fuzzy logic. EURASIP Journal on Advances in Signal Processing 113 (2011)
5. Chakraborti, T., Jha, D.K., Chowdhury, A.S., Jiang, X.: A self-adaptive matched filter for retinal blood vessel detection. Machine Vision and Applications 26(1), 55–68 (2015)
6. Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M., Goldbaum, M.: Detection of blood vessels in retinal images using two-dimensional matched filters. IEEE Transactions on Medical Imaging 8(3), 263–269 (1989)
7. Datta, N.S., Dutta, H.S., De, M., Mondal, S.: An effective approach: image quality enhancement for microaneurysms detection of non-dilated retinal fundus image. International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA). Procedia Technology 10, 731–737 (2013)
8. Fatima, J., Syed, A.M., Akram, M.U.: Feature point validation for improved retina recognition. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS) (2013)
9. Fraz, M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A., Owen, C., Barman, S.: Blood vessel segmentation methodologies in retinal images - a survey. Computer Methods and Programs in Biomedicine 108(1), 407–433 (Oct 2012), http://dx.doi.org/10.1016/j.cmpb.2012.03.009
10. Frejlichowski, D., Gościewska, K.: The application of simple shape measures based on a minimum bounding rectangle to the general shape analysis problem. Journal of Theoretical and Applied Computer Science 7(4), 35–41 (2013)
11. Fukuta, K., Nakagawa, T., Hayashi, Y., Hatanaka, Y., Hara, T., Fujita, H.: Personal identification based on blood vessels of retinal fundus images. In: Medical Imaging 2008: Image Processing, Proceedings of SPIE. vol. 6914 (2008)
12. Hatanaka, Y., Muramatsu, C., Hara, T., Fujita, H.: Automatic arteriovenous crossing phenomenon detection on retinal fundus images. In: Proceedings of SPIE, Vol. 7963, 79633V, Medical Imaging 2011: Computer-Aided Diagnosis. vol. 7963 (2011)
13. Hoover, A.: STructured Analysis of the Retina, [Online]. Available: http://cecas.clemson.edu/ ahoover/stare/diagnoses/all-mg-codes.txt (current May 2018)
14. Hoover, A., Goldbaum, M.: Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. IEEE Transactions on Medical Imaging 22(8), 951–958 (2003)
15. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical Imaging 19(3), 203–210 (2000)

16. Hoover, A., Kouznetsova, V., Goldbaum, M.: STARE image dataset (2000), [Online]. Available: http://cecas.clemson.edu/ ahoover/stare/probing/index.html (current March 2017)
17. Iivarinen, J., Peura, M., Särelä, J., Visa, A.: Comparison of combined shape descriptors for irregular objects. In: Proceedings of the 8th British Machine Vision Conference. pp. 430–439 (1997)
18. Kande, G.B., Subbaiah, P.V., Savithri, T.S.: Unsupervised fuzzy based vessel segmentation in pathological digital fundus images. Journal of Medical Systems 34(5), 849–858 (2010)
19. Marín, D., Aquino, A., Gegúndez-Arias, M.E., Bravo, M.: A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. IEEE Transactions on Medical Imaging 30(1) (2011)
20. Martinez-Perez, M.E., Hughes, A.D., Thom, S.A., Bharath, A.A., Parker, K.H.: Segmentation of blood vessels from red-free and fluorescein retinal images. Medical Image Analysis 11(1), 47–61 (2007)
21. Melinscak, M., Prentasic, P., Loncaric, S.: Retinal vessel segmentation using deep neural networks. In: Proceedings of the 10th International Conference on Computer Vision Theory and Applications. pp. 577–582. Berlin, Germany (2015)
22. Mendonça, A.M., Campilho, A.C.: Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. IEEE Transactions on Medical Imaging 25(9) (2006)
23. Miri, M.S., Mahloojifar, A.: Retinal image analysis using curvelet transform and multistructure elements morphology by reconstruction. IEEE Transactions on Biomedical Engineering 58(5) (2011)
24. Nguyen, U.T.V., Bhuiyan, A., Park, L.A.F., Ramamohanarao, K.: An effective retinal blood vessel segmentation method using multi-scale line detection. Pattern Recognition 46(3), 703–715 (2013)
25. Niemeijer, M., Staal, J., van Ginneken, B., Loog, M., Abràmoff, M.D.: Comparative study of retinal vessel segmentation methods on a new publicly available database. In: Proceedings of SPIE - The International Society for Optic Engineering 5370 (2004)
26. OpenCV team: OpenCV (Open Source Computer Vision), [Online]. Available: http://opencv.org/ (current January 2017)
27. Organization, W.H.: Prevention of blindness and visual impairment, causes of blindness and visual impairment, [Online]. Available: http://www.who.int/blindness/causes/en/ (current December 2016)
28. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics SMC-9(1) (1979)
29. Panchal, P., Bhojani, R., Panchal, T.: An algorithm for retinal feature extraction using hybrid approach. 7th International Conference on Communication, Computing and Virtualization 2016, Procedia Computer Science 79, 61–68 (2016)
30. Petrou, M., Sevilla, P.G.: Image Processing - Dealing with Texture. John Wiley & Sons, Ltd (2006)
31. Rosenfeld, A.: Connectivity in digital pictures. Journal of the Association for Computing Machinery 17(1), 146–160 (1970)
32. Rosin, P.L.: Measuring shape: ellipticity, rectangularity, and triangularity. Machine Vision and Applications 14(3), 172–184 (2003)
33. Russ, J.C.: The Image Processing Handbook, Second Edition. CRC Press, Inc. (1995)
34. Salazar-Gonzalez, A., Kaba, D., Li, Y., Liu, X.: Segmentation of the blood vessels and optic disk in retinal images. IEEE Journal of Biomedical and Health Informatics 18(6), 1874–1886 (2014)
35. Santhi, N., Shajula, G.D., Ramar, K.: A novel algorithm to analyze retinal image using morphological operators. International Conference On Modelling Optimization and Computing (ICMOC-2012). Procedia Engineering 38, 3880–3891 (2012)

36. Sinthanayothin, C., Boyce, J.F., Cook, H.L., Williamson, T.H.: Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images. The British Journal of Ophthalmology 83(8), 902–910 (1999)
37. Smith, S.M., Brady, J.M.: SUSAN - A new approach to low level image processing. Technical report TR95SMS1c, Defence Research Agency, Farnborough, Hampshire, UK (1995)
38. Smith, S.M., Brady, J.M.: SUSAN - A new approach to low level image processing. International Journal of Computer Vision 23(1), 45–78 (1997)
39. Soares, J.V.B., Leandro, J.J.G., Cesar-Jr., R.M., Jelinek, H.F., Cree, M.J.: Retinal vessel segmentation using the 2-D Morlet wavelet and supervised classification. IEEE Transactions on Medical Imaging 25(9), 1214–1222 (2006)
40. Sreejini, K.S., Govindan, V.K.: Improved multiscale matched filter for retina vessel segmentation using PSO algorithm. Egyptian Informatics Journal 16(3), 253–260 (2015)
41. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. IEEE Transactions on Medical Imaging 23(4), 501–509 (2004)
42. Xu, L., Luo, S.: A novel method for blood vessel detection from retinal images. BioMedical Engineering OnLine 9(14) (2010)
43. Yin, Y., Adel, M., Bourennane, S.: Automatic segmentation and measurement of vasculature in retinal fundus images using probabilistic formulation. Hindawi Publishing Corporation, Computational and Mathematical Methods in Medicine, Article ID 260410 2013 (2013)
44. Zhang, B., Zhang, L., Zhang, L., Karray, F.: Retinal vessel extraction by matched filter with first-order derivative of Gaussian. Computers in Biology and Medicine 40(4), 438–45 (2010)
45. Zhao, Y., Rada, L., Chen, K., Harding, S.P., Zheng, Y.: Automated vessel segmentation using infinite perimeter active contour model with hybrid region information with application to retinal images. IEEE Transactions on Medical Imaging 34(9), 1797–1807 (2015)

Maja Braović is a post-doctoral researcher at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture at the University of Split, Croatia. She obtained her PhD in Articifial Intelligence in 2015. Her research interests include image processing, medical and forensic image analysis, machine learning, expert systems, cryptography and artificial intelligence in general.

Darko Stipaničev is a full professor at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture at the University of Split, Croatia. He obtained his PhD in Electrical Engineering in 1987. His research interests include image processing, video surveillance and monitoring, robotics, fuzzy systems, and artificial intelligence in general.

Ljiljana Šerić is an assistant professor at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture at the University of Split, Croatia. She obtained her PhD in Artificial Intelligence in 2010. Her research interests include image processing, video surveillance and monitoring, sensor networks, distributed information systems and artificial intelligence in general.

Received: February 20, 2018; Accepted: July 11, 2018.

Computer Science and Information Systems 16(1):247–271 https://doi.org/10.2298/CSIS180430017D

An Empirical Study of Data Visualization Techniques in PACS Design

Dinu Dragan[1], Veljko B. Petrović[1], Dušan B. Gajić[1], Žarko Živanov[1], and Dragan Ivetić[1]

University of Novi Sad, Faculty of Technical Sciences
Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
{dinud,pveljko,dusan.gajic,zzarko,ivetic}@uns.ac.rs

Abstract. The paper presents an empirical study of multidimensional visualization techniques. The study is motivated by the problem of decision making in PACS (Picture Archiving and Communications System) design. A comprehensive survey of visualizations used in literature is performed and these survey results are then used to produce the final set of considered visualizations: tables (as control), scatterplots, parallel coordinates, and star plots. An electronic testing tool is developed to present visualizations to three sets of experimental subjects in order to determine which visualization technique allows users to make the correct decision in a sample decision making problem based on real-world data. Statistical analysis of the results demonstrates that visualizations show better results in decision support than tables. Further, when number of dimensions is large, 2D parallel coordinates show the best results in accuracy. The contribution of the presented research operates on two levels of abstraction. On the object level, it provides useful data regarding the relative merits of visualization techniques for the considered narrow use-case, which can then be generalized to other similar problem sets. On the meta level above, it contributes an enhanced methodology to the area of empirical visualization evaluation methods.

Keywords: visualization, parallel coordinates, star plots, medical imaging, PACS.

1. Introduction

This paper presents an empirical study of visualization techniques used in the evaluation of medical image compression. It aims to select a dominant visualization technique in the considered context, validate this choice (through analysis of literature and direct testing), and demonstrate that visualization is sufficiently useful to be a dominant component in decision support systems. Further, its results are intended to be generalized to multidimensional data visualization as such. The study employs statistically-analyzed empirical data as well as domain expert use case analysis and provides a comprehensive overview of current work in the field. The paper consists of five sections: introduction, related work, proposed approach, data analysis, and conclusion. This introductory section will cover motivation for the research by presenting the problem of image compression in medicine by way of PACS (Picture Archiving and Communications System) as well as outline the general methodology of the study.

1.1. Motivation

This study, while aiming for generalizability, was motivated by the question of decision making in PACS design. A PACS is a very complex system which covers all the aspects of image workflow in one or more medical institutions [4], including mobile medicine and telemedicine applications [29]. PACS is the dominant paradigm in the medical field and remains so, despite occasional criticism [26]. Designing one, naturally, brings to the forefront the question of compression. Image compression in PACS is, generally speaking, desirable because of the relaxation of storage and transmission requirements and the increase of image turnaround time [19]. Compression also allows for the use of lower-capability devices in medicine and telemedicine applications [16]. Lower-capability devices are still relevant despite the recent explosion in device capabilities because telemedical applications are of great interest in less developed countries. This is due to the need for the efficient use of a small number of doctors and the markedly high market penetration of mobile telephony in those regions [49].

Choosing a suitable image compression technique for a PACS is a complex decision-making problem which must take into account various requirements [15], imposed and evaluated by a heterogenous group of stakeholders. Requirements rely on various metrics of quality [14] with over 250 being attested in the literature, though usually only a subset is used [17]. To simplify this considerable task, the authors of the paper have developed a decision-support system called SICEP (Still Image Compression Evaluation for PACS) which can be used to integrate disparate metrics and requirements into a unified data set amenable to analysis.

No matter how unified this data set is, it is still an intimidating amount of information which needs to be presented to stakeholders in PACS design. Contrary to business intelligence (BI) analytics [10], once the unified data set is defined, not even the pre-processing techniques applied in BI (such as data aggregation, drill-downs, or OLAP) can lead to further dimensionality reduction and a solution for presenting the data to stakeholders needs to be found in data visualization techniques [54]. Stakeholders, while experts in their various domain specialties, are not experts in the evaluation of the metrics used, most of which present interacting tradeoffs in properties which require a technical education to merely understand. Our desire was to permit for control on the part of the stakeholders allowing the decision to be entirely theirs while minimizing the cognitive load. Therefore we chose data visualization [47] as a suitable shortcut, and we began to develop a subsystem for data visualization we named VisSys. This, however, raised a new difficulty: how to select a visualization technique to use as the dominant one in our system and how to validate that (a) that technique is the correct choice and that (b) a visualization helps at all compared to presenting the data directly. This paper presents the studies we have performed to find satisfactory answers to both of those questions.

To summarize, the fundamental contribution of the research operates on two levels of abstraction. On the object level it provides useful data regarding the relative merits of visualization techniques for this narrow use-case which can then be generalized to a limited but substantial degree to other similar problem sets (see 2.1). On the meta level above it contributes an enhanced methodology to the area of empirical visualization evaluation methods.

For the studies to be clear, a quick introduction to SICEP/VisSys is required. The image compression evaluation process is guided by the requirements an image compres-

sion technique must meet in a specific PACS. Individual metrics are grouped into one or more requirement indicators to indicate how well compression performed against a given requirement. The SICEP system can adapt to the needs of any PACS by modifying requirements and requirement indicators. This means that VisSys should express the same flexibility as the SICEP system by enabling easy selection of requirement indicators and their modification.

Each of the individual metrics chosen for the SICEP system is important for reaching valid decisions [14]. Therefore, each of them should be represented with an individual dimension in VisSys visualizations. Also, since a requirement indicator can have an arbitrary number of metrics from a total number of observed metrics which is also arbitrarily large, VisSys should accommodate a variable number of dimensions. Thus, designing VisSys required grappling with the problem of multidimensional evaluation which, further, can operate on arbitrarily-created sub-groups of dimensions.

The method we chose to select an appropriate visualization technique is as follows. First, we would perform a literature review to determine which methods theoretically satisfied our goals of visualization with an indefinite number of dimensions as well as which methods are being used in the literature. Second, we would compare in an empirical test commonly used methods, a table view of the data as a control, and the methods we chose based on theory. Then we would compare whatever performed best in this initial test to the leading multidimensional data visualization solutions and, as a control for basic visualization, bar graphs. This we would do in two tests: one with a low number of dimensions and one with a higher number of dimensions in play. This comparison would be accomplished with a statistical procedure based on robust statistical methods, specifically for the data on the continual level of measurement a dependent robust bootstrapped ANOVA with trimmed means, a *post-hoc* test based on Yuen's modification of Student's T-test with Rom's faimilywise error correction method. For data on the categorical measurement level a combination of McNemar's tests preformed iteratively with Holm's correction, a multilevel logistical regression, and testing proportional confidence intervals. This procedure is explained in detail in section 3.5.

Lastly, we would show the visualizations most successful at the empirical tests to actual stakeholders in a PACS and then observe them as they used it to reach a decision, interviewing them about their experience. This method is outlined in further sections of the paper, specifically 2.1 and 3.

2. Related Work

This section presents all aspects of previous work done in this field and related fields which substantially informed the work presented in this paper. It consists of subsections relating to multidimensional data visualization as such, and a section on visualization evaluation.

2.1. Multidimensional Data Visualization

In this paper, we define multidimensional data visualization (MDV) as visualization of a set of variables measured according to continuous or discrete measurement scales where all variables in the set must be visible simultaneously for relationships to be visualized,

and the number of variables exceeds by a significant margin the number of spatial dimensions available for mapping. In the most common type of visualization two variables are shown at the same time, highlighting the relationship between them, such as in a simple line graph where one variable is time and another something we measure. If interactive computer-based tools or 3D printing is used, one additional spatial dimension can be used. To these two or three spatial dimensions, a number of visual variables–as discussed in [7] (which evaluates the use of variables as a basis of visualization) [25] (which evaluates the use of similar mapping in geographical visualizations) and [33] (which proposes an expansion of the visual variable concept into the dynamic domain)–can be added, each being mapped to a dimension of the data. This gives us an absolute upper limit of nine dimensions present at the same time, though, of course, practical concerns tend to reduce that to no more than five.

To overcome this problem, MDV techniques employ something else apart from spatial dimensions to map data dimensions to. Examples include star plots [48] which use axes distributed radially in order to map data, Fig. 1, as well as parallel plots in 2D which are discussed in [27] (which presents the general concept), [30] (which reviews the literature), and [53] (which discusses as one of their salient features their ability to reveal structure in data). Further, parallel coordinates are available in 3D [27], and there is some use of techniques like 3D glyphs [11, 23], and Chernoff faces [6, 42]. Parallel coordinates are especially commonly used when very large data sets need to be displayed [50]. 3D parallel coordinates are an extension of 2D parallel coordinates into the third dimension, which is difficult because of the necessity of a rule for connecting axes. In the case of 2D coordinates direct comparison only works between axes which are neighbors, and which axes are neighbors is simple to determine. In the 3D case the situation is more complex, because if all axes are to be connected to all other axes the display becomes too cluttered. A method for resolving this issue is the use of axis connecting rules. One commonly used rule is used to generate a type of display known as clustered multi-relational parallel coordinates (CMRP). This rule chooses one axis and compares all the other ones to it by placing the chosen axis in the center and connecting the remaining $n-1$ axes to be visualized around it, forming a regular $n-1$ sided prism, see [22].

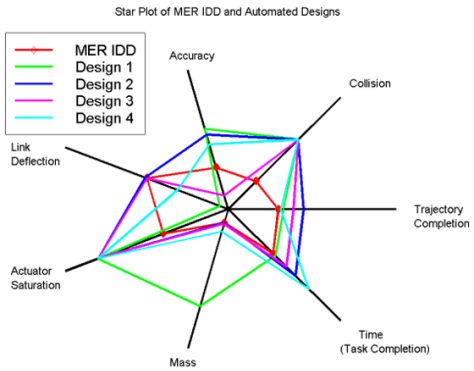


Fig. 1. Star plot, public domain image from NASA for illustrative purposes only

This is much less cluttered and it would suit us to use it but, unfortunately, it did not fulfill our requirements. It shows not the full n dimensions and their relations but, very specifically, the relations of $n-1$ dimensions to one baseline dimension which does not fit our requirements. We still wanted to use a 3D analogue for parallel coordinates, and so extended CMRP coordinates in such a way that there is no central coordinate. Instead, n-dimensional space is presented by mapping n values onto sides of a n-sided regular prism. A point in such a space is then presented by way of a polyline connecting all the axes along the sides forming a regular prism with a slice taken out of it, Fig. 2 (right). This is contrasted with 2D parallel coordinates visible on the left. Clearly, one can only interconnect those values sharing a side of the prism, but this still exceeds the capabilities of non-extended CMRP. These extended CMRP coordinates (ECMRP) are only fully usable using a custom tool to display them interactively [18]. To simplify the nomenclature we will refer to the ECMRP subvariant as simply 3D parallel coordinates.

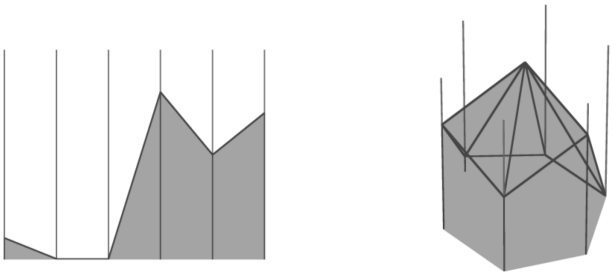


Fig. 2. Parallel coordinates: 2D (on the left) and 3D (on the right) in modified form as extended Clustered Multi-Relational Parallel Coordinates, image from [18]

Star plots, in Fig. 1, and 2D parallel plots in Fig. 2 (left), both use essentially the same method: instead of using spatial dimensions to map data dimensions to, they instead use equally spaced axes, either radially or in parallel arrangement. 3D glyphs and Chernoff faces on the other hand map the dimensions of the data to a specific visual feature of the glyph, in case of 3D glyphs, and on facial features of cartoon faces, in the case of Chernoff faces.

Chernoff faces are a special case of glyph which use facial features specifically in the hope that this will allow access to the preliminal processing characteristic of human facial recognition, thus dramatically increasing data efficiency, especially when presented extremely quickly. This, however, is yet to be shown as actually working, as studies designed to detect this feature have failed to record any effect [42].

Of course, the situation is dramatically altered in the case of dimensionality reduction. It is possible to reduce the number of dimensions shown and, thus, fundamentally alter the factors that affect the choice of a visualization. Normally dimensionality reduction is only to be discussed on the per-case basis, though work has been done on using artificial intelligence [38] to prioritize certain subsets of multidimensional relationships as ones to show to the user as the most 'interesting' where 'interest' is defined as those that provide optimal separation. This, however, is outside the scope of this paper.

2.2. Approaches to Visualization Evaluation

A lot of previous work has gone into analyzing possible approaches to visualization evaluation including Lam's comprehensive taxonomy of approaches [37] buttressed by a literature review [28], as well as a comprehensive review by Carpendale [8]. Using Lam's taxonomy, visualization evaluations can be divided into: understanding environments and work practices (UWP), evaluating visual data analysis and reasoning (VDAR), communication through visualization (CTV), evaluating collaborative data analysis (CDA), evaluating user performance (UP), evaluating user experience (UE), and evaluating visualization algorithms (VA).

Based on this taxonomy and on the already-described requirements of our work on VisSys certain methods are preferable to others. VDAR is simply impossible since it requires a system already used in practice. CTV does not fit our requirements since we do not aim to exclusively communicate via visualization but, instead, wish to facilitate the comparison between things visualized. CDA does not fit the scope of our tests since we must first validate our choice for single users before we consider the impact on the group. VA would suit us well, requirement-wise, but we do not know which objectively analyzable features of visualization algorithm output correspond to insight in our users. We also rejected the, otherwise excellent, laboratory driven approach of Engelke et al. [20] out of fear that a sensitive statistical test such as one we would have to do and which are extensively used in UP-coded visualization scenarios [37] would amplify unconscious biases in the experimental setup which we could not fight with a Bayesian [2, 41] approach due to lacking the data to establish suitable priors.

In the end, the method we chose was a hybrid of several approaches: First, UWP for the initial selection which takes after a tradition of papers like the work by Freitas et al. [24] or the seminal work by Keim and Kriegel [34] who codified the priors of their investigation by introducing a set of criteria based partially on own work and partially on existing literature in the vein of the work by Pillat [44]. Second we employ a rigorous literature review, followed by a multi-stage combination of UP and UE tests as a part of a controlled empirical study following in a tradition of empirical studies typified by the likes of Cawthon [9] and finished with a UP/UE interview with domain experts/PACS stakeholders.

3. The Proposed Approach

This section outlines the proposed approach to the study based on previous work in the field. It outlines how candidate visualization studies were chosen in the first two sections, first by a modified UWP approach (see 2.2) and then by a statistical analysis of published papers in the field. It then describes the test to be undertaken through the outline of the test, a discussion of methodology, and finally a discussion on statistical procedure.

3.1. Requirements-based Analysis

While we did not have sufficient access to PACS stakeholders to execute many field interviews—they are by the nature of their responsible positions people with limited time for other people's research—having designed SICEP we had data on what would be the

subject of VisSys visualization and based our decisions on this. The requirements that VisSys visualization has to fulfil are:

- **Dimensional scalability (DS).** The visualization technique should be such that it easily adapts to very large numbers of data dimensions.
- **Lens support (LS).** The visualization technique should be such that it allows for the easy selection and analysis of regions of interest.
- **Comparison in isolation (CI).** The visualization technique should be such that it can be fully used if comparing only two entities.

Although some of these requirements can be fulfilled in pre-processing stage, such as the ones in business intelligence (BI) data analytics, we previously established that once the unified data set is defined, no further reduction is possible, therefore, these requirements have to be supported using a visualization technique. Based on these requirements and the literature which covers glyphs[11, 23], star coordinates[12, 48], parallel coordinates[22, 27, 53], and Chernoff faces[42], it was relatively easy to formulate an evaluation of these visualization types based on these criteria.

Star plots definitely implement the LS and CI criteria, but implement the DS criterion only partially. Parallel plots implement all three criteria, while glyphs and Chernoff faces implement none of them. The only point of possible contention is saying that the star plot only implements DS partially. We feel this is correct, however, because it is impossible to find more than 360 degrees in a circle. Our preliminary research suggested possible issues in examples with extremely high-dimensionality.

The question that now arises is whether to use 2D or our modified version of 3D parallel coordinates. 2D coordinates have the advantage of simplicity and naturally fit 2D based displays which we would naturally have to use. On the other hand, 3D visualization has a lot to recommend it, especially in reducing cognitive distance in the case of multidimensional data [12, 31]. In the end, we decided to use both 2D and 3D parallel coordinates.

3.2. Statistical Literature Review

A sample of 591 papers drawn from image compression literature was reviewed by the authors and each visualization technique employed was noted including tables which, for the purposes of this survey, were counted as visualization methods. Listing the contents of the sample is far beyond the scope of this paper, but the raw data used is available on request. Of the 591 assessed papers, 26.40% were conference papers, 72.59% were journal papers, 0.51% were monographs, and 0.34% were technical reports. Of the papers in the sample, 4.23% are from the period of 1995-1999, 15.57% from 2000-2004, 44.67% from 2005-2010, and 35.36% from 2010 to 2016. A simple statistical analysis was then performed on the data, computing visualization method frequencies and their 95% confidence intervals (see table **??**).

As can be seen the result is weighed towards tables, scatterlines/scatterplots (two names for, fundamentally, the same plot: scattered points with optional trend or connecting lines), and bar graphs to an extreme degree.

Based on the above, we decided that the final test must include tables (which we already wanted as a control), a scatterplot of some sort, and bar graphs. We also added star

Type	Frequency [%]	Lower Bound [%]	Upper Bound [%]
2D/3D bar graph	0.1692	-0.1621	0.5006
3D bar graph	0.1692	-0.1621	0.5006
Color code	0.1692	-0.1621	0.5006
Pixel map	0.1692	-0.1621	0.5006
Pie chart	0.1692	-0.1621	0.5006
Hosaka plot	0.3384	-0.1298	0.8066
Star plot	0.3384	-0.1298	0.8066
ROC	1.184	0.3122	2.057
3D plane	1.184	0.3122	2.057
Interquartille range	1.861	0.7716	2.951
Scatterplot	8.122	5.919	10.32
Bar graph	16.41	13.43	19.4
Scatterline (2D line)	54.65	50.64	58.67
Table	84.43	81.51	87.36

Table 1. 95% Confidence intervals for visualization technique use frequencies, $n = 591$

plots to the consideration because they are the only plot we have found in the literature at all that can be said to be fulfilling our requirements outlined in subsection 3.1. Therefore, the final selection of visualization methods to test is: table, scatterplot, bar graph, parallel coordinates, and starplot. We specifically decided on two subvariants of the starplot: a 'dense' one and a 'sparse' one. The dense form has multiple measurements on one graph and corresponds to parallel coordinates in information density, while the sparse form has only one measurement per graph and corresponds to the bar graph in information density.

3.3. Test Outline

Four tests were planned: three to be done with large groups, and a final study of the use of the visualization, if any, that proved best in the tests by a panel of PACS-domain experts. The group tests all had the same form: experimental subjects were presented with a web tool for visualization evaluation we developed. They were offered instructions both in text form and in the form of a video presentation lasting 9 minutes 47 seconds. Then, they were presented with four visualizations of the same data set, though they were not told that they would be analyzing the same data set nor what the data represents. They were then asked to choose the 'best' entity and estimate how certain they were of their choice. The choice of 'best' is informed based on multidimensional comparison between entities where all data was industry data, and all axes were scaled to be uniform. This is a task which, in real use, would be covered by the SICEP system. The time they spent on each test page was measured.

The three tests were planned to be the initial test (stage 1), low-impact test (stage 2), and high-impact test (stage 3). The groups used for these tests were planned to be sixty people for each, but technical issues meant that final data analyzed had, respectively, 60, 43, and 59 people involved. The groups were strictly non-overlapping, anonymous, voluntary and comprised of an equal mix of adults ranging from ages of 20 up to 65 and of all walks of life. The members were pre-screened for relevant expertise by asking

before the test if they worked with visualizations in a professional capacity, and the groups are composed of 46.67% men and 53.33% women in the first stage, 62.79% men and 37.21% women in the second, and 47.46% men and 52.54% women in the third stage. No test-takers expressed any other gender identity. In stage 1, the requirement-based choice (parallel coordinates, 2D and 3D) is tested against the baseline (table) as well as the most popular visualization choice found in the literature (scatterplot). This is done on a data set with a relatively small number of dimensions. The idea behind this choice is to make it as fair towards what is used in the literature as possible, and make the requirements-based choice justify the need for it.

In stage 2, the 'winner' of stage 1 was to be pitted against a visualization we knew was not suited to the task as a control (bar graph which was attested in the literature), as well as the best possible competition found in the literature (star plots in both dense and sparse forms). Stage 2 keeps the 'low dimensionality' condition (specifically ten dimensions in three requirement indicators) in which, it is expected, nearly all visualizations will do well.

The techniques tested in stage 3 are the same as in stage 2, but the data is high-dimensional. Specifically, the used data have 25 dimensions in six requirement indicators, as presented in Fig. 3 (top) and (bottom). In this condition, it was expected that the most suitable visualization method would separate itself out from the competition and the control would do markedly worse.

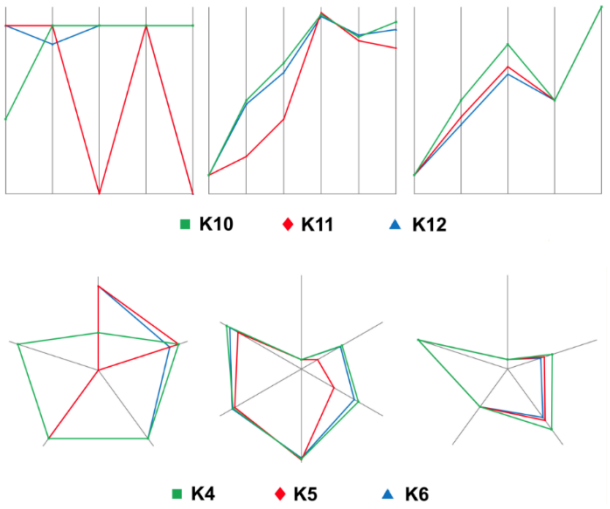


Fig. 3. Cropped screenshot of the visualization evaluation tool showing the high-dimensionality data set visualized using 2D parallel coordinates (top) and dense star plots (bottom)

Fig. 3's values labeled with a K and a number are deliberately anonymized to avoid any possibility of any sort of context outside of the visualization itself affecting the test subject's judgement. To provide a fair comparison between tests all visualizations are based on the same data. K3 and K10, for example, are exactly the same requirement indicator much in the same way K4/K11 and K5/K12 are. They are labeled differently

to stop previous tests from influencing the thought-process of the test subject. The K-values are the outputs of requirement indicators devised as a part of SICEP and their full definition exceeds the scope of this article, and their exact values are wholly irrelevant to the purpose of the study. The only relevant datum involved is that, universally, higher values mean that the choice presented by them is preferable and that the test subjects were informed of this by way of video training and verbal briefing. Details of how requirement indicators are derived are available in [17] and [15] with [18] providing background on the practicalities of how they were displayed.

Expert verification had to be limited to the final stage because our access to actual PACS stakeholders was limited. We resolved to show them only those visualizations which proved themselves the best and gather their experiences in a field-interview setting. This stage could only function to verify that we have not made some sort of mistake in previous stages and therefore produced an unusable result.

3.4. Methodological Caveats and Considerations

When doing empirical studies relying on statistical analysis, the risk of error is considerable. This is especially due to psychological reasons, e.g., that of false positive results either by subconscious interference from the researcher or by amplifying an effect that is orthogonal to that which was to be studied. To forestall these issues, we made sure to deal with boredom and practice effects, and eliminate researcher bias and user bias insofar as that is possible.

We dealt with boredom and practice effects by counterbalancing the design and presenting the experimental subjects with problems in random order with answers likewise randomized using the Fisher-Yates shuffle to ensure fair permutation. We did our best to eliminate researcher bias by double-blinding our research protocol. The data was prepared by one of the authors, but the tests were conducted by another author who did not know what the results were meant to be, nor had the opportunity to see what the user sees. The statistical analysis, too, was conducted on anonymized data meaning that the researcher who performed the analysis did not know what result he 'wanted.' A possible source of bias was the initial choice of data which was limited by only using actual data: all the data presented to the users was data extracted from the literature on PACS design and was, thus, not amenable to distortion. The data was assembled from sources such as [5, 35] that evaluated support for region of interest (ROI) coding, [13, 52] that evaluate error resilience, and [36, 40, 43, 46] that evaluated lossy and lossless compression techniques.

We tried to limit the effect of the peculiarity of individual test subjects by completely anonymizing the data: the subjects did not know what the data represented so domain knowledge, if any, did not interfere with the results. We also made sure that users knew that nobody would ever be able to connect them to individual answers meaning that they did not feel ashamed of not knowing an answer, which proved an issue in pre-study interviews.

3.5. Statistical Procedure

Tests were performed on time and on error rate. For the time measurement, dependent robust bootstrapped ANOVA with trimmed means [39, 51] was used, followed by a robust

post hoc test based on the Yuen modification of Student's T-test corrected for familywise error rate by the approach of Rom according to the work by Wilcox [39, 51]. The setup for the ANOVA is based on the visualization group dummy variable as the predictor (when viewing ANOVA as a special case of the General Linear Model) and the time as the outcome variable. The predictor, therefore, is categorical and the outcome is measured on the continuous interval measurement level.

In the case of error rate, the test is a bit more particular since it represents a robust analysis of dependent categorical data, which is an infrequently explored case. The method employed here is to use three parallel tests. The first is to simply compute the confidence intervals of the proportions and check for overlap (overlap being of course a sign that they are impossible to distinguish). The second is to perform a multilevel logistic regression with mixed effects and a randomly varying β_0 [3] and to estimate the relative quality of techniques by the confidence interval of their fitted parameter as compared to a baseline (which is either a table or a bar graph which were there for control purposes to begin with). The third statistical method is to perform multiple McNemar's χ^2 tests [1] and then correct for the familywise error rate by using Holm's correction.

To avoid the possibility of fishing for p-values, all three tests had to be positive for the result to count as positive. The fact that they all measure the same thing but by very different methods should serve to limit the chance of Type I errors.

4. Results and Data Analysis

This section contains the results of the study and their statistical analysis presented with minimal interpretation. It consists of three sections: an analysis of time taken, an analysis of accuracy, and the results of PACS domain expert interviews upon using the visualization to make decisions. As a part of this section to save space we will be using abbreviations for visualization technique names. Specifically, we will call the bar graph 'bar,' the scatterplot just 'scatter,' the parallel coordinates X2d and X3d for the 2D and 3D version, and as for the starplot we will differentiate between the Star3k and Star9kz version depending on whether it is the dense or sparse variant, as in Section 3.2.

4.1. Analysis of Time Taken

Fig. 4 (a) shows the time taken results for stage one as a mean plot with 95% CI error bars. Predictably, the figure shows that the table is slowest, and that parallel coordinates are the fastest. An unexpected result is that 3D parallel coordinates prove to be much slower than they were expected to be providing no improvement over the table which is meant to be a control.

If these results are subjected to statistical analysis, with the null hypothesis being that the mean of all the visualization times taken is equal, and the alternative being that they differ, an ANOVA test reports a test statistic of F(2.5, 87.66) = 22.3918 and a p-value of 0 i.e. too small to compute, allowing us to reject the null hypothesis. Table **??** shows the results of a post hoc test, testing a set of null hypotheses that all pairs of means are equal.

Fig. 4 (b) shows the results for time-taken in stage two as a mean plot with 95% CI error bars. Quite visibly, there is no real difference among the values.

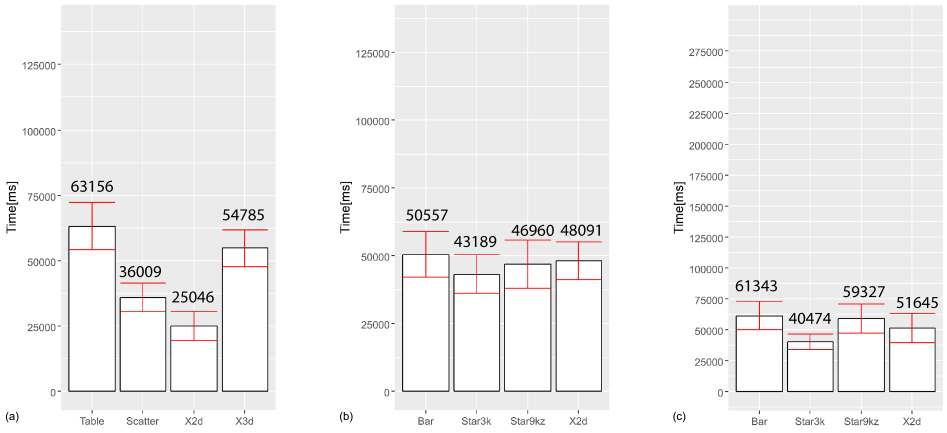


Fig. 4. Mean plot of the time taken to reach a result: (a) stage 1 ($N = 60$), (b) stage 2 ($N = 43$), (c) stage 3 ($N = 59$) all with value labels.

Comparison	P-value	Critical p-value	Significant
Table vs. Scatter	0.00003	0.01270	Yes
Table vs. X2d	0.00000	0.01020	Yes
Table vs. X3d	0.70336	0.05000	No
Scatter vs. X2d	0.00085	0.01690	Yes
Scatter vs. X3d	0.00158	0.02500	Yes
X2d vs. X3d	0.00000	0.00851	Yes

Table 2. Post hoc testing results, stage 1 ($N = 60$)

If these results are subjected to statistical analysis, with the null hypothesis being that the mean of all the visualization times taken is equal, and the alternative being that they differ, an ANOVA test predictably reports a test statistic of F(2.65, 69) = 0.5134 and a p-value of 0.65215. This means we cannot reject the null hypothesis of all values being the same, and there's no call for a post hoc test.

Fig. 4 (c) shows the time-taken results for stage three as a mean plot with 95% CI error bars. Mostly there is no real difference, except in the case of Star3k which seems quite close to being (barely) the fastest.

If these results are subjected to statistical analysis, with the null hypothesis being that the mean of all the visualization times taken is equal, and the alternative being that they differ, an ANOVA test reports a test statistic of F(2.93, 105.38) = 4.2831 and a p-value of 0.00722. This means that not all of the values are the same, which is to say that we may reject the null hypothesis of all means being equal. Table **??** shows the results of a post hoc test, testing a set of null hypotheses that all pairs of means are equal.

Comparison	P-value	Critical p-value	Significant
Bar vs. X2d	0.04678	0.01270	No
Bar vs. Star3k	0.00086	0.00851	Yes
Bar vs. Star9kz	0.94928	0.05000	No
X2d vs. Star3k	0.24238	0.01690	No
X2d vs. Star9kz	0.27513	0.02500	No
Star3k vs. Star9kz	0.00267	0.01020	Yes

Table 3. Post hoc testing results, stage 3 ($N = 59$)

The post hoc test results show that despite the appearance of the graph there is no statistically significant difference between 2D parallel coordinates and a dense star plot.

4.2. Analysis of Accuracy

Fig. 5 (a) shows the relative accuracies of visualization techniques in stage 1 data with blue representing the percentage of correct answers. It is clearly visible that 2D parallel coordinates provided the best result by a significant margin, while 3D parallel coordinates did not display the effectiveness we expected, being no better than the control technique. As predicted, the scatter visualization technique is between the table and parallel coordinates in accuracy.

Table **??** shows the actual value of the proportions, their confidence intervals, and the corresponding coefficients in the logistic model comparing them to the control technique (here table), and their confidence intervals. We will test these values using three separate statistical tests, all of whom take as their null hypothesis that the proportions of accuracy are the same, and as their alternative hypothesis that they differ. In case of the iterated McNemar's test the null hypothesis is expanded to a set where there are several H_0 being considered, each proposing that the proportions of accurate answers are equal between any two techniques studied. The alternative hypotheses are, therefore, that the proportions are not equal. The proportion confidence intervals do not overlap with any other technique

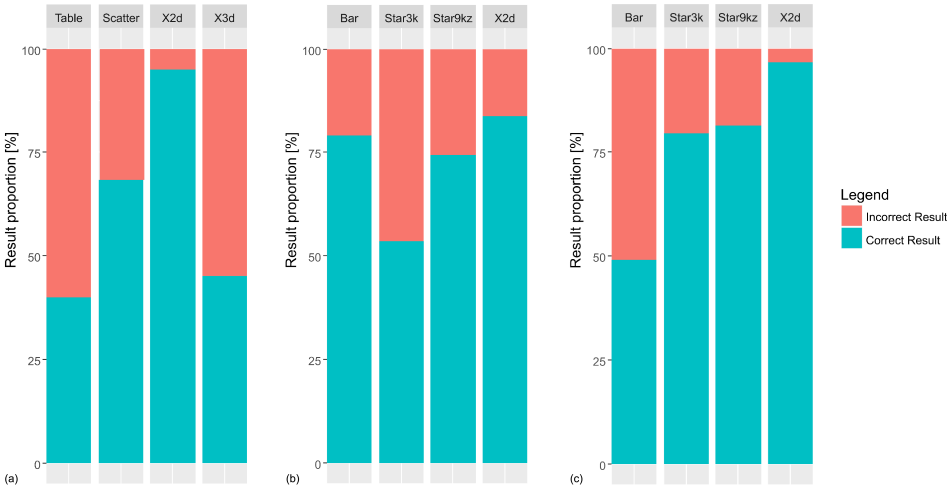


Fig. 5. Proportion of correct answers for techniques: (a) stage 1 ($N = 60$), (b) stage 2 ($N = 43$), (c) stage 3 ($N = 59$)

only for 2D parallel coordinates, so based on this test we can suggest that 2D parallel coordinates are significantly more accurate than any other tested here. As for logistic regression coefficients, compared to the table, the confidence interval does not include 1 (doing so indicates insignificance) for the scatterplot and 2D parallel coordinates, with the 2D parallel coordinates showing the largest effect size. This can be interpreted to mean that the odds that the answer will be valid increase by 28.5 times (compared to the table) if the technique used is 2D parallel coordinates.

Visualization	Proportion			Log. reg. coeff.		
	From	Value	To	From	Value	To
Table	27.604%	40.000%	52.396%	N/A	N/A	N/A
Scatter	56.563%	68.333%	80.104%	1.529	3.237	6.853
X2d	89.485%	95.000%	100%+	7.998	28.500	101.556
X3d	32.412%	45.000%	57.588%	0.594	1.227	2.534

Table 4. Proportion and coefficient confidence intervals, stage 1 ($N = 60$)

Table **??** shows the results of a Holm-corrected McNemar χ^2 test. The results that are marked significant are X2d compared to everything else, while scatter is only not significant compared to X3d. This corresponds nicely to the results in Table **??**.

Based on these results and according to the criterion outlined in 3.5, we can claim with statistical significance that two dimensional parallel coordinates are the most accurate technique.

Fig. 5 (b) shows the relative accuracies of visualization techniques in stage 2 data with blue representing the percentage of correct answers. We will test these values using three

Comparison	Adjusted p value	Significant
Table vs. Scatter	0.0062267	Yes
Table vs. X2d	0.0000002	Yes
Table vs. X3d	0.7277235	No
Scatter vs. X2d	0.0031849	Yes
Scatter vs. X3d	0.0605206	No
X2d vs. X3d	0.0000015	Yes

Table 5. Results of iterated McNemar test with Holm correction, stage 1 ($N = 60$)

separate statistical tests, all of whom take as their null hypothesis that the proportions of accuracy are the same, and as their alternative hypothesis that they differ. In case of the iterated McNemar's test the null hypothesis is expanded to a set where there are several H_0 being considered, each proposing that the proportions of accurate answers are equal between any two techniques studied. The alternative hypotheses are, therefore, that the proportions are not equal. The surprising result here is the success of the bar graph. The interpretation that seems obvious is that the bar graph is the most familiar visualization here and this seems dominant in this low-impact test case.

Table **??** shows the actual value of the proportions, their confidence intervals, and the corresponding coefficients in the logistic model comparing them to the control technique (here bar graph), and their confidence intervals. As can be seen, there are only the slightest indications of significance, chiefly with Star3k being noticeably worse than the others.

Visualization	Proportion			Log. reg. coeff.		
	From	Value	To	From	Value	To
Bar	66.911%	79.070%	91.229%	N/A	N/A	N/A
Star3k	38.580%	53.488%	68.397%	0.102	0.278	0.759
Star9kz	61.377%	74.419%	87.460%	0.268	0.757	2.136
X2d	72.687%	83.721%	94.755%	0.450	1.385	4.263

Table 6. Proportion and coefficient confidence intervals, stage 2 ($N = 43$)

Table **??** shows the results of a Holm-corrected McNemar χ^2 test which is nowhere significant.

The result of the above according to the criterion outlined in section 3.5 is that we cannot claim that any technique is significantly more accurate than any other. The results of the low-impact test show, as was expected, that at this level of dimensionality familiarity outstrips nearly all other factors. It can be noted that, while we cannot claim a statistically significant difference, 2D parallel coordinates did do the best in absolute terms and, crucially, did no worse than any other tested visualization.

Fig. 5 (c) shows the relative accuracies of visualization techniques in stage 3 data with blue representing the percentage of correct answers. We will test these values using three separate statistical tests, all of whom take as their null hypothesis that the proportions of

Comparison	Adjusted p value	Significant
Bar vs. X2d	1.000	No
Bar vs. Star3k	0.185	No
Bar vs. Star9kz	1.000	No
X2d vs. Star3k	0.053	No
X2d vs. Star9kz	1.000	No
Star3k vs. Star9kz	0.381	No

Table 7. Results of iterated McNemar test with Holm correction, stage 2 ($N = 43$)

accuracy are the same, and as their alternative hypothesis that they differ. In case of the iterated McNemar's test the null hypothesis is expanded to a set where there are several H_0 being considered, each proposing that the proportions of accurate answers are equal between any two techniques studied. The alternative hypotheses are, therefore, that the proportions are not equal. Obviously, the star plots are working in a broadly similar fashion, and 2D parallel coordinates are clearly the best or nearly so, almost replicating their result from stage 1.

Table **??** shows the actual value of the proportions, their confidence intervals, and the corresponding coefficients in the logistic model comparing them to the control technique (here bar graph), and their confidence intervals. The only consistently non-overlapping interval is 2D parallel coordinates, and they also increase the odds of a correct answer the most.

Visualization	Proportion			Log. reg. coeff.		
	From	Value	To	From	Value	To
Bar	36.396%	49.153%	61.909%	N/A	N/A	N/A
Star3k	69.390%	79.661%	89.932%	2.571	7.621	22.588
Star9kz	71.418%	81.356%	91.294%	2.898	8.861	27.089
X2d	91.992%	96.610%	100%+	13.958	97.723	684.165

Table 8. Proportion and coefficient confidence intervals, stage 3 ($N = 59$)

Table **??** shows the results of a Holm-corrected McNemar χ^2 test. Nearly all of the differences are significant, the only exception being the difference between the star plots which is to be expected. This fits perfectly with the results in Table **??**, and fig. 5 (c).

The result of the above according to the criterion outlined in 3.5 is that we can only claim that 2D parallel coordinates are consistently more accurate than all other techniques. Star plots are roughly the same and better than the bar graph in a statistically significant manner.

4.3. Expert Use-case Verification

All the results thus far have been achieved using nonexpert users using anonymized data which worked well to control a statistically-analyzed empirical study but also removed

Comparison	Adjusted p value	Significant
Bar vs. X2d	0.0000020	Yes
Bar vs. Star3k	0.0034246	Yes
Bar vs. Star9kz	0.0026600	Yes
X2d vs. Star3k	0.0132796	Yes
X2d vs. Star9kz	0.0317227	Yes
Star3k vs. Star9kz	1.0000000	No

Table 9. Results of iterated McNemar test with Holm correction, stage 3 ($N = 59$)

from consideration crucial elements of how this sort of system would be used in practice. To rectify this, we created three real-world scenarios based on real data and modeled them in SICEP. Then we visualized them using 2D parallel coordinates (because they were the consistent winner in all our tests as shown by statistical analysis including but not limited to ANOVA) and presented them to a panel of three experts. These experts were actual stakeholders and users of PACS but, crucially, while they were domain experts, they had absolutely no experience in image compression in general or in the context of PACS in particular.

The panel consisted of a domain expert in healthcare, a domain expert in medical information systems, and a domain expert in finance. Once the panel used the visualizations and the system to make their decision we interviewed them on their impressions and compared their decisions to the industry consensus.

All the scenarios are based on choosing between some subset of JPEG2000, SPIHT, lossy and lossless JPEG, and JPEG-LS, and differ on the requirements and the context. The scenarios observed are:

- A regional medical center PACS that supports both telemedicine and mobile medicine.
- A local medical institution PACS with limited capacities. It is a system that does not support lossless compression, telemedicine, or mobile medicine.
- A local medical institution PACS with extensive resources. This is a system that supports lossless image compression in order to decrease image turnaround time and for more efficient image transmission [21]. Telemedicine and mobile medicine are not supported.

These scenarios were modeled in SICEP by forming seven requirement indicators (visual acceptance of lossy image compression, lossy compression efficiency, lossy decompression efficiency, lossless compression efficiency, lossless decompression efficiency, acceptability of region of interest coding, and error resilience of image compression) which variously combined a total of fourteen dimensions. We provided the ability to visualize this set either on the indicator level (which shows indicators) or the detail level (which shows details of a single indicator either against an arbitrary condition of acceptability or against other compressions being tested). This level also displays the explicit measured values.

Fig. 6 illustrates the visualization shown to experts and displays all seven indicators because, in the case of a regional PACS, all are relevant. Each sub-graph visualized with 2D parallel coordinates which, the empirical and statistical tests suggested, were optimal

for this task, is labeled with the name of its indicator and the number of vertical axes represent the specific measurements which are a part of this indicator. To give an example, in the case of the visual acceptance of lossy image compression visible in the top left, it consists of four measurements (corresponding to the four axes), indicating the compression ratio, peak signal-to-noise ratio, structured similarity index, and receiver operating characteristic. These are combined because they are all relevant to the decision to be made regarding this particular indicator. Intelligent grouping made using SICEP is what allows us to manage the number of dimensions used to display the data. This is done by grouping those axes, whose interactions interest us most, into convenient indicators representing specific questions in the decision-making process being supported by this visualization.

Fig. 6 (top row) displays a subset of the data because the use case was modeled differently in SICEP prioritizing certain factors and ignoring others. Since capabilities are limited in the PACS studied here, lossy compression is the subject of focus.

Similarly to the earlier case, Fig. 6 (middle row) only shows the subset of the data of interest according to the SICEP decision-supporting model. In the case of the local PACS with extensive capabilities, the trade-off favors increased quality over speed and space, and so only lossless factors are relevant.

The way the panel of experts used these visualizations was to be told, briefly, what scenario they were engaged in and what their priorities are. The experts were then allowed to interact with a visualization solution displaying the visualizations illustrated Fig. 6. Also available was a zoomed-in detail level, which focused on only one indicator rather than the overview of all indicators, as well as, if necessary, access to a table display of all values. They could also choose to compare the data to an acceptability threshold (Fig. 7), or simply view the visualization with an overlay containing the data values (Fig. 8).

During the test, the experts never asked for the table display, which the empirical tests indicated would happen, and mostly focused on the indicator display, sorting their options and identifying candidates to reject out of hand or to consider further. Only once was a detail unclear and one of the panel experts, the domain expert in medical information systems, asked for a zoomed-in detail level in the case of lossless compression efficiency to confirm a suspicion. This corresponds to the criteria derived from the survey of literature and is precisely why 2D parallel coordinates were included in the empirical tests.

In all three cases, the panel reached the 'correct' (industry consensus) choice, serving to strengthen the conclusion reached in the empirical and statistical testing phase. In the first scenario, the choice of JPEG2000 was instant which made sense given how overconstrained the problem was. In the other two, the selection took a while, but the correct response was always found and, afterwards, was held with considerable confidence.

In an interview after the decision was made, user experience was solicited from the panel and the general impressions are that the system is easy to use. This is largely because it manages the amount of information visible at any one time (which is a property of SICEP as well as the visualization being tested), and because the information is easily distinguished without being all over the place, as one of the panel remarked. The fact that all the information they needed to make their decision was within eye's reach to, again, quote a panel member, was the one factor the expert panel found to be of greatest use. These additional observations serve to add a dimension that UP/UE testing scenarios we used for the empirical and statistical tests could not capture by their very nature.

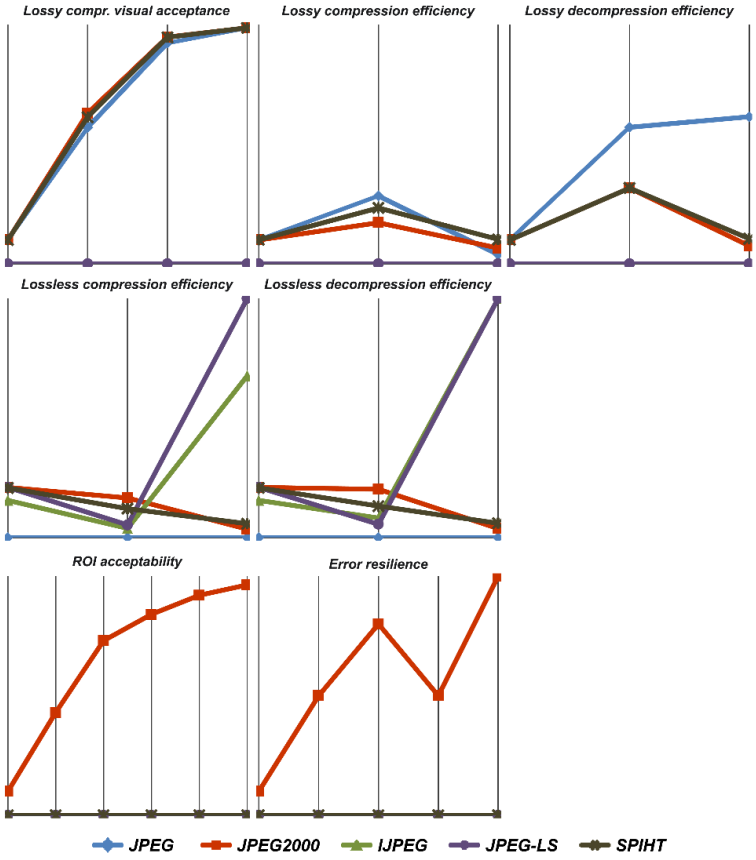


Fig. 6. Indicator level of the regional PACS visualization presented to expert users visualized using 2D parallel coordinates according to the tests performed, demonstrating the end result of the visualization evaluation process

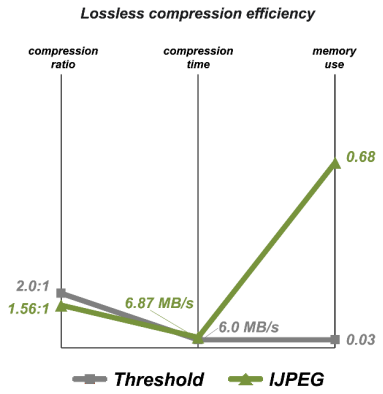


Fig. 7. Comparison of an individual compression technique with a specified threshold of acceptability visualized using 2D parallel coordinates according to the tests performed, demonstrating the end result of the visualization evaluation process

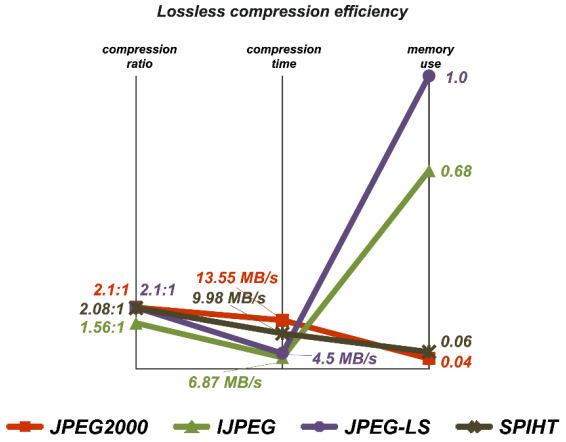


Fig. 8. Comparison of compression techniques with an actual-value overlay visualized using 2D parallel coordinates according to the tests performed, demonstrating the end result of the visualization evaluation process

5. Conclusion

Based on our research we can claim with considerable confidence that:

- The use of visualization as a primary component in this sort of decision support system is justified.
- Out of the considered visualization techniques, the most accurate and the one that requires the least time to produce a result in the considered context of medical image compression is 2D parallel coordinates, followed by a dense star plot. No other visualization technique compares in the examined use-case.
- The choice of visualization techniques in the reviewed literature is nowhere close to optimal.
- Out of the considered visualization techniques, the best choice for the design of the VisSys module is 2D parallel coordinates.

A question that immediately comes to mind is how applicable is this research outside the relatively niche, if not unimportant, field of PACS design. It is our position that essentially all the results presented herein are entirely applicable to any field which faces a problem of using multidimensional data sets to make choices between complexly-described alternatives. Complexly-described alternatives, in this context, mean any entity that

- is described by a large number of attributes, where large is defined by significantly exceeding the capacity of short-term memory [32],
- possesses some sort of structure including those attributes,
- has this structure, in aggregate, measure some sort of desirability of the entity presenting tradeoffs and varying requirements.

So described *desiderata* of complexly-described alternatives must be compared and a decision must be reached selecting one of the proposed alternatives based on the relative values of attributes and arbitrary minimum requirements over those attributes.

Described abstractly it may seem like an unlikely contingency, but, in fact, any purchasing decision one might agonize over is an example of using multidimensional data sets to make choices between complexly-described alternatives: the desirability of a house (corresponding to the quality of a compression technique) depends on a number of attributes (area, price, availability of schools, facilities, and many others a moment's reflection ought to furnish) which are both used to compare and to disqualify (such as a maximum price or minimum area). This problem is complex enough that it is studied through successive hierarchical decomposition [45]. The same could be said for the case of selecting one of several proposals for public works. This is by no means a problem limited to PACS design and we were cognizant of this fact when preparing the tests for the study.

We have also learned that 3D visualization performs considerably worse than we have expected. Based on this, as well as on poor results for bar graphs and tables, the data studied suggests but does not guarantee that compactness of the data is a key feature that allows for insight. The limiting factor in this sort of comparison visualization appears to be short-term memory. This, in turn, suggests the first potential avenue of further research: testing which features of these visualization techniques are salient, by using more user telemetry, especially eye-tracking in order to determine the locus of user attention.

Another area of research that presents itself is considering extremely large numbers of dimensions. Since we have determined star plots and 2D parallel coordinates as the best candidates, we should test how they behave in extreme-impact tests where the dimensionality exceeds 100. In the same vein, a potentially fruitful area of research would be to re-run these tests or, perhaps, only stage 3 tests, using multiple sources of data in order to determine if the same results hold for different data sets or if some feature of the data, even if anonymized, influences the choice of suitable visualization technique.

Lastly, as the body of data gathered in these tests grows and this implementation of visualization testing is refined and validated, it can find a new use by being used 'backward' as it were: this methodology of gauging visualization quality can, if the quality of the visualization is known, be used to evaluate perception. Thus, it would provide a way to quantify the impact of visual perception disorders and disabilities by running tests, much like the ones presented in this paper, using known-quantity visualization methods alongside simulated disabilities and disorders of perception.

Acknowledgments. The research reported in this paper is partially supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia, projects number TR32044 "The Development of Software Tools for Business Process Analysis and Improvement" (2011-2018) and III44006 "Development of New Information and Communication Technologies using Advanced Mathematical Methods with Applications in Medicine, Energy, e-Government, and Protection of National Heritage" (2011-2018).

References

1. Agresti, A., Kateri, M.: Categorical data analysis. Springer (2011)
2. Antonelli, J., Trippa, L., Haneuse, S., others: Mitigating Bias in Generalized Linear Mixed Models: The Case for Bayesian Nonparametrics. Statistical Science 31(1), 80–95 (2016)
3. Bates, D., Mchler, M., Bolker, B., Walker, S.: Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67(1), 1–48 (2015)
4. Bellon, E., Feron, M., Deprez, T., Reynders, R., Van den Bosch, B.: Trends in PACS architecture. European Journal of Radiology 78(2), 199–204 (2011)
5. Bharti, P., Gupta, S., Bhatia, R.: Comparative analysis of image compression techniques: a case study on medical images. In: Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on. pp. 820–822. IEEE (2009)
6. Bruckner, L.A.: On Chernoff faces. Graphical Representation of Multivariate Data pp. 93–121 (1978)
7. Carpendale, M.: Considering visual variables as a basis for information visualisation. Computer Science TR# 2001-693 16 (2003)
8. Carpendale, S.: Evaluating information visualizations. In: Information Visualization, pp. 19–45. Springer (2008)
9. Cawthon, N., Moere, A.V.: The effect of aesthetic on the usability of data visualization. In: Information Visualization, 2007. IV'07. 11th International Conference. pp. 637–648. IEEE (2007)
10. Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: A survey on big data 275, 314–347 (2014)
11. Choi, S.M., Lee, D.S., Yoo, S.J., Kim, M.H.: Interactive visualization of diagnostic data from cardiac images using 3D glyphs. In: International Symposium on Medical Data Analysis. pp. 83–90. Springer (2003)

12. Cooprider, N.D., Burton, R.P.: Extension of star coordinates into three dimensions. In: Electronic Imaging 2007. pp. 64950Q–64950Q. International Society for Optics and Photonics (2007)
13. Dhouib, D., Nat-Ali, A., Olivier, C., Naceur, M.: Performance evaluation of wavelet based coders on brain MRI volumetric medical datasets for storage and wireless transmission. International Journal of Biological, Biomedical and Medical Sciences 3(3), 147–156 (2008)
14. Dragan, D., Ivetić, D.: Quality evaluation of medical image compression: What to measure? In: IEEE 8th International Symposium on Intelligent Systems and Informatics (2010)
15. Dragan, D., Ivetić, D.: A comprehensive quality evaluation system for PACS. In: Ubiquitous Computing and Communication Journal, Special Issue on ICIT 2009 Conference-Bioinformatics and Image. vol. 4, pp. 642–650 (2009)
16. Dragan, D., Ivetić, D.: Request redirection paradigm in medical image archive implementation. Computer Methods and Programs in Biomedicine 107(2), 111–121 (2012)
17. Dragan, D., Ivetić, D., Petrović, V.B.: Introducing an acceptability metric for image compression in PACS-A model. In: E-Health and Bioengineering Conference (EHB), 2013. pp. 1–4. IEEE (2013)
18. Dragan, D., Petrović, V.B., Ivetić, D.: Software Tool for 2D and 3D Visualization of Requirement Indicators in Compression Evaluation for PACS. Proceedings of the 4th Inetrnational Scientific Conference on Geometry and Graphics, MoNGeometrija Vol. 1 p. 315 (2014)
19. Dreyer, K.J., Hirschorn, D.S., Thrall, J.H., Mehta, A.: PACS: A Guide to the Digital Revolution. Springer Science and Business Media (2006)
20. Engelke, U., Vuong, J., Heinrich, J.: Visual Performance in Multidimensional Data Characterisation with Scatterplots and Parallel Coordinates. Electronic Imaging 2016(16), 1–6 (2016)
21. Eskicioglu, A.M.: Quality measurement for monochrome compressed images in the past 25 years. In: Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. vol. 6, pp. 1907–1910. IEEE (2000)
22. Forsell, C., Johansson, J.: Task-based evaluation of multirelational 3D and standard 2D parallel coordinates. In: Electronic Imaging 2007. pp. 64950C–64950C. International Society for Optics and Photonics (2007)
23. Forsell, C., Seipel, S., Lind, M.: Simple 3D glyphs for spatial multivariate data. In: IEEE Symposium on Information Visualization, 2005. INFOVIS 2005. pp. 119–124. IEEE (2005)
24. Freitas, C.M., Luzzardi, P.R., Cava, R.A., Winckler, M., Pimenta, M.S., Nedel, L.P.: On evaluating information visualization techniques. In: Proceedings of the Working Conference on Advanced Visual Interfaces. pp. 373–374. ACM (2002)
25. Garlandini, S., Fabrikant, S.I.: Evaluating the effectiveness and efficiency of visual variables for geographic information visualization. In: Spatial information theory, pp. 195–211. Springer (2009)
26. Gupta, M., Henry, J.K., Schwab, F., Klineberg, E., Smith, J., Gum, J., Polly, D.W., Liabaud, B., Diebo, B., Hamilton, D.K., others: Dedicated Spine Measurement Software Quantifies Key Spino-Pelvic Parameters More Reliably Than Traditional PACS. Global Spine Journal 6(S 01), GO082 (2016)
27. Inselberg, A.: Parallel coordinates: visualization, exploration and classification of high-dimensional data. In: Handbook of Data Visualization, pp. 643–680. Springer (2008)
28. Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., Mller, T.: A systematic review on the practice of evaluating visualization. IEEE Transactions on Visualization and Computer Graphics 19(12), 2818–2827 (2013)
29. Ivetić, D., Dragan, D.: Medical image on the go! Journal of Medical Systems 35(4), 499–516 (2011)
30. Johansson, J., Forsell, C.: Evaluation of parallel coordinates: Overview, categorization and guidelines for future research. IEEE Transactions on Visualization and Computer Graphics 22(1), 579–588 (2016)

31. John, N.W., McCloy, R.F.: Navigating and visualizing three-dimensional data sets. The British Journal of Radiology 77(suppl_2), S108–S113 (Dec 2004), http://www.birpublications.org/doi/abs/10.1259/bjr/45222871
32. Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., Moore, K.S.: The mind and brain of short-term memory. Annual Review of Psychology 59, 193 (2008)
33. Kbben, B., Yaman, M.: Evaluating dynamic visual variables. In: Proceedings of the Seminar on Teaching Animated Cartography, Madrid, Spain. pp. 45–51 (1995)
34. Keim, D.A., Kriegel, H.P.: Visualization techniques for mining large databases: A comparison. IEEE Transactions on Knowledge & Data Engineering (6), 923–938 (1996)
35. Kosheleva, O.M., Cabrera, S.D.: Application of task-specific metrics in JPEG2000 ROI compression. In: Image Analysis and Interpretation, 2002. Proceedings. Fifth IEEE Southwest Symposium on. pp. 163–167. IEEE (2002)
36. Kumar, B., Singh, S.P., Mohan, A., Singh, H.V.: MOS prediction of SPIHT medical images using objective quality parameters. In: 2009 International Conference on Signal Processing Systems. pp. 219–223. IEEE (2009)
37. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical studies in information visualization: Seven scenarios. IEEE Transactions on Visualization and Computer Graphics 18(9), 1520–1536 (2012)
38. Leban, G., Bratko, I., Petrovic, U., Curk, T., Zupan, B.: Vizrank: finding informative data projections in functional genomics by machine learning. Bioinformatics 21(3), 413–414 (2004)
39. Mair, P., Schoenbrodt, F., Wilcox, R.: WRS2: Wilcox robust estimation and testing (2015), 0.4-0
40. Man, H., Docef, A., Kossentini, F.: Performance analysis of the JPEG 2000 image coding standard. Multimedia Tools and Applications 26(1), 27–57 (2005)
41. McNeish, D.M.: Using Data-Dependent Priors to Mitigate Small Sample Bias in Latent Growth Models: A Discussion and Illustration Using M plus. Journal of Educational and Behavioral Statistics 41(1), 27–56 (2016)
42. Morris, C.J., Ebert, D.S., Rheingans, P.L.: Experimental analysis of the effectiveness of features in Chernoff faces. In: 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making. pp. 12–17. International Society for Optics and Photonics (2000)
43. Penedo, M., Souto, M., Tahoces, P., Carreira, J., Villalon, J., Porto, G., Seoane, C., Vidal, J., Berbaum, K., Chakraborty, D., others: FROC evaluation of JPEG2000 and object-based SPIHT lossy compression on digitized mammograms. Radiology 237(2), 450–457 (2005)
44. Pillat, R.M., Valiati, E.R., Freitas, C.M.: Experimental study on evaluation of multidimensional information visualization techniques. In: Proceedings of the 2005 Latin American conference on Human-computer interaction. pp. 20–30. ACM (2005)
45. Saaty, T.L.: Decision-making with the AHP: Why is the principal eigenvector necessary. European Journal of Operational Research 145(1), 85–91 (2003)
46. Santa-Cruz, D., Grosbois, R., Ebrahimi, T.: JPEG 2000 performance evaluation and assessment. Signal Processing: Image Communication 17(1), 113–130 (2002)
47. Speier, C.: The influence of information presentation formats on complex task decision-making performance. International Journal of Human-Computer Studies 64(11), 1115–1131 (2006)
48. Tufte, E.R., Graves-Morris, P.: The visual display of quantitative information, vol. 2. Graphics Press Cheshire, CT (1983)
49. Union, I.T.: World Telecommunication/ICT Indicators Database 17th edition. Tech. rep. (2013)
50. Wang, J., Liu, X., Shen, H.W., Lin, G.: Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots. IEEE Transactions on Visualization and Computer Graphics 23(1), 81–90 (2017)
51. Wilcox, R.R.: Introduction to robust estimation and hypothesis testing. Academic Press (2012)
52. Xiang, W., Clemence, A., Leis, J., Wang, Y.: Error resilience analysis of wireless image transmission using JPEG, JPEG 2000 and JPWL. In: Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on. pp. 1–6. IEEE (2009)

53. Zhao, X., Kaufman, A.: Structure revealing techniques based on parallel coordinates plot. The Visual Computer 28(6-8), 541–551 (2012)
54. Zheng, J.G.: Data visualization in business intelligence. In: Global Business Intelligence, pp. 67–81. Routledge (2017)

Dinu Dragan was born in 1979 in Zrenjanin. He is an Assistant Professor at the Department of Computing and Control Engineering of the Faculty of Technical Sciences, University of Novi Sad, Serbia. He received his Dipl. Ing. degree in Electrical Engineering and Computing in 2003, and the MSc and PhD in Computer Science from the Faculty Technical Sciences, University of Novi Sad, Serbia, in 2008 and 2013, respectively. His research interests include data compression, computer graphics, human computer interaction, data visualization, and computer vision.

Veljko B. Petrović was born in 1986 in Čačak. He completed his studies at the Faculty of Technical Sciences in 2010 and holds BSc and MSc degrees in Computer and Electrical Engineering. He is currently in the final stages of defending his doctoral thesis at the same institution under the title ”Domain specific language for visualization evaluated by the statistical analysis of small data sets.” His research fields include visualization, humancomputer interaction, and large-scale high-performance computing.

Dušan B. Gajić is an Assistant Professor at the Department of Computing and Control Engineering of the Faculty of Technical Sciences, University of Novi Sad, Serbia. He received his MSc and PhD in Electrical Engineering and Computing from the Faculty of Electronic Engineering, University of Niš, Serbia, in 2009 and 2014, respectively. His research interests include parallel and distributed computing, blockchain, GPGPU, algorithm design, signal processing, and computer vision.

Žarko Živanov was born in 1974 in Zrenjanin. He is an Associate Professor at the Department of Computing and Control Engineering of the Faculty of Technical Sciences, University of Novi Sad, Serbia. He received his Dipl. Ing. degree in Electrical Engineering and Computing in 2000, and the MSc and PhD in Computer Science from the Faculty Technical Sciences, University of Novi Sad, Serbia, in 2007 and 2012, respectively. His research interests include computer architecture, operating systems, parallel programming and data visualization.

Dragan Ivetić was born in Subotica and received the Dipl. Ing. degree in Electrical Engineering in 1990, and the M.Sc. and Sc.D. degrees in computer science, in 1994 and 1999 respectively, from the Faculty of Technical Science, University of Novi Sad. Since 2000 he has been Professor of Computer Science at the Department for Computing and Automatics, Faculty of Technical Sciences, University of Novi Sad. His research interests include HCI, multimedia, and computer graphics. Profesor Ivetić received DAAD scholarship in 1997 and ACM scholarship in 1998.

Received: April 30, 2018; Accepted: August 22, 2018.

Computer Science and Information Systems 16(1):273–287 https://doi.org/10.2298/CSIS171001036J

An Experience in Automatically Building Lexicons for Affective Computing in Multiple Target Languages

Francisco Jurado and Pilar Rodriguez

Department of Computer Engineering
Universidad Autónoma de Madrid
Francisco Tomas & Valiente 11, Madrid
{Francisco.Jurado; Pilar.Rodriguez]@uam.es

Abstract. Affective Computing in text attempts to identify the emotional charge reflected in it, trying to analyse the moods transmitted while writing. There are several techniques and approaches to perform Affective Computing in texts, but lexicons are their common point. However, it is difficult to find solutions for specific languages different from English. Thus, this article presents an experience in automatically generating lexicons to perform Affective Computing following a multiple-target languages approach. The experience starts with some initial seeds of words in English that define the emotions we want to identify. It then expands them as much as possible with related words in a bootstrapping process and finally obtains a lexicon by processing the context sentences from parallel translated text where the terms have been used. We have checked the resulting lexicons by conducting an exploratory analysis of the affective fingerprint on a parallel corpus with books translated from and to different languages. The obtained results look promising, showing really similar affective fingerprints in different language translations for the same books.

Keywords. Sentiment Analysis, Affective Computing, Multiple Target Languages, Automatically Building Lexicons

1 Introduction

Sentiment Analysis and *Opinion Mining* can be defined as the *"computational treatment of opinion, sentiment, and subjectivity in text"* [1, 2]. In recent years, it has been applied to many different contexts, such as reviewing customers' products and services, monitoring reputations in social networks, tracking people's feelings about politicians, promoting marketing campaigns, etc. [3]. So it has become a growing research discipline [4]. Moreover, because currently social media texts are widely available and need to be analysed to create knowledge for decision-making processes, new approaches rise to perform sentiment analysis on code repositories like GitHub [5] and social networks like Twitter [6] or YouTube [7]. However, it looks as though *"affective computing and sentiment analysis are still finding their own voice as new interdisciplinary fields"* [8].

Given the variety of contexts and applications where these techniques can be applied, we can note some distinctions and nuances in terminology. Thus, Subjectivity Analysis is used to classify a given text into subjective or objective. Once a text has been classified

as subjective, a Sentiment Analysis (also called Polarity Analysis) can be performed by assigning a score (positive or negative) to the opinion underlying the text.

Regardless of the subjectivity and objectivity of the text, Affective Computing attempts to identify the emotional charge (happiness, sadness, fear, anger-passion, etc.) that is reflected and hence transmitted in the text, depending on the words used.

Accordingly, it is straightforward to see the difference between examining the opinion about something (a product or service) or somebody (an actor or politician) with a Polarity Analysis in the subjective text and analysing the moods reflected and transmitted in a text, either subjective or objective. We will focus our attention on this last kind of analysis, namely, Affective Computing.

Independently of the kind of analysis to be performed, there are several techniques and approaches to computing, including and Affective Computing in the text [1, 2, 9–11], among which we can mention Naive Bayes, Maximum Entropy, Support Vector Machines, Lexicon-based, etc. Nevertheless, the common point is a central piece known as lexicon [3].

Fortunately, there are several lexicons available. As examples, we can mention works like SentiWordNet [12, 13], which gives positivity, negativity and objectivity scores to each synset (synonym set) of WordNet [14]; WordNet-Affect [15], which assigns one or more affective labels (a-labels) for those synsets from WordNet that represent emotions (with emotional categories), emotional valence (positive, negative, ambiguous and neutral), moods, cognitive states, behaviour, etc.; the Affective Norms for English Words (ANEW) [16], which provides valence, arousal and dominance scores for the 1,034 most used words in English; and the NRC Word-Emotion Association Lexicon also known as EmoLex [17], which provides polarity and emotional annotations (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) for 14,182 English words and their corresponding automatic translations to more than 40 languages.

However, the main drawback is that most of them have English as the goal language and it is difficult to find solutions for other languages [18].

While building a lexicon we must make some decisions. In particular, high precision lexicons can be obtained manually but with a low coverage, and in the opposite situation we find those automatically derived from pre-existing knowledge [19]. That is, the more human supervision devoted to classification, the more precision, but the more automatically derived, the more coverage. In particular, developing a supervised approach will be slower but more precise than an unsupervised one. However, because human supervision is not always available and we are focused on working with multiple target languages, we will centre our attention on this last kind of approach.

From this perspective, the main achievements of this paper can be summarised as follows:

- Facing the issue of automatically generating lexicons to perform Sentiment Analysis on texts following a target language independent approach and using automatic translation systems.
- Detailing an automatic approach that makes use of initial seeds of words in English that define the emotions we want to identify (anger, disgust, fear, joy, sadness and surprise). Then applying bootstrapping to expand them as much as possible with related words, and finally obtaining a target independent language lexicon to detect emotions in a text by processing the context sentences where the terms have been used in a parallel text.

- Checking the validity of the approach by conducting an exploratory analysis on a parallel corpus with books translated into different target languages to analyse the affective fingerprint identified in the texts and comparing the results between two different target languages.

Thus, the rest of the article is structured as follows: section 2 shows some related works; section 3 outlines the approach we followed to build the lexicons; section 4 shows some implementation issues; section 5 explains some inspections we carried out to the lexicons; section 6 exposes the conducted exploratory analysis we performed on a parallel corpus; section 7 provides a discussion and some further works; and finally, section 8 ends with the conclusions.

2 Related Works

Although most of the research in Sentiment Analysis has been done for the English language, the literature offers a few examples of multilingual automatic translation approaches to build resources and tools for Sentiment Analysis.

Thus, [18, 20, 21] use bilingual dictionaries and parallel corpora to rapidly create tools for Subjectivity Analysis in the new language. The authors employ resources available for English and apply machine translation to generate resources for subjectivity analysis in other languages different from English. Comparing evaluations of Romanian and Spanish, they conclude that automatic translation is a viable alternative for the construction of resources and tools for Subjectivity Analysis in a new target language.

In the same way, [22, 23] face the issue of sentiment detection in different languages using distinct machine translation systems (Bing Translator, Google Translate and the Moses statistical machine translation system), concluding that those systems can be used to obtain training data and to build Sentiment Analysis tools for languages other than English with comparable performance to the one obtained for English.

All these previous approaches make use of existing English resources and perform the corresponding translation to a target language; however, none of them builds the original resources on their own.

For their part, [24] try to produce a set of lexicons for 136 major languages via graph propagation. They use the Lius' lexicon [2] as the initial seed of words in English and make use of several resources like Wiktionary [25], Google Translate and WordNet [14]. As a result, they are able to build polarity-annotated lexicons for the major languages.

As far as we know, although all these researchers have proved the validity of using automatic machine translations and propagation to build resources for Sentiment Analysis, the above approaches have not been developed with an aim to build an emotional lexicon to assemble tools that allow analysis of the moods reflected in a text, but also the polarity.

The only exception we have found is the NRC Word-Emotion Association Lexicon (also named *EmoLex* by their authors) developed by [17]. These authors used Mechanical Turk (https://www.mturk.com) to crowdsource a lexicon that provides polarity (positive and negative) and emotional annotations (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) for 14,182 English words. Starting from these English words, they provide the corresponding automatic translations for more than 40 languages. In a very

similar way, [26] have used crowdsourcing for harvesting emotional annotations of news articles rather than the words. After processing these annotated articles, they were able to build an English lexicon for Sentiment Analysis. However, they do not provide translation into any other languages.

In these last-mentioned works, although the translation to the non-English languages can be automatically performed, the translation is not validated, and the time and effort consumed by the crowdsourcing process cannot always be tackled depending on the application domain.

Although not designed for multilingual purposes, [6] presents SentiCircles, an interesting lexicon-based approach that takes into account the co-occurrence of words to provide a context-specific sentiment orientation. Their authors argue that *"words that co-occur in a given context tend to have certain relation or semantic influence"*. Thus, the SentiCircles approach modifies the pre-established sentiment value by taking into consideration the context in which the term was used and building lexicons that use a contextual representation of words.

Aside from the fact that SentiClircles was designed to perform polarity analysis on Twitter (not tested in general texts) and only in the English language, what is the reason for limiting this idea of co-occurrence of words to single words but not the associated emotion? If words are used that appear near emotional and affective terms, they are somehow related to that emotion and must be incorporated in the emotional lexicon. That is, we will identify those words that co-occur to emotional related terms, and then we will label them with the emotion with which they co-occur.

From this perspective, we will show an approach that makes use of the co-occurrence of words for emotional terms in sentences retrieved from parallel texts to build lexicons for affective computing in multiple target languages.

3 An Approach to Automatically Building Lexicons for Affective Analysis

To automatically generate lexicons to perform Affective Analysis that annotates words with emotions in different target languages, our approach will start from the initial seeds of words in English and then use linguistic tools to perform expansion and translation. The outline of the approach can be summarized as follows:

1. Build sets of emotional words in English starting from some initial seeds for each emotion.
2. Retrieve parallel sentences in English and in the target language, where the words from the initial sets in English have been used.
3. Compute the most frequently used words in the parallel context sentences obtained for the target language.
4. Build the lexicon with the most frequently used words for each emotion.

Thus, we will first build a set of words in English for each emotion we want to identify. To do this, the set associated with each emotion initially contains only two seeds. These seeds correspond to the emotion noun and its related adjective in English.

In order to identify emotions to work with, we must consider disjoined categories. Although initially conceived to identify facial expressions within the Computer Image Analysis field, one of the most common models is the one proposed by Ekman [27], who defined a six basic emotions model that takes into account these emotions: anger, disgust, fear, happiness, sadness, and surprise.

Within the Natural Language Processing field, inter-annotator agreement studies for each of these six emotions have been performed in order to measure how well the annotators can make the same annotation decision for a certain category [15, 20, 28, 29], showing that the six emotions from Ekman's model are applicable. Thus, Table 1 shows the initial seeds associated with each emotion. As we can see, the seeds contain the emotion noun and its related adjective in the English language.

Table 1. Seeds associated with each emotion (noun and adjective) in English.

Emotion	Seeds (noun + adjective)
Joy	joy, joyful
Sadness	sadness, sad
Anger	anger, angry
Fear	fear, afraid
Disgust	disgust, displeased
Surprise	surprise, surprised

Then, to build the sets of emotional words in the English language, each word from these seeds is expanded with its synonyms from WordNet [14] and terms with similar meaning taking into account the target language from bab.la [30]. So far, we have obtained a set of words for each emotion in the English language. Afterwards, we collect context sentences in English where the emotional words have been used, and the corresponding translation of these sentences into the target language. That is, we get a corpus of parallel sentences where the emotional words are used. To build it, we have made use of the bab.la service [30]. Finally, we annotate the most frequently used words from the context sentences in the target languages with the corresponding emotion from which they come from, and then create the lexicon with those words.

4 Details of the Experience on Applying the Approach

In this section, we will explain how we applied the steps outlined in the previous section in the way detailed in Algorithm 1. Thus, as introduced in the outline section, we started with a seed ω of words in English for each emotion, particularly, the emotion noun and its related adjective in English. Then, we expanded these initial seeds with their related words and synonyms in a bootstrapping process. As a result, we obtained a set of words related to each emotion in English. To perform these tasks, we made use of bab.la [30], a language portal that provides multilanguage dictionaries. These dictionaries provide not only translations but also a full list of terms with a similar meaning for each term used, a list of synonyms based on different thesauri depending on the target language, such as WordNet [14] for English or OpenThesaurus [31] for German. Because bab.la does not provide an API interface, we have developed the necessary routines which request specific terms and scrapes the retrieved HTML in order to extract the relevant information.

1. Bootstrap the initial seeds of emotional words in English
1.1 For each *Emotion*
1.1.1 $\omega_{\{Emotion\}}$ = set {$noun_{\{Emotion\}}$, $adjective_{\{Emotion\}}$}
1.1.2 For each word in $\omega_{\{Emotion\}}$
1.1.2.1 $\omega_{\{Emotion\}} = \omega_{\{Emotion\}} \cup$ Synonyms(word)
2. Retrieve the context sentences from parallel texts
2.1 For each *Emotion*
2.1.1 For each *word* in the $\omega_{\{Emotion\}}$
2.1.1.1 $\Omega_{\{Emotion\}}$= Context_sentences(*word*, lang)
3. Obtain the sets of words for the target language
3.1 π = {entities} $\cup$ {abbreviations} $\cup$ {acronyms}
3.2 For each *Emotion*
3.2.1 For each *cs* in $\Omega_{\{Emotion\}}$
3.2.1.1 $\omega'_{\{Emotion\}}$ = tokenize(*cs*) - π
3.3 $\Upsilon = \varnothing$
3.4 For each *E1, E2* in combination(*Emotions*)
3.4.1 $\Upsilon = \Upsilon \cup (\omega'_{\{E1\}} \cap \omega'_{\{E2\}})$
3.5 For each *Emotion*
3.5.1 $\omega'_{\{Emotion\}} = \omega'_{\{Emotion\}} - \Upsilon$
4. Build the lexicon with the words for each emotion
4.1 $\varphi = \varnothing$
4.2 For each *Emotion*
4.2.1 For each *word* in $\omega'_{\{Emotion\}}$
4.2.1.1 $\varphi = \varphi \cup$ { (word, Emotion) }

Algorithm 1. Steps to implement the approach.

Following Algorithm 1, the next step was to retrieve the context sentences in a target language. To perform this step, we again made use of bab.la, which provides a large list of parallel context sentences to exemplify the usage of the words for both the source and the target languages.

Thus, for each word from the corresponding set ω, we fetched as many context sentences as possible. Bab.la reports these sentences as parallel examples where each word is used in both the original language (English) and its corresponding translation into a target language. Therefore, we filter only the translated sentences. The result is the context sentences Ω in the target language for each emotion.

Because entities, abbreviations and acronyms do not provide affective information, we needed to remove them. To perform this task in an unsupervised way, we simply used regular expressions to identify them. Thus, to find entities we used a regular expression looking for those words that start with only one uppercase letter, are followed by several lowercase letters and are not at the beginning of a sentence. In the same way, to detect abbreviations we used a regular expression looking for those words whose length is not greater than three letters, start with an uppercase letter and are not at the beginning of a sentence. Finally, to identify acronyms, we used a regular expression looking for combinations of uppercase letters, all of them joined by dots. Obviously, the best way to perform this task is to count with a set of well-known entities, abbreviations and acronyms, but because of the automatic nature of our approach, we were looking for a language-

independent approach, and this is an exploratory study, we considered this method a good starting point.

Following those steps, after removing the entities, abbreviations and acronyms that appear in the context sentences, we created a set ω' of words for each emotion in the target language by including the most frequently used words in the context sentence for the corresponding emotion. Because our aim was to obtain exclusive sets of words for each emotion, we removed the intersection $\curlyvee$ between sets of words so that we eliminated stop-words, domain-specific words and ambiguous words that could be associated with more than one emotion. Moreover, we did not include in the sets those words that have appeared only in one context sentence.

Finally, the lexicon φ is composed by the ω' sets containing the words in the target language and their corresponding emotion. To measure the precision of these obtained sets, we use EmoLex [17] as labelled data.

To summarize and help make the rest of the article easier to follow, we can provide these definitions:

- $\omega_{\{Emotion\}}$ is the initial set of emotion-related words in English, including the seeds we use to build the lexicon (i.e., the emotion nouns and their related adjectives), as well as their synonyms.
- $\Omega_{\{Emotion\}}$ contains the context sentences in the target language for the specific emotion, where the words from $\omega_{\{Emotion\}}$ have been used in parallel translated texts.
- $\omega'_{\{Emotion\}}$ is the set of words in the target language for the specific emotion.
- φ is the resulting lexicon containing the aggregation of $\omega'_{\{Emotion\}}$ with all the emotions.

5 Checking the Obtained Lexicons

To validate the set of English words ($\omega_{\{Emotion\}}$ from step one of Algorithm 1, we checked their precision against a gold-labelled lexicon for two different languages with very different origins and roots, namely, Spanish and German. As far as we know, currently the only manually labelled emotional English lexicon is *EmoLex* [17] with 14,182 words. Thus, we will use *EmoLex* as a reference dataset.

Table 2 shows the number of words labelled under certain emotions in *EmoLex* and how many of them can be found in the initial sets of emotional words we obtained in English. As expected, the initial sets contain a reduced number of words compared with those contained in *EmoLex*. For their part, Table 3 shows the precision computed for each emotion after using bab.la for the Spanish and German languages to obtain the initial sets in English. To compute the precision we followed Equation (1).

$$\text{precision}_{\{\text{emotion}\}} = \text{card}(\text{EmoLex}_{\{\text{emotion}\}} \cap \omega_{\{Emotion\}})/\text{card}(\omega_{\{Emotion\}}), \quad (1)$$

where $\text{EmoLex}_{\{\text{emotion}\}}$ is the subset from the reference dataset containing the words for the specific emotion, and $\omega_{\{\text{emotion}\}}$ is the set of words we have obtained for the same emotion.

Table 2. Number of words for each emotion in *EmoLex* and in the emotional words in English for Spanish and German.

		Spanish	German
Emotion	Emolex	Emolex$_{\{emotion\}} \cap \omega_{\{Emotion\})}$	Emolex$_{\{emotion\}} \cap \omega_{\{Emotion\}}$
Anger	1247	54	108
Disgust	1058	29	65
Fear	1476	60	191
Joy	689	57	83
Sadness	1191	68	140
Surprise	534	20	39

Table 3. Precision for each emotion for the initial sets of emotional words in English.

	Precision for Spanish	Precision for German
Anger	0.333	0.315
Disgust	0.379	0.246
Fear	0.317	0.204
Joy	0.404	0.337
Sadness	0.529	0.393
Surprise	0.350	0.231

Table 4. Number of missing words we can find in the obtained initial seeds of emotional words but which do not appear in *EmoLex*.

	Spanish	German
Anger	17	52
Disgust	9	22
Fear	19	68
Joy	28	44
Sadness	20	64
Surprise	5	15

Table 5. Number of retrieved context sentences and tokens per sentence scrapped from bab.la

			Tokens per sentence		
Emotion	Lang.	Nr. of sentences	mean	std	Var
Fear	German	38,303	13.831	4.954	24.541
	Spanish	11,440	15.981	5.599	31.344
Anger	German	6,695	14.563	7.082	50.149
	Spanish	5,326	18.326	8.249	68.047
Surprise	German	5,441	14.947	5.889	34.675
	Spanish	3,396	15.078	4.959	24.593
Disgust	German	5,394	15.561	5.930	35.167
	Spanish	3,134	18.668	7.256	52.654
Sadness	German	12,134	15.230	6.643	44.132
	Spanish	9,875	16.751	6.081	36.974
Joy	German	9,039	15.286	5.475	29.972
	Spanish	4,837	15.822	7.721	59.619

As we can observe, the obtained precision is a bit low. The reason for this can be found by looking inside the lexicon: although the initial sets clearly contain emotional words, those words are not contained in the *EmoLex* lexicon. Table 4 shows the number of seeds per emotion we have obtained but are missed in the *EmoLex* dataset. Details about the whole list of missing seeds can be found in Appendix A.

According to these results, it looks obvious that the initial sets contain a great percentage of emotional words related to their corresponding emotion, but currently there are is not a gold-labelled lexicon that we can use to validate them.

After performing step 2 of Algorithm 1 to obtain the context sentences for each emotion $\Omega_{\{Emotion\}}$ for German and Spanish as target languages, Table 5 summarizes the results. The table shows the number of context sentences we gathered from bab.la in December 2016 grouped by emotion and language. In addition, it presents the mean, standard deviation and variance for the number of tokens (words) per sentence, once we have removed acronyms, entities and abbreviations. As can be seen, there are differences in the number of gathered sentences per sentiment depending on the language. This is really significant because the resulting lexicon quality is directly related to the number of processed sentences.

Continuing applying step 3 of the algorithm to obtain the $\omega'_{\{Emotion\}}$, we retrieved the words for each language from the context sentences after removing the entities, abbreviations, acronyms, stop-words and domain-specific terms. As a result, we obtained the $\omega'_{\{Emotion\}}$ containing the words related to each one of the six emotions for German and Spanish.

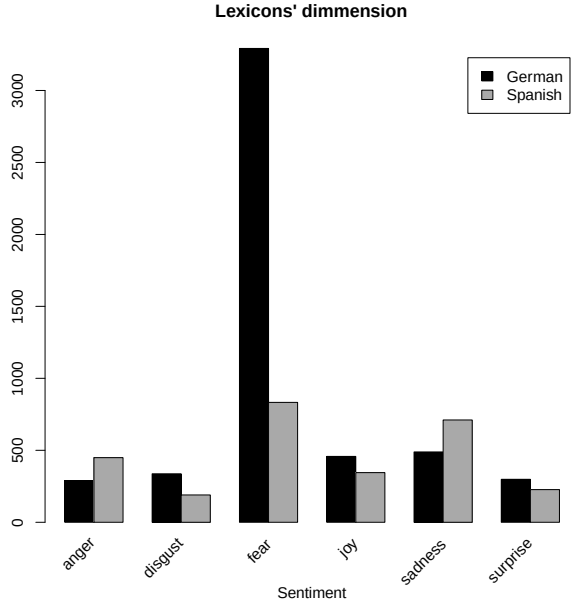


Fig. 1. Number of words per sentiment for the obtained lexicons.

By performing step 4 of the algorithm and obtaining the final lexicons, we got those words that appear at least twice, that is, in two sentences. As a result, Figure 1 shows the number of words per sentiment for the obtained φ lexicons for German and Spanish. As you can see, the number of words we obtained for each emotion may significantly differ among emotions and between languages.

6 Exploratory Analysis in a Parallel Corpora for German and Spanish

As previously stated, there is currently no gold-labelled lexicon that we can use to validate the obtained lexicons. Therefore, taking into account that most of the studies on Sentiment Analysis focused on documents [11], we conducted an exploratory Affective Analysis on a parallel corpus composed of well-known books in order to check the validity of the approach.

Thus, by checking the emotions identified in the texts as well as comparing the affective fingerprints between two different languages for the same books, we will see the viability of the proposal. In the next subsections, we will give the details and the initial results obtained.

6.1 Choosing the Corpus and Target Languages

The corpus we've used was developed in the OPUS project (Open Parallel corpUS) [32]. It constitutes a collection of translated aligned free texts from the web. In particular, from the corpora available in OPUS, we selected *"books"*, a collection of multilingually aligned copyright-free well-known books.

As target languages, we again selected German and Spanish. Accordingly, the books that we can find multilingually aligned in OPUS for those two languages are *Jane Eyre* by C. Bronte, *Alice in Wonderland* by L. Carroll, *Die Verwandlung* by F. Kafka, *The Fall of the House of Usher* by E.A. Poe, and *Anna Karenina* by L. Tolstoy. All of them are well-known titles.

6.2 Performing a Simple Emotional Identification to the Parallel Corpus

In order to check only the computed lexicon but not any classification algorithm, we simply performed a wordspotting approach to perform the Affective Analysis.

Accordingly, to avoid the effect of morphological change of the words, we repeated the algorithm to obtain a lexicon not of words but of stems, including those that appear at least ten times. Thus, we obtained a lexicon with stems categorized under a certain emotion. Later, to compute the affective fingerprint of a book, we simply increased the counter every time a stem categorized under a certain emotion appeared in the corresponding book. Lastly, we computed the percentage the stems with emotional charge represented in the whole book.

The obtained results are summarized in Figure 2. This figure shows a radar chart for each book with the six emotions in both target languages, namely, German (in black) and Spanish (in grey).

Recalling Figure 1, *fear* has over 3,000 words in the generated lexicon for the German language, but fewer than 1,000 for the Spanish language. Therefore, to avoid possible bias introduced by the size of the lexicon, in order to compare the emotions, values have been normalized for the German language following Equations (2), (3) and (4):

$$\%_{\{emotion\}} = 100 \text{ x } (\text{card}(Book_{\{emotion\}})/\text{card}(Book)) \quad (2)$$

$$factor_{\{emotion\}} = \text{card}(\varphi\text{-Spanish}_{\{emotion\}})/\text{card}(\varphi\text{-German}_{\{emotion\}}) \quad (3)$$

$$Normalized\ \%_{\{emotion\}} = \%_{\{emotion\}} \text{ x factor}_{\{emotion\}} \quad (4)$$

where $\%_{\{emotion\}}$ is the computed percentage of words from each book that are identified under a certain *emotion*, and $factor_{\{emotion\}}$ is the scaling factor in order to avoid lexicon size bias for the specific emotion.

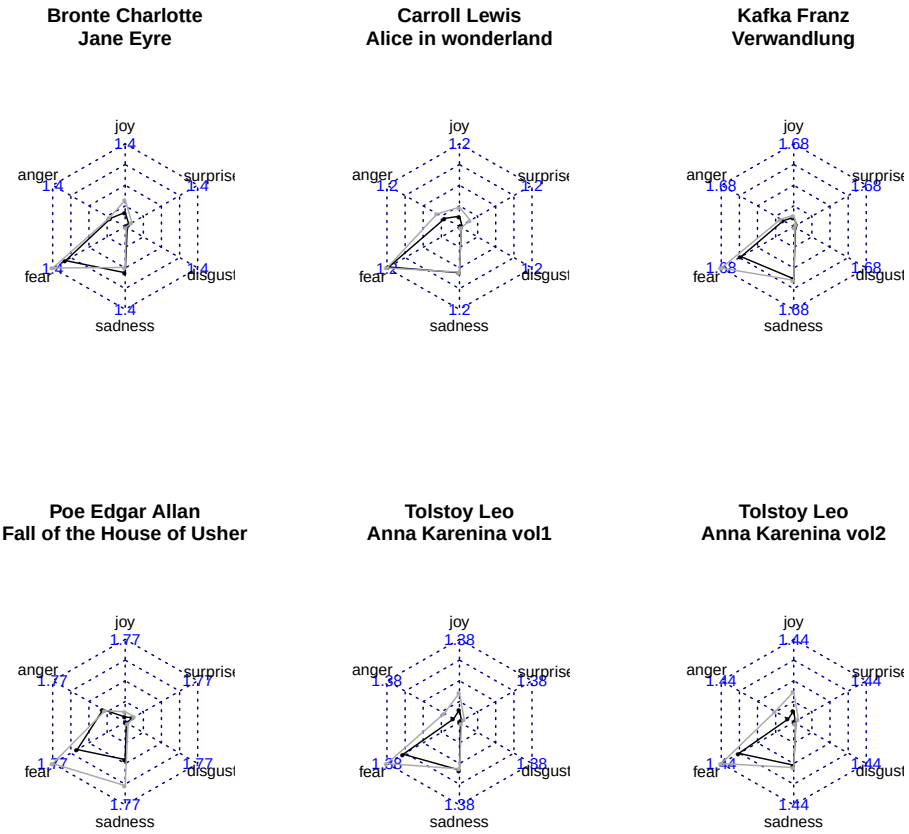


Fig. 2. Radar chart for the six emotions processed for each book in the German (black) and Spanish (grey) languages.

Because they are all well-known books, it is easy to check their emotional sign. Moreover, that affective fingerprint is really similar in both languages, that is, they present the same emotional tendency in both languages. It is interesting to notice that in fact *disgust* and *surprise* have almost no emotional load in the text. This agrees with the work performed by [33], who advise that the basic emotions are only four: *anger, fear, joy* and *sadness*.

6.3 Discussion and Further Work

Along with this article, we have detailed an approach for automatically generating target language independent lexicons to perform Affective Analysis. The approach starts with some initial seeds of nouns and adjectives for each emotion in English; these seeds are expanded by a bootstrapping process; and finally, the affective lexicon for the target language is obtained by processing the context sentence from the parallel text where the words for each emotion have been used.

We implemented the necessary routines that make use of bab.la, a language portal that provides multilanguage dictionaries, translations of common phrases and expressions, etc. in order to check our approach, to automatize the creation of sets of words with those words with similar meanings, to retrieve the context sentences and to perform the translation process.

Later, we conducted an exploratory analysis by applying the approach to a parallel corpus, checking the emotions identified in the texts and comparing the results between two different target languages. This exploratory analysis has shown that the approach provides coherent results and can be used to automatically generate target language independent lexicons to perform Affective Analysis.

However, although the conducted study has shown evidence about the benefits of the approach by providing us with an affective fingerprint in whole texts, it is necessary to improve its accuracy and enhance several critical points. The precision of the resulting lexicons depends on the quality of the context sentences as well as on the domain from which they are extracted. Validating the proposal using a set of context sentences for different specific domains will give us a better idea of the accuracy of our approach. In addition, it is necessary to test the proposed approach by using different bilingual dictionaries and translation tools such as Bing Translator, Google Translate, the Moses statistical machine translation system, and others.

7 Conclusions

In this article, we have presented an experience in automatically building lexicons for affective computing in multiple target languages. Starting with just a few seeds of words in English for each emotion we are interested in, we then bootstrapped them and retrieved context sentences from a parallel translated text where they had been used. By processing the translated sentences corresponding to the target language, we generated the affective lexicon for the target language. Although it is necessary to perform more work to get better precision, the exploratory analysis performed against parallel books in different languages has demonstrated its validity.

Acknowledgement. This research has been partially supported by the Ministry of Economy and Competitiveness (in Spanish *Ministerio de Economía y Competitividad*) through the projects REF: TIN2013-44586-R, TIN2014-52129-R and by the Ministry of Education, Youth and Sports, Community of Madrid (in Spanish *Consejería de Educación, Juventud y Deporte, Comunidad de Madrid*) through the project S2013/ICE-2715.

1. References

1. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr. 2, 1–135 (2008).
2. Liu, B.: Sentiment Analysis and Subjectivity. Presented at the (2010).
3. Feldman, R.: Techniques and Applications for Sentiment Analysis. Commun. ACM. 56, 82–89 (2013).
4. Piryani, R., Madhavi, D., Singh, V.K.: Analytical mapping of opinion mining and sentiment analysis research during 2000-2015. Inf. Process. Manag. 53, 122–150 (2017).
5. Jurado, F., Rodriguez, P.: Sentiment Analysis in monitoring software development processes: An exploratory case study on GitHub's project issues. J. Syst. Softw. 104, 82–89 (2015).
6. Saif, H., He, Y., Fernandez, M., Alani, H.: Contextual semantics for sentiment analysis of Twitter. Inf. Process. Manag. 52, 5–19 (2016).
7. Severyn, A., Moschitti, A., Uryupina, O., Plank, B., Filippova, K.: Multi-lingual opinion mining on YouTube. Inf. Process. Manag. 52, 46–60 (2016).
8. Cambria, E.: Affective Computing and Sentiment Analysis. IEEE Intell. Syst. 31, 102–107 (2016).
9. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. Presented at the (2002).
10. Liu, B.: Sentiment Analysis and Opinion Mining. Presented at the (2012).
11. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Syst. 89, 14–46 (2015).
12. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. Presented at the (2006).
13. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Presented at the (2010).
14. Wordnet: Wordnet, https://wordnet.princeton.edu/wordnet.
15. Strapparava, C., Valitutti, A.: WordNet-Affect: an affective extension of WordNet. Presented at the May (2004).
16. Bradley, M.M., Lang, P.J.: Affective Norms for English Words (ANEW): Instruction manual and affective ratings. (1999).
17. Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word-Emotion Association Lexicon. Comput. Intell. 29, 436–465 (2013).
18. Banea, C., Mihalcea, R., Wiebe, J., Hassan, S.: Multilingual Subjectivity Analysis Using Machine Translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 127–135. Association for Computational Linguistics, Stroudsburg, PA, USA (2008).
19. Gatti, L., Guerini, M., Turchi, M.: SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. IEEE Trans. Affect. Comput. 7, 409–421 (2016).
20. Strapparava, C., Mihalcea, R.: SemEval-2007 Task 14: Affective Text. Presented at the (2007).
21. Banea, C., Mihalcea, R., Wiebe, J.: Porting multilingual subjectivity resources across languages. IEEE Trans. Affect. Comput. 4, 211–225 (2013).
22. Balahur, A., Turchi, M.: Multilingual Sentiment Analysis Using Machine Translation? In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. pp. 52–60. CEUR Workshop Proceedings, Stroudsburg, PA, USA (2012).

23. Balahur, A., Turchi, M.: Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput. Speech Lang. 28, 56–75 (2014).
24. Chen, Y., Skiena, S.: Building Sentiment Lexicons for All Major Languages. In: for Computational Linguistics, A. (ed.) Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 383–389. , Baltimore, Maryland, USA (2014).
25. Wiktionary, the free dictionary, https://www.wiktionary.org/.
26. Staiano, J., Guerini, M.: DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 427–433. , Baltimore, Maryland, USA (2014).
27. Ekman, P.: Handbook of Cognition and Emotion. Presented at the (1999).
28. Strapparava, C., Mihalcea, R.: Learning to Identify Emotions in Text. Presented at the (2008).
29. Bhowmick, P.K., Mitra, P., Basu, A.: An Agreement Measure for Determining Inter-annotator Reliability of Human Judgements on Affective Text. In: Proceedings of the Workshop on Human Judgements in Computational Linguistics. pp. 58–65. Association for Computational Linguistics, Stroudsburg, PA, USA (2008).
30. Babla: Babla, http://bab.la.
31. TheOpenThesaurus: TheOpenThesaurus, https://www.openthesaurus.de.
32. Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In: Chair), N.C. (Conference, Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012).
33. Zinck, A., Newen, A.: Classifying emotion: a developmental account. Synthese. 161, 1–25 (2008).

Francisco Jurado is Lecturer in the Computer Engineering Department of the Universidad Autónoma de Madrid, Spain. He received his Ph.D. degree with honours in Computer Science from the University of Castilla-La Mancha in 2010. His research areas include Natural Language Processing, Intelligent Tutoring Systems, Heterogeneous Distributed eLearning Systems and Computer Supported Collaborative Environments.

Pilar Rodriguez is Associate Professor in the Computer Engineering Department of the Universidad Autónoma de Madrid. She received her Ph.D. degree in 1990, with a thesis on Computational Linguistics. She joined IBM in 1985, working at the IBM-UAM Scientific Center until 1989, when she was assigned to the Instituto de Ingeniería del Conocimiento, IIC. In November 1996 she joined UAM. At present, she is a member of the GHIA group at the UAM. Main research focuses on adaptive systems, especially for learning purposes, both in individual and collaborative environments.

Received: October 1, 2017; Accepted: September 30, 2018

Appendix A. Missing words in *Emolex* from the initial sets of emotional words in English

	Spanish	German
Anger	abuzz, angriness, annoyed, apotheosize, braveness, choler, courageousness, enraged, feral, heated-up, impassionated, involving, lawless, ricochet, tempestuous, thundery, vexed	aggravate, angrily, angriness, annoyed, blowy, blustering, bolshy, bombing, brawling, choler, cock-a-hoop, coltish, dissolute, ebullient, enrage, exasperated, extravagant, fantastic, feral, ferociously, great, groovy, hilariously, incensed, infuriated, infuriates, intemperate, irately, mean, miff, mindless, mustered, obstreperous, peeve, rabidly, rampant, random, rapturous, rioting, riotously, roaring, skittish, speeding, squally, sulfurous, sulphurous, super, tempestuous, vex, wrathful, wrathfully, wroth
Disgust	abhorrence, creeps, displease, nauseate, reluctancy, repel, repulse, sicken, uspet	abhorrence, abominate, appal, appall, cloy, disgruntle, displease, dissatisfy, exasperate, insubordination, insurgency, nauseate, odium, putsch, rebelliousness, repel, repugnance, repulse, scuff, sicken, spurn, vex
Fear	agonized, anguished, attend, belong, bow, brokenhearted, browbeat, bussy, carefulness, enshrine, fearfulness, heedfulness, incumbency, trouble, troubling, tyrannize, venerate, worrier, worrisome	accurateness, affect, agitate, angst, appertain, apprehensibility, apprehensiveness, attentiveness, biz, browbeat, carefulness, cause, cautiousness, chariness, consideration, cure, daunt, devotion, devotions, dignify, disconcertment, disquiet, disturb, edited, egis, elaborateness, enshrine, enshrinement, fearfulness, fosterage, funky, godawful, handling, harassment, heedfulness, honour, idolise, idolize, insurgency, involve, ministration, misgiving, mousey, mousy, neatness, nursing, obeisance, overawe, painstakingness, regard, regardfulness, revers, safe-keeping, scariness, shy, solicitousness, solicitude, therapy, timidly, trouble, turnout, unease, value, venerate, vex, wariness, wish, wonder
Joy	God, buoyant, cock-a-hoop, deity, delectation, elation, enjoyment, enthuse, euphoric, exhilarate, exhilarated, exult, gaiety, gladden, gleeful, gloat, godhead, jolly, joyfulness, joyousness, jubilate, jubilation, lively, luxuriate, mirthful, perky, playfulness, pleasure	blissfulness, bright, buoyant, chirpy, complacence, contentment, convivial, delectation, elate, elates, emboldened, encouraged, encourages, enjoyable, enthral, enthrall, exhilarated, exultant, exulting, gaiety, gay, gayety, gladden, gleeful, gloating, gloatingly, gratification, gratifying, gratifyingly, joyfully, joyfulness, joyousness, jubilantly, jubilate, jubilating, lively, nerved, please, pleasure, prosperousness, regale, sneering, zestful, zestfulness
Sadness	aggrieved, anguished, distressful, disturbing, gloominess, grief-stricken, grieved, heartbroken, lamentable, lugubriousness, paretic, pitiable, pitiful, regretful, sad, shamed, sorrowfulness, sorry, troubling, wistfulness	abominably, afflicting, afflictive, afflictively, afflicts, awkward, bewail, bitchiness, bitter, broody, caitiff, deplorably, depressingness, dim, disconcerting, dolefully, dolorousness, down-at-heel, dreariness, funereal, gloominess, grief-stricken, grilling, heartrending, heavy-hearted, hoodoo, lamentable, lamentably, lamentation, lugubrious, lugubriousness, meagre, mercifully, misadventure, niggling, penitence, piteous, pitiable, pitiably, pitiful, pitifully, pitying, plangent, regretfulness, regrettably, ruefulness, sad, saddening, serious, shattering, smart, sombreness, sorrowfulness, sorry, tearfulness, teariness, tormenting, trouble, troubled, troubling, upsetting, verminous, vexed, woebegone
Surprise	rainstorm, startled, surprisal, wonder, wonderment	amazed, amazes, astonish, astonished, besiege, dumbfound, dumfound, eye-opener, pelt, storminess, surprisal, welter, wonder, wondering, wonderment

Computer Science and Information Systems 16(1):289–311 https://doi.org/10.2298/CSIS180122013L

Improving Sentiment Analysis for Twitter Data by Handling Negation Rules in the Serbian Language

Adela Ljajić[1] and Ulfeta Marovac[1]

[1]State University of Novi Pazar, Vuka Karadžića bb,
36300 Novi Pazar, Serbia
{acrnisanin, umarovac}@np.ac.rs

Abstract. The importance of determining sentiment for short text increases with the rise in the number of comments on social networks. The presence of negation in these texts affects their sentiment, because it has a greater range of action in proportion to the length of the text. In this paper, we examine how the treatment of negation impacts the sentiment of tweets in the Serbian language. The grammatical rules that influence the change of polarity are processed. We performed an analysis of the effect of the negation treatment on the overall process of sentiment analysis. A statistically significant relative improvement was obtained (up to 31.16% or up to 2.65%) when the negation was processed using our rules with the lexicon-based approach or machine learning methods. By applying machine learning methods, an accuracy of 68.84% was achieved on a set of positive, negative and neutral tweets, and an accuracy of as much as 91.13% when applied to the set of positive and negative tweets.

Keywords: Sentiment Analysis, Serbian Language, Twitter, Negation Detection, Negation Rules, Machine Learning.

1. Introduction

Sentiment Analysis (SA) belongs to one of the sub-fields of Natural Language Processing (NLP). NLP is a technology that deals with ubiquitous human language that appears on websites, product descriptions, newspaper articles, social media, and scientific articles, all in thousands of languages and linguistic variations. SA belongs to the field of language technology which is on the rise, and with the appearance of machine processing of natural language and machine learning algorithms for classification, greater progress in this area occurred. The opinions of people expressed and written in the given language are not easy to analyze and it is not easy to determine their sentiment, given the complexity of the linguistic expressions and the different way of expressing oneself in a particular language. Some other problems in SA are the ambiguity of words or syntagms, non-standard writing, use of slang, neologisms, segmentation problems, irony, metaphor, and negation. Regardless of the inability of a computer to automatically determine sentiment, under certain assumptions the complexity of such a task can be simplified. The simplest systems determine the sentiment based on the occurrence of a word in the positive and negative sentiment lexicon. A few more advanced systems take into account POS tags, apply the rules of negation and detect irony. Although the determination of polarity, as one of the tasks of

SA, is an area which has been studied to some extent, some aspects that are linguistically specific (negation, irony, metaphor) still pose challenges and areas where improvements are expected. In this paper, we deal with the processing of grammatical rules of negation in the Serbian language. The main contribution of this paper is to show whether the treatment of these rules of negation and their integration into the method of classification of the sentiment can improve the accuracy of the prediction of short texts - in our case, tweets. The rules that are processed represent a subset of grammatical rules for the processing of negation in the Serbian language, i.e. the rules that are most commonly encountered in this type of short text. The applied rules and their influence on the polarity classification will be tested by an unsupervised lexicon-based method and a supervised machine learning method. It will be shown that, by integrating the rules of negation, there is a statistically significant improvement in the application of the method for the prediction of sentiment.

The rest of the paper is organized as follows. We begin with a description of related work in SA with negation in Section 2. Section 3 contains information about the dataset structure, sentiment lexicon, negation signals lexicon and other lexicons used. Section 4 describes the proposed method. The experiment is included in Section 5. Section 6 contains an evaluation and analysis of the results. Finally, we conclude and present directions for future work in Section 7.

2. Related work

In recent years there has been great interest in the research done in the field of SA. When determining text sentiment, researchers usually assume that a speaker has some attitude or sentiment about the subject, object or person, that the sentiment has a fixed value (good or bad) and that the sentiment in the text is represented by a word or combination of a small number of words. However, in the literature, there is little discussion of what a "sentiment" or "opinion/attitude" actually is [1]. In some review papers [2,3], the authors give an overview of the different approaches of determining sentiment. Sentiment can be analyzed at the document level, at the sentence level, or at the aspect level. Since SA is used to determine the subjective attitude about a phenomenon (object, person), some systems for SA include, in addition to positive and negative, neutral (objective) attitudes. Therefore, SA mainly involves research within three classes: positive, negative, and neutral. In [4] Pang et al. define the baseline method for sentiment analysis based on the sentiment lexicon. Sentiment lexicons contain words that express a positive or negative sentiment. By counting words that carry sentiment and using them as features for the machine learning methods, the sentiment of the sentence is determined. The extension of this method depends on the type of text, as well as on the specificity of the language on which this method is applied. Most papers deal with SA in English. Multilingual SA was introduced in [34] where authors compare their own implementation of existing approaches. The emphasis of the work is on the methods used for the Serbian language, and we will deal with them in more detail later on. The popularity of Twitter grows with the number of tweets, so an increasing number of authors have decided to test sentiment of this type of short text. An overview of the methods for analyzing sentiment data in Twitter is given in [19]. Most papers use supervised learning methods to determine sentiment, although not an

insignificant number of approaches provide analysis by methods of unsupervised or combined semi-supervised learning.

2.1. Sentiment Analysis in Serbian

SA in the Serbian language was mostly dealt with by authors in [23] and [21]. Mladenović et al. in [23] used morphological dictionary rules for inflection expansion to build sentiment lexicons and a stop word list that contains lemmas and inflectional forms of the lemma in order to reduce feature space for machine learning classification. They used the Maximum entropy classifier method for the prediction of text sentiment in the Serbian language using a set of newspaper articles for the application of a 10-fold CV. To confirm the obtained results, the method has been tested for two more special sets: a set of newspaper articles and a set of movie reviews. The authors used rich lexical resources and worked with two classes (positive and negative). They used the attribute reduction method for attributes (words) found in sentiment lexicons and mapped them to a new set of attributes. They received an accuracy for a 10-fold CV of up to 95.6%, of up to 78.3% for movie reviews and of up to 79, 2% for the news test set. In their work, negation is not particularly addressed.

Batanović et al. in [21] created the Serbian movie review dataset – SerbMR using a dataset balancing algorithm that minimizes the sample selection bias by eliminating irrelevant systematic differences between the sentiment classes. The authors gave a SA using various combinations of n-grams, stems, lemmatizers, and different types of normalization of the attribute. Finally, applying optimal attribute settings over the NBSVM classifier, they achieved an accuracy of up to 85.55% for two, and 62.69% for three classes. Some authors, for example in [20] examine the influence of 2 modes of morphological normalization of the text (using a stemmer and lemmatizer for the Serbian language) on the influence of the sentiment classification on a data set of movie reviews. Using a combination of unigrams and bigrams, they achieved a statistically significant improvement compared to the baseline method for two classes (accuracy 86.11%) and for three classes (accuracy 63.02%). A comparative overview of the method of analyzing sentiment for the Croatian language is given in [22]. Due to the great similarity of the Croatian and the Serbian language in the complex-morphological sense and in other characteristics, we can consider that the results achieved are similar to those in the Serbian language. The authors used the "word embedding" method and "string kernels" comparison over three sets of short texts (game reviews, tweets from a specific domain and general-topic tweets). They showed that the "word embedding" technique and "string kernel" technique achieved an improvement over the bag of words (BOW) baseline method.

2.2. Negation in Sentiment Analysis

The most commonly used method for processing negation is to add the suffix _NEG to a sentiment word that occurs in the negated scope and to treat the negation from the negation signal to the first punctuation mark [4, 5, 6], or to the first positive or negative sentiment word found [7]. Although most negation processing systems take into account

all types of sentiment words, some take only adjectives [8] or adjectives and adverbs [9]. In [6], the authors examined the effect of the negation signal, as one form of valence shifters, on the change in the polarity of the word in the negated part. Most authors simply change the polarity of the negated word when processing negation (from n to -n and vice versa) [7, 10]. Sophisticated negation systems specifically treat sentiment words that appear in the scope of negation and show that negation with different intensities changes the polarities of the positive and negative sentiments of the word [11]. The authors who advocate this approach have shown in [12] that the production of such specific vocabulary for words that occur in the negation scope gives a significant improvement in the system for predicting sentiment. In addition to works dealing with negation using the negation signal, in [14] phenomena reminiscent of negation is processed because they change the polarity, but is not negation in the classical sense, as they do not contain words that are negation signals (no, not, did not ...). In [15], the authors suggested the detection of the scope of negation using the reinforcement learning method. Evaluation of the effects of the detected negation on the SA was done by measuring the correlation between the sentiment sentences and the corresponding "daily stock market return" as the measure of the prediction, and a correlation of 10.63% was obtained between the sentiment value and the stock market return. In [16], the authors detected negation signals and the scope of negation. They applied detected negation to the SA classifier for the entire set and the set containing negations. Their sophisticated method for predicting sentiment (when using negation treated attributes) in the whole set achieved an improvement in the accuracy of 0.01. Although most of the methods for evaluating the impact of negation on the sentiment of the text are based on supervised learning, the authors [10, 16] proposed to calculate the intensity of negation and its impact on polarity by applying its methods of calculating the intensity of negation and polarity of the text. In [10] they dealt with the analysis of negations in the Spanish language and their influence on SA. A corpus of tweets in Spanish was used along with an unsupervised method for predicting sentiment that, in addition to other resources, includes a processed negation. They obtained results that showed that by treating the negation, they achieved a significant improvement in accuracy in determining the polarity of tweet. In [18], the rules for identifying negation and calculating its intensity are described. The rules were created in order to improve the sentiment text analysis. The method that predicts sentiment is an unsupervised method that calculates the polarity and intensity of the words and phrases. They showed that there is a positive correlation between the polarity determined by the system and the one assigned by the 5 participants in the experiment.

It is difficult to say that there is a versatile, best classifier of text sentiment. Systems that achieve good results for long texts can be inappropriate for short ones. An example is the work done by Cerezo-Costas er al. [17], whose method used on a dataset containing sarcasm (Tweet Sarcasm 2014) took first place, while for the SemEval 2015 group (Analysis Task 10) it took 16th place. A review paper [13] provides an overview of most of these approaches in processing negation from the earliest papers published on the topic; represented ways of processing negation, the scope of negation, the selection of training attributes, and the boundaries in the negation model for SA.

For Serbian, there are no studies dealing with the detection of grammatical rules of negation in a text, and therefore the influence of such processing on SA. Batanović et al. applied a classical method of processing negation, by changing the polarity of words that appear after the negation signal [21]. By dealing with negation on a set of movie

reviews, they achieved the following improvements in relation to the initial method [4]: for three classes they achieved the greatest improvement by means of the MNB method of 0.94% (accuracy of 54.72 to 55.66) marking two words after the negation; for two classes they achieved the greatest improvement with the SVM method of 0.66% (75.98 to 76.64) by marking only the first word after the negation.

3. Dataset and lexicons

This section presents the preparation of dataset and describes specialized lexicons for the SA and negation analysis. SA requires the use of a lexicon with the annotated sentiment of a word (sentiment lexicon). Negation signals, negative quantifiers, and particle intensifiers are needed to process negation as an indispensable part of the SA. The use of a stop words lexicon is widespread in text analysis and must be adapted to negation analysis so as not to lose the negation signals.

3.1. Dataset preparation and structure

So far, Serbian language sentiment analyses have been carried out on datasets of long texts (newspaper articles, movie reviews). In short texts, especially tweets, the sentiment is expressed more concisely than in long texts, but it is more difficult to determine, given the small number of words, ambiguity, informal writing, and often the presence of irony and sarcasm that significantly distort the classification. There is no annotated corpus of short texts for the analysis of sentiment in the Serbian language. For the purposes of this research, we collected 9059 tweets and annotated them. The tweets were collected using twitter streaming API from profiles of 15 official mass medias, 10 musicians, 5 public figures from the entertainment world, 5 athletes and 7 actors. The sentiments are grouped, according to sentiment, into three classes: negative, neutral and positive. The dataset contains 4334 negative, 2507 neutral and 822 positive tweets. In addition to positive, negative and neutral, irrelevant tweets were detected. These are tweets that do not contain clear information, contain only a link, or contain an informal phrase that cannot be attributed even to a neutral tweet. Irrelevant tweets were discarded. The data was annotated by 3 independent people, two men, and one woman. One person holds a master's degree in electrical engineering, one is doctor of medicine, and one is a student of Serbian language and literature. Only tweets which all three annotators agreed on (7663 tweets) were taken into account. In 15.41% of the cases (1396 tweets) , the annotators did not agree. An imbalanced number of tweets by class are expected because the topic of the tweets we have collected is more negative than positive. Regardless of the variety of media and personalities, tweets that carry a negative sentiment prevail, while the number of tweets with a positive sentiment is far smaller.

3.2. Lexicons used

Sentiment lexicon

The most general sentiment lexicons are for the English language. The "General Inquirer" sentiment lexicon dates back to 1966 [27] and consists of 1915 positive and 2291 negative words. LIWC [28] contains 2300 words and 70 classes, such as negative emotions (hatred, anger, problematic issues...) and positive emotions (love, beauty, kindness ...). The MPQA lexicon [29] contains 6885 words from 8221 lemmas, of which 2718 are positive and 4912 are negative. The Bing Lui [30] Opinion Lexicon consists of 6786 words, of which 2006 are positive and 4783 negative. The SentiWordNet [31] is a lexicon where every word has an associated coefficient that determines how positive, negative, or objective it is. To the best of our knowledge, there is no publicly available sentiment lexicon for Serbian. For the needs of our analysis, a sentiment lexicon was created; the translation of the Opinion Lexicon [30] was taken as the starting version of our sentiment lexicon.

Negation signals lexicon

We named the words involved in creating a syntactic negation "negation signals". The negation signal lexicon is derived from a list of all the word forms in the Serbian language that are involved in creating a syntactic negation. Negation terms ("ne" (no) and "ni" (no)) are included in the negation signal lexicon. In the Serbian language, the verbs "biti" (to be), "hteti" (to want) and "imati" (to have) have an irregular negation form, and the negation is written merged with the verb. The forms of these verbs are: "ne jesam"(to not be) in the present tense: "nisam" (I am not), "nisi" (you are not), "nije" (he/she/it is not), "nismo" (we are not), "niste" (you are not), "nisu" (they are not);"ne biti" (to not be) in the imperative mood: "nemoj" (do not), "nemojmo" (let's not), "nemojte" (do not); "ne hteti" (do not want) in the present tense: "neću" (I do not want), "nećeš" (you do not want), "neće" (he does not want), "nećemo" (we do not want), "nećete" (you do not want), "neće" (they do not want); "ne imati" (do not have) in the present tense: "nemam" (I do not have), "nemaš" (you do not have), "nema" (he does not have), "nemamo" (we do not have), "nemate" (you do not have), "nemaju" (they do not have). These verbs in the negative form (negation) mingle with negatives and become special negative signals. All forms of negations of the verbs "jesam" (to be), "hteti" (to want) and "imati" (to have) are also included in negation signals lexicon. The total number of negation signals in this lexicon is 25.

Negative quantifiers lexicon

The negative quantifiers in the rules of negation in the Serbian language play a completely different role than negation signals, so it would be wrong to equate them. Thus, we have singled out negative quantifiers in a separate lexicon. Universal negative quantifiers are morphological "ni-" (no) negations of negative pronouns ((niko, ništa, nikakav... (nobody, nothing, no ...)), negative adverbs (nikad, nigde, nikako... (never,

nowhere, noway ...)) and a small number of negative pronominal adverbials (nimalo, nijedanput... (not at all, never ...)) [32]. The negative quantifiers always stand with negation signal (almost always in front of it) and confirm the negation that lies with them. For example, in the sentence "Nikad nije lagao." ("He has never lied."), "Nikad" ("Never") is a negative quantifier that agrees with the negation signal "nije" ("is not") that changes the polarity of the word "lagao" ("lied") from negative to positive. The agreement of negative quantifiers with negation means either the confirmation or intensification of the negating part.

Particle Intensifiers Lexicon

In the Serbian language, there are types of words that are called particles, and their scope is such that they express their relation only to the content of one term in a sentence [32]. Particles that affect negation are considered by intensifying or extending the scope of negation validity. We used particle intensifiers of negation by extending the scope of the validity of the negation even after the punctuation mark if the particle intensifier is behind it. The particle intensifier lexicon contains: "ni" ("no"), "nit" ("neither") and "niti" ("neither").

3.3. Stop Words Lexicon

The role of stop words is to remove those words (potential attributes) that do not bring meaning to the text. A stop words lexicon can be general (universal) or specific to the field. Specific lexicons are created mainly by extracting words that have the highest TF or the smallest TF/IDF. The authors in [23] experimentally confirmed that a universal lexicon gives approximately the same results as a specific lexicon. In the paper, we used the general stop words prepared in [24], which mostly consist of adverbs, prepositions, conjunctions. The stop word lexicon does not contain negation signals and words that participate in any rule of syntax negation.

4. The proposed method

4.1. Motivation

In sentences that have a certain construction in which a negation occurs, there is sometimes a neutralization of negation, sometimes there is an intensification and sometimes a change in polarity. The different ways in which negation can act are introduced in [25]. The idea is to process the rules determining when the negation changes the polarity of the word, and in which part of the sentence, and when the negation is actually a pseudo-negation and does not simply change the polarity but expresses the affirmative. In this paper, we want to show that by taking into account the specific rules of negation, by detecting the scope to which the signals of negation act, or

by ignoring the negation signal, it is possible to improve the quality of the classification of tweets based on sentiment.

4.2. Normalization

Languages that are morphologically rich and highly inflected require the application of linguistic rules for the processing and classification of texts. In order for a text to be classified, it must first be normalized. Each text classification requires a specific normalization of the text. Normalization of tweets was done using:

- the tokenizer adapted to the specific structure of tweets,
- converting text from one alphabet to another – to Latin
- removing the stop word
- a Serbian language stemmer.

Tokenization and parsing of tweets

Tokenization is the first task in the preprocessing (normalization) of tweets. Tokenization was done by tokenizer built into the Python programming language using the *nltk.tokenize* module. The RegexpTokenizer was used to split the tweet into tokens and to process dates, various number formats and hashtags. Terms used by Twitter which are not relevant to our analysis (via, RT ...) were removed using the *re* python module and the regular expression. Since the Serbian language has two official alphabets (Cyrillic and Latin), it was necessary to process tweets written in Cyrillic and Latin, but separately. In order to avoid duplicate analyses, tweets written in Cyrillic were converted into Latin using the *cyrtranslit 0.4* python package. Following conversion into the Latin alphabet, stemming was performed.

Stemming

In the study we used a stemmer that encodes special Serbian Latin characters ‘č’, ‘ć’, ‘š’, ‘đ’, ‘ž’ to ‘cx’, ‘cy’, ‘sx’, ‘dx’, ‘zx’ [26]. It creates stems longer than 2 characters and has a dictionary of irregular verbs and their flections. Other words are stemmed using generalized stemmer rules. The tasks that the stemmer does not deal with are incorrect flections, sound changes, and short words. In addition to the above deficiencies, the accuracy of the stemmer is 90%, so we decided that it is sufficient to use it in the normalization of our text. An accuracy of 90% was obtained by machine stemming, after which the human read the stems and compared them. The main problem was noted in the words that have sound changes. The rules are designed to cover as many words as possible, and as sound changes occur in a fewer number of words, it is not possible to make an algorithmic rule that supports both words with a sound change and words without it. The other problem that came up with short words consisted of 3 or 4 characters, and which after stemming had only two, and the original meaning was in many cases difficult to extract. Adjusting the stem for the experiment was possible because the code is publicly available in the Python programming language. We tested stemming from a stem longer than 2 letters, irrespective of the length of the word, and

stemming in a stem longer than 2 letters, but only for words longer than 4 letters. The experiment showed that stemming in a stem longer than 2 letters, but only for words longer than 4 letters, gives a better result. The Serbian language is morphologically rich; therefore, the use of a stemmer is expected to improve the classification of sentiment. By sticking together, different forms of the same word are reduced to the same stem, so the number of occurrences of that stems increases. The authors in [21] showed that using a stem on their data set increases the average number of occurrences of each word by approximately 50% and decreases the size of sentiment lexicons for each class by approximately 30-35%. After stemming, our dataset is ready to apply the rules for processing negation.

4.3. Negation rules in the Serbian language

Negation is an area studied both by logic and linguistics. The linguistic analysis of negation must include the logic rules of negation [32]. The question to be answered in the negation is to determine the scope to which logical negation will be applied in accordance with linguistic rules. The Serbian language generally has complex rules. Negation can be lexical, syntactic and morphological, depending on the type of language unit in which it appears. Depending on the range of the unit whose content is negated, negation can be partial (the content of the part of the sentence is negated) or total (the entire sentence content is negated).

Morphological negation is accomplished by using prefixes such as “ne-” (non-), “bez-” (no-), “ni-” (not-), “a-” (a-), “dis-” (dis-) and “in-” (in-). Morphological negation is exhausted only at the lexical level. It has an effect on sentence negation only if the lexemes appear in the role of sentence members (as negative words, i.e. negative lexemes). The morphological negation in this paper was processed in a way that negated lexemes were inserted into the sentiment lexicon. Lexical negation is related to the use of a word whose meaning has a negative component (“sumnja” (doubt) and “nedostatak” (a lack of)). Lexical negation was not processed in this paper.

Syntactic (sentence) negation is accomplished by negation signals “ne” (not) and “ni” (not). The verbs which have an irregular negation form behave like negation signals. In this paper, we analyzed the syntactic (sentence) negation, i.e. the effect of the negation signal on a part of sentence or on the whole sentence.

The created negation rules

The rules of syntactic negation in the Serbian language are shown in [32]. The author presents the large numbers of negation and noted the rules that rarely appear. By analyzing the dataset of 3000 unannotated tweets, the rules of negation that most commonly appear in tweets were processed. For the initial scope of the negation, a set of words was taken from the first negation signal to the first punctuation mark. In some cases, it was necessary to change the scope of the negation. The following rules of negation improve the detection of the scope to which the negation relates and also improves overall SA. The rules of negation are:

1. The treatment of the negation type “Nije samo …. nego/već” (Not only .... but/instead) – the word that is the negation signal, which stands in front of the word “only”, is omitted. The scope of the negation is empty.
2. The treatment of the negation type “Nije…. nego/već” (Not .... but/instead) – in this case, the scope of the negation ends with the word “nego/već” (but/ instead). The negated part of the sentence has a smaller impact on SA than the following part of the sentence, so it should be omitted from SA.
3. The treatment of the negation in the questions of the type “Zar nije lepo?” (“Is it not nice?”) - in this case, the negation is neutralized because the sentiment after the negation signal does not change the polarity of the sentiment words. The scope of the negation is empty.
4. Treatment of the negation in the presence of an intensifier - If the intensifier appears after the negated word, then the next part of the sentence is negated, too. The scope of the negation is extended even after intensifier to the next punctuation mark.
5. Intensification of negation by negative quantifiers - agreement between negative quantifiers and the negation signal. Negative quantifiers increase the intensity of the negation.

The rules are applied in the order they are given. The correct order of the application of the rules is important because the Rule1 is, in fact, a pseudo-negation (the negation signal does not affect the following part which remains affirmative), and it excludes the application of the Rule2. Rule3 refers to a special question type that contains the negation signal preceded by the word “zar” (isn’t, aren't) or the word “jel” (isn’t, aren't). “Jel” is not grammatically correct, but we included it because it is used in informal speech and on Twitter for the same purpose as “zar”. These words belong to the group of negation neutralizers in questions. Rule4 is more general and can be applied to determine the scope of the negation independently of the language. Its use necessitates a change in the polarity of the negated word. Rule5 identifies negative quantifiers which agree with the negation (they confirm the negation) and in this way intensify it. When determining the sentiment of one tweet, the negation is processed in such a way that sentiment words from the scope of negation change the polarity and the number of negative quantifiers that affect the negated part assigned to that tweet. The rules are shown on the examples of tweets in Table 1.

Table 1. Tweets with the negation rules examples (highlighted - the initial scope of the negation, underlined - negation signal, bolded - the words that participate in the rule, framed - the final scope of negation)

Rule	The tweet containing the rule
1	**Ne** **samo** bezobrazluk, **već** i gaženje i ismevanje naroda. / **Not** **only** disgrace **but** also the violating and the mocking of the people.
2	Ovo se dešava kad se novac **ne** investira **nego** se troši. / This happens when money is **not** invested **but** consumed.
3	**Zar** vas nije sramota? / **Are** **not** you embarrassed?
4	Nije vredan, pametan **niti** lep. / He is not hardworking, smart **nor** nice.
5	***Nikad niko*** nije loše pričao o njemu. / ***Nobody*** ever (***never***) talked (did not talk) about him badly.

In the last rule (Rule 5), with the translation of Serbian into English, the double negation disappears because it does not exist in English (for clarification, word/words that would exist in the tweet if English had the same rule as Serbian are given in brackets).

4.4. Baseline methods

Method0 – This one performs SA without processing the negation, only based on the number of positive and negative words from the sentiment lexicon.

Method1 – This one processes negation by a simple change in polarity of the sentiment of the first word following the negation. The attributes are the following ones:

- the number of negation signals
- the number of positive words
- the number of negative words
- the number of negated positive words - one word after negation
- the number of negated negative words - one word after negation.

The negation can be processed simply by a change in the polarity of the word following the negation, for the words before a punctuation mark [4]. The authors in [21] analyzed the influence of negation on the words following it, and obtained the best results when there was a change in the polarity of the first word following the negation. This is why Method1 was taken as the baseline method of comparison which processes the negation, and which changes the polarity of the sentiment of only the first word following the negation.

4.5. The method

The method we propose (Method2) needs to determine the sentiment of the tweet, taking into account the rules of negation that we have processed. The method uses previously described resources such as the sentiment lexicon, the negation signal lexicon, the particle intensifier lexicon, and the lexicon of the negative quantifiers.

After normalizing the text, it was determined whether the text contains negation and if it does, which group of rules of negation the negation belongs to. After processing the rules of negation, attributes important for the determination of the sentiment were extracted. The attributes are the following ones:

- the number of negation signals
- the number of negative quantifiers
- the number of positive words
- the number of negative words
- the number of negated positive words - one word after negation
- the number of negated negative words - one word after negation
- the number of negated positive words – in the negation scope
- the number of negated negative words - in the negation scope

By using the selected attributes, prediction of the sentiments of tweets was performed using:

- an unsupervised classifier lexicon-based method (LBM)
- supervised machine learning classifiers (MLM).

In the experiment, these attributes are used for the application of LMB or MLM in relation to the two baseline methods (Fig. 1).

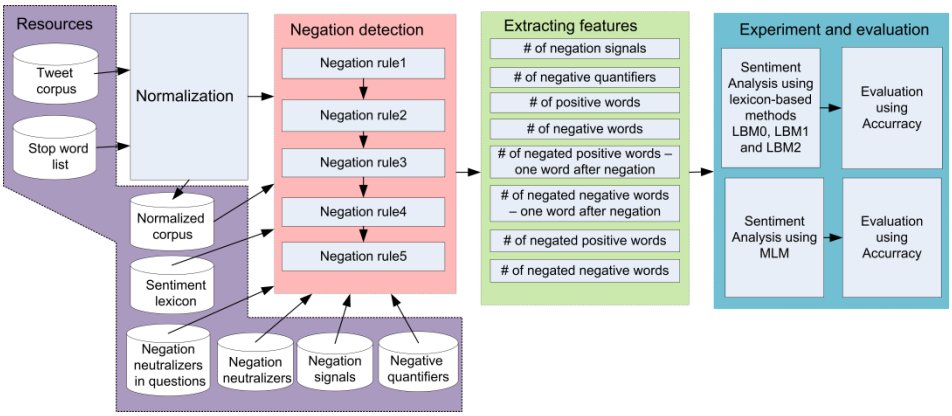


Fig 1. The architecture of the system

5. The experiment

Even though they are not structured texts, and are written with some informality, tweets can still be processed, and by implementing specific rules of negation the determination of the sentiment of these texts can be improved.

The aim of this paper was to determine the extent to which the processing of negation by means of the proposed rules can improve the SA of short texts, or in our case, selected tweets. Thus, a comparison was carried out of the obtained accuracy of the SA of the text, which processes negation by means of the proposed method (Method2) and the accuracy of the baseline method (Method0 and Method1).

For all three cases of negation processing by means of Method0, Method1, and the proposed Method 2, two techniques of sentiment prediction were implemented:

- LBM
- MLM.

In addition, both techniques of sentiment prediction (LBM, MLM) were performed on:

- the entire set (ALL)
- the set which includes only negation (OnlyNeg)
- the set which includes only negations which were included by the rules (OnlyRuleNeg).

The analysis which included the complete set, the set with negation signals and the one in which the rules were applied is the same methodology that was used by Jiménez-Zafra et al. [10]. Statistical analysis by means the unsupervised lexicon-based method was used on a set of three classes. A set of three classes (3-class) consists of positive, negative and neutral tweets, and a set of two classes (2-class) consists of positive and negative tweets. Methods of machine learning were also implemented on 3-class and 2-class. Table 2 shows the number of tweets for 2-class and 3-class for ALL and OnlyNeg and OnlyRuleNeg.

Table 2. The number of tweets in the entire set, the set with negation signals and the set of negations included by the rules

	3-class	2-class
ALL	7664	5156
OnlyNeg	3747	2726
OnlyRuleNeg	2313	1733

6. Evaluation and analysis of the results

6.1. Statistical analysis by means of the lexicon-based method

In order to justify the application of the method which includes the processed rules of negation, the quality of classification of the tweets by sentiment was analyzed by applying LBM. The polarity of the tweet was determined according to the following formula (1):

$$Tweet\ sentiment = \begin{cases} positive\ if\ sumPos > sumNeg \\ neutral\ if\ sumPos = sumNeg \\ negative\ if\ sumPos < sumNeg \end{cases} \quad (1)$$

This method was applied in three ways, depending on the selected method for processing the negation (Method0, Method1 and Method2). The first method (LBM0) classifies words from the sentiment lexicon only based on the positive and negative sentiment. The second method (LBM1) classifies tweets by including the negation of only the first word following the negation signal. The third method (LBM2) includes everything included in the second and first, in addition to the rules which we included for detecting and processing negation.

In LBM1 and LBM2, the number of positive and negative words is corrected by the number of words that changed polarity by processing the negation. The analysis was performed for all three methods and in three cases: for the whole set (ALL), for the set within which negation occurs at least once (OnlyNeg), and for the set with a negation which is included by the rules which we processed (OnlyRuleNeg). An analysis was given in the case of 3-class. The absolute improvements and relative improvements of

LBM1 and LBM2 were calculated in relation to LBM0 (baseline) and are calculated according to formula (1) and (2), respectively.

$$AbsChange\ in\ accuracy = \ compared\ method - baseline \tag{2}$$

$$RelChange\ in\ accuracy = \frac{compared\ method - baseline}{baseline} * 100 \tag{3}$$

In the sequel, the improvement will be considered as the relative improvement. The results are shown in Table 3.

Table 3. The result of applying SA on different datasets using lexicon-based methods

		LBM0	LBM1	LBM2
ALL	Accuracy	48.57%	51.07%	53.73%
	Improv.		2.50%	5.16%
	Rel. improv.		5.15%	10.62%
OnlyNeg	Accuracy	39.66%	44.76%	50.23%
	Improv.		5.09%	10.56%
	Rel. improv.		12.86%	26.63%
OnlyRuleNeg	Accuracy	38.86%	46.89%	50.97%
	Improv.		8.03%	12.11%
	Rel. improv.		20.66%	31.16%

Table 4.The statistical analysis

Data set	Analyzed method	Compared method	P(3 classes)
ALL	LBM2	LBM0	<0.0001
ALL	LBM2	LBM1	<0.0001
OnlyNeg	LBM2	LBM0	<0.0001
OnlyNeg	LBM2	LBM1	<0.0001
OnlyRuleNeg	LBM2	LBM0	<0.0001
OnlyRuleNeg	LBM2	LBM1	<0.0001

Table 3 indicates a significant improvement following the application of the method which includes our rules of negation. These results are encouraging and justify the application of the MLM for training and predicting sentiment, which will be shown in subsection 6.4. What is of great significance is the large improvement in the method LBM2 for the analysis of the set which consists only of negation, and even more importantly, when the set analyzed is the one which contains only negations included by the rules we have processed. Greater accuracy was expected and achieved for the entire set (ALL) than it was for the set which consists only of negations (OnlyNeg) and the set

which consists only of the processed negations (OnlyRuleNeg). This is an indicator of the bad influence of the presence of negation on the prediction of sentiment. However, the LBM1 method (the baseline method for processing negation), and especially our proposed LBM2 method which includes the processed negation, provide a greater improvement for the set OnlyNeg and OnlyRuleNeg.

6.2. Statistical justification of the results

In order to verify the statistical significance of the obtained improvement in the classification, we will test the hypothesis that the method of text classification based on sentiment is significantly better at classifying text if the LBM2 rules of negation are used compared to the methods LBM0 and LBM1. In order to prove it, we will apply the MC Nemar test. When applying this method, it is necessary to construct a 2x2 matrix which consists of the ratio of accurately and inaccurately classified tweets by implementing two different methods. In all these cases, by comparing the results obtained by the method LBM2 and the methods LBM1 and LBM0, a statistically significant improvement was determined (Table 4).

6.3. Testing the influence of various means of attribute selection on the accuracy of predicting sentiment

The previous chapter contains a lexicon-based classification of tweets by means of three methods, and it was proven that the LBM2 method creates a significant improvement. The next step is for the same set of attributes used in the lexicon-based methods to be used in the application of the ML method. Since ML methods do not work well with a small number of attributes, especially numerical attributes, the baseline set of attributes should be expanded. To the set of attributes listed in subsection 4.5, we added more attributes. Additional attributes (word attributes), were obtained by a transformation of the normalized text into a vector representation of words. Text to word vector transformation was performed in Weka software using the StringToWordVector method for transformation to word attributes and the AttributeSelection method for selecting attributes.

In order to test the influence of the word attributes, we used three methods of ML: Naïve Bayes, Logistic regression and SVM. ML methods were applied only at word attributes on a 3-class dataset. The reason behind the selection of these three methods is that they mutually differ quite significantly and have proven to provide good results in the classification of texts and the determination of sentiment. For each case of text transformation into vector representation of words, we also performed a reduction of attributes by the application of the technique of information gain. Information entropy (gain) takes values ranging from 0 to 1 depending on how much information the attribute brings. All of the attributes which have a value greater than 0 were used, that is, all of the ones which carry any kind of information will be found on the list of attributes. The results of the influence of various means of transforming text into vector representation of words and the selection (reduction) of attributes from the text on the accuracy of the prediction of sentiment are shown in the Table 5. A 5-fold cross

validation was carried out by applying the Naïve Bayes, Logistic and SVM methods. A 5-fold cross-validation was performed because a large number of attributes were used, and a 10-fold would be time and memory limited.

From Table 5 we can conclude that the attributes used for training which make the greatest contribution are unigrams (U), especially in the case when only the presence or absence in the tweet is being detected (the greatest improvement to the NB-row 1 in Table 5). If the vector representation of words which contains the presence or absence of a word is replaced by the number of occurrences of a word, the accuracy of the prediction decreases, which can be explained by the fact that the text of the tweet is rather short and that the words rarely repeat several times, both in the same tweet and in the entire group. In addition, the normalization of the IDF brings no improvement – quite the contrary; the reason is the same as the one for the application of the TF normalization.

Table 5. Accuracy of the ML method in the determination of sentiment by transforming text into vector representations of words

	Feature type	Filter applied	# of features	NB	Log	SVM
1	U	word presence	408	57.54%	64.67%	63.26%
2	U	word count	404	56.36%	64.00%	62.83%
3	U	TF	404	54.54%	64.21%	63.01%
4	U	TF/IDF	404	54.54%	64.21%	63.00%
5	U	TF+ Normalized tweet length	265	39.26%	62.31%	60.92%
6	U	TF/IDF+ Normalized tweet length	255	28.77%	62.04%	60.82%
7	U+B+T	word presence	1023	57.40%	64.87%	63.88%
8	U+B+T	word count	1015	55.94%	64.52%	63.61%
9	U+B	word presence	756	57.49%	65.30%	64.31%
10	U+B	word count	750	56.21%	65.16%	63.97%

The normalization of the length of the tweet has a negative effect on the prediction of all three methods. The negative influence of normalization of the length of the tweet is especially clear in the application of the Naïve Bayes method, as a result of the assumption on the independence of the variables, which contributes to poor results. Unigrams (U) combined with bigrams (B) and trigrams (T) offer less improvement compared to unigrams combined with bigrams. Unigrams combined with bigrams (highlighted in Table 5, row 9) offer the second highest improvement (the greatest improvement for Log and SVM) and will also be tested during the evaluation, along with unigrams for Naïve Bayes (highlighted in Table 5. row 1).

The results presented in Table 5 do not have great values of accuracy, but the aim of this analysis was to determine which selection of attributes from the text is the most suitable as an addition to the basic attributes which we have described above.

6.4. Machine learning methods

In order to evaluate the proposed method for handling negation, the following machine learning methods were used: Naïve Bayes, Logistic Regression, SVM, and J48-decision tree method. The attributes used to train contain all the attributes from the lexicon-based methods (LBM0, LBM1, and LBM2), and additional textual attributes selected in the manner presented in sub-section 6.3. Additional attributes would certainly bring better results for each method, but our aim is not to present the best system of classification of tweets by means of a SA, but to determine whether the attributes obtained by the rules of detecting negation influence the improvement in the classification of tweets in the Serbian language. Table 6 shows the accuracy of the proposed method for NB, LOG, J48 and SVM for:

- the entire set: with unigrams (ALL U); with unigrams and bigrams (ALL U+B),
- the set which contains only negations: with unigrams (OnlyNeg-U); with unigrams and bigrams (OnlyNeg-U+B),
- the set which contains only negations included by the rules: with unigrams (OnlyRuleNeg-U); with unigrams and bigrams (OnlyRuleNeg-U+B)
- all this on the 3-class set.

Table 6. Accuracy of the proposed method for the various groups and the different ML methods, for 3-class

MLM2	ALL U	ALL U+B	OnlyNeg U	OnlyNeg U+B	OnlyRule Neg U	OnlyRule Neg U+B
NB	57.82%	57.80%	62.53%	61.97%	62.17%	62.77%
LOG	68.46%	68.84%	69.25%	69.76%	68.96%	69.69%
J48	61.88%	61.92%	61.57%	61.65%	61.95%	62.04%
SVM	67.45%	65.98%	65.86%	67.65%	67.83%	68.61%

Table 7 presents the same information as in Table 6, but for 2-class.

Table 7. Accuracy of the proposed method for various groups and different ML methods, for 2-class

MLM2	ALL U	ALL U+B	OnlyNeg U	OnlyNeg U+B	OnlyRule Neg U	OnlyRule Neg U+B
NB	86.49%	86.47%	84.63%	85.03%	85.46%	85.98%
LOG	91.15%	91.13%	90.46%	90.53%	90.82%	90.45%
J48	86.88%	86.80%	85.29%	84.92%	84.30%	83.32%
SVM	89.56%	89.36%	88.74%	88.92%	88.29%	88.75%

The results we have achieved by applying all three aforementioned methods (two baseline: MLM0 and MLM1 and the proposed MLM2) for the 3-class and 2-class groups are given in the following tables.

Table 8 shows the results of the application of the methods for the entire set, for all three classes. The table also indicates the accuracy for the “Only words” (instances when only the textual attributes of the vector representation of words are used). The

improvements were calculated in relation to the baseline MLM0 method. From Table 8 it can be concluded that the proposed method which includes the rules of negation we have processed (MLM2) provides better results than all the three previous cases. In order to better present the effects of the application of the processed rules, in Table 9 we present the results for the set of tweets which obligatorily contain at least one negation.

Table 8. The results and improvement for the entire group, for 3-class

	ALL-U Accuracy	Improvement	ALL-U+B Accuracy	Improvement
Only words	64.6660%		65.3053%	
MLM0	67.4843%		68.3586%	
MLM1	68.2281%	1.1022%	68.7500%	0.5726%
MLM2	68.4629%	1.4501%	68.8413%	0.7061%

Table 9. The results and improvements for the set OnlyNeg, for 3-class

	OnlyNeg-U Accuracy	Improvement	OnlyNeg-U+B Accuracy	Improvement
Only words	67.0224%		68.7033%	
MLM0	67.4673%		68.6683%	
MLM1	68.5615%	1.6218%	69.7892%	1.6323%
MLM2	69.2554%	2.6503%	69.7625%	1.5935%

Table 10. The results and improvements for the set OnlyRuleNeg, for 3-class

	OnlyRuleNeg-U Accuracy	Improvement	OnlyRuleNeg-U+B Accuracy	Improvement
Only words	67.7182%		68.6690%	
MLM0	68.4825%		69.4336%	
MLM1	68.2663%	-0.3157%	68.9148%	-0.7472%
MLM2	68.9581%	0.6945%	69.6930%	0.3736%

Table 10 shows the accuracy of the prediction of sentiment of only those tweets which contain a negation, or to be precise, the type of negation which was processed by the proposed rules (OnlyRuleNeg). The negative improvement achieved with the MLM1 method indicates that the simple change in the polarity of the word following the negation signal in the case of our dataset leads to worse results than does the exclusion of this rule. The reason for this is that the set OnlyRuleNeg contains tweets which are the most difficult to process. The proposed method, MLM2, offers an improvement even in the case when we use only unigrams, and in the case when both unigrams and bigrams are used together. The results which were achieved for the set 2-class, for all the tweets, are given in Table 11.

Table 11. The results and improvements for the entire set, for 2-class

	ALL-U Accuracy	Improvement	ALL-U+B Accuracy	Improvement
Only words	90.1474%		90.8362%	
MLM0	90.4134%		90.9179%	
MLM1	91.0926%	0.7512%	91.1120%	0.2135%
MLM2	91.1508%	0.8156%	91.1314%	0.2348%

We can see that the accuracy achieved for the set 2-class is a lot greater compared to the one for 3-class. This indicates the problematic nature of neutral tweets.

Table 12. The results and improvements for the set OnlyNeg, for 2-class

	OnlyNeg-U Accuracy	Improvement	OnlyNeg-U+B Accuracy	Improvement
Only words	89.2268%		90.3994%	
MLM0	88.9949%		90.2421%	
MLM1	89.3984%	0.4534%	90.4622%	0.2439%
MLM2	90.4622%	1.6487%	90.5356%	0.3252%

For the set which includes only negation (OnlyNeg), in Table 12 we find a lower percentage of accuracy for all methods. The reason for this is that all the tweets which contain negation are more difficult to classify. However, the improvement which was achieved by using the method MLM2 is greater (1.648%) compared to the improvement for the same method achieved for the entire set (0.8156%).

Table 13. The results and improvements for the set OnlyRuleNeg, for 2-class

	OnlyRuleNeg-U Accuracy	Improvement	OnlyRuleNeg-U+B Accuracy	Improvement
Only words	88.9273%		89.5040%	
MLM0	89.6711%		89.4403%	
MLM1	90.3058%	0.7078%	89.5557%	0.1290%
MLM2	90.8252%	1.2870%	90.4512%	1.1303%

Table 13 shows improvements for the set of tweets that contain the rules of negation we have processed (OnlyRuleNeg). The set of neutral tweets significantly disrupts the quality of the classification of sentiments both for the entire set and for the set with negations. This can be seen by comparing the results of the classifications for the sets 3-class and 2-class. Neutral tweets contain words with sentiment and they determine the sentiment of that tweet. The incorrect impact of sentiment words in neutral tweets can be solved by the introduction of the degree of polarity, which might significantly influence the quality of the classification of the neutral tweets. The accuracy obtained for the set 2-class is satisfactory.

The lexicon-based methods (LBM0, LBM1, LBM2) offer lower accuracy than the ML method since in addition to the attributes which contain the number of occurrences of positive and negative terms, only the attributes which are a direct consequence of the

processing of rules of negation were used. However, an improvement in Method1 and Method2 compared to Method0 are expressed better by a comparison of the lexicon-based methods (LBM0, LBM1, LBM2), precisely for the aforementioned reasons.

Using 10-fold cross validation, no qualitative data is obtained because there are no classification data for the individual element from the test set. Therefore, we cannot use the a MC Nemar test for a comparison of the methods, but in this case, we will use T-test. By applying the T-test we analyzed whether the results were obtained by various methods (Only words, MLM0, MLM1, MLM2) for all the sets (All, OnlyNeg and OnlyRuleNeg) in the case when we analyzed 3-class and in the case when we analyzed 2-class. (Table 14)

Table 14. The results of the T-test

Analyzed method	Compared method	P (3-class)	P (2-class)
Only words	MLM0	0.028954	0.250274
Only words	MLM1	0.015784	0.042767
Only words	MLM2	0.004331	0.008634
MLM0	MLM1	0.089519	0.006044
MLM0	MLM2	0.007086	0.004997
MLM1	MLM2	0.017721	0.033535

In all of the studied cases, except for the comparison of the MLM0 and MLM1 for 3-class and Only words and MLM0 for 2-class, a statistically significant improvement was proven. In all the other cases we obtained significance at the $p<0.05$ level. What is of special importance is that in the case of the comparison of the proposed MLM2 method with Only words and MLM0 methods (in the case of 3-class and in the case of 2-class) we obtain significance at the $p<0.01$ level, based on which we can confirm that the proposed MLM2 method can be used to achieve a statistically significant improvement compared to the other two methods. The improvement achieved by applying the MLM2 method is statistically significant compared to the MLM1 method ($p<0.05$).

7. Conclusion and future work

In this paper we have provided an overview of the group of rules for processing negation which occurs in tweets. It has been shown that tweets containing negation are more complicated for SA and that the application of the proposed rules contributes to improvement in determining sentiment. For a more detailed processing of the rules of negation, it is necessary to ensure the existence of an corpus with annotated negation scopes. The existence of this type of resource could enable an additional analysis of the phenomenon of negation and the determination of additional rules which might help the process of negation to be included more successfully into the system of determining the sentiment of the short texts it is contained in. The application of a morphological dictionary might greatly improve the results of the SA, but the aim of this paper was to prove that with the application of the rules of negation there is an improvement in the prediction of sentiment.

The next planned step is the creation of a special corpus which consists of annotated scopes of negation (the scope of the signal of negation in the part of the text). Furthermore, we plan to use the existing morphological dictionary and expand the dictionary of synonyms from the Serbian WordNet [33]. In addition, we also plan to analyze the influence of intensifiers on general sentiment. We also plan to create a special lexicon that will be based on the corpus and which will contain the coefficient of terms that occur in the presence of negation. Using these additional resources, we expect to obtain a significant improvement in training the negation itself, as well as an improvement in the general system for detecting the polarity of short texts.

Acknowledgment. This work was partially supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, Project III-44007. The authors are grateful to Professor Milena Stanković, PhD, for her support in writing this article.

References

1. Hovy, E.H.: What are Sentiment, Affect, and Emotion? Applying the Methodology of Michael Zock to Sentiment Analysis. In: Gala N., Rapp R., Bel-Enguix G. (eds) Language Production, Cognition, and the Lexicon. Text, Speech and Language Technology, Springer, Cham, Vol. 48, 13-24. (2015)
2. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr., Vol. 2, No. 1-2, 1-135. (2008)
3. Liu, B., Zhang, L.: A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. (2012)
4. Pang, B., Lee, L., Rd, H., Jose, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Vol. 10, 79–86. (2002)
5. Das R., S., Chen Y., M.: Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web. Management Science, Vol. 53, No. 9, 1375–1388. JSTOR. (2007)
6. Polanyi, L., Zaenen, A.: Contextual Valence Shifters. In: Shanahan J.G., Qu Y., Wiebe J. (eds) Computing Attitude and Affect in Text: Theory and Applications. The Information Retrieval Series, Vol 20. Springer, Dordrecht. (2006)
7. Kennedy, A., Inkpen, D.: Sentiment classification of movie and product reviews using contextual valence shifters. In Proceedings of the Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations, Ottawa, Ontario, Canada. (2005)
8. Hu, M., Liu, B.: Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). ACM, New York, NY, USA, 168-177. (2004)
9. Benamara, F., Cesarano, C., Picariello, A., Recupero, D.R., Subrahmanian, V.S.: Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. ICWSM. (2007)
10. Jimenez-Zafra, S. M., Martin Valdivia, M. T., Martinez Camara, E., Urena-Lopez, L. A.: Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter. IEEE Transactions on Affective Computing, IEEE, Vol. PP, Issue: 99. (2017)
11. Zhu, X., Guo, H., Mohammad, S., Kiritchenko, S.: An Empirical Study on the Effect of Negation Words on Sentiment. In ACL (1), 304-313. (2014).
12. Kiritchenko, S., Mohammad, S., Zhu, X.: Sentiment Analysis of Short Informal Texts. J. Artif. Intell. Res., Vol. 50, 723-762. (2014)
13. Wiegand, M., Balahur, A., Roth, B., Klakow, D., Montoyo, A.: A survey on the role of negation in sentiment analysis. In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (NeSp-NLP '10), Roser Morante and Caroline Sporleder (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 60-68. (2010)

14. Sharif, W., Samsudin, N. A., Deris, M. M., Naseem, R.: Effect of negation in sentiment analysis. 2016 Sixth International Conference on Innovative Computing Technology (INTECH), 718-723. IEEE. (2016)
15. Pröllochs, N., Feuerriegel, S., Neumann, D.: Negation scope detection in sentiment analysis, Decis. Support Syst., Vol. 88, No. C, 67-75. (2016)
16. Reitan, J., Faret, J., Gambäck, B., Bungum, L.: Negation Scope Detection for Twitter Sentiment Analysis. Association for Computational Linguistics. (2015)
17. Cerezo-Costas, H., Celix-Salgado, D.: Gradiant-analytics: training polarity shifters with CRFs for message level polarity detection. In Proceedings of the 9th international workshop on semantic evaluation (SemEval-2015), 539–544. Denver, Colorado: Association for Computational Linguistics. (2015, June).
18. Asmi, A., Ishaya, T.: Negation identification and calculation in sentiment analysis. In The Second International Conference on Advances in Information Mining and Management, 1-7. (2012)
19. Giachanou, A., Crestani, F.: Like it or not: A survey of Twitter sentiment analysis methods, ACM Computing Surveys (CSUR), Vol. 49, No. 2, 28:1–28:41. (2016)
20. Batanović, V., Nikolić, B.: Sentiment classification of documents in Serbian: The effects of morphological normalization. In Proceedings of the 24th Telecommunications Forum (TELFOR), Belgrade, Serbia, 1-4. (2016)
21. Batanović, V., Nikolić, B., Milosavljević, M.: Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 2688-2696. (2016)
22. Rotim, L., Šnajder, J.: Comparison of Short-Text Sentiment Analysis Methods for Croatian. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics, 69-75. (2017)
23. Mladenović, M., Mitrović, J., Krstev, C., Vitas, D.: Hybrid Sentiment Analysis Framework for A Morphologically Rich Language. Journal of Intelligent Information Systems, Vol. 46:3, 599–620. (2016)
24. Marovac, U., Pljaskovic, A., Crnisanin, A., Kajan, E.: N-gram analysis of text documents in Serbian language, In Proceedings of the 20th Telecommunications Forum (TELFOR), Belgrade, Serbia, 1385-1388. (2012)
25. Jimenez-Zafra, S. M., Taule, M., Martin Valdivia, M. T., Urena-Lopez, L. A., Marti M., A.: SFU ReviewSP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. Language Resources and Evaluation, Springer, Vol. 52, No. 2, 533-569. (2018)
26. Milosevic, N.: Stemmer for Serbian language. arXiv preprint arXiv:1209.4471. (2012)
27. Stone, P. J., Dunphy, D. C., Smith, M. S.: The General Inquirer: A Computer Approach to Content Analysis. M.I.T. press. (1966)
28. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., Booth, R. J.: The Development and Psychometric Properties of LIWC2007. The University of Texas at Austin, 1–22. (2007)
29. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 347–354. (2005)
30. Hu, M., Liu, B.: Mining and Summarizing CustomerReviews. In Proceedings of the Tenth SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '04. New York, NY, USA: ACM, 168–177. (2004)
31. Baccianella, S., Esuli, A. Sebastiani, F.: SENTI WORD NET 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation, Vol. 0, 2200–2204. (2008)
32. Kovačević, M.: Sintaksička negacija u srpskome jeziku. Izdavačka jedinica Univerziteta u Nišu. (2002)
33. Krstev, C., Pavlović-Lažetić, G., Vitas, D., Obradović, I.: Using Textual and Lexical Resources in Developing Serbian Wordnet. Romanian Journal of Information Science and Technology, Vol.7, No. 1-2, 147—161. (2004)

34. Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Y. A. Hawalah, A., Gelbukh, A., Zhou, Q.: Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. Cognitive Computation, Springer, Vol. 8, 757-771. (2016)

Adela Ljajić is a teaching assistant at the State University of Novi Pazar (Serbia). She received an M.S. degree in Information Systems and Technologies from the University of Belgrade (Serbia). She is currently a Ph.D. student in Electrical Engineering and Computer Science at the University of Niš (Serbia). Her current research interests are related to Natural Language Processing, Text Mining, Machine Learning and Sentiment Analysis with a special focus on treatment of negation. She has authored/co-authored a significant number of publications in international conferences in the field of NLP and Sentiment Analysis.

Ulfeta Marovac is a teaching assistant with PhD at the State University of Novi Pazar, Serbia. She holds a PhD in Computer Science from the University of Belgrade, Serbia. Her thesis "Mining sequential patterns for determination of protein characteristics" proposed the model for classification of proteins in COG categories based on amino acid n-grams (sequences of n length). Her research interests focus on, biological sequences analysis, text mining, natural language processing, logic and its applications in computer science. She is authored and coauthored research publications in Journal of Multiple-Valued Logic and Soft Computing, Match Communications in Mathematical And In Computer Chemistry, Publications de'lInstitut Mathematique.

Received: January 22, 2018; Accepted: June 26, 2018

Computer Science and Information Systems 16(1):313–332 https://doi.org/10.2298/CSIS180328016B

Rejecting the Death of Passwords: Advice for the Future

Leon Bošnjak[1] and Boštjan Brumen[1]

[1] Faculty of Electrical Engineering and Computer Science, University of Maribor,
Smetanova ulica 17, 2000 Maribor, Slovenia
{leon.bosnjak, bostjan.brumen}@um.si

Abstract. Passwords have been a recurring subject of research ever since Morris and Thompson first pointed out their disadvantages in 1979. Several decades later, textual passwords remain to be the most used authentication method, despite the growing number of security breaches. In this article, we highlight technological advances that have the potential to ease brute-force attacks on longer passwords. We point out users' persistently bad password creation and management practices, arguing that the users will be unable to keep up with the increasingly demanding security requirements in the future. We examine a set of real, user-generated passwords, and compare them to the passwords collected by Morris and Thompson. The results show that today's passwords remain as weak as they were nearly four decades ago. We provide insight on how the current password security could be improved by giving recommendations to users, administrators, and researchers. We dispute the reiterated claim that passwords should be replaced, by exposing the alternatives' weaknesses. Finally, we argue passwords will remain widespread until two conditions are met: First, a Pareto-improving authentication method is discovered, and second, the users are motivated to replace textual passwords.

Keywords: authentication, password security, comparison.

1. Introduction

Technology has seen tremendous growth and advancement since Morris and Thompson highlighted password security vulnerabilities back in 1979 [1]. Since the mid-1970s, global computer sales have increased from 50,000 to a record of 355 million units per year in 2011. With a rapid growth in the smartphone market in the last decade, cellular phone sales quickly overtook computer sales, reaching over 1.4 billion smartphones sold in 2015 alone. This exponential advancement has been influenced greatly by the Internet, the user base of which has grown to nearly half of the world's entire population. It became the predisposition for the data explosion, with global Internet traffic exceeding one zettabyte of data per year by the end of 2016. It is difficult to estimate the amount of sensitive data stored and processed online, but frequent data breaches and the emerging challenges we face to protect this data remind us of the growing importance of security.

To this day, textual passwords remain one of the most common authentication methods. Their continuous dominance can be attributed to their diverse advantages, such

as low cost, simplicity of implementation, and convenience. In computing, they have been used since the earliest days, with MIT's time-sharing system CTSS being a prime example from the early 1960s. Several possible threats have been identified in the following decades, often through accidentally discovered lapses in system security, such as poorly implemented access control lists or vulnerabilities in password encryption algorithms. While some of these problems were dealt with throughout the years, others persisted (many systems keep their passwords in a plaintext format), or even escalated (constantly increasing computer processing power is decreasing the amount of time necessary to perform a successful key search).

Ultimately, the level of security provided depends largely on the underlying security scheme (or lack thereof) implemented by the system in question. Unfortunately, systems often do not enforce sufficient levels of security. In the past, these decisions were based on the knowledge and awareness of security officers and administrators, allowing room for human error. To mitigate emerging threats, security experts have been proposing stricter password policies gradually over the years. The extent to which these recommendations are implemented, however, still depends on system designers: A conscious attempt to shift the balance between security and memorability is often made in favor of user convenience.

Thus, a considerable part in the endeavor to protect their own personal data still lays on the shoulders of the users. Morris and Thompson were among the first authors to expose users' bad password creation habits. In response, experts in the growing field of Password Security began to raise awareness among the general population. Despite being educated about password security risks and requirements, however, users remained slow and reluctant to adapt these good practices. In the end, nothing prevents the user from selecting a password that barely meets the minimal requirements enforced by the system.

Parallel to textual passwords' multi-dimensional problems, countless novel authentication methods are being introduced in research papers. They approach the password dilemma with an argument that the textual passwords have reached the end of their useful life, and must be replaced. They expose the passwords' inability to withstand offline cracking, and the users' tendency to create and manage bad passwords, while highlighting their method's advantages. As a result, stakeholders are advocating for alternatives, despite limited empirical evidence to support their superiority in real-world applications.

The problem with this approach is that all of the above points have already been well-documented, and at least some version of the argument has been advanced over and over for the last few decades. Nonetheless, textual passwords have withstood all of these claims and do not appear in danger of being replaced any time soon. The challenges thus lie in ensuring a sufficient level of security to mitigate privacy-related incidents, while continuing the search for Pareto optimal authentication solutions.

2. Related Work

User-chosen passwords have been evaluated empirically many times over the course of four decades. In their seminal paper, Morris and Thompson were the first to expose bad

password creation habits [1]. Their work was important, not only for providing insight into the extent of the password security problem as it appeared in 1979, but primarily for asserting that the situation will get worse if no improvements are made in the future.

The next twenty years saw very few empirical studies of user-generated passwords, until the steady growth of data breaches over the years prompted Zviran and Haga to conduct a study on password characteristics and usage patterns [2]. They found out that passwords had not improved at all (almost 50% of users had passwords composed of 5 characters or less, and an alarming 80% used only alphabetic characters), proclaiming the lack of user education and decreased concern with information security as the primary reasons. They provided several directions for future research that were explored in the following years.

In 2007, Florencio and Herley conducted the first large-scale password study, analyzing over 500,000 user-generated passwords [3]. They found that, albeit the minimum password length had increased to 6 characters on average, the overwhelming majority of users still created passwords that contained only lowercase characters. Three years later, Dell'Amico et al. analyzed three sets of real-world passwords from the standpoint of several password guessing techniques, including dictionary attacks, mangling rules, probabilistic context-free grammars, and Markov models [4]. Aside from finding out that no strategy prevails over the others, they once again pointed out the users' weak passwords, stating that password policies are not guaranteed to prevent the users from making them. In a 2012 study, Bonneau developed partial guessing metrics, and analyzed a corpus of 70 million passwords [5]. They showed that, against an optimal attacker performing an unrestricted brute force attack, textual passwords provide security roughly equivalent to only 20-bit random strings. Additionally, different population appears to have a minimal effect on the distribution of passwords. On the contrary, a 2016 study by Shen et al. suggests the average password length had increased to 9.46, and that the users tended to select random characters more often [6]. They also reported that the users preferred alphanumeric over alphabetic characters in their passwords, and were more likely to select combo-meaningful data rather than single-meaningful data as their passwords. In one of the most recent big-scale studies of over 145 million passwords, however, Ji et al. reported shorter passwords (composed of less than 8 characters on average) following predictable patterns (e.g. L, D, LD, and LDL) [7]. They managed to crack up to 65% of passwords in some of the datasets, and showed that several password meters classify passwords as stronger than they really are against modern password cracking algorithms.

These inconsistent findings on the (lack of) improvement of password security over the last decades have been the cause of many disagreements in regard to whether textual passwords should be replaced or not. This question is discussed extensively by Bonneau et al. [8], who offer a multi-dimensional perspective on the password-related dilemma, including the problems with password research as a whole. One of our main contributions is expanding on the work of Bonneau et al. by providing a historical overview from the standpoint of technological advances in both password creation and brute-force cracking. We support our findings by analyzing a sample of real-world, student passwords, and offer a look into the future of authentication.

Two previous studies have observed password selection habits and password strength of university students [9][10]. However, neither had made an attempt to compare the then-current passwords with older passwords examined in previous studies. In general,

comparison of past and present password security has been relatively scarce in literature, despite the oversaturation of the password security field. Feldmeier and Karn conducted a follow-up study ten years after the seminal paper by Morris and Thompson, focusing on the improvements in hardware and software [11]. Shen et al. compared the considered passwords with the passwords from several previous studies, though they did not include the study by Morris and Thompson [6]. More importantly, they did not evaluate the past and present passwords' resilience against different types of attacks, while taking user education and hardware improvements into account. Our work aims to close that gap.

3. Textual Passwords Through Time

The history of data breaches is as old as the idea of storing and protecting the digital data itself. Initially, incidents began to occur in companies that kept sensitive, business-oriented data in shared systems – one of the first reported dates back to the mid-1960s, when the CTSS operating system exposed all users' passwords as a daily welcome message [12]. In November 1978, the largest bank theft in U.S. history at the time occurred due to bad password practices [13].

Security officials and administrators of these systems were the first ones to warn about password security vulnerabilities. They suggested stronger and computationally slower encryption algorithms, salted passwords, and encouraged the users to create more complex passwords [1]. Nonetheless, reports of breaches from the early days of computing come mostly in the form of anecdotal knowledge – public awareness of password security threats did not begin to rise until the early 1990s, when the development and expansion of the Internet began to lead to the data explosion and reliance on Information Technology we are witnessing today.

While records of data breaches exist from that time [14], the majority of the largest data breaches have occurred and, consequently, been reported, in the last decade. One of the first most notable leaks in the recent years happened in 2009, when hackers managed to obtain more than 32 million plaintext passwords from the gaming website RockYou. In 2011 alone, over 70 million unencrypted passwords were leaked from various Chinese websites. It should be noted that such publicly known data breaches affect only a small margin of password-protected systems online, meaning that a much larger number of stored unencrypted passwords likely still exists in the wild.

Worryingly, even hashing does not warrant full safety of the stored passwords. Upon gaining access to a database of hashed passwords, an adversary only needs to determine the hash function used before she can attempt an offline brute-force attack. With continuous advances in processing speeds, it is becoming increasingly easier to find matching hashes. One of the persistent problems is that many systems continue to use deprecated, fast hashing algorithms. For example, around 2 million Ubuntu forum users' e-mails and MD5-hashed passwords were compromised in a 2013 hack. In 2016, the business-oriented social network service LinkedIn released a statement in which they acknowledged that a large set of hashed passwords was being sold on the black market. It turned out that at least 117 million passwords must have been leaked in 2012, of which 85% of SHA-1 hashes had already been cracked. Despite SHA-1 being

announced deprecated in 2011, industry remains too slow to move to safer alternatives. Variably iterative and memory-hard hashing algorithms were also proposed to render a brute-force attack too time-consuming to attempt. Nonetheless, even this solution knows examples of incidents. In 2016, over 43 million passwords were stolen from the web design platform Weebly, despite having been protected by bcrypt.

Security breaches have affected even large and well-known companies such as Adobe, Dropbox, eBay, MySpace, Tumblr, and Yahoo, to name a few.[1] The last is also an example of the extent of damage such incidents can cause: Yahoo reported that an overwhelming 1.5 billion user account` credentials had been stolen in hacking incidents that occurred in 2014 and 2015. The main reason for the sudden increase in the frequency and severity of these attacks stems mainly from the exponential growth of data processed online, which offers the potential attackers a realistic opportunity to expose large amounts of sensitive data by targeting a single online service. Additionally, developing technologies, such as IoT, pose many yet unresolved challenges that can be exploited [16].

It is safe to assume, and even expect, that this upward trend will continue in the future. Even more concerning is the prospect of such attacks becoming increasingly successful with the advances in technology. Hardware improvements have mostly been in accordance with Moore's Law ever since its establishment more than 50 years ago. While continuous exponential growth of processing power has already begun to slow down due to physical limitations, most semiconductor forecasters predict that Moore's Law will remain relevant for at least another ten years. By that point, factors other than transistor size might prolong the longevity of the projection: Beyond Intel's tri-gate transistors, we could be expecting monolithic 3D chips in the next decade, which would allow multiple layers of components built up vertically on a single silicon die. Alternative materials, such as indium or arsenide alloys, as well as different forms of carbon such as graphene or nanotubes, are promising as well, due to their highly conductive properties and lower power consumption.

Farther down the line, we might be looking at the more radical successors of the silicon era. One possibility is spintronic semiconductors, which exploit electrons' rotational energy rather than their charge to represent bits. While this technology has been in development for more than 15 years, there are still challenges that need to be overcome before it can reach the production stage [17].

Quantum computing is another revolutionary approach that has the potential to mark one of the most impactful leaps in computer history since the establishment of the Von Neumann architecture. Due to the nature of their operation, quantum computers excel at particular types of mathematical tasks. Prime factorization and discrete logarithm problems are two types of hard mathematical problems, used widely in cryptography, that could be solved in polynomial time using Shor's algorithm [18]. As fundamental questions such as what to build qubits out of and how to maintain quantum superposition are answered, we might be looking at drastic increases in processing power, equivalent to several orders of magnitude, which would make password breaking a trivial task.

Despite the aforementioned and other possible future improvements in hardware that are still in development, there are other ways for potential adversaries to speed up

[1] InformationIsBeautiful collects data about recent data breaches based on media reports, illustrating the growing trend through meaningful visualization [15].

password cracking times. The benefit of parallel processing over sequential searches allows modern GPUs to offer more than 10-fold increases in processing speeds when compared to general processors, making them a viable choice for executing brute-force attacks; increasing the number of nodes in the cluster provides linear scalability with very little to no overhead. Dedicated FPGA cores can also be employed to improve password guessing speeds when specialized password-hashing functions (i.e. bcrypt or scrypt) are used to slow down the computation time intentionally [19]. Recently, more advanced techniques that attempt to optimize the password guessing order have begun to emerge. Weir et al. suggested the use of probabilistic context-free grammars, which can be derived from training sets of previously revealed passwords. These grammars provide a basis to generate word-mangling rules, which determine the guessing order of individual passwords, based on the estimated probability of their occurrence [20]. Exploiting the distribution of letters within the passwords, Markov chain models work on the assumption that most users choose easy-to-remember passwords that follow the rules of their native language. Unlikely passwords are, thus, removed to reduce the initial search space, while the remaining passwords can be ordered by the probability of a given password's characters appearing in a specific order. This approach provides a substantial improvement over the classic, straightforward dictionary attack [21].

The growing frequency and effectiveness of security breaches cannot be attributed only to computing advances and improvements in cracking techniques, however. Passwords can also be retrieved by exploiting system vulnerabilities, such as the COUPLE attack [22]. According to [15], some of the largest leaks happened because passwords were lost, stolen, or published accidentally. Furthermore, numerous studies have observed users' password creation and management practices, and have concluded that humans are the weakest link in password security [23][24][25]. Their password selection methods are influenced by the characteristics of the human memory. First, cognitive burden increases with the number of elements that a user is required to memorize. Second, humans are decisively better at remembering meaningful strings they can recall based on associations rather than random sequences. Due to these limitations, humans generally gravitate toward generating short, simple, and predictable passwords. While such passwords might be easy to remember for humans, they are, unfortunately, also easy to guess for machines: Short length and low password complexity allows them to traverse the entire available search space and 'find' the correct password in a very short time.

On the other hand, forcing the users to authenticate using a computer-generated, random password might not be the optimal solution. A study on the memorability of high-entropy passwords has shown that the longer the character string is, the more difficult it is for humans to remember and recall it later. At the same time, it takes longer for them to input the password and they are more likely to make an error [26].

One of the primary challenges in password security has, thus, been finding the balance between security and memorability. Password policies are one of the most well-known and established approaches to reach the middle ground. They were introduced with the intention of ensuring a sufficient level of security while still allowing the users to retain some originality when creating their own passwords. As expected, stricter password policies generally provide a higher level of security, at the expense of user convenience. Lee and Choong recommend simple rule sets, arguing that significant drops in usability do not outweigh the minimal increases in security provided by the

more complex rule sets [27]. Nonetheless, studies suggest that the relationship between security and perceived usability by the users is not dependent only on the strictness of password policies [28]. Well-crafted policies are supposed to offer an equal or higher degree of security, with slight, or even insignificant, decreases in usability, when compared to sub-par policies. However, there are no absolute guidelines to creating good password policies; most authors agree that optimal policies are difficult to create [27][28]. Furthermore, poorly formed policies that do not take into account users' psychological profiles, their work practices, and perceived importance of their accounts, can actually be more damaging than beneficial [29]. The inclusion of the surrounding context is recommended not only for textual passwords, but access control models as a whole [30]. In practice, policies often impose high constraints for limited benefit; Bonneau et al. suggests they might not actually reduce harm [8].

Users' lack of security knowledge is yet another problem that has been largely pointed out in scientific literature. It has been suggested that, in order to improve password composition and management practices, it is crucial to raise awareness of security-related threats and educate the users [29][31]. However, recent studies show that users know what constitutes a good password and are well aware of the consequences of bad password mismanagement [23], but are willing to trade security for comfort [31]. Consequently, they exhibit lax security behavior: They are disinterested and frustrated with password policies, and resort to bad password management practices out of laziness and convenience [27]. In order not to forget their passwords, users avoid changing them [32] or write them down [31][32]. If they are forced to change their passwords upon expiry, they tend to select similar passwords [33]. To reduce the memory load, they are also known to reuse their passwords across several services [32][34][35], or share them with their family members or friends [31][34].

The extent of these bad practices reveals that it is not enough for the users only to be aware of the importance of password security – they must be motivated to protect their personal data. Because of that, they are more likely to use stronger passwords for accounts they perceive to be of high importance (e.g. bank accounts) [25][35], though the actual strength of the chosen passwords still depends on their understanding of what constitutes a good password [32]. Furthermore, their attitude towards security-related issues is influenced by their past experiences. If they are using their accounts continuously without any security incidents affecting them, users feel less threatened, or even begin to doubt that their accounts can be breached. Studies suggest that, in order for them to become security-conscious, they need to feel a personal loss as a result of their own data being compromised [24][32].

4. Comparison of Past and Present

37 years ago, Morris and Thompson conducted a study in which they collected and analyzed 3,289 user-generated passwords. They divided them into several categories based on their length and composition and found that 2,831 (or 86.07%) of those passwords could be recovered in a matter of days. Finally, they provided several suggestions on how to improve the system in order to prevent an adversary from being able to execute a brute-force attack in a reasonable amount of time [1]. Several of their

propositions (such as improved hashing functions, salted passwords, and password policies) later became a widely-used means to combat the ever-growing computer processing power.

Over the next few decades, many studies observed the characteristics of users' passwords leaked at the time, and pointed out that they are alarmingly bad. However, to the best of our knowledge, none of them provided a direct comparison of current and past password strength while also taking into consideration advances in technology, user education, and security procedures (e.g. password policies). In the current era of exponentially growing data, and the corresponding increase in security-related incidents, it is essential to ask ourselves whether there have been any improvements in the field of Password Security.

For the purposes of this study, we evaluated password strength based on the amount of time it would take to try all possible passwords of a given length and composition. In the original study by Morris and Thompson, it was estimated that it would take a second to check about 800 encrypted passwords on a UNIBUS system PDP-11/70, containing what was, at the time, a high-performance CPU [1]. To make a valid comparison, we chose a contemporary high-end processing unit. In a recent article, Qui et al. measured performance of AMD GPUs generating MD5 and SHA-1 password hashes. After optimizing the hashing algorithm and password generation, they were capable of achieving the speed of 6,877 million and 2,615 million tries per second for MD5 and SHA-1 respectively, on a high-end desktop graphic card AMD HD7970 [36]. Additionally, we compared the cracking times of an 8 nVidia GeForce GTX 1080 Ti rig generating much slower bcrypt hashes. According to the latest benchmark results, the rig achieved up to 185,000 tries per second for bcrypt running 5 iterations (32 rounds) [37].

We calculated the running times for an AMD HD7970 unit generating SHA-1 hashed passwords, as well as a rig of eight nVidia GeForce GTX 1080 Ti units generating bcrypt hashes. Bcrypt's adjustable work factor allowed us to explore the running times for three configurations: Low (5), medium (12), and high (20) numbers of iterations. The choices of work factors are not arbitrary: Five iterations reproduce the setup used in all benchmark tests, twelve iterations were chosen because they are often used in practice and currently accepted as the security standard, while twenty iterations represent a high-end alternative for security-sensitive applications.

A direct comparison of SHA-1 running times with the running times reported by Morris and Thompson revealed that there has been a substantial decrease (of over 326 million percent) in the amount of time it takes for a state-of-the-art computer to traverse the same given search space of available passwords. That change is roughly in line with Moore's law; faster computation was balanced by slower and more complex hashing algorithms that were developed over the years. Hash generation times indicate that much longer SHA-1 passwords can be retrieved nowadays as opposed to encrypted passwords 37 years ago. A potential adversary with access to a commercially available AMD HD7970 could easily test strings up to 7 characters or lowercase alphanumeric passwords up to 9 characters in length. By utilizing more powerful processing units and then distributing the task of password hashing across several of them, it would be possible to find even longer and more complex passwords.

However, recent improvements in password hashing (such as Argon2, bcrypt and scrypt) once again limit a potential adversary's endeavors. These iterative, memory-hard functions are designed to slow down password hashing, even with constant increases in

computer processing power. When comparing the running times of a single GPU AMD HD7970 generating SHA-1 hashes with eight nVidia GeForce GTX 1080 Ti computing bcrypt hashes with work factor 5, we found the latter were computed more than 14 times slower, despite the much larger processing power dedicated to the password hashing task. Adding just one additional iteration already doubles the password hashing time. While such hashing speeds are by no means slow enough to protect against brute-force attacks, they illustrate the potential of memory-hard functions: As computer processing power continues to increase, security experts will merely have to increment the work factor to compensate for the decreased time necessary to perform simple operations when computing password hashes.

To determine the hashing algorithms that are resistant to offline brute-force attacks, we devised two scenarios. In both scenarios, the attacker has obtained a full list of hashed passwords, and decides to perform a brute-force attack to uncover passwords. The difference between the scenarios is in the attacker's objective. In the first scenario, the adversary simply wants to get access to the system, meaning that she will be trying multiple passwords one after another, expecting to find a particularly weak password that can be cracked in a relatively short time. For that purpose, we chose one full day as a cut-off point for cracking a single password before the attacker moves onto the next one. In the second scenario, the attacker is targeting a particular user's password, and is willing to direct all available resources to uncover it. We assume, however, that she can spend up to a year trying to crack the password, before giving up under the assumption that the said password cannot be recovered in a reasonable amount of time.

Table 1. Password length and estimated cracking times for two hashing algorithms: SHA-1 ran on an AMD HD7970 GPU, and bcrypt with low (5), medium (12), and high (20) work factor, ran on 8x nVidia GTX 1080 Ti rig

	pool size	10	26	36	52	62	95	128
cutoff point	algorithm	digits	lowercase	lowercase & digits	alphabetic	alphanumeric	printable ASCII	all ASCII
One day	SHA-1	14 chars (10h 37m)	10 chars (15h)	9 chars (10h 47m)	8 chars (5h 41m)	8 chars (23h 12m)	7 chars (7h 25m)	6 chars (28m 1s)
	bcrypt (5)	10 chars (15h 2m)	7 chars (12h 4m)	6 chars (3h 16m)	5 chars (34m 17s)	5 chars (1h 23m)	5 chars (11h 38m)	4 chars (24m 12s)
	bcrypt (12)	8 chars (19h 14m)	5 chars (2h 17m)	5 chars (11h 38m)	4 chars (1h 24m)	4 chars (2h 51m)	4 chars (15h 40m)	3 chars (24m 12s)
	bcrypt (20)	5 chars (4h 56m)	4 chars (22h 30m)	3 chars (2h 18m)	3 chars (6h 56m)	3 chars (11h 44m)	2 chars (26m 40s)	2 chars (48m 25s)
One year	SHA-1	17 chars (1.21Y)	12 chars (1.16Y)	11 chars (1.59Y)	10 chars (1.75Y)	10 chars (10.17Y)	9 chars (7.64Y)	9 chars (111.77Y)
	bcrypt (5)	13 chars (1.71Y)	10 chars (24.21Y)	9 chars (17.42Y)	8 chars (9.17Y)	8 chars (37.44Y)	7 chars (11.97Y)	7 chars (96.53Y)
	bcrypt (12)	11 chars (2.19Y)	8 chars (4.58Y)	7 chars (1.72Y)	7 chars (22.56Y)	6 chars (1.25Y)	6 chars (16.13Y)	6 chars (96.51Y)
	bcrypt (20)	9 chars (5.62Y)	6 chars (1.74Y)	6 chars (12.23Y)	5 chars (2.14Y)	5 chars (5.15Y)	5 chars (43.48Y)	4 chars (1.51Y)

Table 1 summarizes the estimated cracking times for passwords hashed by either SHA-1 or bcrypt. In each cell, we list password length, and the amount of time it would take to crack that password. In particular, we are interested in cracking times for both

listed scenarios. In the first scenario, we are looking for the longest passwords of various compositions that can be cracked in less than a day. For example, any SHA-1 hashed alphanumeric password of 8 characters or less could be cracked in more than a day (an 8 character long password would take approximately 23 hours and 12 minutes). That allows us to identify passwords that are considered too weak to be protected by a given hashing algorithm. Ideally, we want none of the stored passwords to fall into that category. If there are some, or even the majority of passwords of a given composition that are of the same, or shorter length than the ones listed for a specific hashing algorithm, we encourage system administrators strongly to hash all of their passwords with a slower hashing algorithm.

The second scenario lists the lengths of the shortest passwords of several compositions that cannot be cracked in less than a year. For instance, a 13-digit password hashed with bcrypt (work factor 5) would take more than a year (1.71 years) to crack. We characterize passwords of that length and longer as strong; if most of the passwords of a given composition can be classified into that category, we can assume that the hashing algorithm used to protect these passwords is sufficiently secure. Naturally, with further advances in computer processing speeds, cracking times should be re-evaluated to determine whether the passwords should be re-hashed with a stronger hashing algorithm.

Aside from advances in computing, we should also consider decades of user education and enforced security procedures, which were employed primarily to increase security and slow down brute-force attacks. Back in the 1970s, at the time of Morris and Thompson's study, users were still unaware of the relevance of their role in protecting their own accounts. Nowadays, people are mostly aware of good password creation practices, resulting in them making longer and more complex passwords; many web sites also limit their bad password choices by enforcing password policies. Consequently, it would take longer to crack these passwords on the same machines, which affects the choice of hashing algorithms.

To determine the average present-day password strengths, we collected a total of 7,408 university passwords. These passwords were generated by the users between the years 2008 and 2014, and were used to protect real, sensitive data, such as students' personal information, e-mails, course materials and grades. An adversary gaining access to the accounts protected by these passwords could access any Wi-Fi network within the Eduroam system, download software and licenses free of charge, register the account owner for an exam, or cancel an existing exam registration.

The plaintext passwords were obtained from the university under a strict security policy and as a result of a several months long negotiating process. A computer unit with restricted physical access was devoted exclusively to storing and processing the data. The unit's web access was physically disabled at all times, and the operating system was protected by a username and password. The password data stored on the device was encrypted; the processing and statistical analysis was performed on temporarily unencrypted data, which was erased after processing.

As we were operating with real data, ethical concerns were addressed as well. The obtained data was anonymized by the university security service's personnel, and contained only plaintext passwords. The use of such personal data without prior or written consent of a subject (e.g. student) is allowed for the purpose of research under Article 11 (2), Article 13 (2), and Article 32 (3) of the Data Protection Directive [38].

Establishing a link between a student and their password would require access to their personal information (such as name, surname or student ID), which is kept in the university's databases; however, if a malicious party gained access to the database through any means, reverse engineering the obtained passwords would be unnecessary as the records could be accessed directly. Furthermore, we believe other factors assist in reducing a potential adversary's chance of an unauthorized breach of security and/or privacy to a minimum. The majority of passwords belong to students that no longer use them (graduates or drop-outs), and whose accounts are thus no longer active.

Our main objective was to determine how vulnerable individual passwords are to current password cracking. For that purpose, we considered the first cracking scenario, in which the adversary would attempt to crack each password for a full day before moving onto the next one. Similarly to Morris and Thompson, we modeled a two-stage attack: In the first part, we estimated the amount of time it would take to find a given password in an exhaustive brute-force attack, and in the second part, we checked the passwords against various dictionaries.

Due to time and processing power constraints, we did not execute the brute-force attack on all passwords. Instead, each password was classified into one of several categories based on their length and composition. In the original article by Morris and Thompson, the categories included passwords that could be retrieved in a matter of days on a PDP-11/70. For comparison, we compiled a new list of categories for every considered hashing algorithm, to benefit current-age security officers. Each individual password that was classified into one of these categories could be retrieved in less than a day on modern-age GPUs.

Next, we performed a dictionary attack on the collection of SHA-1 hashed passwords, using several available wordlists. Altogether, we recovered 668 distinct passwords (approximately 9% of the set), however, only 47 passwords have not already been found in the brute-force attack. Such a simple attack could be executed in just ten minutes; by including diverse wordlists, and applying advanced techniques such as word mangling rules, we believe a more substantial percent of passwords could have been recovered.

In total, each out of 5,992 SHA-1 hashed passwords or 81% of the analyzed collection could be found in less than a day on a modern high-end graphics card using a simple brute-force and dictionary attack. This figure is alarmingly close to the 86% of user-generated passwords uncovered by Thompson and Morris back in 1979. Even if the average password length and complexity has increased over the years, the change is too slight to counter the rapid advances in processing power. Considering the vast number of websites that still use the SHA-1, the deprecated MD5, or even save their passwords in plaintext, we can conclude that that there has not been much improvement in password security in the past three and a half decades.

On the other hand, should the websites move to stronger hashing algorithms, such as bcrypt, the amount of recovered passwords would decrease dramatically. In our case, simply hashing all passwords with a bcrypt algorithm with the work factor 5 would mean that less than half (47.8%) of the passwords could be recovered in less than a day on a powerful GPU rig. Increasing the work factor to the conventional 12 would mean only about 3.6% less passwords could be cracked, though a further breakdown indicates why that is the case; while 501 (6.76%) less passwords could be recovered in a brute-force attack as a result of the higher work factor, 236 (3.19%) more passwords were instead found in dictionaries. Most of these passwords were simple, lowercase passwords of 6

and 7 characters in length. They illustrate that even a fortified defense against brute-force attacks does not yet guarantee that the passwords cannot be compromised via another attack channel; all of these passwords are still vulnerable to dictionary, social engineering, and other types of attacks on passwords.

Table 2. Number of passwords that could be cracked in less than a day for a given hashing algorithm, based on password length and composition

	Morris & Thompson (1979)	Student passwords (2008-2014) SHA-1	bcrypt (5)	bcrypt (12)	bcrypt (20)
1: Brute-force	2339 (71.11%)	5945 (80.25%)	3194 (43.12%)	2693 (36.35%)	97 (1.31%)
ASCII	551 (16.8%) 1-3 chars	3305 (44.6%) 1-7 chars	190 (2.6%) 1-5 chars	94 (1.3%) 1-4 chars	2 (0.03%) 1-2 chars
Alphanumeric	477 (14.5%) 4 chars	1737 (23.4%) 8 chars			3 (0.04%) 3 chars
Alphabetic	706 (21.5%) 5 chars				
Lowercase & digits		763 (10.2%) 9 chars	2751 (37.1%) 6 chars	96 (1.3%) 5 chars	
Lowercase	605 (18.4%) 6 chars	132 (1.8%) 10 chars	163 (2.2%) 7 chars		86 (1.2%) 4 chars
Digits		8 (0.1%) 11-14 chars	90 (1.2%) 8-10 chars	2503 (33.8%) 6-8 chars	6 (0.08%) 5 chars
2: Dictionary	492 (14.96%)	47 (0.63%)	347 (4.68%)	583 (7.87%)	638 (8.61%)
Total	2831 (86.07%)	5992 (80.89%)	3541 (47.80%)	3276 (44.22%)	735 (9.92%)

Nonetheless, bcrypt with work factor 12 still allowed for any out of more than a third of user-generated passwords to be recovered in less than a day. The majority of these were digits of 6 to 8 characters in length; most of the alphabetic and alphanumeric passwords were long enough (above 5 characters) to withstand a day-long brute-force attack. Even so, if a large portion of the password set is vulnerable to brute-force attacks, administrators are encouraged to protect the passwords with a stronger hashing algorithm. We considered the bcrypt algorithm with work factor 20, which would greatly increase the resilience against brute-force attacks, at the expense of performance. Only 97 out of 7,408 passwords (a little over 1%) could potentially be recovered in less than a day, though a simple dictionary attack could find 638 more passwords (all were lowercase passwords of 5 characters and more), bringing the total to 735, or nearly 10% of the password set.

5. Advice to the Partakers

Much of the blame for the password security problem over the past decades has been placed on the shoulders of the users. Extensive research has focused on mitigating the issue from the user viewpoint, but, at the end of the day, even if the users can be educated to adopt good password creation and management habits, the human memory capacity will remain limited. As technology continues to progress, it is unreasonable to expect humans will be able to follow such a pace, even if they are familiar with the often clashing, limited-benefit, or even counter-productive requirements, they will eventually become unable to fulfil them.

That does not, however, imply that the users do not have to be aware of the password-related problems. While it should not be on them to devise passwords capable of sustaining prolonged and dedicated brute-force attacks, their efforts when choosing and managing their own passwords can greatly diminish the effectiveness of some other types of attacks, particularly social engineering. Actively avoiding the choice of words, conventional symbol replacements, and personal information in passwords are still good password practices to adhere to, but users remain slow to adopt them because they have no motivation and see no immediate gain. Rather than giving out advice (which is often ignored), or enforcing minimum requirements (which can lead to frustration), we believe users should be stimulated by positive feedback (such as visual indicators of password strength), and user choice. In particular, a system could provide several suggestions to improve a password during the account creation, allowing the user to choose the most preferred one. Similarly, a system could encourage the user to incorporate personal associations into their passwords by increasing the score when password choice includes syllables derived from multiple words, combined with symbols.

At the same time, as more users are persuaded to trade convenience for greater security (e.g. password managers, or two-factor authentication), we are working towards a global shift in the way of thinking. Just like users have shifted gradually towards longer and more complex passwords over the last few decades as a result of education, once a certain threshold of security-conscious users is reached, prioritizing security over convenient use should become the norm, rather than an exception.

While continuous and constant user education would certainly help, we identify it as only a part of, rather than a wholesome solution, to the long-lasting password security problem. Proportionally, a much larger responsibility lies on the shoulders of security officers and administrators: While users need only to manage their personal accounts, administrators should ensure the security of the entire system. Given that many users are not expected to realize the security requirements for their own accounts, administrators are required to implement sufficient security measures to protect all accounts against a wide variety of possible attacks, ranging from brute-force and dictionary, to traffic interception and system vulnerability exploitation [22]. Since many of these accounts and their passwords might be particularly poor, it is all the more important for the administrators to ensure a high level of security. Unfortunately, there are no all-encompassing rules or guidelines that all administrators should conform to. The appropriate level of security is unique to the system, and largely dependent on the profile of a typical user (e.g. customers vs. CEOs), sensitivity of stored data (e.g. consumer data vs. healthcare data), as well as the scope and purpose of the system (e.g. small retail systems vs. social networks). Security officials are strongly advised to assess security requirements based on these variables and then design, implement, and continuously review a comprehensive security policy tailored to the system in question.

Current approaches to authentication security remain simple and often ad-hoc. Most services attempt to achieve a sufficient level of security by enforcing password policies on the users. However, too few of them stop to consider exactly which attacks they are defending against and what the implemented policies are trying to accomplish. Often, that leads to administrators implementing strict password policies which are attempting to protect the accounts from several independent attacks. However, as a result, they limit users' password choices greatly, which can, in turn, affect their motivation, and cause them to fall back into bad password habits. At the same time, overly strict policies might

actually benefit the attacker – since she is aware of the requirements, and can reasonably suspect many users will try only just barely to match them, which allows her to tailor her guesses in a way to check the most likely passwords first. Instead, administrators should consider using password policies as a way to mitigate one, or perhaps just a few targeted attacks. Dictionary attack comes up as the most viable option, as it prevents the users from choosing simple, and predictable passwords that could otherwise withstand a brute-force attack due to their lengths.

At the same time, different measures should be implemented to safeguard the passwords against other types of attacks. Already Morris and Thompson had suggested for stored passwords to be hashed and salted as a countermeasure to brute-force attacks. In our experiment, we have shown that hashing passwords with deprecated algorithms can be as inefficient as storing them in plaintext. Administrators are, thus, strongly urged to keep updated on the improvements in password hashing. A good approach is to evaluate the passwords as they are created, then choose a hashing algorithm based on the general strength of the passwords stored. If the weakest passwords in the database could be cracked in a reasonable amount of time (i.e. less than a day), administrators should consider re-hashing all of the passwords with a stronger hashing algorithm. The general idea is not to wait until the hashing function in use becomes obsolete, but instead always remain a step ahead.

6. The Future of Authentication

The argument that improvements in password cracking imply the end of passwords, has been made many times over the past few decades. For example, St. Clair et al. predicted password exhaustion based on the prevailing and predicted computer processing power, shown to be capable of cracking most real passwords used by the university students [39]. Various tabulations of cracking speeds have been offered repeatedly either to cajole users into stronger password choices, or convince them that passwords must be replaced. In the last decade especially, numerous parties advocating for alternatives to passwords made the claims that the passwords are dead.

However, it remains to be answered whether deprecating textual passwords and moving onto another authentication method entirely, is really the best answer. Research on authentication methods is bloated with proposals on novel authentication schemes aimed at replacing textual passwords. Many promote their improvements over passwords, but then fail to address weaknesses that often render them unusable in practice. Bonneau et al. examine closely many diverse authentication methods proposed in recent years from several angles, showing that all of the suggested alternatives suffer from serious drawbacks that make them less suitable than textual passwords for many purposes [40]. To name just a few:

- Graphical passwords may increase memorability because of their visual aspect, but suffer from most of the problems that their textual counterparts do, such as guessing (brute-force, dictionary) and capture (malware, phishing, social engineering) attacks, and a few more besides. Most prominently, the increased security of these schemes generally means decreased usability, [27]

- Hardware tokens provide higher security while reducing the need to memorize additional information (with the exception of the verifier PIN). However, they also create a significant inconvenience for the user, as she cannot authenticate without the token on her person. A stolen or lost token would also need to be revoked and re-issued. Because of that, it is unlikely to expect that users would want to use tokens for a multitude of everyday accounts in everyday situations,
- Biometry allows for an easy-to-use and compelling way of authentication, while eliminating the need for the user to memorize information or safeguard physical tokens. For the time being, the high cost and poor accuracy of this technology remains a problem. Even if there is room for improvement in both cases, biometric solutions still have poor deployability when compared to passwords. By far the most concerning problem, however, comes into play with a possible breach of the database of secrets. While biometric credentials can be stolen, extorted or forged relatively easily, they cannot be revoked without blocking both malicious and legitimate use in the process,
- Cognitive schemes share several security and usability benefits with textual passwords, even if it would be generous to say they are on par. For example, while textual passwords can be typed from muscle memory, cognitive passwords are input slower because they require effort and attention. Furthermore, minimizing secret leakage increases cognitive workload, resulting in longer login times and higher login errors,
- Password managers, while offering advantages over classical passwords in terms of security and selected usability aspects, lack in terms of deployability (accessing the database of secrets from another device requires Internet connection and a proprietary application; alternatively, it can be stored on a physical device, but it can be lost or stolen) which affects convenience of use. Most notably, however, they still rely on the underlying technology of textual passwords. Even so, Bonneau et al. argue they could become a common coping strategy, should the passwords remain widespread. [26]

To understand the complexity of the age-old password security dilemma, we can define it as a multi-objective optimization problem. In this context, security is just one of several contradicting parameters that we are trying to optimize. If we visualize existing authentication methods as points in a multi-dimensional space in which axes represent the parameters, we are essentially looking for methods that have one, or preferably several, parameters maxed out (i.e. the Pareto front).

Naturally, researchers should be encouraged to keep investigating new authentication possibilities, while striving to improve the already existing ones. The main goal of the former is to search the security-usability-deployability continuum thoroughly for any authentication methods that appear on or beyond the Pareto front. The latter aims to push the existing schemes closer toward, and hopefully past, the current front. Both research directions are important for us to gain a comprehensive understanding of the field, so that a more educated approach toward better authentication can be taken in the future.

The current predicament is that the researchers do not take a step back to assess objectively what has already been discovered and try to fit their own piece into the authentication mosaic. Instead, they work under the assumption that the existing research on authentication mechanisms is conclusive. They often assume a biased view of the

measures that highlight their method's strengths, and are focused primarily on trying to advocate for their new or improved methods to either complement or replace textual passwords. Such approach can distort academia's view of the authentication problem, further enforcing the idea that a shift in the field of Authentication is necessary and imminent.

However, our heavy reliance on textual passwords stands in stark disagreement with this argument. Were any of the alternatives scoring better than textual passwords in several competing aspects, they would likely have replaced them over the years. Still, at least for the time being, textual passwords undoubtedly remain on the Pareto front of authentication mechanisms. The users have no incentive to switch them for any other alternative on the front, as that would effectively be trading one set of advantages for another.

It is largely true that user-chosen passwords cannot withstand brute-force attacks. However, mechanisms such as salting and strong, up-to-date hashing algorithms, can diminish the effectiveness of these attacks. The means to defend against this weakness exist – it is on the shoulders of system administrators to ensure the safety of users' accounts from this point of view. It is idealistic to expect that all of them will do so; consequently, we are likely to keep hearing about new security breaches in the future. However, in the grand scheme of things, these are more or less isolated cases, and researchers suggest that offline attacks (which are most often studied in password literature) are not as common as we might think [8].

It is particularly difficult to assess textual passwords' susceptibility to other types of attack in real-world scenarios. When examining the world's largest data breaches in the last decade, most of the passwords were either exposed in accidental or deliberate leaks, or obtained from information system hacks and, subsequently, cracked [15]. Successful cracking (where necessary) was made possible due to the deprecated and/or insecure hashing algorithms used to protect them. That further supports the claim that the way passwords are stored matters quite a bit, as Florencio et al. discuss [42]. Regardless, password guessing has moved beyond targeted brute-force and dictionary attacks. For example, social engineering techniques can be used to take advantage of the human factor in password creation. The current state-of-the-art cracking employs recurrent neural networks, which provide a much more accurate sense of password vulnerability [43]. Further research should examine how feasible these attacks are in real-world scenarios, and attempt to assess how often they actually appear in practice.

When predicting textual passwords' longevity and the future of authentication, we believe two points should be considered. The first is the users' motivation to switch to another authentication scheme. Despite reported attacks and breaches of password-protected accounts, as well as constant cajoling by both stakeholders and the media that the passwords should be replaced, users are not compelled to replace them. Most are convinced a breach of privacy cannot happen to them, and will not have a shift in their way of thinking unless they experience personal loss. Therefore, the very small percentage of users affected by data breaches is nowhere near the threshold necessary for the entire population to consider switching textual passwords for another authentication method. Their motivation is also affected partially by the lack of competitive alternatives. That highlights the second problem: Users are unlikely to favor another authentication method if it would mean nothing but trading different sets of

advantages. Until there are no decisively better alternatives for the users to shift towards, textual passwords are likely to remain in widespread use.

7. Conclusion

Already in 1979, Morris and Thompson inspected a set of real, user-generated passwords, and concluded that they were inadequate. In an endeavor to increase its security, they suggested several improvements to the original authentication scheme. More than three and a half decades later, we are witnessing an upsurge in data breaches, while the security officers contemplate whether to replace textual passwords or not.

In this article, we make several important contributions to the resolution of this dilemma and the field of Password Security as a whole. First, we make a direct comparison of recent, real-world passwords and the passwords collected by Morris and Thompson. Through historical analysis, we show that there has been no efficient change in password security in the last couple of decades. The main reason why that is the case stems from the fact that neither users nor system administrators conform to Morris and Thompson's advice diligently. As a result, textual passwords are more vulnerable to the attackers, prompting more individuals from academia and the industry alike to advocate for a shift in security.

Our second contribution is, therefore, an advice to administrators to implement password checkers that will forbid the use of dictionary words in textual passwords. In our article, we establish that the composite of technological solutions, security procedures, and user education, as proposed by Thompson and Morris, can still be effective even today. We demonstrate that employing memory-hard hash functions like bcrypt can protect passwords despite technological advancements – but against brute-force attacks only. Nothing prevents users from choosing simple, predictable passwords that are vulnerable to dictionary attacks. Nearly forty years have taught us that users will not change. It is, therefore, vital for administrators to realize that fact and act upon it. It will be easier to change the administrators' behavior than the users'.

Finally, we believe that the currently available alternatives to passwords cannot solve the security problem. Their few advantages over textual passwords often come with an extensive list of disadvantages that are poorly documented and validated in literature. Currently, no existing solution outperforms textual passwords in terms of security, usability and deployability. Ultimately, we do not expect textual passwords will be replaced in the foreseeable future. While an imperfect form of authentication, their shortcomings will be overlooked as long as they are considered "good enough" by the typical user.

In this article, we assessed the current state of textual passwords critically and provided insight on how password security can be improved from the perspective of both users and system administrators. We believe such approaches could reinforce their position as the dominant authentication method in the future. Regardless, we do not dismiss the existence of a better authentication scheme. Thus, we urge researchers to keep developing new and improving existing authentication schemes in an effort to expand our understanding of authentication security and identify Pareto-improving authentication methods.

Acknowledgement. The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0057).

References

1. Morris R., Thompson K.: Password Security: A Case History. Communications of the ACM, Vol. 22, No. 11, 594–597. (1979)
2. Zviran M., Haga W. J.: Password Security: An Empirical Study. Journal of Management Information Systems, Vol. 15, No. 4, 161–184. (1999)
3. Florencio D., Herley C.: A Large-Scale Study of Web Password Habits. In Proceedings of the 16th International Conference on World Wide Web. Banff, AB, Canada, 657–666. (2007)
4. Dell'Amico M., Michiardi P., Roudier Y.: Password Strength: An Empirical Analysis. In Proceedings of the 29th IEEE International Conference on Computer Communications. San Diego, CA, USA, 983–991. (2010)
5. Bonneau J.: The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In Proceedings of the 33rd IEEE Symposium on Security & Privacy, San Francisco, CA, USA, 538–552. (2012)
6. Shen C., Yu T., Xu H., Yang G., Guan X.: User Practice in Password Security: An Empirical Study of Real-life Passwords in the Wild. Computers & Security, Vol. 61, 130–141. (2016)
7. Ji S., Yang S., Hu X., Han W., Li Z., Beyah R.: Zero-Sum Password Cracking Game: A Large-Scale Empirical Study on the Crackability, Correlation, and Security of Passwords. IEEE Transactions on Dependable and Secure Computing, Vol. 14, No. 5, 550–564. (2017)
8. Bonneau J., Herley C., van Oorschot P. C., Stajano F.: Passwords and the Evolution of Imperfect Authentication. Communications of the ACM, Vol. 58, No. 7, 78–87. (2015)
9. Mazurek M. L. et al.: Measuring Password Guessability for an Entire University. In Proceedings of the 13th Conference on Computer & Communications Security, Berlin, Germany, 173–186. (2013)
10. Awad M., Al-Qudah Z., Idwan S., Jallad A. H.: Password Security: Password Behavior Analysis at a Small University. In Proceedings of the 5th International Conference on Electronic Devices, Systems and Applications, Ras Al Khaimah, UAE. (2016)
11. Feldmeier D. C., Karn P. R.: UNIX Password Security - Ten Years Later. In Proceedings of the 9th Conference on the Theory and Applications of Cryptology. Santa Barbara, CA, USA, 44–63. (1989)
12. Corbató F. J.: On Building Systems That Will Fail. In: ACM Turing Award Lectures. ACM, New York, 72–81. (2007)
13. Mitnick K. D.: The Art of Deception – Controlling the Human Element of Security. John Wiley & Sons, Indianapolis, USA. (2002)
14. Hancock B.: Network Breaches: They are Real. Computer Fraud & Security, Vol. 1998, No. 10, 8–11. (1998)
15. McCandless D.: World's Biggest Data Breaches. (2018). [Online]. Available: http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/ (current July 2018). Archived by WaybackMachine® at http://web.archive.org/web/20180711024143/http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/
16. Mavropoulos O., Mouratidis H., Fish A., Panaousis E., Kalloniatis C.: A Conceptual Model to Support Security Analysis in the Internet of Things. Computer Science and Information Systems, Vol. 14, No. 2, 557–578. (2017)

17. Ranmohotti K. G. S. et al.: Coexistence of High-Tc Ferromagnetism and n-Type Electrical Conductivity in FeBi2Se4. Journal of the American Chemical Society, Vol. 137, No. 2, 691–698. (2015)
18. Shor P. W.: Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer. SIAM Journal on Computing, Vol. 26, 1484–1509. (1997)
19. Wiemer F., Zimmermann R.: High-Speed Implementation of bCrypt Password Search using Special-Purpose Hardware. In Proceedings of the 2014 International Conference on Reconfigurable Computing and FPGAs. Cancun, Mexico. (2014)
20. Weir M., Aggarwal S., De Medeiros B., Glodek B.: Password Cracking using Probabilistic Context-Free Grammars. In Proceedings of the 30th IEEE Symposium on Security and Privacy. Berkeley, CA, USA, 391–405. (2009)
21. Narayanan A., Shmatikov V.: Fast Dictionary Attacks on Passwords using Time-Space Tradeoff. In Proceedings of the 12th ACM Conference on Compututer and Communications Security. Alexandria, VA, USA, 364–372. (2005)
22. Cho B., Lee S., Xu M., Ji S., Kim T., Kim J.: Prevention of Cross-update Privacy Leaks on Android. Computer Science and Information Systems, Vol. 15, No. 1, 111–137. (2018)
23. Tarwireyi P., Flowerday S., Bayaga A.: Information Security Competence Test with regards to Password Management. In Proceedings of the 2011 Conference on Information Security South Africa. Johannesburg, South Africa. (2011)
24. Tam L., Glassman M., Vandenwauver M.: The Psychology of Password Management: A Tradeoff between Security and Convenience. Journal of Behaviour & Information Technology, Vol. 29, No. 3, 233–244. (2010)
25. Notoatmodjo G., Thomborson C.: Passwords and Perceptions. In Proceedings of the 7th Australasian Conference on Information Security. Wellington, New Zealand, Vol. 98, 71–78. (2009)
26. Stanton B. C., Greene K. K.: Character Strings, Memory and Passwords: What a Recall Study Can Tell Us. In Proceedings of the 2nd International Conference on Human Aspects of Information Security, Privacy and Trust. Heraklion, Crete, Greece, 195–206. (2014)
27. Lee P. Y., Choong Y. Y.: Human Generated Passwords – The Impacts of Password Requirements and Presentation Styles. In Proceedings of the 3rd International Conference on Human Aspects of Information Security, Privacy and Trust. Los Angeles, CA, USA, 83–94. (2015)
28. Komanduri S. et al.: Of Passwords and People: Measuring the Effect of Password-Composition Policies. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Vancouver, BC, Canada, 2595–2604. (2011)
29. Adams A., Sasse M. A.: Users Are Not the Enemy. Communications of the ACM, Vol. 42, No. 12, 40–46. (1999)
30. Milosavljević G., Sladić G., Milosavljević B., Zarić M., Gostojić S., Slivka J.: Context-sensitive Constraints for Access Control of Business Processes. Computer Science and Information Systems, Vol. 15, No. 1, 1–30. (2018)
31. Lorenz B., Kikkas K., Klooster A.: 'The Four Most-Used Passwords Are Love, Sex, Secret, and God': Password Security and Training in Different User Groups. In Proceedings of the 1st International Conference on Human Aspects of Information Security, Privacy and Trust. Las Vegas, NV, USA, 276–283. (2013)
32. Grawemeyer B., Johnson H.: Using and Managing Multiple Passwords: A Week to a View. Interacting with Computers, Vol. 23, No. 3, 256–267. (2011)
33. Zhang Y., Monrose F., Reiter M. K.: The Security of Modern Password Expiration. In Proceedings of the 17th ACM Conference on Computer and Communications Security. Chicago, IL, USA, 176–186. (2010)
34. Shay R. et al.: Encountering Stronger Password Requirements: User Attitudes and Behaviors. In Proceedings of the 6th Symposium on Usable Privacy and Security. Redmond, WA, USA. (2010)

35. von Zezschwitz E., De Luca A., Hussmann H.: Survival of the Shortest: A Retrospective Analysis of Influencing Factors on Password Composition. In: Kotze, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler M. (eds.): Human-Computer Interaction - Interact 2013. Lecture Notes in Computer Science, Vol. 8119. Springer-Verlag, 460–467. (2013)
36. Qiu W., Gong Z., Guo Y., Liu B., Tang X., Yuan Y.: GPU-Based High Performance Password Recovery Technique for Hash Functions. Journal of Information Science and Engineering, Vol. 32, 97–112. (2016)
37. Gosney J. M.: 8x Nvidia GTX 1080 Ti Hashcat Benchmarks. (2018). [Online]. Available: https://gist.github.com/epixoip/ace60d09981be09544fdd35005051505/ (current July 2018). Archived by WebCite® at http://www.webcitation.org/70sbkefP5/
38. EUR-Lex. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. (1995) [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML (current July 2018). Archived by WebCite® at http://www.webcitation.org/70sc3rROv/
39. St. Clair L. et al: Password Exhaustion: Predicting the End of Password Usefulness. In Proceedings of the 2nd International Conference on Information Systems Security. Kolkata, India, 37-55. (2006)
40. Bonneau J., Herley C., Van Oorschot P. C., Stajano F.: The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. 2012 IEEE Symposium on Security and Privacy. San Francisco, CA, USA, 553-567. (2012)
41. Biddle R., Chiasson S., Van Oorschot P. C.: Graphical Passwords: Learning From the First Twelve Years. ACM Computing Surveys, Vol. 44, No. 4, 1-25. (2012)
42. Florencio D., Herley C., Van Oorschot P. C.: Pushing on String: The 'Don't Care' Region of Password Strength. Communications of the ACM, Vol. 59, No. 11, 66-74. (2016)
43. Melicher W. et al.: Fast, Lean, and Accurate: Modeling Password Guessability Using Neural Networks. In Proceedings of the 25th USENIX Security Symposium. Austin, TX, USA, 175-191. (2016)

Leon Bošnjak is an assistant and researcher at the Faculty of Electrical Engineering and Computer Science of University of Maribor, Slovenia. In 2014, he received his master's degree in Informatics and Technologies of Communication, and is currently a doctoral student of Computer Science and Information Technologies. His research interests include system security, textual and graphical passwords, and authentication methods.

Boštjan Brumen received doctor’s degree in informatics in 2004. He is an associate professor at University of Maribor, Faculty of Electrical engineering and computer science. He was Secretary General (Provost) of University of Maribor for two consecutive terms between 2004 and 2011. His research interests include intelligent data analysis, automated learning and learning models, data security and data quality. He has published several articles about passwords in prestigious journals, among them Journal of Medical Internet Research.

Received: March 28, 2018; Accepted: August 25, 2018

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

004

COMPUTER Science and Information Systems : the International journal /
Editor-in-Chief Mirjana Ivanović. – Vol. 16, No 1 (2019) - . – Novi Sad (Trg D. Obradovića 3): ComSIS Consortium, 2018 - (Belgrade : Sigra star). –30 cm

Polugodišnje. – Tekst na engleskom jeziku

ISSN 1820-0214 (Print) 2406-1018 (Online) = Computer Science and Information Systems
COBISS.SR-ID 112261644

Cover design: V. Štavljanin
Printed by: Sigra star, Belgrade