Business Information Systems
1 | 2021

24th International Conference on Business Information Systems

14-17 June 2021 – Hannover, Germany

Conference Proceedings

Editors:

Witold Abramowicz

Sören Auer

Elżbieta Lewańska

TIB OPEN PUBLISHING

Business Information Systems

Business Information Systems (BIS) are the proceedings of the International Conference on Business Information Systems. The International Conference on Business Information Systems is a renowned international conference on business information systems.

ISSN (online): 2747-9986

Business Information Systems is published by by TIB Open Publishing (Technische Informationsbibliothek, Welfengarten 1 B, 30167 Hannover) on the behalf of 24th International Conference on Business Information Systems.

24th International Conference on Business Information Systems

15-17 June 2021 Hannover, Germany

“Enterprise Knowledge and Data Spaces”

Preface

Auer et al.	Preface	1

Big Data

Pondel et al.	Deep Learning for Customer Churn Prediction in E-Commerce Decision Support	3
Kruse et al.	Developing a Legal Form Classification and Extraction Approach For Company Entity Matching: Benchmark of Rule-Based and Machine Learning Approaches	13
Bernardo & Della Valle	Predict COVID-19 Spreading With C-SMOTE	27
Loye et al.	Post-Brexit Power of European Union From the World Trade Network Analysis	39
Schneider & Kusturica	Towards a Guideline Affording Overarching Knowledge Building in Data Analysis Projects	49
Uysal et al.	Optimization-Based Business Process Model Matching	61
Bano et al.	Database-Less Extraction of Event Logs from Redo Logs	73
Shakir et al.	Towards a Concept for Building a Big Data Architecture with Microservices	83

Smart Infrastructures

Ruhkamp & Schönig	Execution of Multi-Perspective Declarative Process Models Using Complex Event Processing	95
Richter & Anke	Exploring Potential Impacts of Self-Sovereign Identity on Smart Service Systems: An Analysis of Electric Vehicle Charging Services	105

Knowledge Graphs

Klessascheck et al.	Domain-Specific Event Abstraction	117
Tkachenko et al.	Ontological Modeling of the State Economic Development Policy for Cultural Industries	127
Flotyński et al.	Semantic Representation of Domain Knowledge for Professional VR Training	139
Filipiak et al.	Mapping of ImageNet and Wikidata for Knowledge Graphs Enabled Computer Vision	151

Artificial Intelligence

Szmydt	Contextual Personality-Aware Recommender System Versus Big Data Recommender System	163
Parfait & Totohasina	Generating a Condensed Representation for Positive and Negative Association Rules: A Condensed Representation for Association Rules	175
Zahmatkesh et al.	Supporting an Expert-centric Process of New Product Introduction With Statistical Machine Learning	187
Zarifis & Cheng	Evaluating the New AI and Data Driven Insurance Business Models for Incumbents and Disruptors: Is there Convergence	199
Witte et al.	Evaluation of Deep Learning Instance Segmentation models for Pig Precision Livestock Farming	209
Pegoraro et al.	Text-Aware Predictive Monitoring of Business Processes	221

Social Media

Zhang & Hara	Predicting E-commerce Item Sales With Web Environment Temporal Background	233
Bukar et al.	Social Media Crisis Communication Model for Building Public Resilience: A Preliminary Study	245

Applications

Wecel et al.	Stream Processing Tools for Analyzing Objects in Motion Sending High-Volume Location Data	257
Eichler et al.	Enterprise-Wide Metadata Management: An Industry Case on the Current State and Challenges	269
Van Veldhoven et al.	A Scoping Review of the Digital Transformation Literature Using Scientometric Analysis	281
Ebner et al.	Innovating in Circles: A Qualitative Analysis on Cycles of IT Feature Recombinations for Performative and Creative Outcomes	293
Borissova & Dimitrova	An Integrated Group Decision-Making Approach Considering Uncertainty Conditions	307
Lazuardi et al.	Interoperability of Health Digitalization: Case Study on Use of Information Technology for Maternal and Child Health Services in Indonesia	317
Bollweg et al.	The Digitalization of Local Owner-Operated Retail Outlets: How Environmental and Organizational Factors Drive the Use of Digital Tools and Applications	329
Bhargava et al.	A Novel Example-Dependent Cost-Sensitive Stacking Classifier to Identify Tax Return Defaulters	343

ICT Project

Seeba et al.	Development of the Information Security Management System Standard for Public Sector Organisations in Estonia	355
Bueechl et al.	Potentials and Barriers of Agility in Small and Medium Sized Enterprises: Insights From Qualitative Research in Germany	367

24th International Conference on Business Information Systems (BIS 2021)

Preface

https://doi.org/10.52825/bis.v1i.70

Published: 02 July 2021

24th International Conference on Business Information Systems

Preface

Witold Abramowicz[1][https://orcid.org/0000-0001-5464-9698] and Sören Auer [2][https://orcid.org/0000-0002-0698-2864]

[1] Poznań University of Economics and Business, Poland

[2] Technische Informationsbibliothek, Germany

The theme of the 24th BIS conference was Enterprise Knowledge and Data Spaces. Both concepts are relevant for data organization and reuse. One of the contemporary ways to represent knowledge in the enterprises are enterprise knowledge graphs or just knowledge graphs. They are a flexible way to represent interlinked information about virtually anything. From the modelling point of view they are graphs consisting of concepts, properties, and entity descriptions. So, we distinguish here schema data, instance data, and metadata. What is important is that they are fully compliant to FAIR Data Principles. The principles characterize the desired features of digital assets: Findable, Accessible, Interoperable, Reusable.

Data spaces are more about information technology architecture. They are considered a virtual space for safeguarding data where data remains with the data owner and it is first shared with trusted business partners. The focus is on the information model for describing data assets and also providing the standardized interface for interoperability. More often than in the case of enterprise knowledge we emphasize the economic valuation of data, including pricing of data transactions. It is noteworthy that data spaces are considered a core of the implementation of the European Strategy on Data. It foreseen to regulate the flow of data within European Union and across sectors and is also based on the FAIR principles. As the ultimate goal of FAIR is to optimise the reuse of data we expect papers covering all aspects of data collection, standardization, integration, exchange, reuse, and valuation.

The 24th edition of International Conference on Business Information Systems was held in Hannover, Germany. Due to the COVID-19 pandemic, the conference was held as an online event. Since its first edition in 1997, the BIS conference became a well-respected event and it gathered a wide and active community of researchers and business practitioners.

The BIS 2021 proceedings includes 32 articles divided into 7 parts that reflect main areas of interests of BIS community. Those are: Big Data, Smart Infrastructure, Knowledge Graphs, Artificial Intelligence, Social Media, Applications and ICT Projects. Authors submitted 93 papers total thus the acceptance rate was 34%.

The Program Committee consisted of 105 members from 30 countries, who carefully evaluated all the submitted papers.

We would like to sincerely thank all people who are involved in the BIS community. The reviewers, who dedicate their time and prepare insightful comments. Keynote Speakers, for their interesting presentations that started a vivid discussion among the conference participants. Last but not least, we would like to thank all authors who submitted their papers.

June, 2021, Witold Abramowicz & Sören Auer

24th International Conference on Business Information Systems (BIS 2021)

Big Data

https://doi.org/10.52825/bis.v1i.42

Published: 02 July 2021

Deep learning for customer churn prediction in e-commerce decision support

Maciej Pondel[1][https://orcid.org/0000-0002-1978-6571], Maciej Wuczyński[2][https://orcid.org/0000-0001-7376-1933], Wiesława Gryncewicz[3][https://orcid.org/0000-0003-1208-4099], Łukasz Łysik[4][https://orcid.org/0000-0002-7886-5130], Marcin Hernes[5][https://orcid.org/0000-0002-3832-8134], Artur Rot[6][https://orcid.org/0000-0002-7281-8253], and Agata Kozina [7][https://orcid.org/0000-0003-0447-4038]

[1-7] Wroclaw University of Economics and Business, Komandorska 118/120, 53345 Wrocław, Poland

Abstract. Churn prediction is a Big Data domain, one of the most demanding use cases of recent time. It is also one of the most critical indicators of a healthy and growing business, irrespective of the size or channel of sales. This paper aims to develop a deep learning model for customers' churn prediction in e-commerce, which is the main contribution of the article. The experiment was performed over real e-commerce data where 75% of buyers are one-off customers. The prediction based on this business specificity (many one-off customers and very few regular ones) is extremely challenging and, in a natural way, must be inaccurate to a certain ex-tent. Looking from another perspective, correct prediction and subsequent actions resulting in a higher customer retention are very attractive for overall business performance. In such a case, predictions with 74% accuracy, 78% precision, and 68% recall are very promising. Also, the paper fills a research gap and contrib-utes to the existing literature in the area of developing a customer churn prediction method for the retail sector by using deep learning tools based on customer churn and the full history of each customer's transactions.

Keywords: Churn prediction, deep learning, machine learning, e-commerce, decision support

Introduction

E-commerce has provided many new opportunities to consumers, and it is still open-ing new ones. The rapid expansion of IT technologies and the Internet has resulted in this sector's rapid growth. According to [1], the value of the e-commerce market in Poland exceeds 51 billion PLN, and 28 million Poles use online shopping in various forms. The global turnover of the e-commerce industry currently amounts to 3 trillion USD per year, and the fastest-growing market is the Asian market, whose dynamics reaches almost 30%. For comparison, the value of the European e-commerce channel was estimated at over 602 billion EUR, of which nearly 60% was generated by three markets: British, German, and French.

The success of companies hugely depends on how well they can analyze the data on their clients' behavior. Customers' churn may be considered as a lost opportunity for profit. The costs of gaining new customers are usually five to even six times higher than the costs of retaining an existing customer [2]. As a result, efforts made by marketing specialists to sustain market share have switched from focusing on acquiring new customers to retaining existing ones – reducing customer churn. For this reason, customer churn, also known as customer turnover, customer attrition, or customer deflection, is a major concern for a number of industries. This is particularly important in the e-commerce context, where

consumers are able to com-pare products or services and change the vendor with minimal effort.

Churn prediction is a Big Data domain, one of the most demanding use cases of recent time. It is also one of the most critical indicators of a healthy growing business, irrespective of the size or channel of sales. Customer attrition allows specialists to estimate the number of customers who will give up on the company's product or ser-vice subscription in a given time frame.

According to Ph. Kotler [3], companies annually lose 10 to 30 per cent of customers while acquiring new customers is about ten times costlier than maintaining existing ones. This information-rich sentence indicates how valuable customers are for a business. Research done by Amy Gallo [4] states that depending on the industry, acquiring a new customer is anywhere from 5-25% more expensive than retaining an existing one. So, it is essential to keep customers happy. For exam-ple, the telecommunications industry experiences an average 30-35 per cent annual churn rate; additionally, it costs 5-10 times more to gain a new customer than to re-tain an existing one [5].

It is crucial for a contemporary business to start analyzing why customers abandon relationships with a company by cancelling services or ceasing to buy products. This type of analysis allows e-commerce specialists to modify their current activities and adjust offers so that customer's needs are better covered, resulting in a lower churn rate.

This paper aims to develop a deep learning model for customers' churn prediction in e-commerce. The study pertains to the prediction of customer churn in B2C e-commerce. It also fills a research gap and contributes to the existing literature in the area of developing a customer churn prediction method for the retail sector by using deep learning tools based on customer churn and the full history of each customer's transactions, which is the major contribution of the article.

Related works

In the field of e-commerce, customer churn can be placed among the most critical problems that need to be addressed and thoroughly examined.

Compared to traditional shopping in retail stores, e-commerce has a significant ad-vantage: instant and accurate track of records and in-depth data collection (shopping activities, order information, delivery information, billing address, etc.). This data collection allows multidimensional analysis of both customers and their buying hab-its, additionally helping businesses to treat customers as individuals and in a personal-ized manner. With the support of gathered data, it is possible to create customer-centric business intelligence based on the following business concerns [6]:

- Which subpages did the customer visit? How long did they stay there? What was the sequence in which they browsed a given web page?
- Who are the most/least valuable customers? What are their distinctive characteristics?
- Who are the most/least loyal customers, and how are they characterized?
- What are customers' purchase behavior patterns?
- Which types of customers are more likely to respond to a particular promotion?
- And so on

There are many both academia researchers and practitioners who have been actively trying to predict customer churn with the help of gathered data – statistics, data mining, or machine learning strategies.

Customer churn is a buzzword that has been used for a long time in the field of e-commerce, and early determination of consumers that might be lost should be identified

accurately through data mining and data analysis and related in time with effective marketing measures [7].

Churn models are made to detect, as soon as possible, signals of potential churn and help to identify customers willing to abandon a given company voluntarily.

Customer churn prediction is a very demanding and challenging process aimed at identifying consumers willing to abandon a company or a service. Decision-makers and machine learning specialists focus on designing models which can help to identify early churn signals and recognize consumers on the verge of a decision between leaving or continuing. Therefore, to retain customers, academics, as well as practitioners, find it crucial to build a churn prediction model that is as accurate as possible in order to minimize the risk of customer churn [8]. Also, researchers have confirmed that customer churn prediction models can improve a company's revenue and its reputation in the market [9], [10]. Reducing the rate of churn and retaining current customers are the most cost-effective marketing approaches that will maximize the shareholder's value [11], [12]. Today, companies have enough information, of every kind, about the behaviour of their customers - this has created an opportunity for the machine learning (ML) community to develop predictive modelling techniques to handle the customer churn prediction. [13].

During the last decade, customer churn prediction has received a growing consideration in order to survive in an increasingly competitive and global marketplace [14]. Companies should strive for models that can accurately identify potential churners, and this becomes even more important in the digital economy context. Over the last decade, this issue has been mentioned and researched by many practitioners and academics. In contemporary literature, we can observe two main trends concerning customer churn. According to [15], the first branch includes traditional classification methods such as decision tree (DT) and logistic regression (LR) [16] [17] [18] [19] [20 [21] [22] [23]). The second mentioned line of thought is based on artificial intelligence methods such as neural networks [24] [17] [23]; [25], evolutionary learning [17], genetic algorithms [17][18], random forests [26]; [27], improved balanced random forests [28], K-nearest neighbour [29], fuzzy logic Systems [30], and support vector machines [28]. The decision tree and logistic regression are dedicated to the analysis of continuous data; they cannot, however, guarantee the accuracy of constructed models for large scale, nonlinearity, and high-dimensionality [31].

All of the presented models of customer churn prediction are very helpful in creating measures which can help a company to prevent customers from attrition. Worth mentioning is that customer churn predictive models are usually solely evaluated based on their predictive performance in which the models show the ability to correctly identify customer churns and non-churns separately and accurately [21].

For the customer churn prediction problem, most of the related academic works focus on the socalled post-paid industries. This means that the contract with the customer ends or is terminated (e.g. banks, Internet service providers, insurance companies, and telephone service companies) [32] [13] [33]. The subject of this paper, as it was mentioned in the introduction, is the use of deep learning algorithms for churn prediction in the retail industry. A characteristic feature of this sector is the uncertainty surrounding the return of a customer to the same seller. As they are not bound by any contract, they can easily abandon the existing relationship. That is why the purpose of this article is to create a model that calculates the probability that a customer will return to the same vendor and in how many days they will return.

It becomes very significant in the e-commerce context, where competitors are only a few 'mouse clicks' away, and consumers can compare and contrast competing products and services with minimal expenditure of personal time or effort and move from one company to another [34]. In our deliberations, we will narrow down the research area even further, focusing on only one part of e-commerce, namely the retail sector. We are going to develop a useful churn prediction model for B2C context outperforming the commonly used methods because of two reasons. It is a model capable of capturing the specific characteristics of B2C

e-commerce relations, and – the second thing – it can predict when the customer will return to the same vendor.

In most domains, churning is usually referred to as losing a client. For example, [35] predicts the churn probability for prepaid clients of a cellular telecommunication company. In financial services (banking and insurance), churn is usually seen as closing accounts [32]. [36] predict the switching probability of an insured person to another auto insurance company. As far as retail is concerned, most studies also focus on the customer's ability to leave to identify the exact moment when customers will discontinue their relationship with companies. In the retail sector literature, churn has also been considered as the partial or progressive defection of customers. [37] used several classification techniques and proposed predictive models for partial customer turnover in retail. Most customers exhibit partial defection, which may subsequently lead to a complete switch. Also, Buckinx and Van den Poel [26] used the concept of partial churn to identify customers that the company should focus on if concerned with customer retention. The costs of gaining new customers are usually five to even six times higher than the costs of retaining an existing customer [2]. Nevertheless, they still talked about attrition. Also churn models based on risks models has been developed [38].

Our study pertains to the prediction of customer churn in B2C e-commerce. In contrast to existing research, we developed a deep learning model based on the full history of each customer's transactions, which can be useful in existing customer segmentation mechanism.

Materials and methods

Dataset and data processing

The original dataset consists of 626,275 rows and 131 columns. Each row concerns a single purchase and an aggregated history of all previous purchases of the customer who made it. The target variable, churn, indicates whether another purchase will be made by the same customer in the future. Preprocessing, conducted using Pandas 0.25.1 library installed under Python 3.7.4, included the removal of duplicates, redundant columns, and outliers. Principal component analysis was used as a dimensionality reduction technique to represent highly correlated variables (abs(Spearmann correlation) > 0.8). After those steps, class imbalance was very high, as the data consisted of 79% of rows with churn=1. Thus, random undersampling was used to achieve class balance. Finally, the data, having 152,456 rows and 113 columns, (112 predictors: *base_price,discount,n_products,previous_Winter_hats,previous_Football_accessories,previous_Dresses,previous_DKNY,previous_Polo_shirts,previous_Training_shoes,previous_Stripes,previous_Lifestyle_shoes,previous_Swimsuits,previous_Care_products,previous_Wallets,previous_Gloves_and_scarves,previous_Running_shoes,previous_Backpacks,previous_Gucci,previous_Hilfiger,previous_Winter_coats,previous_Bags,previous_Shirts,previous_Casual_shoes,previous_Skirts,previous_Tops,previous_Trainers,previous_Balls,previous_Vests,previous_Basketball_shoes,previous_Jeans,previous_Sandals,previous_Underwear,previous_Tennis_shoes,previous_Outdoor_shoes,previous_Autumn_jackets,previous_Accessories,previous_Tracksuits,previous_Slippers,previous_Hiking_boots,previous_Trousers,previous_Glasses,previous_Training_accessories,previous_Sweaters,previous_Sweatshirts,previous_Shoes,previous_Shorts,previous_Clothes,previous_Hats,previous_Football_boots,previous_Armani,previous_Football_clothing,previous_Fleece,previous_Versace,previous_Socks,previous_Calvin_Klein,previous_Lauren,Shoes,Lauren,Running_shoes,Calvin_Klein,DKNY,Hilfiger,Clothes,Hats,Gucci,Lifestyle_shoes,Accessories,Tops,Sweatshirts,Trousers,Shorts,Versace,Casual_shoes,Armani,Stripes,Shirts,Skirts,Tennis_shoes,Glasses,Backpacks,Trainers,Socks,Slippers,Winter_hats,Autumn_jackets,Sandals,Bags,Football_accessories,Balls,Training_shoes,Outdoor_shoes,Football_boots,Basketball_shoes,Hiking_boots,Care_products,Underwear,Wallets,Winter_coats,Fleece,Tracksuits,Gloves_and_scarves,Swimsuits,Polo_shirts,Football_clothing,Sweaters,Dresses,Jeans,Training_accessories,Vests,transaction_date,days_since_fir*

st_purchase,value_previous_transaction, and target variable: churn), was standarized to be centered at zero and to have a unit variance.

In order to use a recurrent network topology, the data needed to be represented as a time series. Therefore, each row was transformed into a two-step series, with the first step including data on previous purchases and the second – on the current purchase. Some variables, such as the date of the purchase, were inadequate for such a transformation and were represented as two steps with identical values. The target variable was stored as a separate, 1-dimensional vector. The new dataset consisted of 152,456 time series with 2 steps and 58 features.

Model tuning

The prediction of customer churn was performed using two base artificial neural network topologies. A multilayer perceptron (MLP), with one or two fullyconnected dense layers was used. Also recurrent layer as a first hidden layer (RNN), optionally supported by an additional dense layer was used. Particular numbers of neurons were preliminarily selected by comparing the accuracy and F1 scores of models of different widths. When multiple models performed similarly, the simpler one was selected. Each network was optimized using a binary cross-entropy loss function. The output layer used a sigmoid activation function. For the purpose of overfitting prevention, each model was augmented with an extra dropout after every hidden layer. Both versions of the model, with and without dropout, were trained and compared. In case the “dying ReLU” problem appeared, each model was trained in two versions – using the standard rectified linear unit activation function and using the Leaky ReLU activation function. All models were prepared and trained using Keras 2.3.1 library with TensorFlow 2.0.0 backend [www.tensorflow.org].

Learning

A 10-fold split was performed over a dataset. Each model was trained independently 10 times, using consecutive data sections as validation sets and the remaining parts as training sets. The batch size amounted to 10,000 randomly selected rows. The models consisting of only fully-connected layers were trained over 40 epochs. The models containing recurrent layers were trained over 60 epochs. Model accuracy on the training and validation set was measured after each epoch. After the last epoch, additional metrics were calculated.

Experimental results

The trained models were used to predict samples from the validation set. Based on these predictions, a set of metrics was calculated and presented in Table 1. Accuracy, precision, recall, and confusion matrices were calculated for the prediction threshold equaling 0.5. For each row, the metrics were averaged over 10 independent models of the same architecture (but different dataset split using the 10 folds). The first column describes the number, type, and width of hidden layers. The second column indicates the probability of dropout for each hidden layer. The third column concerns the activation function used. The next columns contain the averaged metric values and their standard errors, in parentheses. The last column presents an averaged confusion matrix of the model, denoted using estimated probability values in percentages.

Table 1. Topologies of the models and corresponding metric values.

Network architecture	Dropout	Activation function	Average metric value ± SE					Averaged confusion matrix (%)
			Train accuracy	Test accuracy	Precision	Recall	AROC	
Dense(4)	0.0	ReLU	0.737 (0.001)	0.736 (0.001)	0.769 (0.003)	0.674 (0.004)	0.803 (0.002)	40 10 16 34
Dense(4)	0.3	ReLU	0.734 (<0.001)	0.733 (0.001)	0.742 (0.001)	0.714 (0.004)	0.807 (0.001)	38 12 14 36
Dense(4)	0.0	Leaky ReLU	0.735 (0.001)	0.735 (0.001)	0.765 (0.003)	0.678 (0.005)	0.807 (0.001)	40 10 16 34
Dense(4)	0.3	Leaky ReLU	0.732 (<0.001)	0.732 (0.001)	0.775 (0.002)	0.655 (0.003)	0.806 (0.001)	40 10 17 33
Dense(4), Dense (2)	0.0	ReLU	0.736 (0.001)	0.735 (0.001)	0.771 (0.002)	0.669 (0.002)	0.807 (0.002)	40 10 17 33
Dense(4), Dense (2)	0.3	ReLU	0.731 (<0.001)	0.730 (0.001)	0.733 (0.006)	0.726 (0.008)	0.805 (0.001)	37 13 14 36
Dense(4), Dense (2)	0.0	Leaky ReLU	0.737 (0.001)	0.736 (0.001)	0.770 (0.002)	0.672 (0.002)	0.810 (0.001)	40 10 16 33
Dense(4), Dense (2)	0.3	Leaky ReLU	0.733 (<0.001)	0.733 (0.001)	0.752 (0.002)	0.694 (0.002)	0.807 (0.001)	39 11 15 35
Recurrent(6)	0.0	ReLU	0.740 (0.001)	0.739 (0.001)	0.779 (0.002)	0.668 (0.002)	0.813 (0.001)	40 10 17 33
Recurrent(6)	0.3	ReLU	0.736 (0.001)	0.735 (0.001)	0.797 (0.003)	0.631 (0.002)	0.810 (0.002)	42 08 18 32
Recurrent(6)	0.0	Leaky ReLU	0.739 (0.001)	0.739 (0.002)	0.767 (0.003)	0.685 (0.002)	0.811 (0.002)	40 10 16 34
Recurrent(6)	0.3	Leaky ReLU	0.733 (<0.001)	0.732 (0.001)	0.780 (0.001)	0.647 (0.003)	0.807 (0.001)	41 09 18 32
Recurrent(6), Dense(4)	0.0	ReLU	0.739 (0.001)	0.739 (0.001)	0.778 (0.004)	0.668 (0.006)	0.812 (0.001)	40 10 17 33
Recurrent(6), Dense(4)	0.3	ReLU	0.739 (0.001)	0.737 (0.001)	0.777 (0.003)	0.665 (0.004)	0.811 (0.001)	40 10 17 33
Recurrent(6), Dense(4)	0.0	Leaky ReLU	0.739 (0.001)	0.739 (0.001)	0.777 (0.003)	0.669 (0.004)	0.812 (0.001)	40 10 17 33
Recurrent(6), Dense(4)	0.3	Leaky ReLU	0.733 (<0.001)	0.733 (0.001)	0.776 (0.002)	0.655 (0.002)	0.807 (0.001)	40 10 17 33

In order to compare the results of the 16 models, a series of statistical tests was performed, all with significance level α=0.05. The introductory testing for normality, performed using Shapiro-Wilk test, revealed non-normality in some groups for every metric. A Levene test suggested a heterogeneity of variance of the training accuracy between the results of the models. The results of those tests implied further use of non-parametric methods. A Kruskal-Wallis test was used to check the equality of medians between the groups, while A Dunn's test with a Holm adjustment was applied for the post-hoc analysis. A Levene test with a Bonferonni correction was used for pairwise comparisons of variance. Simultaneous inference of multiple metrics was performed using a Friedman's test, with a Nemenyi test used for the post-hoc analysis.

The resulting models were of similar quality, with the global average accuracy of 73.6%. Accuracy was significantly less volatile than precision (Nemenyi p=1.4x103) and recall (p=2.6x105). AROC had lesser variance than recall, with marginal significance (p=0.02).

The Friedman test revealed a marginally significant difference of the combined metrics between the models (p=0.01). The Nemenyi post-hoc test suggested only one relevant difference (p=0.05), as the one-layer recurrent model without dropout and with ReLU activation performed better than one-layer MLP with 30% dropout and Leaky ReLU, considering all the measured metrics.

Although only two models were compellingly different in overall performance, there were 92 significant differences between particular metrics. The Dunn's test sug-gested 30 differences in the training accuracy between models. They overlapped with the only 5 significant differences in test accuracy. There were 27 relevant differences between models in precision and 26 – in recall. AROC values consistently differed between 4 models. Table 2 presents the results of the Dunn's test. Each pair of values indicate the number of models that performed, respectively, worse and better than the competitive model.

Table 2. Topologies of the models and aggregated results of the Dunn's post-hoc analysis.

Network architecture	Dropout	Activation function	Average metric value ± SE				
			Train accuracy	Test accuracy	Precision	Recall	AROC
Dense(4)	0.0	ReLU	+2, -0	+0, -0	+0, -1	+1, -0	+0, -2
Dense(4)	0.3	ReLU	+0, -3	+0, -0	+0, -8	+6, -0	+0, -0
Dense(4)	0.0	Leaky ReLU	+0, -0	+0, -0	+0, -1	+2, -0	+0, -0
Dense(4)	0.3	Leaky ReLU	+0, -7	+0, -0	+2, -0	+0, -4	+0, -0
Dense(4), Dense (2)	0.0	ReLU	+1, -0	+0, -0	+0, -1	+1, -0	+0, -0
Dense(4), Dense (2)	0.3	ReLU	+0, -8	+0, -5	+0, -8	+6, -0	+0, -2
Dense(4), Dense (2)	0.0	Leaky ReLU	+2, -0	+0, -0	+0, -1	+1, -0	+0, -0
Dense(4), Dense (2)	0.3	Leaky ReLU	+0, -5	+0, -0	+0, -6	+4, -0	+0, -0
Recurrent(6)	0.0	ReLU	+6, -0	+1, -0	+3, -0	+0, -0	+2, -0
Recurrent(6)	0.3	ReLU	+0, -0	+0, -0	+8, -0	+0, -9	+0, -0
Recurrent(6)	0.0	Leaky ReLU	+4, -0	+1, -0	+0, -1	+4, -0	+0, -0
Recurrent(6)	0.3	Leaky ReLU	+0, -5	+0, -0	+3, -0	+0, -5	+0, -0
Recurrent(6), Dense(4)	0.0	ReLU	+6, -0	+1, -0	+3, -0	+1, -2	+2, -0
Recurrent(6), Dense(4)	0.3	ReLU	+4, -0	+1, -0	+3, -0	+0, -2	+0, -0
Recurrent(6), Dense(4)	0.0	Leaky ReLU	+5, -0	+1, -0	+3, -0	+1, -0	+0, -0
Recurrent(6), Dense(4)	0.3	Leaky ReLU	+0, -2	+0, -0	+2, -0	+0, -4	+0, -0

Most models had a similar variance of all measured metrics. Precision was significantly less volatile ($p=0.03$) in the model with one recurrent and one dense layer, with 30% dropout and Leaky ReLU activations, than in the model with one dense layer, without dropout and with ReLU activation function. The model with two dense layers, 30% dropout, and Leaky ReLU has a lower variance of training accuracy than the same model without dropout ($p=0.03$), and than the recurrent model without a dense layer and dropout, with Leaky ReLU activation ($p=0.04$). Also, the model with one dense layer, 30% dropout, and Leaky ReLU activation had a significantly lower variance than the model with two dense layers, no dropout, and Leaky ReLU, with $p=0.04$.

Conclusions

The experiment was performed over real e-commerce data in an industry where 75% of buyers are one-off customers. It means that such a number of customers made a purchase only once and they have never returned to the store. In contrast, the number of regular customers (with more than 5 purchases) accounts for only 2% in the whole population. Such conditions in the e-commerce business make the input dataset unbalanced, which was mentioned in the specification of the method. It makes the churn prediction much more challenging than in any other line of business. The prediction basing on this business specificity (many one-off customers and very little regular ones), churn prediction is extremely challenging and in a natural way must be inaccurate to a certain level. Looking from another perspective, correct prediction and subsequent actions resulting with a higher customer retention are very attractive for the overall business performance. In such a case, prediction with 74% accuracy, 78% precision, and 68% recall is very promising. Even though in other business cases similar results could be considered insufficient, the achieved results

are significantly promising. The presented research has a preliminary status. The main disadvantage is using only a filter method of feature selection. The application of wrapper methods is needed for the reduction of the input attributes set. Also instead of using random sampling to generate the training and test datasets it might be interesting to develop approach to ensure that all transactions of a customer can exist only in one dataset. A very important issue also consists in the identification of the point in time at which the customer will return to the same retailer. Such an approach can better address the churn problem in a retail business due to the unclear definition of the churned customer. The research in these areas will be performed in the future works.

Funding: This research was funded by the Ministry of Science and Higher Education in Poland under the program "Regional Initiative of Excellence" 2019–2022, project number 015/RID/2018/19, total funding amount 10,721,040.00 PLN.

References

[1] E-COMERCE Homepage. https://media.pl.cushmanwakefield.com.pl/pr/444970/deweloperzy-magazynowi-i-operatorzy-logistyczni-sa-zgodni-e-commerce-r. Accessed 2021 January 04.

[2] Bhattacharya CB. When Customers Are Members: Customer Retention in Paid Membership Contexts. *Journal of the Academy of Marketing Science*. 1998 01 01;26(1):31-44. https://doi.org/10.1177/0092070398261004

[3] Kotler P, Keller KL. *Marketing Management*. 14th Edition. Pearson; 2012.

[4] Gallo A. The Value of Keeping the Right Customers. *Harvard Business Review (https://hbr.org/2014/10/the-value-of-keeping-the-right-customers)*. 2014;

[5] Lu J. Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS . SUGI 27., Paper 114-27.

[6] Chen D, Sain SL, Guo K. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*. 2012 08 27;19(3):197-208. https://doi.org/10.1057/dbm.2012.17

[7] Li X, Li Z. A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm. *Ingénierie des systèmes d information*. 2019 Nov 26;24(5):525-530. https://doi.org/10.18280/isi.240510

[8] Gordini N, Veglio V. Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*. 2017 04;62:100-107. https://doi.org/10.1016/j.indmarman.2016.08.003

[9] Oskarsdottir M, Bravo C, Verbeke W, Sarraute C, Baesens B, Vanthienen J. A comparative study of social network classifiers for predicting churn in the telecommunication industry. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2016 08. https://doi.org/10.1109/asonam.2016.7752384

[10] Amin A, Anwar S, Adnan A, Khan MA, Iqbal Z. Classification of cyber attacks based on rough set theory. *2015 First International Conference on Anti-Cybercrime (ICACC)*. 2015 First International Conference on Anti-Cybercrime (ICACC). 2015 Nov. https://doi.org/10.1109/anti-cybercrime.2015.7351952

[11] Ekinci Y, Uray N, Ülengin F. A customer lifetime value model for the banking industry: a guide to marketing actions. *European Journal of Marketing*. 2014 04 08;48(3/4):761-784. https://doi.org/10.1108/ejm-12-2011-0714

[12] Ngai E, Xiu L, Chau D. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*. 2009 03;36(2):2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021

[13] Amin A, Al-Obeidat F, Shah B, Adnan A, Loo J, Anwar S. Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*. 2019 01;94:290-301. https://doi.org/10.1016/j.jbusres.2018.03.003

[14] Gordini N. Market-Driven Management: A Critical Literature Review. *Symphonya. Emerging Issues in Management*. 2010 Dec 01;(2). https://doi.org/10.4468/2010.2.08gordini

[15] Çelik Ö, Usame OO. Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*. 2019;4(1).

[16] Burez J, Van den Poel D. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*. 2007 02;32(2):277-288. https://doi.org/10.1016/j.eswa.2005.11.037

[17] Gordini N, Veglio V. Using neural networks for customer churn predictionmodeling: preliminary findings from the italian electricity industry. *Proceedings de X° Convegno Annuale della Società Italiana Marketing: "Smart Life. Dall'Innovazione Tecnologica al Mercato", Università degli Studi di Milano-Bicocca, Italy*. 2013, 1–13.

[18] Gordini N, Veglio V. Customer relationship management and data mining: A classification decision tree to predict customer purchasing behavior in global market. In: *Handbook of Research on Novel Soft Computing Intelligent Algorithms: Theory and Practical Applications*. Vol. 1-2. 2013:1–40.

[19] Verbeke W, Martens D, Mues C, Baesens B. Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*. 2011 03;38(3):2354-2364. https://doi.org/10.1016/j.eswa.2010.08.023

[20] Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*. 2012 04;218(1):211-229. https://doi.org/10.1016/j.ejor.2011.09.031

[21] De Caigny A, Coussement K, De Bock KW. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*. 2018 09;269(2):760-772. https://doi.org/10.1016/j.ejor.2018.02.009

[22] Deng Z, Lu Y, Wei KK, Zhang J. Understanding customer satisfaction and loyalty: An empirical study of mobile instant messages in China. *International Journal of Information Management*. 2010 08;30(4):289-300. https://doi.org/10.1016/j.ijinfomgt.2009.10.001

[23] Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*. 2006 05;43(2):204-211. https://doi.org/10.1509/jmkr.43.2.204

[24] Xu S, Lai S, Qiu M. Privacy preserving churn prediction. *Proceedings of the 2009 ACM symposium on Applied Computing - SAC '09*. the 2009 ACM symposium. 2009. https://doi.org/10.1145/1529282.1529643

[25] Sharma A, Kumar Panigrahi P. A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. *International Journal of Computer Applications*. 2011 08 31;27(11):26-31. https://doi.org/10.5120/3344-4605

[26] Buckinx W, Van den Poel D. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*. 2005 07;164(1):252-268. https://doi.org/10.1016/j.ejor.2003.12.010

[27] Coussement K, Van den Poel D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques.

Expert Systems with Applications. 2008 01;34(1):313-327. https://doi.org/10.1016/j.eswa.2006.09.038

[28] Xie Y, Li X, Ngai E, Ying W. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*. 2009 04;36(3):5445-5449. https://doi.org/10.1016/j.eswa.2008.06.121

[29] Ahmed AA, Maheswari D. Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal*. 2017 Nov;18(3):215-220. https://doi.org/10.1016/j.eij.2017.02.002

[30] Abbasimehr H, Setak M, Tarokh MJ. A Neuro-Fuzzy Classifier for Customer Churn Prediction. *International Journal of Computer Applications*. 2011 Apr;19(8):35-41.

[31] Yu X, Guo S, Guo J, Huang X. An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications*. 2011 03;38(3):1425-1430. https://doi.org/10.1016/j.eswa.2010.07.049

[32] Huang Y, Kechadi T. An effective hybrid learning system for telecommunication churn prediction. *Expert Systems with Applications*. 2013 Oct;40(14):5635-5647. https://doi.org/10.1016/j.eswa.2013.04.020

[33] Pełka M, Rybicka A. Identification of factors that can cause mobile phone customer churn with application of symbolic interval-valued logistic regression and conjoint analysis. The 13th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. 2019, 187–195.

[34] Tamaddoni Jahromi A, Stakhovych S, Ewing M. Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*. 2014 Oct;43(7):1258-1268. https://doi.org/10.1016/j.indmarman.2014.06.016

[35] Owczarczuk M. Churn models for prepaid customers in the cellular telecommunication industry using large data marts. *Expert Systems with Applications*. 2010 06;37(6):4710-4712. https://doi.org/10.1016/j.eswa.2009.11.083

[36] Hur Y, Lim S. Customer Churning Prediction Using Support Vector Machines in Online Auto Insurance Service. In: Wang J, Liao XF, Yi Z, eds. *Advances in Neural Networks – ISNN 2005*. 3497. Berlin, Heidelberg: Springer; 2005. https://doi.org/https://doi.org/10.1007/11427445_149

[37] Miguéis V, Van den Poel D, Camanho A, Falcão e Cunha J. Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*. 2012 09;39(12):11250-11256. https://doi.org/10.1016/j.eswa.2012.03.073

[38] Farquad M, Ravi V, Raju SB. Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing*. 2014 06;19:31-40. https://doi.org/10.1016/j.asoc.2014.01.031

[39] Slof D, Frasincar F, Matsiiako V. A competing risks model based on latent Dirichlet Allocation for predicting churn reasons. *Decision Support Systems*. 2021 07;146:113541. https://doi.org/10.1016/j.dss.2021.113541

24th International Conference on Business Information Systems (BIS 2021)

Big Data

https://doi.org/10.52825/bis.v1i.44

Published: 02 July 2021

Developing a legal form classification and extraction approach for company entity matching

Benchmark of rule-based and machine learning approaches

Felix Kruse [1][https://orcid.org/0000-0001-8033-0840], Jan-Philipp Awick [1], Jorge Marx Gómez [1], and Peter Loos [2]

[1] University of Oldenburg

[2] DFKI & Saarland University

Abstract. This paper explores the data integration process step record linkage. Thereby we focus on the entity company. For the integration of company data, the company name is a crucial attribute, which often includes the legal form. This legal form is not concise and consistent represented among different data sources, which leads to considerable data quality problems for the further process steps in record linkage. To solve these problems, we classify and ex-tract the legal form from the attribute company name. For this purpose, we iteratively developed four different approaches and compared them in a benchmark. The best approach is a hybrid approach combining a rule set and a supervised machine learning model. With our developed hybrid approach, any company data sets from research or business can be processed. Thus, the data quality for subsequent data processing steps such as record linkage can be improved. Furthermore, our approach can be adapted to solve the same data quality problems in other attributes.

Keywords: Record Linkage, Company Entity Matching, Data Integration, Data Quality, Data Preparation.

Motivation and problem statement

Companies try to integrate data in their decision-making processes in the most efficient way to achieve corporate added value. The cyclical process of the information value chain describes this approach of the companies. First, data is transferred into information and then into knowledge. This knowledge is used in decision-making processes and subsequent actions to generate added value for the company [1]. The information advantage becomes a crucial part of a company's economic success. Heinrich and Stühler [2] showed that companies that integrate relevant data directly into their decision-making processes are more competitive [2]. For example, data about competitors, suppliers, or corporate customers may contain such company and competition relevant information. However, this information is often hidden in multiple external and internal data sources [3–5]. In many cases, only the combination of external and internal data sources leads to interesting, novel, valuable, and unexpected insights that provide a competitive advantage [4–6].

The data integration goal is to provide unified access to these external and internal data sources [6]. The data integration process consists of the process steps (1) schema matching, (2) record linkage (RL), and (3) data fusion to achieve this goal. The schema matching step serves to identify the attributes that have the same meaning [6]. The RL step matches data records from different data sources that refer to the same real-world entity such as companies, products, or persons [7, p. 3-4]. The data fusion step determines the valid values

of the respective attributes of the matched record [6]. While the data integration goal is easy to formulate it is still hard to achieve [6]. RL is a crucial task in the data integration process and has become a sub-discipline of data science due to its complexity and similarity to classical data science tasks [8–10]. The complexity is caused when unique identification numbers among the data sources are missing and other existing attributes are very heterogeneous. If no identification number exists, RL must be performed using additional attributes in the data sources. For competitors, suppliers, or enterprise customers, these are the company name, the address, or the company description [11]. For RL, market participants such as competitors, suppliers, or corporate customers represent the real-world entity company. RL is still very messy in practice since there are many data sources containing different real-world entities, this leads to many RL scenarios with several challenges [8]. Köpcke *et al.* [12] try to reduce the multitude of RL scenarios' complexity by focusing on the real-world entity product [12]. Our RL research focuses on the real-world entity company. We define this as company entity matching. We have identified several RL challenges within the existing attributes company name, address data and company description [13]. We identified these challenges through our data-driven inductive research method [14]. This method describes our approach to analysing our eleven existing data sources (see table 1) and integrating various of them through a RL process to find general RL challenges for the real-world entity company.

Table 1: Company data sources for inductive data-driven research

Data source	Source
Handelsregister	https://offeneregister.de/
OpenCorporates	https://opencorporates.com/
Crunchbase ODM	https://data.crunchbase.com/docs/open-data-map
Crunchbase Snapshot	https://data.crunchbase.com/docs/2013-snapshot
GLEIF	https://www.gleif.org/en
USPTO	https://developer.uspto.gov/
Wikidata	https://www.wikidata.org/
Uscompanylist - Company	https://www.uscompanieslist.com/
Uscompanylist - Business	https://www.uscompanieslist.com/
AlphaVantage	https://www.alphavantage.co/
Owler	https://corp.owler.com/

One of the most relevant attributes in company entity matching is the company name [15, 16] which we will focus on in this paper. The legal form of a company is also an important attribute, as it is discriminatory when comparing companies [15]. In our eleven data sources (see table 1), the company legal form is always contained in the company name attribute, as the nine examples in table 2 show. The company name contains the company's legal form and thus is not atomic. This leads to the problem that the attribute legal form cannot be directly analysed without further data preparation efforts. Wang and Strong [17] formalize this as a not concise representation of the data and thus as a data quality problem. Besides, the company legal form is often represented inconsistently. Table 2 shows nine different representations for the German company legal form "GmbH". The nine records show punctuation problems, upper- and lower-case problems, abbreviation problems, and umlaut problems. The legal form is not always at the end of a company name (ID 5), and the legal form tokens can be separated by tokens of the company name (ID 9). The consistent representation is also defined as a data quality dimension by Wang and Strong [17]. The inconsistent representation of the legal form leads to the problem that, for example, the analysis of a particular legal form like the "GmbH" requires much effort. In addition, company names can be represented differently in various databases due to the inconsistent representation of the legal form. This makes the company entity matching more difficult. The two data quality problems of concise and consistent representation of the legal form are even more complicated as various legal forms exist for each country in the world.

Table 2: Different representation of the legal form "GmbH" in the company name

ID	Company_name
1	Selbstfahrer Union **G.m.b.H.**
2	GIANT Weilerswist g21 **GmbH**
3	FABIUS Vermietungs**gesellschaft mbH**
4	Infrastrukturentwicklungs**gesellschaft** Hilden **mbH**
5	ITM & C **GmbH** International Trade Marketing & Consulting
6	FHS Gabelstapler **Gesellschaft mit beschränkter Haftung**
7	bunse aufzuege **gesellschaft mit beschraenkter haftung**
8	alint 458 grundstueckverwaltung **gesellschaft m.b.h.**
9	**gesellschaft** zur verwertung von leistungsschutzrechten **mit beschraenkter haftung** gvl

Figure 1 shows the company entity matching challenges when the legal form is included in the company name, which we identified through our inductive data-driven experiments, and the matching when the company name and legal form are split into different attributes. First, we discuss the matching problems when the company name and legal form are within the same attribute. There are two data sources, A and B, with two companies that differ only in their legal form. As a human being, it is obvious that the tuples with the ID's C100 and 2 and C101 and 1 belong to the same entity. The classic string similarity measures such as normalized Levenshtein, Jarowinkler, Jaccard, or Soft TF/IDF [18] do not provide exact results to determine match and no-match tuples. The highest values of the normalized Levenshtein distance would classify the two non-match tuples as matches. The highest values of the Jarowinkler Distance would combine a match (ID C101 and 1) and a non-match (ID C100 and 1). The Jaccard distance does not distinguish the tuples. The Soft TF/IDF distance does not distinguish the tuples for the company with the ID C100. For the company with the ID C101, the higher Soft TF/IDF would be the match. With this small example, the problem for company entity matching with the legal form within the company name attribute is shown. However, with the second example shown in figure 1 where company name and legal form are split into two separate attributes, all string similarity measures show a similarity of 100% for the cleaned name, but the legal form is only the same for the matches. This allows the correct tuples to be selected as matches. This shows that a data preparation approach is needed to split the company name into the attributes company name without legal form (cleaned name) and company legal form (legal form) to achieve our goal.

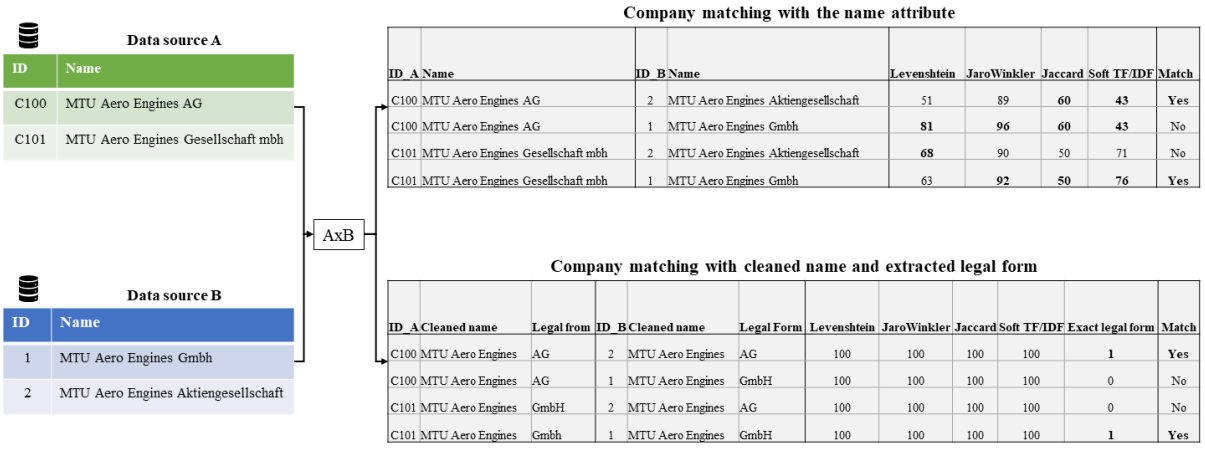


Figure 1: Example of the legal form problem with company entity matching

Our paper therefore aims to develop an approach that classifies and extracts the company name's legal form to improve the data quality and support further data processing steps such as the RL. Note that we focus on the German legal forms as a starting point for our research. Based on this introduction and problem statement, we address the research question:

"Which approach is appropriate to classify and extract the legal form in the company name?" To answer the research question, we follow the inductive data-driven research approach according to [14]. This approach seems to us to be most suitable for data science research, because Maass *et al.* [19] defines data-driven research as "an exploratory approach that analyses data to extract scientifically interesting insights (e.g., patterns) by applying analytical techniques and modes of reasoning". To carry out the research approach, we iteratively identified, implemented, and evaluated potential approaches and analysed the data to extract scientifically insights about the best performing approach. We have tried to improve the approaches or identify new approaches until the results were acceptable. We present the results of the developed approaches in a summarising benchmark.

The paper is structured as follows. Section 2 describes the related work. In section 3, we present our four identified and implemented approaches for the benchmark. In section 4, we describe and analyse the results of our conducted benchmark. In section 5, we present the theoretical and practical implications and the limitations of our paper. The paper ends in section 6 with a conclusion and outlook.

Related Work

The process steps (1) data preparation, (2) blocking, (3) record pair comparison, (4) classification, and (5) evaluation perform RL [7, p. 24, 20]. The literature review by Kruse *et al.* [20] shows that the focus of current research in the field of RL is primarily on the process step classification and the entire RL process for a given data source pair.
In general RL research, there are RL approaches that achieve high F1-scores [21–23] on existing RL datasets provided mainly by the Magellan project [8, 24]. Nevertheless, we cannot compare our research with these results because the used datasets do not consider the RL challenge of company legal form that we identified. Most of the datasets are used to link the real-world entities product or person in which the company legal form does not exist. Only one dataset is used to link the real-world entity company[1]. Mudgal *et al.* [22] classify the dataset as a textual dataset because the dataset has only a unstructured company description as an attribute. This attribute does not have the RL problem we identified with the company legal form, as we narrow this problem down to the structured attribute company name.

Since no benchmark dataset exists to test our approach against other RL processes, we initially classify the work in the research area of company entity matching and data preparation in RL. This paper deals with the process step data preparation, which has been little researched in the context of company entity matching. The papers identified in the areas of company entity matching and data preparation for RL are presented below.

Company Entity Matching

We have identified the papers of Schild and Schultz [15], Cuffe and Goldschlag [25], and Gschwind *et al.* [16] as research papers that focus specifically on company entity matching. Schild and Schultz [15] present in their paper a self-developed RL process to integrate different data sources containing companies for research purposes of the Deutsche Bundesbank. In the paper, seven data sources are used. Two data sources are provided by external data providers Bureau van Dijk and Hoppenstedt/Bisnode. Five are internal data sources of the Bundesbank. The attributes company name, legal form, postal code, city, and street were used for RL. Schild and Schultz [15] describe the company name as the most important attribute to distinguish company entities. The company name's distinctiveness can be enhanced by geographical additions or the legal form in the company name. For Schild and Schultz [15], the most important attribute for comparing companies is their legal form. They have developed a set of rules consisting of regular expressions to classify the legal form. The set of rules classifies only the german legal forms. Schild and Schultz (2017) research results show the importance of company matching for subsequent analytical use

[1] https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md#company

cases and the impact of the legal form. In the paper, a very static set of rules for the classification of legal forms was implemented. The approach cannot extract the legal form from the company name. Our approach aims to classify the legal form and to generate a company name without the legal form.

Cuffe and Goldschlag [25] address the problem that there are many individual RL methods and approaches and try to consolidate them in a framework called MAMBA (Multiple Algorithm Matching for Better Analytics). They focus on the entity company and use Census microdata. MAMBA's focus is on applying different string similarity measures in machine learning methods to improve RL results. Cuffe and Goldschlag [25] do not focus on data preparation and thus do not deal with the classification and extraction of the company name's legal form.

Gschwind *et al.* [16] focus on company entity matching to integrate data sources needed for further processing, such as data analytics. The attributes company name, location, and industry are used. The result of the paper is a practical end-to-end system. They used a rule-based approach for the RL process step (4) classification. They developed a machine learning (ML) method for the data preparation process step, which generates a short company name from the company name. For example, the ML method extracts the short company name "Aston Martin" from the company name "Aston Martin Lagonda Limited". The authors defined this problem as a sequence labeling task and trained a conditional random field algorithm to identify and extract the company short name. Gschwind *et al.* [16] do not deal with the company legal form in their ML method to extract the short company name. The company's legal form rarely appears in the company's short name in their case. Their first approach to the RL process deleted the legal form like "inc." or "ltd. However, they observed that some companies only differ in their legal form. As a solution, they have given less weight to the legal form in their scoring process. The paper does not describe how the legal form is identified to give it a lower weight. Neither is the legal form extracted to use it as an additional attribute in the RL process for the record pair comparison.

Data Preparation for Record Linkage

Randall *et al.* [26] and Koumarelas *et al.* [27] focus on the RL process step data preparation. Randall *et al.* [26] examine the effect of data preparation on RL quality. Based on a review by Linkage Software, they identified a set of different data preparation procedures. They applied them to a synthetic dataset and a real administrative dataset to compare RL quality with and without data preparation. The results show that data preparation has little impact on RL quality. The paper does not consider the entity company. The data preparation methods used were very general. The authors themselves say that additional data sets need to be evaluated to make a final statement about the negative or positive impact of data preparation on RL quality. Randall *et al.* [26] call in their outlook that further research should be conducted in more specialized processing of name and address attributes.

Koumarelas *et al.* [27] present a process to select the data preparation methods that improve RL quality the most. The process should contribute to the comparability of future evaluation results in RL research since the description of data preparation is not yet sufficient for this purpose, according to Koumarelas *et al.* [27]. The data preparation procedures considered by Koumarelas *et al.* [27] do not refer to the company name or legal form. General data preparation methods such as "split attributes", "remove special characters" or address related data methods are considered. There is no focus on the entity company for the RL process.

Approaches to classify and extract the legal form

The related work shows no data preparation approach to classify the legal form of a company and extract it from the company name. Furthermore, it shows that specific data preparation can lead to a general increase in data quality and an increase in RL quality. We

have adapted and further developed the approaches Bundesbank and Cleanco and developed two completely new approaches, we called them Deep Learning approach and Hybrid approach, and benchmarked them to classify and extract German company legal forms from company names. In the following, these four developed approaches are described.

Adapted rule-based approach from Bundesbank

The Bundesbank approach is a rule-based approach to classify the company legal form based on [15]. Despite the restriction that the approach only classifies the legal form and does not extract it, it should be included in the benchmark. Schild and Schultz [15] have described the regular expressions in the appendix of their paper and the set of rules to combine certain regular expressions to determine the legal form. The syntax of the regular expressions described in the paper is PERL. They implemented the following German legal forms "GmbH", "AG", "SE", "KG", "OHG", "UG", "GbR", "e.V.", "e.G.", "KGaA", "VVaG.", "GmbH & Co. KG", "GmbH & Co. KGaA", "GmbH & Co. OHG" and "SE". We have implemented the regular expressions and the set of rules in Python. Figure 2 shows the regular expression in Python syntax for the legal form "GmbH". This example illustrates how complicated the regular expressions are to read and extend. For example, regular expressions were developed to classify the legal form "GmbH" and "KG" and the addition "& Co.". If the three regular expressions are found together in a company name, the set of rules classifies a "GmbH & Co. KG". Due to the high modeling effort required to extend the Bundesbank approach e.g. with the legal form extraction function and to add more legal forms, we focused on improving the necessary manual effort in the next approaches.

```
patternGMBH = "(GMBH)|(GmbH)|(G|g)?(E|e)?(S|s)?\.?(ELL|ell)?\.?(SCH|sch)?\.?(AFT|aft)?
              ?(M|m)\.?(IT|it)? ?(B|b)(ESCHR|eschr)?\.?(Ä|ä)?(AE?|ae?)?(NKTER|nkter)?
              ?(H|h)( |,|\.|$|AFTUNG|aftung)"
```

Figure 2: Regular expression for the classification of legal form

Adapted rule-based approach by Cleanco

The Cleanco approach is based on the Github project Cleanco. We identified this project through an internet search for approaches to classify and extract legal forms. Cleanco is a Python-based package that identifies the legal form, removes it, and returns a cleaned company name. Cleanco is based on a rule-based approach. The Bundesbank approach presented in section 3.1 could only classify German legal forms. The Cleanco set of rules contains legal forms from 66 different countries. In our benchmark. Since Cleanco contains German legal forms it was considered in the benchmark. By default, Cleanco has only implemented the German legal forms 'gmbh & co. kg', 'gmbh & co. kg', 'e.g.', 'e.v.', 'gbr', 'ohg', 'partg', 'kgaa', 'gmbh', 'g.m.b.h.' and 'ag'. Besides, Cleanco standardizes all legal forms from different countries to the English legal forms. For example, a "GmbH" is classified as its English equivalent "Limited".

To enable a benchmark with the other approaches, we have adapted and expanded the German legal forms in the Cleanco rules. The legal form "gmbh & co. kg" is implemented in the Cleanco Standard package but is missing in our Cleanco rule set. Since the current implementation of Cleanco cannot classify legal forms consisting of several tokens like "gmbh & co. kg" caused by the technical implementation. For this reason, we had to remove all legal forms that consist of several tokens. Due to the high modelling effort required to remove the technical restriction of classifying legal forms that only consist of one token ("GmbH" works but "GmbH & Co. KG" does not) we have focused on another approach.

Deep Learning (DL) Approach

To implement the deep learning (DL) approach we define the legal form classification and extraction as a sequence labeling problem, such as part-of-speech tagging or named entity recognition [28]. A Sequence Labeling Problem exists if a label from a defined label set is

assigned to each token of a sequence [28]. In our case, the sequence of tokens is the company name. These tokens are to be labeled whether they belong to a specific legal form or not. For sequence labeling, a tagging scheme has to be chosen [29]. In our case, we have chosen the conventional BIO tagging scheme [29, 30]. A starting tag (B), an inner tag (I), and an outside tag (O) is defined. Each of the 27 legal forms (see table 5) is provided with a beginning tag (B-legal form) and an inner tag (I-legal form). All tokens which do not belong to a legal form are assigned the tag (O). An example is shown in table 3.

Table 3: Example of the Sequence Labeling Schema

Name	kuhn	gmbh	facilities	management		
Label	O	B-Gmbh	O	O		

Name	agl	maschinenbau	gesellschaft	mit	beschraenkter	haftung
Label	O	O	B-Gmbh	I-Gmbh	I-Gmbh	I-Gmbh

We created a sample of 10,000 company names based on the GLEIF, Crunchbase ODM, and OpenCorporates databases (see Table 1) to create the training data set. We filtered the databases for German companies. We used the labeling tool *doccano* to label the 10,000 company names with our BIO tagging scheme [31]. We have identified other legal forms such as "Stiftung" or "EK" during the labeling process that are not implemented in the previous approaches Bundesbank and Cleanco. We have created a balanced labelled data set with 18.300 company names. The neural network architectures is a classical bi-directional LSTM (BI-LSTM), often used for sequence labeling problems [28, 30]. The labeled company data set was divided into 80% training data and 20% test data. The BI-LSTM with the best parameter settings achieved an F1 score of about 99.2% on the test data. The DL approach delivers good results but has problems with some legal forms, such as the "gGmbH" and the "PartG", which are often wrongly classified as "GmbH". In addition, sequence labeling according to the BIO tagging scheme requires a high manual effort, as each token in the company name has to be tagged with a label. In order to solve the problems mentioned and expand other legal forms in the future, a high manual labeling effort is necessary. For these reasons, we have developed another approach that should achieve the same or better results, involves less label effort and allows the input of domain knowledge to solve the problems with legal forms such as the "gGmbH".

Hybrid: Rule-based with Machine Learning

The Hybrid approach consists of a rule-based and a supervised ML component to perform the classification and extraction of legal forms from the company name. In the past, rule-based systems were used for the classification of texts. Today, ML approaches are increasingly used. The rules of the rule-based approaches need to be set up manually, which often results in high effort and complexity [32, 33]. In contrast, supervised ML algorithms enable the automated creation of complex sets of rules based on massive amounts of data. However, the algorithms require sufficiently labeled data to learn the rules, which is a one-time manual effort. One advantage of rule-based approaches is that humans can apply their domain knowledge directly when creating a set of rules. This makes the set of rules easy to understand and extensible for humans [32]. The legal form classification and extraction problem demonstrate that legal forms' inconsistent representation and diversity require special domain knowledge. While analysing and labeling the data, we identified other legal forms such as "EK" or "Stiftung", which are not implemented in the rule-based approaches Bundesbank (section 3.1) and Cleanco (section 3.2). Also, we identified other representations for the individual legal forms such as "g.m.b.h." or "o.h.g." that are not implemented in the existing approaches. To solve the classification and extraction problem of legal forms, we combine rule-based components and ML methods to take advantage of both. For this purpose, we divide the legal form classification and extraction problem into the subtasks: (1) identification of legal form relevant tokens, (2) classification of the legal form based on the legal form relevant tokens, and (3) extraction of the legal form relevant tokens

from the company name. The data flow and the solution approach for the hybrid approach's subtasks are shown in figure 3. They are described below:

(1) Identification of legal form relevant tokens: For the legal form's classification, only the legal form tokens in the company name are relevant. For the company name "Example gesellschaft mbh", these are the tokens "gesellschaft" and "mbh". Since we have already established that the diversity of existing legal forms and the inconsistent representation of the individual legal forms requires domain knowledge, we implemented an identification rule set to solve this subtask. The rule set consists of a list of all tokens that are part of a legal form, such as "ek", "eg", "ag" "aktiengesellschaft" or "gmbh". Experts can easily extend this list. With the list's help, all tokens relevant to the legal form of a company name are extracted. The legal form is classified based on the extracted tokens relevant to the legal form in the next step.

(2) Classification of the legal form based on the legal form relevant tokens: For the classification of the legal form based on the extracted legal form relevant tokens we use ML approaches, since the manual creation of a rule set would be very complex and time consuming.

We used and compared the ML methods of Random Forest Tree and Support Vector Classifier (SVC). The labeled data set of the DL approach (section 3.3) is used as training data. The dataset was extended by 500 samples, in which no legal form is included in the company name. Also, the represented legal form was extracted from the BIO labels as a single label. The entire data set thus comprises 18800 training samples. The extracted components are encoded by multi-label binarization. This results in a vector that contains a 1 for each recognized component of the created list and a 0 for all others. With this vector, the two methods were trained with 80% of the data and evaluated with 20% of the data. The results are shown in table 4.

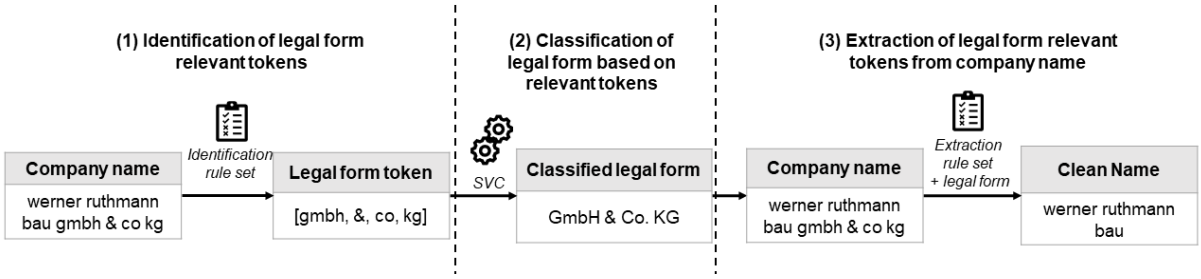


Figure 3: Data flow of the hybrid approach

Table 4 shows that the quality of the models is very high, with over 99%. The SVC shows with a weighted F1 score of 99.7% the better performance than the random forest tree with 99.57%. That both models achieve excellent results shows that the classification of a legal form from the extracted legal form relevant tokens works very reliably. However, the most correct and complete extraction of the legal form components from the company name is decisive for the good performance, which has a significant influence on the classification's success. This shows that combining a rule-based extraction of the legal form relevant tokens and the classification with an ML method is very successful for this application. Finally, we have chosen the SVC as the classifier of the hybrid approach.

Table 4: Results Models for legal form classification

Model	Precision	Recall	F1-Score
Random Forest	0.9963	0.9952	0.9957
SVC	0.9969	0.9971	0.9970

Furthermore, this approach's extensibility is easy to implement by domain experts extending the list of legal-form-relevant tokens. Besides, the effort for labeling new data sets is less than with the DL approach, since it not necessary to label according to the BIO tagging scheme, but rather it is sufficient to label only the legal form belonging to a company name.

(3) Extraction of the legal form relevant tokens from the company name: A rule-based approach does the extraction of the legal form relevant tokens. We have maintained an extraction rule set containing a list of tokens for each legal form that should be searched for and removed from the company name. The rule-based approach needs the previously classified legal form as input. The dependence of the extraction on the classified legal form is an essential condition for the hybrid approach. For example, the following company, "Meyer Gesellschaft Stiftung" is classified as a "Stiftung". In this case, the token "gesellschaft" belongs to the company name and is not a legal form relevant token. If all current legal form relevant tokens would be extracted during the extraction, the token "gesellschaft" beside the "Stiftung" would be erroneously removed from the company name. In our approach, the tokens to be removed depending on the classified legal form. Thus, we ensure that only the tokens belonging to the legal form are removed. In our example, the token “Gesellschaft” is not included in the extraction rule set for the legal form “Stiftung”, so only the token "Stiftung" is removed.

Benchmark of the Approaches

We have created a new labeled data set out of our data sources (see table 1) containing 3733 company names (see table 5). This dataset is used to benchmark the four different approaches for classifying and extracting the company's legal form. This data set is unknown for all approaches to evaluate the quality of the four approaches.

Benchmark data set

When creating a real data set for evaluation, we made sure that the evaluation data set does not contain any company names that have already been used for training the approaches. It was also essential for us to create a real evaluation data set and ensure that all legal forms appear in the data set. Our final dataset contains 3733 company names. These were manually labeled again with the defined BIO tagging scheme (see table 3) in the labeling tool *doccano* [31]. The frequency of each legal form in the evaluation dataset is shown in table 5 (column amount).

Table 5: Benchmark results for every approach with the exact match ratio

		Classification				Extraction		
Legal form	**Amount**	Bundesbank	Cleanco	DL	Hybrid	Cleanco	DL	Hybrid
OHG	345	0.919	0.971	0.980	**0.997**	0.910	**0.962**	0.933
GmbH	324	0.919	0.809	0.969	**0.997**	0.895	**0.966**	0.935
GmbH & Co. KG	307	0.694	0.000	0.958	**0.977**	0.697	**0.945**	0.926
No legal form	287	0.526	**0.969**	0.892	0.937	**0.969**	0.857	0.937
Aktiengesellschaft	257	0.743	0.728	0.969	**0.981**	0.938	0.961	**0.965**
EG	247	0.194	0.854	0.988	**1.000**	0.883	**0.980**	0.960
SE	237	0.958	**1.000**	**1.000**	0.992	0.987	**0.996**	0.987
EK	220	0.000	0.768	0.959	**0.991**	0.900	**0.995**	0.973
Stiftung	216	0.000	0.977	0.958	**0.972**	0.403	0.958	**0.972**
EV	202	0.787	0.832	0.842	**0.990**	0.787	**0.911**	0.896
GbR	198	0.727	0.899	0.934	**1.000**	0.854	**0.949**	0.939
UG	153	0.850	0.935	0.980	**0.987**	0.046	**0.980**	0.974
VVaG	133	0.023	0.571	0.609	**0.774**	0.609	0.602	**0.632**
UG & Co. KG	103	0.660	0.000	0.971	**0.990**	0.000	**0.893**	0.796
KG	84	0.905	0.929	0.976	**1.000**	0.929	0.929	**0.976**
GmbH & Co. KGaA	81	0.790	0.000	0.852	**0.877**	0.000	**0.889**	0.864
gGmbH	79	0.000	0.443	0.418	**0.835**	0.468	0.418	**0.810**

PartG	73	0.000	0.178	0.589	**0.849**	0.096	0.452	**0.521**
GmbH & Co. OHG	59	0.814	0.000	0.831	**0.932**	0.000	0.797	**0.831**
SE & Co. KG	49	0.000	0.000	**1.000**	0.959	0.000	**0.980**	0.959
Stiftung & Co. KG	28	0.000	0.000	**1.000**	**1.000**	0.036	**1.000**	0.929
KGaA	22	0.818	**0.864**	**0.864**	**0.864**	**0.818**	**0.818**	**0.818**
AG & Co. KGaA	11	**0.727**	0.000	**0.727**	**0.727**	0.000	**0.818**	**0.818**
Limited & Co. KG	6	0.833	0.000	0.833	**1.000**	0.000	**0.833**	**0.833**
SE Co. KGaA	5	0.000	0.000	**0.800**	**0.800**	0.000	**0.800**	**0.800**
AG & Co. KG	4	0.750	0.000	**1.000**	0.750	0.000	**1.000**	0.750
AG & Co. OHG	2	**1.000**	0.000	**1.000**	**1.000**	0.000	**1.000**	**1.000**
SE & Co. OHG	1	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
Summary	3733	0.589	0.696	0.919	**0.962**	0.705	0.913	**0.916**

Execution and analysis of the benchmark

The result of the benchmark is shown in table 5. It was divided into legal form classification (classification of the correct legal form) and extraction (were all legal form components identified in the company name). The Bundesbank approach (section 3.1) is only included in evaluating the classification, as it does not extract the legal form. The benchmark was performed for each legal form. The ratio of correctly classified companies to the total number of companies per legal form was calculated as the so-called exact match ratio.

The exact match ratio for the classification was calculated using the ratio of correctly classified companies to the total number of companies per legal form. The legal form's extraction was evaluated as correct if all legal form elements were extracted from the company name. For the extraction, the exact match ratio was thus calculated from the ratio of correctly extracted legal forms to the respective total number of companies.

Overall, the hybrid approach for the classification and extraction of the legal form is the best of the four approaches. It achieves an exact match ratio of 0.962 for classification and 0.916 for extraction. The DL approach is slightly worse. With an exact match ratio for the classification of 0.919, the DL approach is 4.3% behind the Hybrid approach. For extraction, the DL approach is only 0.3% behind the Hybrid approach. The Bundesbank approach achieved an exact match ratio of 0.589 for the classification. Cleanco achieved an exact match ratio of 0.696 for the classification. For the extraction, Cleanco is with an exact match ratio of 0.705 over 20% behind the DL and Hybrid approach.

In general, the Hybrid approach has a 2-3% better Exact Match Ratio per legal form than the DL approach for the task classification. For the legal forms "EV" or "gGmbH" the Hybrid approach has a 14.8% and 41.7% better exact match ratio. The DL approach often classifies the legal form "EV" as "No legal form" or "EK", which results in a difference of 14.8% of the Exact Match Ratio. For the "gGmbH" legal form, the DL approach often classifies a "GmbH", which results in the 41.7% worse exact match ratio. With the legal forms "SE" and "SE & Co. KG", the DL approach has a 0.8% and 4.1% better exact match ratio than the hybrid approach. In some cases, the hybrid approach classifies an "SE" as "SE & Co. KG" and vice versa, which results in the difference of the exact match ratio. The Bundesbank and Cleanco approach only achieve for the legal forms "KGaA", "AG & Co. KGaA" and "AG & Co. OHG" the same exact match ratio as the DL or hybrid approach. The Bundesbank and Cleanco approaches' rules do not reflect the diverse representations within a legal form to the same extent as the DL and Hybrid approaches. From this, it can be concluded that the legal forms "KGaA", "AG & Co. KGaA" and "AG & Co. OHG" do not show a high diversity in the evaluation data since the approaches have the same exact match ratio. The legal form "Stiftung" is represented very consistently in the evaluation data, as the Cleanco approach has the best exact match ratio of 0.977. The hybrid approach has a 0.5% lower exact match ratio for the same legal form. For the label "No legal form" Cleanco has a 3.2% better exact match ratio than the Hybrid approach. The Cleanco approach covers fewer legal form

variants (see table 3). Cleanco generally classifies more records as "No legal form", which explains the difference.

In the extraction, the difference in the exact match ratio between the DL approach and the Hybrid approach is 0.3%. The differences in the exact match ratio per legal form are minimal as well. In 14 cases, the exact match ratio of the DL approach is better than the Hybrid approach. In 7 cases, the Hybrid approach is better than the DL approach. In 5 cases, the exact match ratios of the two approaches are equal. For the legal form "gGmbH", the exact match ratio difference between the DL approach and the Hybrid approach is 39.2%, which is significantly higher than the others. The DL approach classifies some records with the legal form "gGmbH" as "GmbH" and therefore does not extract the legal form correctly. As a result, the DL approach has a 39.2% worse exact match ratio. The only case where neither the DL approach nor the Hybrid approach has the best exact match ratio is for the label "No legal form". For this label, Cleanco shows the best exact match ratio with 0.969.

Discussion

Theoretical and Practical Implications

The results of our benchmark, which approach is suitable for classifying the company legal form and extracting it from the company name, directly influences theory and practice. First, our developed Hybrid approach increases the data quality of company names and company legal forms in company databases. Our application example for our developed data preparation approach is the company entity matching. Here we show with our research that the classification and extraction of the company legal form is a general problem and exists in many data sources. So far, no benchmark dataset for company entity matching exists which contains this RL challenges. We show that this problem can be solved with our Hybrid approach consisting of a set of rules and a supervised ML method, or our DL approach. Thus, we confirm and support the statements of Govind *et al.* [8] and Gschwind *et al.* [16] that ML procedures should be used for subtasks in RL and thus support the automation of the RL process. This statement is confirmed by our approach and encourages us to identify further general problems in RL and data preparation and investigate suitable ML solutions for these problems. For example, the standardization and matching of company address data. Furthermore, the results of our paper show that it is appropriate for RL to consider problems for the respective real-world entities such as products, persons or companies.

Our developed data preparation approaches, Hybrid and DL, can be used for any new scientific and company data source. We show that general data quality problems with the concise and consistent representation of attributes could be solved with such approaches. The approach could be further explored theoretically and practically and applied to other attributes with similar data quality problems as concise and consistent representation.

Limitations

Our research has limitations that lead to potential future research opportunities. In our paper, we focus on German legal forms. Further research should investigate the extension of the DL and Hybrid approach to include other legal forms. In doing so, the extendibility requirement and performance should be measured. A different approach may be necessary for each country and its legal forms in the future.

We have selected the listed data sources due to our focus on German legal forms (table 1). Future research should investigate additional data sources and additional evaluation data sets. The evaluation of the hybrid and DL approach with other data sources could provide further insights into which approach is the better one under which conditions.

In the future, the approach should be also used in real RL experiments to investigate how much influence this data preparation procedure has on company entity matching results. In addition, a benchmark dataset for company entity matching should be created in order to benchmark existing RL approaches.

Conclusion and Outlook

The entity company is present in many internal and external data sources and is often required in analytical use cases. Therefore, the different internal and external data sources need to be integrated. The integration of the data sources is enabled by the data integration process, which consists of the process steps (1) schema matching, (2) record linkage (RL), and (3) data fusion [7]. In this paper, we focus on the RL of the real-world entity company and define this as company entity matching. In company entity matching, the company name is crucial and presents several challenges. The legal form is often included in the company name and is also an important discriminative attribute. Since the legal form is not a separate attribute in most data sources, it cannot be directly analysed for further data processing steps.

Moreover, the legal form lacks data quality, as it is often not concise and consistent represented in the company name. For the German legal form "GmbH" we show 9 different representations (see table 2). Our goal to solve the data quality problems is to classify and harmonize the legal form and to split the company name and legal form into two separate attributes. To achieve this goal, we answer the following research question in our paper*: "Which approaches are suitable to automatically classify and extract the legal form in the company name?"*. We answer the research question through our inductive data-driven research procedure, according to Grover and Lyytinen [14]. As a result, we have iteratively developed four approaches to solve the problem, which we present and evaluate in a summarising benchmark. The first approach, called Bundesbank, is rule-based and is adapted by the paper by Schild and Schultz [15]. The second approach, called Cleanco, is also rule-based and is adapted on the Github project Cleanco. The third approach, called Deep Learning (DL), defines the legal form classification and extraction problem as a sequence labeling problem and solves it with a Bi-LSTM deep learning model. The fourth approach, called Hybrid, is a combination of a rule set for identification and extraction of legal form relevant tokens and a supervised ML algorithm for the classification of the legal form. The benchmark data set contains 3733 records. The Hybrid approach achieves the best values in the benchmark with an exact match ratio of 96.2% for the legal form classification and 91.6% for the legal form extraction. The DL approach achieved the second-best values with 91.9% for classification and 91.3% for extraction. Thus, the Hybrid shows the best performance in the benchmark. Further, experts can easily extend the developed rule sets, meaning the Hybrid approach is easier to expand than the DL approach. Likewise, additional training data sets can be labeled with new legal forms to extend the classification model. The labeling of new training data sets for the DL approach is more complex since all tokens of the company name must be labeled. In contrast, the supervised ML method in the Hybrid approach requires only one label for the company name.

Our approach and results show that general problems exist for the individual real-world entities such as companies represented in different data sources. For these general-entity-specific problems, generic solutions can be created to improve the data quality, such as concise and consistent representation of attributes. Furthermore, our results show that using hybrid ML methods or DL approaches is successful for these problems and should be further researched. In future research, the developed data preparation approach will be used in RL processes to measure the impact in company based RL case studies.

References

[1] A. Abbasi, S. Sarker, and R. Chiang, “Big Data Research in Information Systems: Toward an Inclusive Research Agenda,” *JAIS*, vol. 17, no. 2, pp. I–XXXII, 2016, doi: 10.17705/1jais.00423.

[2] C. Heinrich and G. Stühler, “Die Digitale Wertschöpfungskette: Künstliche Intelligenz im Einkauf und Supply Chain Management,” in *Fallstudien zur Digitalen Transformation :*

Case Studies für die Lehre und praktische Anwendung, Wiesbaden, Germany: Springer Gabler, 2018, pp. 77–88. https://doi.org/10.1007/978-3-658-18745-3_4

[3] M. Stonebraker and I. Ilyas, "Data Integration: The Current Status and the Way Forward," *IEEE Data Eng. Bull.*, vol. 41, no. 2, 3-9, 2018.

[4] P. Christen, "Data Linkage: The Big Picture," *Harvard Data Science Review*, 2019, doi: 10.1162/99608f92.84deb5c4.

[5] F. Kruse, C. Schröer, and J. Marx Gómez, "Data Source Selection Support in the Big Data Integration Process - Towards a Taxonomy," in *Internationale Tagung Wirtschaftsinformatik (WI)*, Universität Duisburg-Essen, 2021.

[6] X. L. Dong and D. Srivastava, "Big Data Integration," *Synthesis Lectures on Data Management*, vol. 7, no. 1, pp. 1–198, 2015, doi: 10.2200/S00578ED1V01Y201404DTM040.

[7] P. Christen, *Data Matching*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 10.1007/978-3-642-31164-2

[8] Y. Govind *et al.,* "Entity Matching Meets Data Science: A Progress Report from the Magellan Project," 2019, https://doi.org/10.1145/3299869.3314042

[9] N. Barlaug and J. Atle Gulla, "Neural Networks for Entity Matching: A Survey," 2020, arXiv:2010.11075

[10] Y. Govind *et al.,* "Cloudmatcher: a hands-off cloud/crowd service for entity matching," *Proc. VLDB Endow.*, vol. 11, no. 12, pp. 2042–2045, 2018, doi: 10.14778/3229863.3236255.

[11] P. Christen and W. E. Winkler, "Record Linkage," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer US, 2016, pp. 1–10.

[12] H. Köpcke, A. Thor, S. Thomas, and E. Rahm, "Tailoring entity resolution for matching product offers," in *Proceedings of the 15th International Conference on Extending Database Technology - EDBT '12*, Berlin, Germany, 2012, p. 545.

[13] P. Behnen, F. Kruse, and J. Marx Gómez, "Enhancement of Record Linkage by Using Attributes containing Natural Language Text," in *AAAI-MAKE 2021 Combining Machine Learning and Knowledge Engineering*, Stanford University, Palo Alto, California, USA, 2021, pp. 1–14.

[14] V. Grover and K. Lyytinen, "New State of Play in Information Systems Research: The Push to the Edges," *MISQ*, vol. 39, no. 2, pp. 271–296, 2015, doi: 10.25300/MISQ/2015/39.2.01.

[15] C.-J. Schild and S. Schultz, "Linking Deutsche Bundesbank Company Data using Machine-Learning-Based Classification," 2017, doi: 10.1145/2951894.2951896.

[16] T. Gschwind, C. Miksovic, J. Minder, K. Mirylenka, and P. Scotton, "Fast Record Linkage for Company Entities," in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 623–630., 10.1109/BigData47090.2019.9006095

[17] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996, doi: 10.1080/07421222.1996.11518099.

[18] N. Gali, R. Mariescu-Istodor, D. Hostettler, and P. Fränti, "Framework for syntactic string similarity measures," *Expert Systems with Applications*, vol. 129, pp. 169–185, 2019, doi: 10.1016/j.eswa.2019.03.048.

[19] W. Maass, J. Parsons, S. Purao, V. C. Storey, and C. Woo, “Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research,” *JAIS*, pp. 1253–1273, 2018, doi: 10.17705/1jais.00526.

[20] F. Kruse, A. P. Hassan, J.-P. Awick, and J. Marx Gómez, “A Qualitative Literature Review on Linkage Techniques for Data Integration,” in *53nd Hawaii International Conference on System Sciences, HICSS 2020, Grand Wailea, Maui, Hawaii, USA, January 7-10, 2020*, 2020, pp. 1063–1073. [Online]. 10.24251/HICSS.2020.132

[21] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan, “Deep Entity Matching with Pre-Trained Language Models,” arXiv:2004.00584, 2020.

[22] S. Mudgal *et al.,* “Deep Learning for Entity Matching,” in *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18*, Houston, TX, USA, 2018, pp. 19–34.

[23] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, and N. Tang, “Distributed representations of tuples for entity resolution,” *Proc. VLDB Endow.*, vol. 11, no. 11, pp. 1454–1467, 2018, doi: 10.14778/3236187.3236198.

[24] A. Doan *et al.,* “Magellan: Toward Building Ecosystems of Entity Matching Solutions,” *Commun. ACM*, vol. 63, no. 8, pp. 83–91, 2020, doi: 10.1145/3405476.

[25] J. Cuffe and N. Goldschlag, “Squeezing More Out of Your Data: Business Record Linkage with Python,” in 2018.

[26] S. M. Randall, A. M. Ferrante, J. H. Boyd, and J. B. Semmens, “The effect of data cleaning on record linkage quality,” *BMC medical informatics and decision making*, vol. 13, pp. 1–10, 2013, doi: 10.1186/1472-6947-13-64.

[27] I. Koumarelas, L. Jiang, and F. Naumann, “Data Preparation for Duplicate Detection,” *Journal of Data and Information Quality (JDIQ)*, vol. 1, no. 1, pp. 1–24, 2020, https://doi.org/10.1145/3377878

[28] A. Akhundov, D. Trautmann, and G. Groh, “Sequence Labeling: A Practical Approach,” *CoRR*, arXiv:1808.03926, 2018.

[29] S. Liu, B. Tang, Q. Chen, and X. Wang, “Drug Name Recognition: Approaches and Resources,” *Information*, vol. 6, no. 4, pp. 790–810, 2015, doi: 10.3390/info6040790.

[30] X. Zhong, E. Cambria, and A. Hussain, “Extracting Time Expressions and Named Entities with Constituent-Based Tagging Schemes,” *Cogn Comput*, vol. 12, no. 4, pp. 844–862, 2020, doi: 10.1007/s12559-020-09714-8.

[31] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang, *doccano: Text Annotation Tool for Human.* [Online]. Available: https://github.com/doccano/doccano

[32] Julio Villena Roman, Sonia Collada-Perez, Sara Lana-Serrano, and Jose C. González-Cristobal, “Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization,” 2011.

[33] Fabrizio Sebastiani, “Machine Learning in Automated Text Categorization,” *ACM computing surveys (CSUR)*, 2002. [Online], https://doi.org/10.1145/505282.505283 1

24th International Conference on Business Information Systems (BIS 2021)
Big Data
https://doi.org/10.52825/bis.v1i.45

Published: 02 July 2021

Predict COVID-19 Spreading with C-SMOTE

Alessio Bernardo[1][https://orcid.org/0000-0002-3492-0345], Emanuele Della Valle[1][https://orcid.org/0000-0002-5176-5885]

[1] DEIB, Politecnico di Milano, Italy

Abstract. Data continuously gathered monitoring the spreading of the COVID-19 pandemic form an unbounded flow of data. Accurately forecasting if the infections will increase or decrease has a high impact, but it is challenging because the pandemic spreads and contracts periodically. Technically, the flow of data is said to be imbalanced and subject to concept drifts because signs of decrements are the minority class during the spreading periods, while they become the majority class in the contraction periods and the other way round. In this paper, we propose a case study applying the Continuous Synthetic Minority Oversampling Technique (C-SMOTE), a novel meta-strategy to pipeline with Streaming Machine Learning (SML) classification algorithms, to forecast the COVID-19 pandemic trend. Benchmarking SML pipelines that use C-SMOTE against state-of-the-art methods on a COVID-19 dataset, we bring statistical evidence that models learned using C-SMOTE are better.

Keywords: SML, Evolving Data Stream, Concept Drift, Balancing, COVID-19

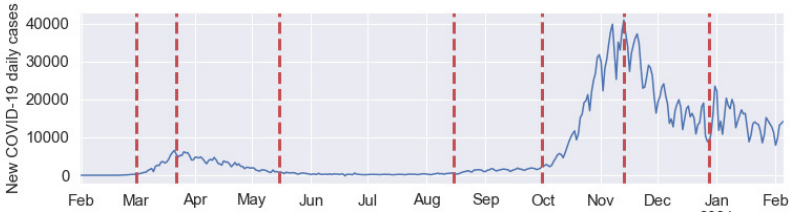
Introduction

Nowadays, a multitude of smartphones, wearables, computers, and Internet of Things (IoT) sensors produce massive, continuous, and unbounded flows of data, namely *data streams*. They pose several challenges to Machine Learning (ML) since they are impossible to load as a whole in memory, and they are often non-stationary (i.e., they present concept drifts [1]). Moreover, when they are the input for a classification problem, they present class imbalance. This is the case of data streams related to the pandemic of COVID-19.

For instance, several ML models used for forecasting sales are failing during COVID-19 because people's behaviour keeps changing. Before the disease, a large part of the people spent all the day at the office, supermarkets, bars, restaurants while only a small part of the population (minority class) stayed at home. When COVID-19 spread, it was the opposite. The majority of the people stayed at home, and only a small percentage of them went outside (new minority class). During 2020 summer/autumn and 2021 winter, such a change in people's behaviour was often observed in many countries worldwide. Those sale forecasting solutions failed because they were unable to detect concept drifts and manage minority instances. Indeed, the traditional ML techniques are not designed to monitor concept drifts, so their models are prone to introduce classification errors when they happen.

In recent years, the *Streaming Approach* (a.k.a. *Data Stream Mining* or *Online Learning*) approach was introduced to tackle those problems. Streaming Machine Learning (SML) methods can detect when concept drifts occur and adapt the model accordingly.

This paper focuses on the COVID-19 case study, aiming at predicting the daily trend in the spreading of COVID-19. It is a *streaming binary classification problem* that requires addressing



(a) New daily COVID-19 cases.

(b) COVID-19 increment/decrement cases.

Figure 1. COVID-19 pandemic spreading in Italy. Red lines are concept drifts.

both concept drift and class imbalance [2]. Indeed, when the pandemic spreads, we may simply forecast an increment trend ignoring signs of decrement. On the over way round, when the pandemic contracts or is stable, we may ignore early signs of increment. Due to the concept drifts occurrences, classes can swap, i.e. all the samples labelled as minority (majority) class before a concept drift occurrence get labelled as majority (minority) class after it. Fig. 1 shows the COVID-19 spreading in Italy. In particular, Fig. 1a shows the number of new daily cases while Fig. 1b shows COVID-19 increment/decrement cases respect to the day before. *0* means that the COVID-19 cases decreased, while *1* means that the COVID-19 cases increased. The red lines represent the concept drifts occurrences[1]. Table 1 shows the percentage of COVID-19 increment/decrement cases in each concept drift. We can see that there is a continuous class swapping.

To address this problem, we applied our novel C-SMOTE [4] SML meta-strategy, inspired by SMOTE [5] class rebalancing algorithm. Considering that in literature there are SML algorithms *natively able* to rebalance streams in presence of concept drifts (let's denote them with SML+) and algorithms *unable* to do so (say, SML-), we formulated the following **research questions**:

Q1 *does prepending* C-SMOTE *to SML- algorithms improve their performances*?
Q2 *are there pipelines of* C-SMOTE *and a SML- algorithms that outperform SML+ models*?

In more detail, the *main contributions* of this paper are statistical evidence that, also in this particular case study, prepending C-SMOTE to SML- algorithms improves the minority class performances w.r.t. both the SML-'s and SML+'s performances, and, hence, better predicts the daily trend of COVID-19 spread.

The remainder of this paper is organized as follows. Section Sampling Techniques for Class Imbalance describes the investigated problem and presents techniques able to handle it. Section C-SMOTE describes the C-SMOTE meta-strategy. Section Related Work introduces the related works. Section Experimental Settings introduces the dataset, metrics, and algorithms used in the experiments on a COVID-19 dataset and presents our research hypotheses. Section Results and Discussion shows and discusses the evaluation results. Finally, Section Conclusions discusses the conclusions and outlines directions for future research.

Table 1. % of COVID-19 increment/decrement cases in each concept drift.

Case	CD1	CD2	CD3	CD4	CD5	CD6	CD7	CD8
Increment (1)	64.52%	38.10%	50.00%	41.30%	42.55%	51.16%	64.44%	46.15%
Decrement (0)	35.48%	61.90%	50.00%	58.70%	57.45%	48.84%	35.56%	53.85%

[1]The concept drifts occurrences are calculated using the ADWIN [3] strategy.

Sampling Techniques for Class Imbalance

Imbalanced data are characterized by an unequal distribution between the classes. Since the minority class(es) instances rarely occur, the models only focus on patterns for correctly classifying the majority class(es) samples, so avoiding the ones for predicting the minority class(es) ones. The resulting problem is that the model will tend to predict all the samples as majority class ones, without really examining any of their features.

In the literature, there are four approaches for handling class imbalance [2]: sampling techniques, cost-sensitive learning, kernel-based methods, and active learning methods. In particular, sampling techniques allow changing the data distribution before the training phase. As a consequence, the algorithms can focus on cases that are more relevant to the user. For this reason, this work focuses on them: they allow a sort of meta-strategy development to prepend to any SML- model chosen by the user.

The most popular balancing technique is SMOTE [5]. For each minority class sample s, it synthetically generates new minority class points lying on the segments that join s to any/all of its *k* minority class nearest neighbours. In general SMOTE has been shown to improve classification, but it may also show drawbacks related to the way it creates synthetic samples. Specifically, SMOTE introduces new samples without considering how the majority class points are distributed into the space region, and so risking to place the new points in a majority class region and to increase the overlapping between classes. To this end, more than a hundred SMOTE variants have been proposed [6] to overcome the overlapping between classes. The most well-known are Borderline-SMOTE [7], ADASYN [8], DBSMOTE [9], MDO [10], SWIM [11], and G-SMOTE [12].

However, SMOTE it is still considered the ”de-facto” balancing technique [6]. This is the reason why C-SMOTE is based on it. Moreover, as a sampling technique, SMOTE caches the entire dataset in memory. This approach is against the basic principles of the data stream paradigm that states that a sample can be inspected only once, as fast as possible, and then discarded. In the next section we explain how to overcome this problem.

C-SMOTE

This section recalls the C-SMOTE description, inspired by SMOTE, originally presented in [4]. C-SMOTE is designed to rebalance an imbalanced data stream and, as Fig. 2 shows, it can be pipelined with any SML- technique. C-SMOTE stands for Continuous-SMOTE, meaning that the new SMOTE version is applied continuously. Its pseudocode is presented in Algorithm 1, while its implementation is available in the MOA GitHub repository[2].

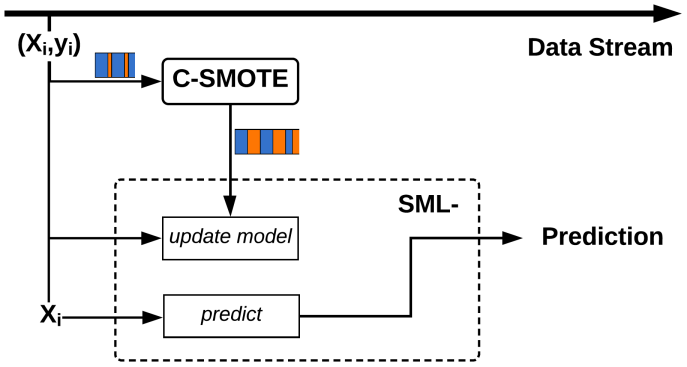


Figure 2. Architecture of SML pipelines using C-SMOTE meta-strategy.

[2] https://github.com/Waikato/moa

Algorithm 1:

```
1  Function C-SMOTE (minSizeMinority, l, rebalanceThreshold, S):
2      W, W_label ← ∅
3      S_0, S_1, S_0, S_1 ← 0
4      S_gen ← ∅
5      imbalanceRatio ← 0
6      adwin ← ∅
7      while hasNext(S) do
8          X, y ← next(S)
9          prequentialEvaluation(X, l)
10         train(X, y, l)
11         W ← add(X, y)
12         updateWindows(X, y, W_label)
13         updateCounters(y, S_0, S_1)
14         adwin ← add(y)
15         checkConceptDrift(adwin, W, W_label, S_0, S_1, S_0, S_1, S_gen)
16         W_min, S_min, S_min ← selectMinorityClass(W, W_label, S_0, S_1, S_0, S_1)
17         W_maj, S_maj, S_maj ← selectMajorityClass(W, W_label, S_0, S_1, S_0, S_1)
18         if checkMinSize(minSizeMinority, S_min) then
19             imbalanceRatio ← ratio(S_min, S_maj, S_min, S_maj)
20             while rebalanceThreshold > imbalanceRatio do
21                 X̂, ŷ ← newSample(W_min, S_gen)
22                 S_min ← S_min + 1
23                 train(X̂, ŷ, l)
24                 imbalanceRatio ← ratio(S_min, S_maj, S_min, S_maj)
25             end
26         end
27     end
28 End Function
```

As said before, the real problem is the lack of the entire data during the rebalance phase. Moreover, it is impossible to store every new sample in memory until the data stream ends for two reasons: 1) the data streams are assumed infinite, and 2) this would be against the stream paradigm approach. The solution is to save the instances into a sliding window W (Line 11) and use ADWIN [3] to keep in W only the i) necessary recently-seen and ii) consistent to the current concept samples (Lines 14-15). ADWIN keeps a variable-length window of recently seen items, with the property that the window has the maximal length statistically consistent with the hypothesis "there has been no change in the average value inside the window" [3]. Then, after having found the real minority and majority classes (Lines 17-18) and the actual $imbalanceRatio$ between the number of minority, minority generated and majority instances stored in W (Line 20), C-SMOTE is ready to rebalance the stream introducing new synthetically generated samples (Line 22). It uses the minority class samples stored in W_{min} to apply an online version of SMOTE. In this way, this one is always applied to data that are consistent with the current concept, and the newly generated samples will be consistent with it, too. SMOTE, before starting to generate new instances, calculates the number of instances to introduce for each minority sample in the batch and then it starts to execute. Instead, C-SMOTE randomly chooses one sample (X, y) from W_{min} and uses it to apply SMOTE. Then, the function, in this rebalance phase, does not use any more (X, y) to generate other synthetic instances. So, in our continuous version, not all the minority class instances are used to generate new instances. After that, at Line 24, the new instance $(\hat{X}, \hat{y})$ is used to train the learner l and the new $imbalanceRatio$ is calculated.

Related Work

To the best of our knowledge, in the literature exist only a few computing models to be adopted to predict the COVID-19 pandemic spreading [22]. Moreover, only a couple of them addresses the problem as a continuously evolving task [23], [24] without however referring to any concept drift detection or class imbalance. Instead, the methods we propose in this paper (i.e. SML+),

Table 2. The principal characteristics of the related works compared to C-SMOTE.

Method	Classifier type	Approach type	Approach for CD	Approach for Class Imbalance
RLSACP [13]	Single	Passive	Forgetting factor	Cost weight
ONN [14]	Ensemble	Passive + Active module	Forgetting factor	Cost weight
ESOS-ELM [15]	Ensemble	Passive	Weighted ensemble	Cost weight
NN [16]	Ensemble	Passive	Weighted ensemble	Cost weight
OnlineUnderOverBagging [17]	Ensemble	Passive	Weighted ensemble	Undersampling + Oversampling
OnlineSMOTEBagging [17]	Ensemble	Passive	Weighted ensemble	SMOTE
OnlineAdaC2 [17]	Ensemble	Passive	Weighted ensemble	Cost weight
OnlineCSB2 [17]	Ensemble	Passive	Weighted ensemble	Cost weight
OnlineRUSBoost [17]	Ensemble	Passive	Weighted ensemble	Undersampling
OnlineSMOTEBoost [17]	Ensemble	Passive	Weighted ensemble	SMOTE
ARE_{RE} [18]	Ensemble	Active	ADWIN	Cost weight
RB [19]	Multiple (4)	Active	ADWIN	SMOTE
OOB [20]	Ensemble	Left to pipelined algorithm	Left to pipelined algorithm	Oversampling + Cost weight
UOB [20]	Ensemble	Left to pipelined algorithm	Left to pipelined algorithm	Undersampling + Cost weight
WEOB1/WEOB2 [21]	Ensemble	Left to pipelined algorithm	Left to pipelined algorithm	Cost weight
C-SMOTE [4]	**Meta-strategy**	**Left to pipelined algorithm**	**Left to pipelined algorithm**	**SMOTE + ADWIN**

wrapped up in Table 2, are able to learn from imbalanced data stream and dealing with concept drift changes that perfectly suit this task. They are commonly categorized into two major groups: passive versus active approaches, depending on whether an explicit drift detection mechanism is employed. Passive approaches train a model continuously without an explicit trigger reporting the drift, while active approaches determine whether a drift has occurred before taking any actions. Examples of passive approaches are RLSACP [13], ONN [14], ESOS-ELM [15], an ensemble of neural network [16], OnlineUnderOverBagging, OnlineSMOTEBagging, OnlineAdaC2, OnlineCSB2, OnlineRUSBoost and OnlineSMOTEBoost [17], while ARF_{RE} [18], RebalanceStream [19] are considered active approaches. Then, there are other models like OOB and UOB [20], WEOB1 and WEOB2 [21] that, similarly to C-SMOTE, propose only an online rebalance strategy, leaving to the pipelined algorithm the concept drift management (approach type and the concept drift approach).

From Table 2, we can notice that the major part of the SML+ methods combine more learners together (ensemble strategy). Only RLSACP uses a single learner and RB uses four learners in parallel. About the approach to manage the concept drift occurrence, we can see that the most used one is to assign a weight to each learner of the ensemble and to discard the one having the lowest weight (weighted ensemble). In this way, the ensemble is composed only of the learners that perform well in the underlying concept. Only a couple of methods use forgetting factor strategies, giving, gradually, less importance to the past data and more importance to the new data in input. There are also two methods that adopt the ADWIN [3] strategy, while the last four leave this task to the pipelined algorithm. Instead, about the class imbalance management, there are different option used. Some algorithms use the cost weight strategy in which the minority class sample importance is increased in comparison to a sample from the majority class, others use SMOTE [5], while others combine together different strategies such as undersampling and oversampling or undersampling/oversampling and cost weight. The last important thing to notice is the difference with C-SMOTE. Being a meta-strategy to be pipelined with any other SML- technique, the approach type and the concept drift approach used by C-SMOTE are the ones used by the model to which is prepended by (SML-). Instead, as class imbalance approach, C-SMOTE uses the combination of ADWIN and SMOTE. In particular, in addition to the different rebalance strategy proposed, OOB, UOB, WEOB1 and WEOB2 differ from C-SMOTE in the classifier type. They all use an ensemble strategy, while C-SMOTE, being a meta-strategy, can be prepended to any type of classifier (single, multiple or ensemble).

Experimental Settings

This section has five parts. The first four ones discuss i) the dataset, ii) the algorithms, iii) the metrics, and iv) the various experimental settings used to test C-SMOTE, while the last part introduces the hypotheses to test.

Table 3. Concept drift and imbalance level summary for several countries.

Country	Concept Drift	% Increment (1)	% Decrement (0)
ITA	8	51.08%	48.92%
FRA	12	46.97%	53.03%
ESP	10	44.47%	55.53%
GBR	9	47.85%	52.15%
USA	9	53.81%	46.19%
BRA	8	38.44%	61.56%
CHN	7	32.02%	67.98%
IND	5	45.31%	54.69%

Table 4. Worldwide summary of concept drift, increment and decrement.

Measure	Min	Max	Avg	Median
Concept drift	0	12	5.42	5
% Increment (1)	0.00%	58.27%	40.96%	44.92%
% Decrement (0)	41.73%	100.00%	59.04%	55.08%

Dataset

To empirically evaluate the C-SMOTE meta-strategy in forecasting the COVID-19 pandemic trend, we used a COVID-19 dataset gathered monitoring the worldwide spreading of the pandemic [25]. The original dataset, updated with the number of new cases every day, had *59 attributes* (54 numerical and 5 nominal) indicating the country, the number of new and total cases and deaths both as absolute numbers and per million of people, the number of tests done, the number of patients in intensive care units and vaccinated people both as absolute numbers and per million of people and some population markers i.e. population number, population density, median age, hospital beds number. We replaced the original *date* attribute with the related *day, month, year, day of week, week of year and is-holiday* attributes. The last attribute states if that day is a holiday or not in that country. We also removed the *tests-units* attribute since it is only a unit of measurement. So, the final version of the dataset has *63 attributes* (54 numerical and 9 nominal). Moreover, we added a label that states if the number of new COVID-19 cases in a day are more or less than the ones that occurred in the previous day (*0* if they are less or equal, *1* if they are greater). For convenience, the *minority* class is always the class *1*, while the *majority* one is the class *0*. The overall imbalance ratio calculated at 05/02/2021 was *41.45%*. Table 3 shows, for the most common country, the number of concept drifts and the imbalance level, while Table 4 shows some statistics of all the countries.

Algorithms

As SML- models, we tested the Adaptive Random Forest (ARF) [26], Naive Bayes (NB), Hoeffding Adaptive Tree (HAT) [27], K-Nearest Neighbor (KNN) and Temporally Augmented Classifier (SWT) [28] with ARF as base learner techniques. We pipelined these algorithms with the C-SMOTE meta-strategy and compared them against the stand-alone versions. Instead, as SML+ models, we tested the ARF_{RE} [18], RB [19], OOB and UOB [20] techniques. Unfortunately, the implementations of the other algorithms cited in the Related Work section were unavailable or did not work.

Metrics

We evaluated the predictive performances using the prequential evaluation approach [29]. Following He and Garcia [2], we used the most adopted metrics in the literature to address class imbalanced learning problem. They are the *Recall* (R), *F1-Measure* (F1), and *G-mean* (GM). In particular, we computed the first two metrics separately for the minority (R[1], F1[1]) and the majority (R[0], F1[0]) class, while the latter metric is across classes and measures the balance

between the minority and majority classes performances [30]. We avoided using the *Accuracy* metric because it is not reliable in a scenario where the interest is to accurately predict the minority class occurrences. This metric would always score a high value due to the majority class samples high occurrences, deceiving the users about the goodness of the result achieved. In fact, *Accuracy* fails to reflect that all the minority class samples were misclassified. Last, we did not show the *Precision* metric results because they can be derived from the comparison of *Recall* and *F1-Measure* metrics results. We performed 10 runs, so the results proposed are the average.

Settings

All the experiments were made using the MOA framework[3] with default hyperparameters values for all the techniques involved. The only parameter that we set in C-SMOTE was $minSizeMinority$, the minimum number of minority class samples stored into the window to allow the rebalancing procedure. We used C-SMOTE with $minSizeMinority = 10$ as it was the top performer among (10,100,500,1000) in our hyper-parameter analysis. In particular, RB used four SWT models, while OOB and UOB, as a pipelined algorithm, used a Hoeffding Tree (HT) [31].

All the tests were run in a machine with 2 virtual CPUs Intel Skylake P-8175 at 2.5 GHz and 8 GiB of RAM.

Research Hypotheses

We formulate our hypotheses as follows:

- *Hp. 1*: We assume that the minority class results of binary classifiers (SML-) pipelined with C-SMOTE are statistically better than those without it.
- *Hp. 2*: We assume that the minority class results of at least one binary classifier (SML-) pipelined with C-SMOTE are statistically better than those achieved by the state of the art techniques (SML+).

Results and Discussion

In the first part of this section, we discuss the results achieved by the comparison between the pipelines of C-SMOTE and some SML- algorithms and the SML- models alone, while in the second part we compare the C-SMOTE pipelines with some SML+ techniques.

SML- Comparison

In this section, we discuss the comparison among the ARF, HAT, NV, KNN and SWT algorithms pipelined with and without the C-SMOTE meta-strategy both in term of performances comparison over time and statistical tests, checking the *Hp. 1* validity.

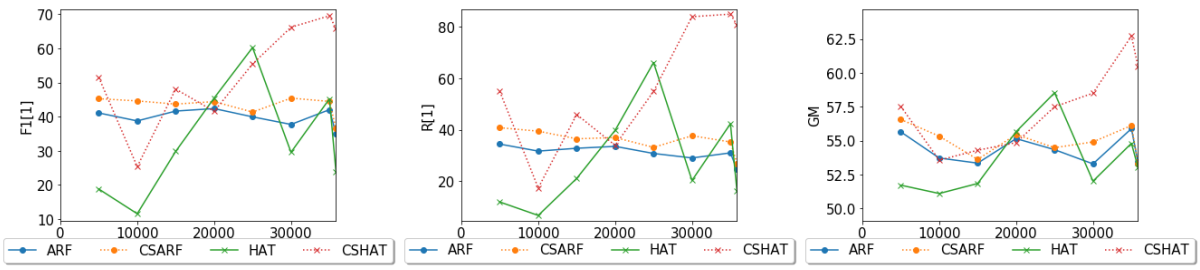
Performances Comparison

Fig. 3 shows some minority class performances comparisons over time among the ARF and HAT techniques pipelined with and without C-SMOTE. In all the minority class metrics, we can notice that both SML- models pipelined with C-SMOTE outperform their respective baselines, already verifying the *Hp. 1* hypothesis. We can also point out that there are multiple concepts drifts occurrences as noticed by the continuous performances ups and downs. However, to validate the performances of all the SML- models pipelined with and without C-SMOTE all over the metrics, we decided to perform a statistical test analysis.

Statistical Tests

To statistically prove that prepending C-SMOTE to one of the SML- methods presented in Section Algorithms improves the performances of each method, we used the one-tailed *T-Student*

[3] `https://moa.cms.waikato.ac.nz/`



(a) F1 minority class results. **(b)** Recall minority class results. **(c)** G-mean results.

Figure 3. Comparison between ARF and HAT pipelined with and without C-SMOTE.

	F1[1]					F1[0]					R [1]					R [0]					GM				
STREAM	**ARF**	**HAT**	**KNN**	**NV**	**SWT**	**ARF**	**HAT**	**KNN**	**NV**	**SWT**	**ARF**	**HAT**	**KNN**	**NV**	**SWT**	**ARF**	**HAT**	**KNN**	**NV**	**SWT**	**ARF**	**HAT**	**KNN**	**NV**	**SWT**
Covid-19																									

Ha_1 Ha_2 H_0

Figure 4. One-tail T-Student test comparing C-SMOTE pipelines with baselines.

test with the significance level $\alpha = 0.05$. We checked the *p-value*, in order to determine if there were significant differences between the results obtained with and without C-SMOTE. The null hypothesis H_0 is that, for each dataset and algorithm, the means of R[0], F1[0], R[1], F1[1] and GM are equal to the means of C-SMOTE R[0], C-SMOTE F1[0], C-SMOTE R[1], C-SMOTE F1[1] and C-SMOTE GM. We define two alternative hypotheses: Ha_1 is that the C-SMOTE means are greater than baseline ones, while Ha_2 is that the baseline means are greater than the C-SMOTE ones. The H_0 hypothesis is rejected in favor of the Ha_1 one if $\frac{\textit{p-value}}{2} < \alpha$ and the *t-statistic* < 0, while H_0 is rejected in favor of Ha_2 if $\frac{\textit{p-value}}{2} < \alpha$ and *t-statistic* > 0. Otherwise, both Ha_1 and Ha_2 are rejected in favor of H_0.

Fig. 4 shows the T-Students test results. The columns show the five metrics used and for each of them, there are five more sub-columns, indicating the comparison between the base version and the C-SMOTE version of that algorithm. The single-cell shows which hypothesis is accepted. Green cells tell that we rejected the H_0 hypothesis in favour of the Ha_1 one, red ones tell that we rejected H_0 in favour of Ha_2, while light green ones tell that both Ha_1 and Ha_2 are rejected in favour of H_0.

Definitively, we can say that *Hp. 1* is almost always verified. The use of the C-SMOTE meta-strategy improves both R[1] and GM results, meaning that the R[1] gain is bigger than the R[0] loss. Looking at the F1[1] results, in more than half of the cases, the C-SMOTE F[1] results are better than the base algorithms ones. This means that the P[1] increased too or that the R[1] gain is bigger than the P[1] loss.

SML+ Comparison

In this section, we discuss the comparison among the ARF_{RE}, RB, OOB and UOB SML+ algorithms and the ARF, HAT, NV, KNN and SWT algorithms pipelined with the C-SMOTE meta-strategy (from now on called C-SMOTE*) both in term of performances comparison over time and statistical tests, checking the *Hp. 2* validity.

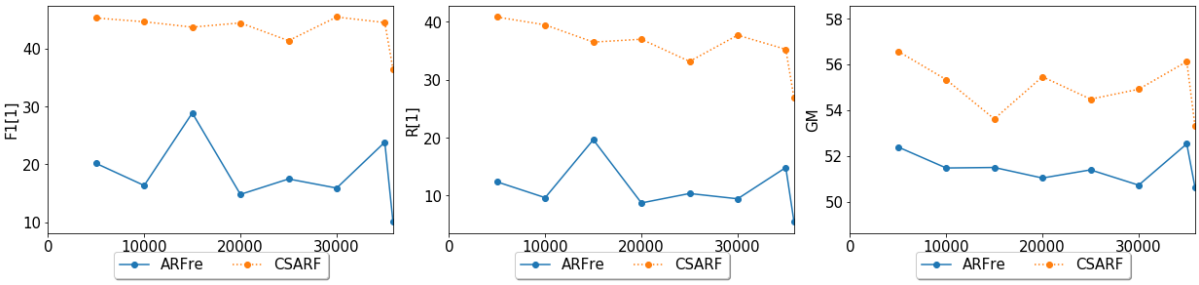
Performances Comparison

Fig. 5, Fig. 6, and Fig. 7 show some minority class performances comparisons over time among, respectively, the ARF_{RE} and ARF pipelined with C-SMOTE techniques, the OOB, UOB and HAT pipelined with C-SMOTE techniques, and the RB and SWT pipelined with C-SMOTE

techniques. The reason why we decided to show the results in this way is the similarity among the algorithms. ARF_{RE} is a new ARF version properly introduced to deal with class imbalance; OOB and UOB used the HT, the first version of HAT, as their baseline; and RB used four SWT models. Both in Fig. 5 and Fig. 6, in all the minority class metrics, we can notice that the SML- models pipelined with C-SMOTE outperform, respectively, the ARF_{RE} and the OOB and UOB techniques. Only the RB technique is better than the SWT pipelined with C-SMOTE one (Fig. 7). Also, in this case, to validate the performances of all the SML- models pipelined with C-SMOTE w.r.t. the SML+ techniques all over the metrics, we decided to perform a statistical test analysis.

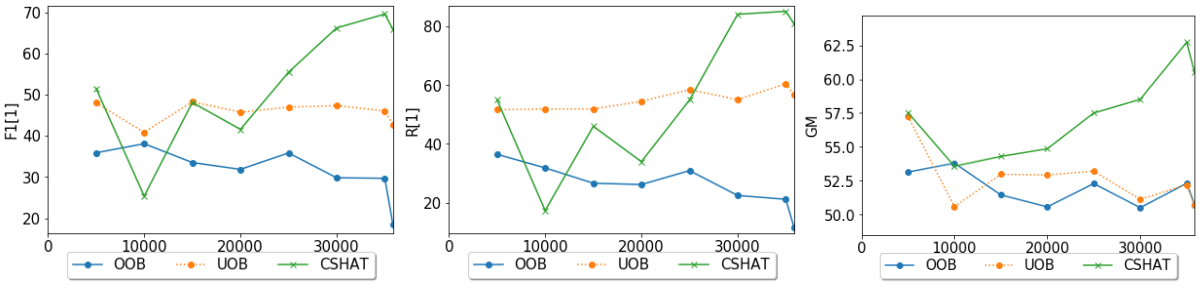
Statistical Tests

To statistically compare the performances of the SML+ algorithms with the performances achieved by C-SMOTE*, we used the same T-Student tests with the same hypotheses used before. As Fig. 8 shows, comparing C-SMOTE* with RB algorithm, HAT pipelined with C-SMOTE is the only algorithm to outperform RB in R[1], F1[1] and GM. We can notice a similar behaviour comparing C-SMOTE* with UOB algorithm, with the considerable difference that, here, C-SMOTE* always improves the GM results. Finally, comparing C-SMOTE* with ARF_{RE} and OOB algorithms, in general, the C-SMOTE* results are better. So, we can conclude that there is at least one algorithm pipelined with C-SMOTE that outperforms all the SML+ models, so *Hp. 2* is validated.



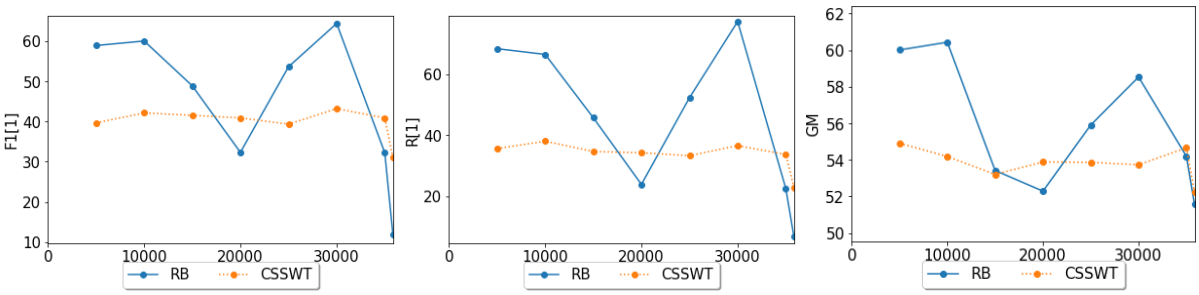
(a) F1 minority class results. **(b)** Recall minority class results. **(c)** G-mean results.

Figure 5. Comparison between ARF_{RE} and ARF pipelined with C-SMOTE (CSARF).



(a) F1 minority class results. **(b)** Recall minority class results. **(c)** G-mean results.

Figure 6. Comparison among OOB, UOB and HAT pipelined with C-SMOTE (CSHAT).



(a) F1 minority class results. **(b)** Recall minority class results. **(c)** G-mean results.

Figure 7. Comparison between RB and SWT pipelined with C-SMOTE (CSSWT).

	F1[1]					F1[0]					R [1]					R [0]					GM				
STREAM	CSARF	CSHAT	CSKNN	CSNV	CSSWT	CSARF	CSHAT	CSKNN	CSNV	CSSWT	CSARF	CSHAT	CSKNN	CSNV	CSSWT	CSARF	CSHAT	CSKNN	CSNV	CSSWT	CSARF	CSHAT	CSKNN	CSNV	CSSWT
RB Covid-19	Ha_2	Ha_1	Ha_2	Ha_2	Ha_2	Ha_1	Ha_2	Ha_1	Ha_1	Ha_2	Ha_2	Ha_1	Ha_2	Ha_2	Ha_2	Ha_1	Ha_2	Ha_1	Ha_1	Ha_1	Ha_2	Ha_1	Ha_2	Ha_2	Ha_2
OOB Covid-19	Ha_1	Ha_1	Ha_1	Ha_2	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_2	Ha_1	Ha_1	Ha_2	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1
UOB Covid-19	Ha_2	Ha_1	Ha_2	Ha_2	Ha_2	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_2	Ha_1	Ha_2	Ha_2	Ha_2	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1
ARFre Covid-19	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_2	Ha_2	Ha_2	Ha_2	Ha_2	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1	Ha_2	Ha_2	Ha_2	Ha_2	Ha_2	Ha_1	Ha_1	Ha_1	Ha_1	Ha_1

Ha_1 Ha_2 H_0

Figure 8. One-tail T-Student test comparing C-SMOTE* with RB, OOB, UOB and ARF_{RE} models.

Conclusions

In this work, we presented the application of the meta-strategy called C-SMOTE, based on the popular SMOTE technique that allows balancing an evolving data stream one sample at a time and it can be used as a data filter with all the streaming techniques, to forecasting the COVID-19 pandemic trend. Concerning SMOTE, C-SMOTE does not need a static batch during the pre-processing phase, but it saves every time the new sample in a window and, accordingly to ADWIN, it uses the minority class samples contained in it to introduce new synthetic samples.

We tested the meta-strategy pipelined with both SML- and SML+ models on a COVID-19 dataset, to demonstrate that C-SMOTE can be useful also in this emergency. Accurately forecasting if the number of infections will increase or decrease has a high impact on the economy and society overall.

The results summarized in Table 5 demonstrate that the C-SMOTE pipelines minority class results are, in most cases, better than both the ones of the SML- models alone (Q1) and the SML+ algorithms (Q2) i.e., C-SMOTE can magnify signs of decrement during the spreading periods and signs of increment in the contraction periods. We also proved that, in general, the recall of the minority class gain is bigger than the recall of the majority class loss i.e., the improvement in the ability to correctly forecast decrements (increments) when the infection is spreading (contracting) is larger than the error introduced. In particular, we can notice that the KNN, ARF, and SWT models, in this case study, are the best SML- all over the metrics to be pipelined with C-SMOTE. Thus, in light of the results achieved, we can affirm that in real-world scenarios presenting multiple concepts drifts occurrences and class imbalance, like the COVID-19 case study, C-SMOTE can enhance the performances of some SML- algorithms better than other SML+ models to the point that they can make an important statistical impact.

For future works, our principal goal is to perform memory- and time-consuming analysis. We also want to improve even more the C-SMOTE performance, reducing as much as possible the trade-off between improving the minority class performances and decreasing the majority class ones. The solution can be using rebalance techniques that consider the overlapping between classes (i.e. Borderline-SMOTE, ADASYN, DBSMOTE). Other improvements can be to adapt C-SMOTE to multiclass and regression tasks. In the long term, the aim is to investigate other meta-strategies based on different rebalance techniques and to compare them with C-SMOTE.

Table 5. Times that C-SMOTE prepended to the SML- models outperformed both the SML- models alone and the SML+ models tested. **Bold** means that it performed better in more than half of the total occurrences (5). The *Rank* column is the average of each model's rankings.

SML-	F1[1]	F1[0]	R[1]	R[0]	GM	Rank
ARF	**3 (2)**	**3 (3)**	**3 (2)**	**3 (2)**	**4 (2)**	2.2
HAT	**5 (1)**	2 (5)	**5 (1)**	1 (5)	**5 (1)**	2.6
KNN	2 (4)	**4 (1)**	**3 (2)**	**4 (1)**	**4 (2)**	**2**
NV	2 (4)	**4 (1)**	2 (5)	**3 (2)**	**4 (2)**	2.8
SWT	**3 (2)**	**3 (3)**	**3 (2)**	**3 (2)**	**4 (2)**	2.2

References

[1] A. Tsymbal, "The problem of concept drift: Definitions and related work," *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[3] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *SDM*, SIAM, 2007, pp. 443–448.

[4] A. Bernardo, H. M. Gomes, J. Montiel, B. Pfahringer, A. Bifet, and E. Della Valle, "C-smote: Continuous synthetic minority oversampling for evolving data streams," in *BigData*, In press, IEEE, 2020.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[6] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018.

[7] H. Han, W. Wang, and B. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *ICIC (1)*, ser. LNCS, vol. 3644, Springer, 2005, pp. 878–887.

[8] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *IJCNN*, IEEE, 2008, pp. 1322–1328.

[9] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "DBSMOTE: density-based synthetic minority over-sampling technique," *Appl. Intell.*, vol. 36, no. 3, pp. 664–684, 2012.

[10] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling and boosting techniques," *Soft Comput.*, vol. 19, no. 12, pp. 3369–3385, 2015.

[11] C. Bellinger, S. Sharma, N. Japkowicz, and O. R. Zaïane, "Framework for extreme imbalance classification: SWIM - sampling with the majority class," *Knowl. Inf. Syst.*, vol. 62, no. 3, pp. 841–866, 2020.

[12] G. Douzas and F. Bação, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," *Inf. Sci.*, vol. 501, pp. 118–135, 2019.

[13] A. Ghazikhani, R. Monsefi, and H. S. Yazdi, "Recursive least square perceptron model for non-stationary and imbalanced data stream classification," *Evol. Syst.*, vol. 4, no. 2, pp. 119–131, 2013.

[14] ——, "Online neural network model for non-stationary and imbalanced data stream classification," *Int. J. Machine Learning & Cybernetics*, vol. 5, no. 1, pp. 51–62, 2014.

[15] B. Mirza, Z. Lin, and N. Liu, "Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift," *Neurocomputing*, vol. 149, pp. 316–329, 2015.

[16] A. Ghazikhani, R. Monsefi, and H. S. Yazdi, "Ensemble of online neural networks for non-stationary and imbalanced data streams," *Neurocomputing*, vol. 122, pp. 535–544, 2013.

[17] B. Wang and J. Pineau, "Online bagging and boosting for imbalanced data streams," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3353–3366, 2016.

[18] L. E. B. Ferreira, H. M. Gomes, A. Bifet, and L. S. Oliveira, "Adaptive random forests with resampling for imbalanced data streams," in *IJCNN*, IEEE, 2019, pp. 1–6.

[19] A. Bernardo, E. Della Valle, and A. Bifet, "Incremental rebalancing learning on evolving data streams," in *ICDM (Workshops)*, IEEE, 2020, pp. 844–850.

[20] S. Wang, L. L. Minku, and X. Yao, "A learning framework for online class imbalance learning," in *CIEL*, IEEE, 2013, pp. 36–45.

[21] ——, "Resampling-based ensemble methods for online class imbalance learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1356–1368, 2015.

[22] I. E. Agbehadji, B. O. Awuzie, A. B. Ngowi, and R. C. Millham, "Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of covid-19 pandemic cases and contact tracing," *International journal of environmental research and public health*, vol. 17, no. 15, p. 5330, 2020.

[23] I. Arpaci, S. Alshehabi, M. Al-Emran, M. Khasawneh, I. Mahariq, T. Abdeljawad, and A. E. Hassanien, "Analysis of twitter data using evolutionary clustering during the covid-19 pandemic," *Computers, Materials & Continua*, vol. 65, no. 1, pp. 193–204, 2020.

[24] J. Farooq and M. A. Bazaz, "A novel adaptive deep learning model of covid-19 with focus on mortality reduction strategies," *Chaos, Solitons & Fractals*, vol. 138, p. 110 148, 2020.

[25] J. Hasell, E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, and H. Ritchie, "A cross-country database of covid-19 testing," *Scientific data*, vol. 7, no. 1, pp. 1–7, 2020.

[26] H. M. Gomes, A. Bifet, J. Read, J. P. Barddal, F. Enembreck, B. Pfharinger, G. Holmes, and T. Abdessalem, "Adaptive random forests for evolving data stream classification," *Mach. Learn.*, vol. 106, no. 9-10, pp. 1469–1495, 2017.

[27] A. Bifet and R. Gavaldà, "Adaptive learning from evolving data streams," in *IDA*, ser. Lecture Notes in Computer Science, vol. 5772, Springer, 2009, pp. 249–260.

[28] A. Bifet, J. Read, I. Zliobaite, B. Pfahringer, and G. Holmes, "Pitfalls in benchmarking data stream classification and how to avoid them," in *ECML/PKDD (1)*, ser. Lecture Notes in Computer Science, vol. 8188, Springer, 2013, pp. 465–479.

[29] J. Gama, R. Sebastião, and P. P. Rodrigues, "Issues in evaluation of stream learning algorithms," in *KDD*, ACM, 2009, pp. 329–338.

[30] J. Akosa, "Predictive accuracy: A misleading performance measure for highly imbalanced data," in *Proceedings of the SAS Global Forum*, vol. 12, 2017.

[31] P. M. Domingos and G. Hulten, "Mining high-speed data streams," in *KDD*, ACM, 2000, pp. 71–80.

24th International Conference on Business Information Systems (BIS 2021)

Big Data

https://doi.org/10.52825/bis.v1i.48

Published: 02 July 2021

Post-Brexit power of European Union from the world trade network analysis

Justin Loye[1,2], Katia Jaffrès-Runser[3], and Dima Shepelyansky[2]

[1]Institut de Recherche en Informatique de Toulouse,
Université de Toulouse, UPS, 31062 Toulouse, France

[2]Laboratoire de Physique Théorique, Université de Toulouse, CNRS, UPS, 31062 Toulouse, France

[3]Institut de Recherche en Informatique de Toulouse,
Université de Toulouse, Toulouse INP, 31071 Toulouse, France

Abstract. We develop the Google matrix analysis of the multiproduct world trade network obtained from the UN COMTRADE database in recent years. The comparison is done between this new approach and the usual Import-Export description of this world trade network. The Google matrix analysis takes into account the multiplicity of trade transactions thus highlighting in a better way the world influence of specific countries and products. It shows that after Brexit, the European Union of 27 countries has the leading position in the world trade network ranking, being ahead of USA and China. Our approach determines also a sensitivity of trade country balance to specific products showing the dominant role of machinery and mineral fuels in multiproduct exchanges. It also underlines the growing influence of Asian countries.

Keywords: International trade, Google matrix, Complex networks

Introduction

The European Union (EU) is now composed from 27 countries and is considered as a major world leading power [1]. January 2021 has seen Brexit officially taking place, triggering the withdrawal of the United Kingdom (UK) from EU [2]. This event has important political, economical and social effects. Here we project and study its consequences from the view point of international trade between world countries. Our analysis is based on the UN COMTRADE database [3] for the multiproduct trade between world countries in recent years. From this database we construct the world trade network (WTN) and evaluate the influence and trade power of specific countries using the Google matrix analysis of the WTN. We consider 27 EU countries as a single trade player having the trade exchange between EU and other countries. Our approach uses the Google matrix tools and algorithms developed for the WTN [4, 5, 6, 7] and other complex directed networks [8, 9]. The efficiency of the Google matrix and PageRank algorithms is well known from the World Wide Web network analysis [10, 11].

Our study shows that the Google matrix approach (GMA) allows to characterize in a more profound manner the trade power of countries compared to the usual method relying on import and export analysis (IEA) between countries. GMA's deeper analysis power originates in the fact that it accounts for the multiplicity of transactions between countries while IEA only takes into account the effect of one step (direct link or relation) transactions. In this paper, we show that the world trade network analysis with GMA identifies EU as the first trade player in the world, well ahead of USA and China.

This paper is structured in the following way. Section 1 introduces first the UN COMTRADE dataset, and then gives a primer on the tools related to Google matrix analysis such as the trade balance metric and the REGOMAX algorithm. In Section 2, the central results of this papers are presented, which are discussed in 3.

1 Data sets, algorithms and methods

We use the UN COMTRADE data [3] for years 2012, 2014, 2016 and 2018 to construct the trade flows of the multiproduct WTN following the procedure detailed in [5, 6]. This paper gives the results for year 2018 only, others are to be found at [12]. Each year is presented by a money matrix, $M^p_{cc'}$, giving the export flow of product p from country c' to country c (transactions are expressed in USD of current year). The data set is given by $N_c = 168$ countries and territories (27 EU countries are considered as one country) and $N_p = 10$ principal type of products (see the lists in [4, 6]). These 10 products are: Food and live animals (0); Beverages and tobacco (1); Crude materials, inedible, except fuels (2); Mineral fuels etc (3); Animal and vegetable oils and fats (4); Chemicals and related products, n.e.s. (5); Basic manufactures (6); Machinery, transport equipment (7); Miscellaneous manufactured articles (8); Goods not classified elsewhere (9) (product index p is given in brackets). They belong to the Standard International Trade Classification (SITC Rev. 1) Thus the total Google matrix G size is given by all system nodes $N = N_c N_p = 1680$ including countries and products.

The Google matrix G_{ij} of direct trade flows is constructed in a standard way described in detail at [5, 6]: monetary trade flows from a node j to node i are normalized to unity for each column j thus given the matrix S of Markov transitions for trade, the columns of dangling nodes with zero transactions are replaced by a column with all elements being $1/N$. The weight of each product is taken into account via a certain personalized vector taking into account the weight of each product in the global trade volume. We use the damping factor $\alpha = 0.5$. The Google matrix is $G_{ij} = \alpha S_{ij} + (1-\alpha)v_i$ where v_i are components of positive column vectors called personalization vectors which take into account the weight of each product in the global trade ($\sum_i v_i = 1$). We also construct the matrix G^* for the inverted trade flows.

The stationary probability distribution described by G is given by the PageRank vector P with maximal eigenvalue $\lambda = 1$: $GP = \lambda P = P$ [8, 10, 11]. In a similar way, for the inverted flow, described by G^*, we have the CheiRank vector P^*, being the eigenvector of $G^*P^* = P^*$. PageRank K and CheiRank K^* indexes are obtained from monotonic ordering of probabilities of PageRank vector P and of CheiRank vector P^* as $P(K) \geq P(K+1)$ and $P^*(K^*) \geq P^*(K^*+1)$ with $K, K^* = 1, \ldots, N$. The sums over all products p gives the PageRank and CheiRank probabilities of a given country as $P_c = \sum_p P_{cp}$ and $P^*{}_c = \sum_p P^*{}_{cp}$ (and in a similar way product probabilities $P_p, P^*{}_p$) [5, 6]. Thus with these probabilities we obtain the related indexes $K_c, K^*{}_c$. We also define from import and export trade volume the probabilities $\hat{P}_p$, $\hat{P}^*_p$, $\hat{P}_c$, $\hat{P}^*_c$, $\hat{P}_{pc}$, $\hat{P}^*_{pc}$ and corresponding indexes $\hat{K}_p$, $\hat{K}^*_p$, $\hat{K}_c$, $\hat{K}^*_c$, $\hat{K}$, $\hat{K}^*$ (these import and export probabilities are normalized to unity by the total import and export volumes, see details in [5, 6]). It is useful to note that qualitatively PageRank probability is proportional to the volume of ingoing trade flow and CheiRank respectively to outgoing flow. Thus, we can approximately consider that the high import gives a high PageRank P probability and a high export a high CheiRank P^* probability.

As in [5, 6], we define the trade balance of a given country with PageRank and CheiRank probabilities given by $B_c = (P^*_c - P_c)/(P^*_c + P_c)$. Also we have from ImportRank and ExportRank probabilities as $\hat{B}_c = (\hat{P}^*_c - \hat{P}_c)/(\hat{P}^*_c + \hat{P}_c)$. The sensitivity of trade balance B_c to the price of energy or machinery can be obtained from the change of corresponding money volume flow related to SITC Rev.1 code $p = 3$ (mineral fuels) or $p = 7$ (machinery) by multiplying it by $(1+\delta)$, renormalizing column to unity and computing all rank probabilities and the derivatives $dB_c/d\delta$.

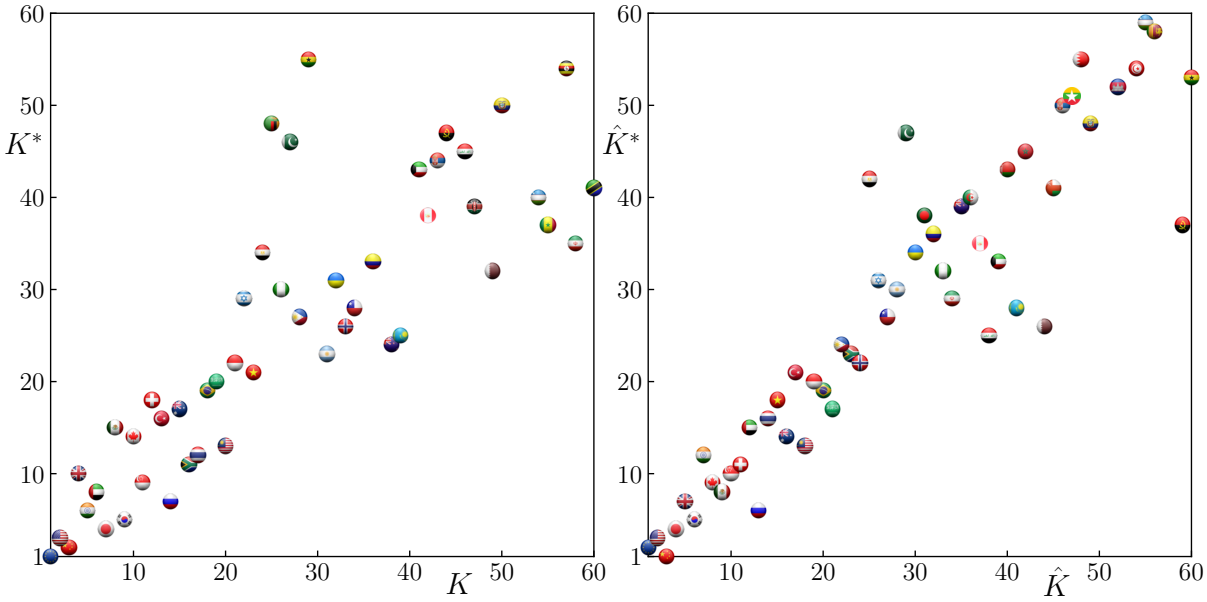


Figure 1. Circles with country flags show country positions on the plane of PageRank-CheiRank indexes (K, K^*) (summation is done over all products) (left panel) and on the plane of ImportRank-ExportRank $\hat{K}$, $\hat{K}^*$ from trade volume (right panel); data is shown only for index values less than 61 in year 2018 .

We also use the REGOMAX algorithm [9, 6] to construct the reduced Google matrix G_R for a selected subset of WTN nodes $N_r \ll N$. This algorithm takes into accounts all transitions of direct and indirect pathways happening in the full Google matrix G between N_r nodes of interest. We use this G_R matrix to construct a reduced network of most strong transitions ("network of friends") between a selection of nodes representing countries and products.

Even if Brexit enter into play in 2021, we use UN COMTRADE data of previous years to make a projecting analysis of present and future power of EU composed of 27 countries.

Finally we note that GMA allows to obtain interesting results for various types of directed networks including Wikipedia [13, 14] and protein-protein interaction [15, 16] networks.

2 Results

2.1 CheiRank and PageRank of countries

The positions of countries on the PageRank-CheiRank (K, K^*) and ImportRank-ExportRank $(\hat{K}, \hat{K}^*)$ planes are shown in Fig. 1 and in Table 1. These results show a significant difference between these two types of ranking. Indeed, EU takes the top PageRank-CheiRank position $K = K^* = 1$ while with Export-Import Ranking it has only $\hat{K} = 1; \hat{K}^* = 2$, with USA at $\hat{K} = 2, \hat{K}^* = 3$ and China at $\hat{K} = 3, \hat{K}^* = 1$. Thus EU takes the leading positions in the GMA frame which takes into account the muliplicity of trade transactions and characterizes the robust features of EU trade relations. Also GMA shows that UK position is significantly weakened compared to IEA description (thus UK moves from $K^* = 7$ in IEA to $K^* = 10$ in GMA). From this data, we see also examples of other countries that significantly improve there rank positions in GMA frame compared to IEA: India ($K = 5$, $K^* = 6$, $\hat{K} = 7$, $\hat{K}^* = 12$), United Arab Emirates ($K = 6$, $K^* = 8$, $\hat{K} = 12$, $\hat{K}^* = 15$), South Africa ($K = 16$, $K^* = 11$, $\hat{K} = 23$, $\hat{K}^* = 23$). We attribute this to well developed, deep and broad trade network of these countries which are well captured by GMA in contrast to IEA. Indeed, IEA only measures the volume of direct trade exchanges, while GMA characterises the multiplicity of trade chains in the world.

Table 1. Top 20 countries of PageRank (K), CheiRank (K^*), ImportRank and ExportRank in 2018.

Rank	PageRank	CheiRank	ImportRank	ExportRank
1	EU	EU	EU	China
2	USA	China	USA	EU
3	China	USA	China	USA
4	United Kingdom	Japan	Japan	Japan
5	India	Repub Korea	United Kingdom	Repub Korea
6	U Arab Emirates	India	Repub Korea	Russia
7	Japan	Russia	India	United Kingdom
8	Mexico	U Arab Emirates	Canada	Mexico
9	Repub Korea	Singapore	Mexico	Canada
10	Canada	United Kingdom	Singapore	Singapore
11	Singapore	South Africa	Switzerland	Switzerland
12	Switzerland	Thailand	U Arab Emirates	India
13	Turkey	Malaysia	Russia	Malaysia
14	Russia	Canada	Thailand	Australia
15	Australia	Mexico	Viet Nam	U Arab Emirates
16	South Africa	Turkey	Australia	Thailand
17	Thailand	Australia	Turkey	Saudi Arabia
18	Brazil	Switzerland	Malaysia	Viet Nam
19	Saudi Arabia	Brazil	Indonesia	Brazil
20	Malaysia	Saudi Arabia	Brazil	Indonesia

2.2 Trade balance and its sensitivity to product prices

The trade balance of countries in IEA and GMA frames is shown in Fig. 2. The countries with 3 strongest positive balance are: Equatorial Guinea ($B_c = 0.732$), Congo ($B_c = 0.645$), Turkmenistan ($B_c = 0.623$) in IEA and China ($B_c = 0.307$), Japan ($B_c = 0.244$), Russia ($B_c = 0.188$) in GMA. We see that IEA marks top countries which have no significant world power while GMA marks countries with real significant world influence. For EU and UK we have respectively $B_c = -0.015; 0.020$ (EU) and $B_c = -0.178; -0.187$ (UK) in IEA; GMA. Thus the UK trade balance is significantly reduced in GMA corresponding to a loss of network trade influence of UK in agreement with data of Fig. 1 and Table 1. (We note that the balance variation bounds in GMA are smaller compared to IEA; we attribute this to the fact of multiplicity of transactions in GMA that smooth various fluctuations which are more typical for IEA).

The balance sensitivity $dB_c/d\delta_s$ to product $s = 3$ (mineral fuels (with strong petroleum and gas contribution)) is shown in Fig. 3. The top 3 strongest positive sensitivities $dB_c/d\delta_s$ are found for Algeria (0.431), Brunei (0.415), South Sudan (0.411) in IEA and Saudi Arabia (0.174), Russia (0.161), Kazakhstan (0.126) in GMA. The results of GMA are rather natural since Saudi Arabia, Russia and Kazakhstan are central petroleum producers. It is worth noting that GMA ranks Iraq at the 4th position. The 3 strongest negative sensitivities are Zimbabwe (-0.137), Nauru (-0.131), Japan (-0.106), in IEA and Japan (-0.066), Korea (-0.062), Zimbabwe (-0.058), in GMA. For China, India we have $dB_c/d\delta_s$ values being respectively: -0.073, -0.086 in IEA and -0.056, 0.010 in GMA. This shows that the trade network of India is more stable to price variations of product $s = 3$. These results demonstrate that GMA selects more globally influential countries.

The balance sensitivity $dB_c/d\delta_s$ to product $s = 7$ (machinery) is shown in Fig. 4. Here the top 3 strongest positive sensitivities $dB_c/d\delta_s$ are found in both IEA and GMA for Japan (respectively 0.167, 0.151), Repub. Korea (0.143, 0.097), Philippines (0.130, 0.091). The 3 strongest negative sensitivities are Brunei (-0.210), Iran (-0.202), Uzbekistan (-0.190) in IEA

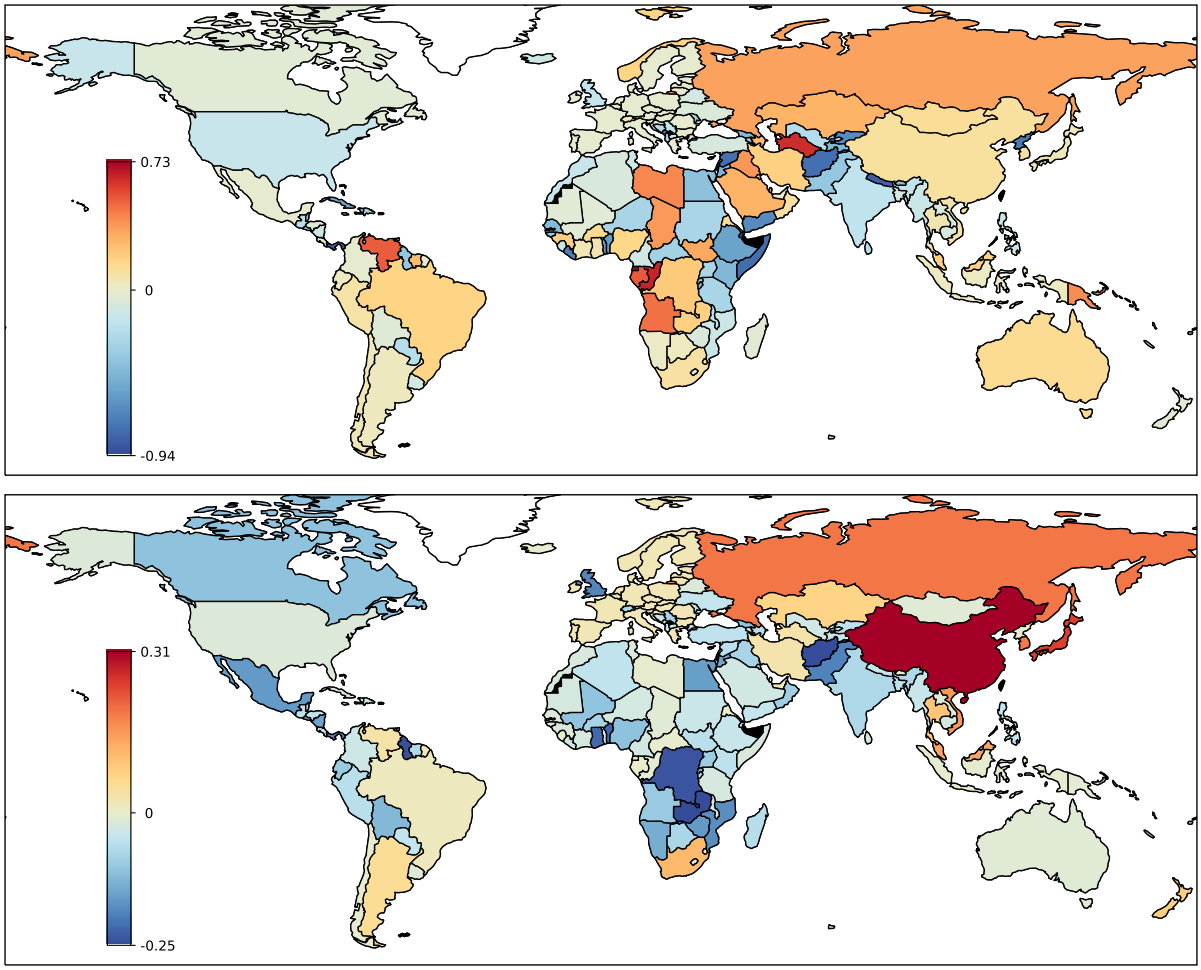


Figure 2. World map of trade balance of countries $B_c = (P_c^* - P_c)/(P_c^* + P_c)$. Top: trade balance values are computed from the trade volume of Export-Import; bottom: trade balance values are computed from PageRank and CheiRank vectors; B_c values are marked by color with the corresponding color bar marked by j; countries absent in the UN COMTRADE report are marked by black color (here and in other Figs).

and Russia (-0.138), Kazakhstan (-0.102), Argentina (-0.097) in GMA. Thus we again see that GMA selects more globally influential countries. The sensitivity $dB_c/d\delta_s$ values for EU, UK, China, Russia, USA are: EU (0.048), UK (0.006), China (0.065), Russia (-0.170), USA (-0.027) in IEA; EU (0.000), UK (0.024), China (0.077), Russia (-0.138), USA (-0.018) in GMA. Latter GMA results show that even if machinery product ($s = 7$) is very important for EU the network power of trade with this product becomes dominated by Asian countries Japan, Repub. Korea, China, Philippines; in this aspect the position of UK is slightly better than EU.

In Figs. 3 and 4, we have considered the sensitivity of country balance to a global price of a specific product (mineral fuel $s = 3$ or machinery $s = 7$). In contrast, with GMA, we can also obtain the sensitivity of country balance to the price of products originating from a specific country. Such results are shown in Fig. 5. They show that machinery ($s = 7$) of EU gives a significant positive balance sensitivity for UK and negative for Russia. This indicates a strong dependence of Russia from EU machinery. Machinery of USA gives strong positive effect for Mexico and Canada with a negative effect for EU, Russia, Brazil, Argentina. Machinery of China gives positive sensitivity for Asian countries (Repub. Korea, Japan, Philippines) and significant negative effect for Mexico. Mineral fuels ($s = 3$) of Russia gives positive effect for Kazakhstan, Uzbekistan, Ukraine (former USSR republics) and negative effect for competing petroleum and gas producers Norway and Algeria.

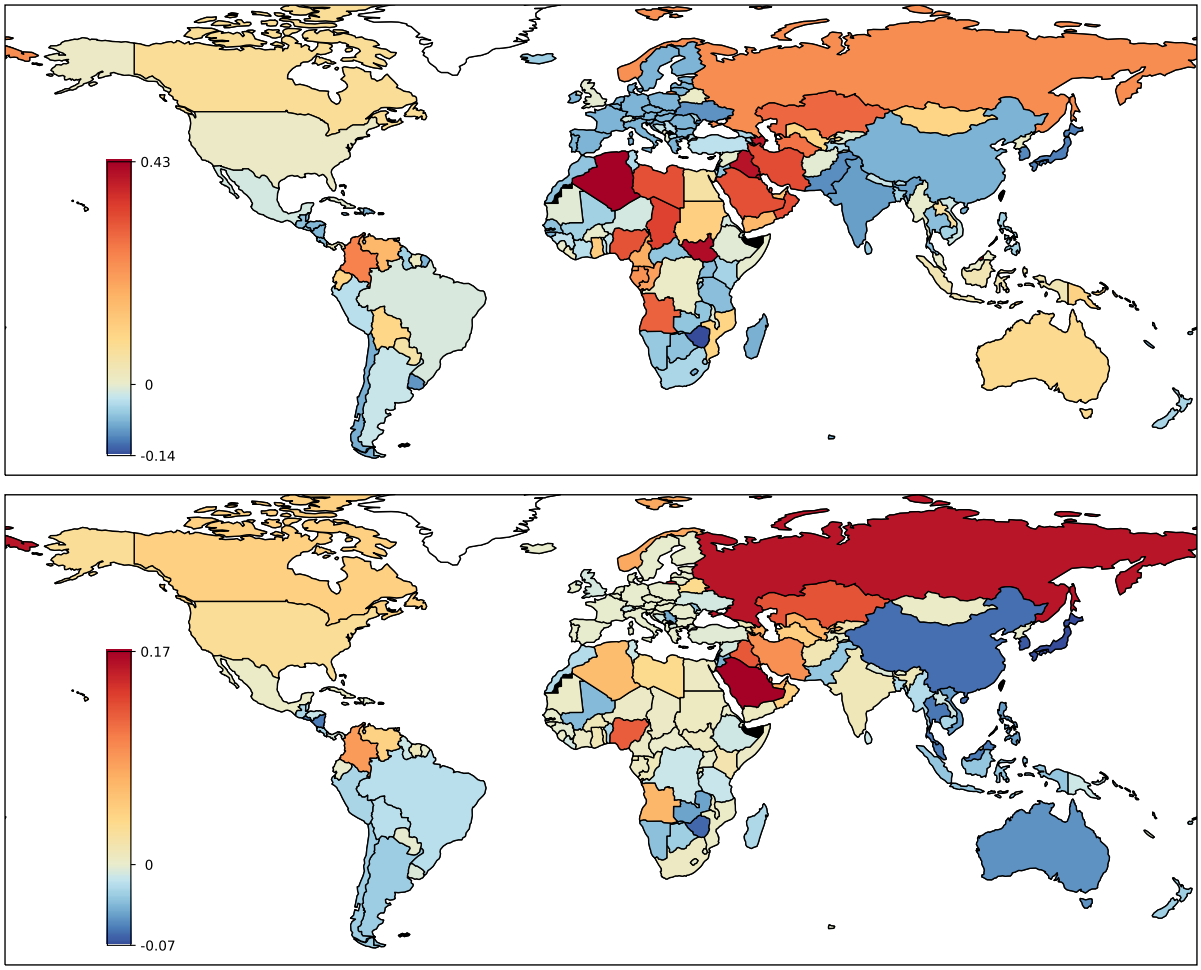


Figure 3. Sensitivity of country balance $dB_c/d\delta_s$ for product $s = 3$ (mineral fuels). Top: probabilities are from the trade volume of Export-Import; bottom: probabilities are from PageRank and CheiRank vectors. Color bar marked by j gives sensitivity.

2.3 Network structure of trade from reduced Google matrix

The network structure for 40 nodes of 10 products of EU, USA, China and Russia is shown in Fig. 6. It is obtained from the reduced Google matrix of $N_r = 40$ nodes of global WTN network with $N = 1680$ nodes on the basis of REGOMAX algorithm which takes into account all pathways between N_r nodes via the global network of N nodes. The networks are shown for the direct (G matrix) and inverted (G^* matrix) trade flows. For each node we show only 4 strongest outgoing links (trade matrix elements) that heuristically can be considered as the four "best friends". The resulting network structure clearly shows the central dominant role of machinery product. For ingoing flows (import direction) of G_R the central dominance of machinery for USA and EU is directly visible while for outgoing flows (export direction), machinery of EU and China dominate exports.

It is interesting to note that the network influence of EU with 27 countries is somewhat similar to the one constituted by a kernel of 9 dominant EU countries (KEU9) (being Austria, Belgium, France, Germany, Italy, Luxembourg, Netherlands, Portugal, Spain) discussed in [7]. This shows the leading role played by these KEU9 countries in the world trade influence of EU.

Finally we note that additional data with figures and tables is available at [12].

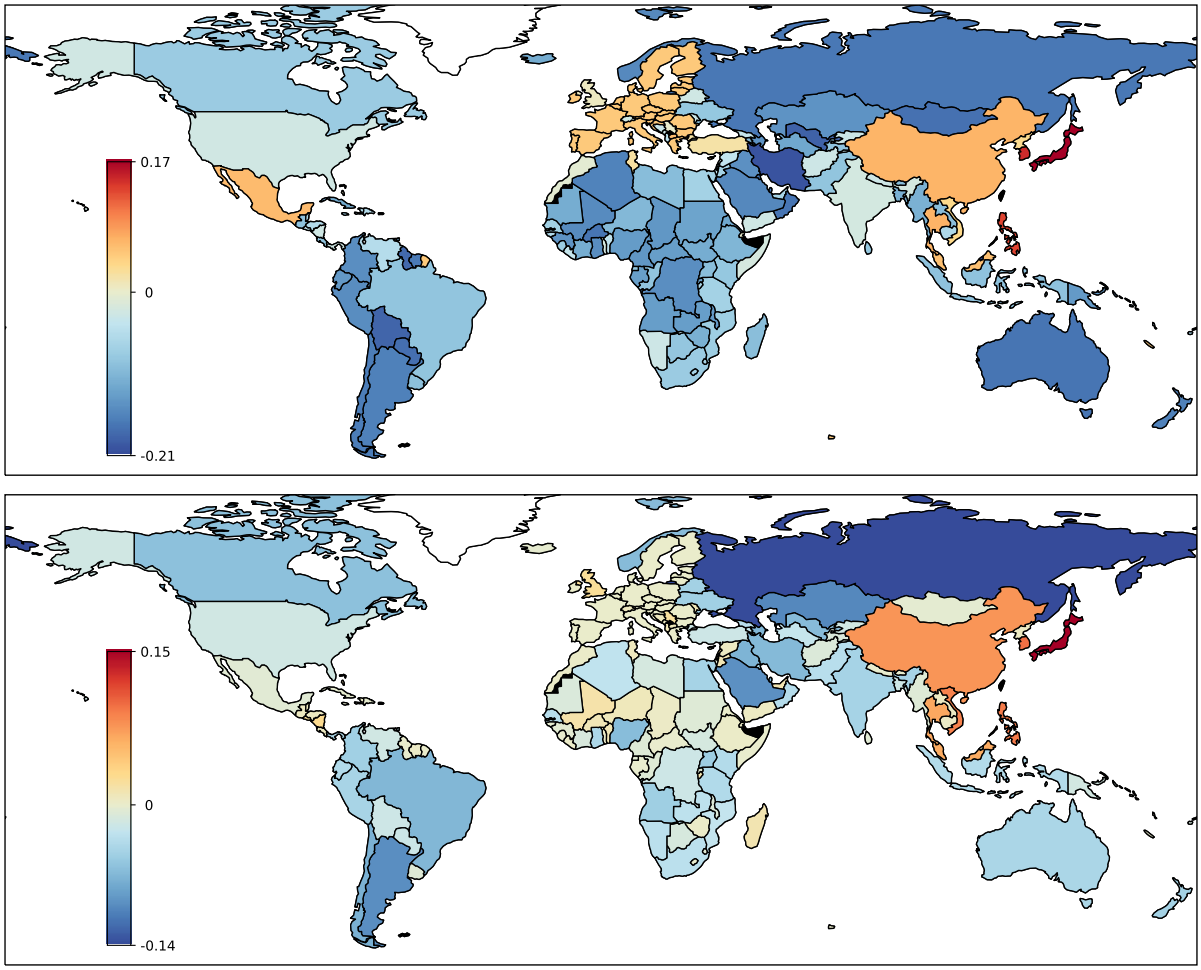


Figure 4. Same as in Fig. 3 but for product $s = 7$ (machinery).

3 Discussion

We presented the Google matrix analysis of multiproduct WTN obtained from UN COMTRADE database in recent years. In contrast to the legacy Import-Export characterization of trade, this new approach captures multiplicity of trade transactions between world countries and highlights in a better way the global significance and influence of trade relations between specific countries and products. The Google matrix analysis clearly shows that the dominant position in WTN is taken by the EU of 27 countries despite the leave of UK after Brexit. This result demonstrates the robust structure of worldwide EU trade. It is in contrast with the usual Import-Export analysis in which USA and China are considered as main players. We also see that machinery and mineral fuels products play a dominant role in the international trade. The Google matrix analysis stresses the growing dominance of machinery products of Asian countries (China, Japan, Republic of Korea).

We hope that the further development of Google matrix analysis of world trade will bring new insights in this complex system of world economy.

Acknowledgments: We thank Leonardo Ermann for useful discussions. This research has been partially supported through the grant NANOX N° ANR-17-EURE-0009 (project MTDINA) in the frame of the *Programme des Investissements d'Avenir, France* and in part by APR 2019 call of University of Toulouse and by Region Occitanie (project GoIA). We thank UN COMTRADE for providing us a friendly access to their detailed database.

Statement on competing interests: The authors declare that there is no conflict of interest.

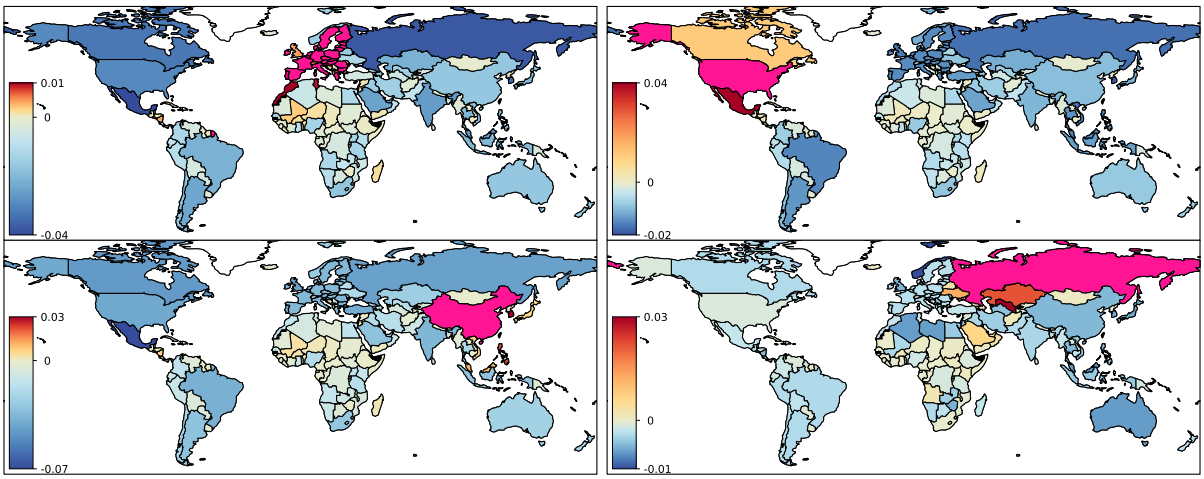


Figure 5. Sensitivity of country balance $dB_c/d\delta_{cs}$ for product price $s = 7$ (machinery) of EU (top left), USA (top right), China (bottom left) and $s = 3$ (mineral fuel) of Russia (bottom right); B_c is computed from PageRank and CheiRank vectors; sensitivity values are marked by color with the corresponding color bar marked by j. For EU, USA, China, Russia we have $dB_c/d\delta_{cs} = 0.11, 0.11, 0.14, 0.12$ respectively, these values are marked by separate magenta color to highlight sensitivity of other countries in a better way.

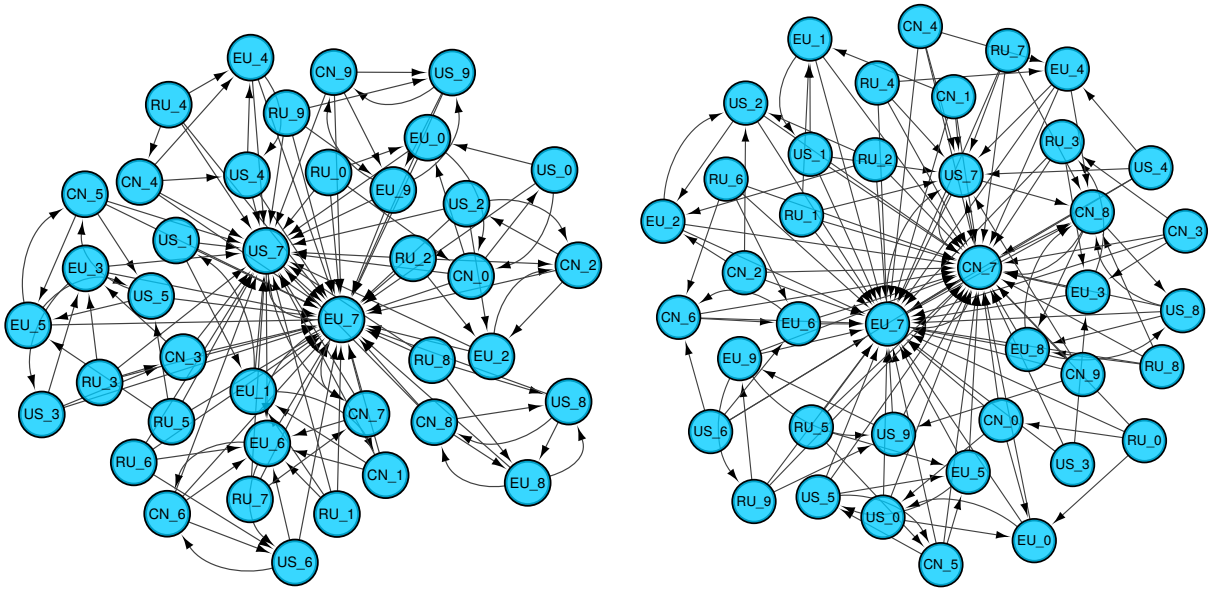


Figure 6. Network trade structure between EU, USA (US), China (CN), Russia (RU) with 10 products. Network is obtained from the reduced Google matrix G_R (left) and $G^*{}_R$ (right) by tracing four strongest outgoing links (similar to 4 “best friends”). Countries are shown by circles with two letters of country and product index listed in Section 2. The arrow direction from node A to node B means that B imports from A (left) and B exports to A (right). All 40 nodes are shown.

References

[1] *European Union*, https://europa.eu/european-union/about-eu/figures/economy_en#trade (Accessed February (2021)).

[2] *Brexit*, https://en.wikipedia.org/wiki/Brexit (Accessed February (2021)).

[3] *UN Comtrade database*, https://comtrade.un.org/ (Accessed February (2021)).

[4] Ermann L. and Shepelyansky D.L.: *Google matrix of the world trade network*, Acta Physica Polonica A **120**, A158 (2011).

[5] Ermann L. and Shepelyansky D.L.: *Google matrix analysis of the multiproduct world trade network*, Eur. Phys. J. B **88**, 84 (2015).

[6] Coquide C., Ermann L., Lages J. and Shepelyansky D.L.: *Influence of petroleum and gas trade on EU economies from the reduced Google matrix analysis of UN COMTRADE data*, Eur. Phys. J. B **92**, 171 (2019).

[7] Loye J., Ermann L. and Shepelyansky D.L.: *World impact of kernel European Union 9 countries from Google matrix analysis of the world trade network*, arXiv:2010.10962[cs.SI] (2020).

[8] Ermann L., Frahm K.M. and Shepelyansky D.L.: *Google matrix analysis of directed networks*, Rev. Mod. Phys. **87**, 1261 (2015).

[9] Frahm K.M., Jaffres-Runser K. and Shepelyansky D.L.: *Wikipedia mining of hidden links between political leaders*, Eur. Phys. J. B **89**, 269 (2016).

[10] Brin S. and Page L.: *The anatomy of a large-scale hypertextual Web search engine*, Computer Networks and ISDN Systems **30**, 107 (1998).

[11] Langville A.M. and Meyer C.D.: *Google's PageRank and beyond: the science of search engine rankings*, Princeton University Press, Princeton (2006).

[12] *Post-Brexit trade power of EU*, https://www.quantware.ups-tlse.fr/QWLIB/euwtn (Accessed February (2021)).

[13] Frahm K.M., El Zant S., Jaffres-Runser K. and Shepelyansky D.L.: *Multi-cultural Wikipedia mining of geopolitics interactions leveraging reduced Google matrix analysis*, Phys. Lett. A **381**, 2677 (2017).

[14] Coquide C. and Lewoniewski W.: *Novel version of PageRank, CheiRank and 2DRank for Wikipedia in multilingual network using social impact*, In: Abramowicz W., Klein G. (eds) *Business Information Systems BIS*, Lecture Notes in Business Information Processing **389**, 319 (2020).

[15] Lages J., Shepelyansky D.L. and Zinovyev A.: *Inferring hidden causal relations between pathway members using reduced Google matrix of directed biological networks*, PLoS ONE **13(1)**, e0190812 (2018).

[16] Frahm K.M. and Shepelyansky D.L.: *Google matrix analysis of bi-functional SIGNOR network of protein-protein interactions*, Physica A **559**, 125019 (2020).

24th International Conference on Business Information Systems (BIS 2021)

Big Data

https://doi.org/10.52825/bis.v1i.56

Published: 02 July 2021

Towards a Guideline Affording Overarching Knowledge Building in Data Analysis Projects

Dorothea Schneider [1][https://orcid.org/0000-0002-6633-5052] and Wibke Kusturica[2][https://orcid.org/0000-0001-6131-2620]

1 Technische Universität Dresden, Institute of Mechatronic Engineering, Dresden, Germany

2 University of Applied Sciences Zwickau, Institute for Management and Information, Zwickau, Germany

Abstract. Tight and competitive market situations pose a serious challenge to enterprises in the manufacturing industry domain. Competing in the use of data analytics to enhance products and processes requires additional resources to deal with the complexity. On the contrary, the possibilities afforded by digitization and data analysis-based approaches make for a valuable asset. In this paper we suggest a guideline to a systematic course of action for the data-based creation of holistic insight. Building an overlaying corpus of knowledge accelerates the learning curve within specific projects as well as across projects by exceeding the project-specific view towards an integrated approach.

Keywords: Data Mining, Knowledge Bases, Reference Model, Methodological Knowledge, Domain-Specific Information Systems Development.

1 Introduction

Demand and supply for insights derived from all kinds of accessible data sources in enterprises are higher than ever before as the pressure to keep up with global competitors meets the ever-growing possibilities of data acquisition and exploitation. A plethora of methods and tools is available to deal with and make use of these resources: from sensors to algorithms, from Industrial Internet of Things (IIoT) solutions to programming libraries and software. [1]

While all business sectors face this situation equally and therefore must deal with similar challenges, the complexity of the task is particularly high in the manufacturing industry domain. [2] [3] This holds true especially for tasks within data-driven enhancement projects (EP) in the manufacturing industry domain which require a high level of innovation and are conducted in a project-based manner like one-of-a-kind production, research and development (R&D), customer-specific machinery and plant engineering or the design of cyber-physical production systems. [4]

First and foremost, conducting successful data analysis projects does not only include the activities directly associated with analyzing data but involves the execution of several elaborate steps as well as strategic measures. To systematically align all relevant aspects affecting the analysis outcome in a wider sense will result in distinct quality improvement. [3]

In our research we aim at providing the means to support achieving strategic goals by conducting data analysis projects which systematically connect relevant information fragments on all levels of aggregation from all relevant sources. Therefore, our research is driven by the following research question (RQ):

RQ: How can a reference model be provided for complex tasks in the industrial domain which provides methodological support for the data-driven construction and utilization of an overlaying corpus of knowledge?

To answer this question, we developed an artifact in the form of a reference model to equip the user with a wide range of methodological support for conducting informed data analyses. The goal of the suggested framework is to not only derive insight about the examined topic of an active data mining project but to preserve and build on the findings exceeding project boundaries. The reference model aims to inspire rigorous and holistic investigation, to provide the means for communication, project management and documentation and to build the foundation for future software applications to support this holistic project-exceeding data mining approach thus also paving the way for an analysis and optimization of the activities undertaken within data mining projects themselves.

Following this approach this paper is structured as follows: In Section 2, we describe our motivation, we then sum up foundations and basic concepts in Section 3. Derived from the key activities of the sensemaking approach as described by [5] and more specifically by [6] a set of design principles is suggested, as will be described in Section 4. In fulfillment of the defined design principles a framework is presented in Section 5 to structure necessary methodological measures and to allocate useful activities within five layers of information aggregation. By presenting the reference model we advocate for a systematic course of action aiming at the creation of holistic insight. Finally, we draw a conclusion and give an outlook for further research in Section 6.

2 Motivation

The major purpose of the presented long-term design science research project is to elaborate methodological support for data-driven knowledge extraction projects in the manufacturing industry domain. Therefore, our main objective is to help artifact users gain a sophisticated understanding of the principles by which to conduct data-driven knowledge extraction projects, to reduce the associated hurdles for manufacturing companies and to create a basis to address and solve them in the future in a repeatable manner. The application of the presented reference model enables domain experts to derive cumulative knowledge, rather than re-inventing technical concepts and methodological procedures under new labels in every new project setting. [7]

Specialists dealing with data analysis projects in the industrial domain face the necessity to cover the methodological skillset required in data science as well as a deep understanding of the domain fundamentals to consider relevant causalities and interactions and to purposefully derive and interpret results according to their context. Hence throughout all industrial sectors on the one hand domain experts successfully gain and apply data analytics knowledge while on the other hand data analysts engage in various domain contexts and oftentimes both have to team up with each other and with additional professionals like computer scientists and mathematicians to derive the desired outcome. While tremendous progress is underway in the domain-specific training of and proficient cooperation with data scientists and in the successful realization of data analytics projects the potential for even better outcome is huge. [8] [9] The main hurdles are the intricate communication between domain experts and data scientists, the scarcity of human resources for data analytics projects and the lack of domain-specific standardized procedures which lead to a singular quality of the execution and the use of results of data-driven analyses. These shortfalls especially hold true where a limited number of experts must realize data analytics projects next to rivaling work tasks as is the case in small and medium sized companies (SME), start-ups and R&D or planning departments. [3]

A pre-study in the form of an exploratory study with six qualitative expert interviews aimed to identify the challenges that occur while setting up a data-driven knowledge extraction project confirmed these hurdles. The interviews were designed as partially standardized interviews using open to semi-open questions as initial starting points for the conversation and took between 70 and 180 minutes. The complete listing of the formulated questions and results will be provided by the authors upon request. The answers showed that practitioners tend to rely on traditional procedures and experience-based knowledge.

Their understanding of Data Mining (DM) mainly focused on the core analysis activities like the application of algorithms and often underestimated the effort and importance of peripheric aspects like the determination of target-aimed questions, data preparation to produce structured evaluable data sets, conclusive feature engineering and context-sensitive model building. The interviewees expressed their wish for more structure and guidance in data analytics projects while they found existing standard processes too generic to apply for their domain as well as not sufficiently considering real-life problems like data acquisition, data quality and operational data processing.

3 Foundation

Pursuing a long-term research project in the field of information systems (IS) aiming at the design of an artifact in the form of a reference model we comply with the design science paradigm stated by [10]. We furthermore adopt the three-cycle view of design science research (DSR) presented in [11] to address the relevance, design and rigor of the developed artifact. Additionally we rely on the steps for DSR research recommended by [12] to apply the paradigm to our research as follows: The ***problem identification and motivation*** for our research is constituted by the experience from numerous research projects and a pre-study in the form of expert interviews as described in Section 2. We then derived theory-based research ***goals and objectives*** by the definition of design principles as described in Section 4 followed by the ***design and development*** of the artifact, the outcome of which is presented in Section 5. While applying the findings in practice the derivation of a context-specific model should then be ***demonstrated*** and ***evaluated*** within future research. In an iterative manner the insights from an initial implementation within an example scenario should be used to further enhance the artifact and undergo subsequent evaluation phases to then be ***transferred to the community***.

When attempting to represent and reduce reality to fulfill a subjective purpose like the understandable formulation of complex facts [13] for a class of similar problems a ***reference model*** is provided by introducing a model which is of recommendatory and universal character and allows for the derivation of application-specific models.[14] Consequently reference models are a generic type of model representing the essence of a common-practice or best-practice view on a class of similar problems intended for re-use and acting as a blueprint for the derivation of specific models. [15]

The addressed application field of the presented reference model comprises tasks in the industrial domain which require a high level of innovation and are conducted in a project-based manner. When attempting to support such tasks there are various user roles and artifacts to take account of, notwithstanding that more than one user role can be fulfilled by one individual. These roles and artifacts are depicted in figure 1.

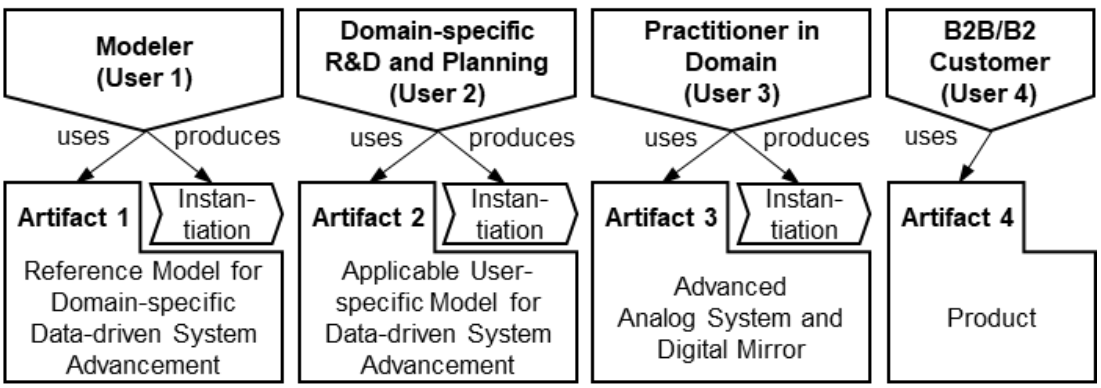


Figure 1. Addressed users and artifacts

As drawing conclusions by the statistical or algorithms-based study of large amounts of data today is widely established throughout all disciplines, numerous attempts have been made to ***standardize the data mining process*** especially in the field of computer science and economic analyses. Such procedure models generally consist of generic steps to structure and guide the planning and execution of DM projects.[20] Prominent standard operating models are subsequently named. Knowledge discovery in databases (KDD) is a description of the central building blocks of the overall multi-step procedure for complex real-world analysis tasks aiming at the discovery of knowledge in large amounts of data.[17] [18] Subsequent approaches like SEMMA and CRISP-DM emerged from the basic concept of KDD. The cross-industry standard process for DM (CRISP-DM) comprises the steps business understanding, data understanding, data preparation, modeling, evaluation and deployment, thus adding a more strategic perspective to the KDD core concept [19] [20]. The sample, explore, modify, model, and assess (SEMMA) methodology was developed by the SAS Institute to methodically organize the functions of its statistical and business intelligence software SAS Enterprise Miner, its constituent phases naming the concept in the form of an acronym. The analytics solutions unified method (ASUM) draws on a combination of agile and traditional implementation principles to achieve set solution goals and therefore complements the defined analysis phases by an additional project management stream to support the organizational realization. [21]

4 Design Principles

The concept of ***sensemaking*** originated in social psychology and was set in an organizational context by [5]. The approach describes how human beings in a social setting derive understanding of their surroundings by combining various information, creating connections and finally adding their own reasoning to it. The concept is described extensively in [22]. [6] sums up relevant literature and derives five key activities found in previous work as listed in table 1 which constitute the making of sense and thereby act as design goals for the developed reference model.

As the developed framework is supposed to not only support the understanding of facts and the creation of insight but also its utilization for the in-project and project-exceeding enhancement of the target-system, one more key activity is needed to complement the sensemaking key activities. By including the creation and utilization of a knowledge base we want to create a linkage to the field of knowledge management and thereby create the concept of ***knowledge making***. By coining the term, we want to emphasize a creative, intuitive and iterative character of the approach, orienting on human behavior and the cognitive and social processes it originates in.

In DSR the concept of ***design principles*** (DP) provides the means to specify prescriptive design knowledge in a way that allows for a precise formulation to describe how the mechanisms of a technology or approach help to achieve particular aims.[23] According to [24] design principles should describe which actions are made possible through the use of an artifact and explain the material properties which make that action possible while naming the boundary conditions under which this description holds true. More precisely [24] suggests the formulation of a DP in the following form: “Provide the system with [material property—in terms of form and function] in order for users to [activity of user/group of users—in terms of action], given that [boundary conditions—user group’s characteristics or implementation settings].” Following this suggestion, we formulated design principles for the presented reference model based on the derived knowledge making key activities as shown in table 1.

Table 1. Sensemaking [6] and knowledge making key activities with derived DPs

Sensemaking key activities	**Knowledge making key activities**	**Design Principles: provide the reference model with features**
(S1) Triggered by disruptive ambiguity	(K1) Open up new possibilities: Provoke action	…providing input to provoke action, in order for users to proactively advance their course of action in EPs, according to their given resources.
(S2) Acts of noticing and bracketing, create initial sense	(K2) Inform and classify: Provide a structure and give information impulses	…for structuring the user's course of reasoning and action, in order for users to gain a comprehensive understanding of relevant aspects to perform EPs, regarding the given domain and project context.
(S3) Requires labelling and categorizing, find common ground	(K3) Label and categorize: Assign methods and prior selections to structure	…to provide a catalogue of relevant action methods and tools, previously chosen alternatives or standard configurations and suitable search terms, in order for users to gain awareness of actionable alternatives.
(S4) Involves presumption to guide action, connect the abstract with the concrete	(K4) Assume and iterate: Intuitive selection of solution options, heuristic procedure	…allowing for heuristic solution approaches, in order for users to initialize solution finding in an intuitive way and vary solution configuration easily, considering established solution approaches and insight of previous data analysis projects.
(S5) Involves communication, draws on the resources of language	(K5) Provide communication basis: Provide means to discuss solution alternatives	…to visualize structure and solution alternatives, in order for users to apply the visual and textual representation of the reference model as means to debate and cooperatively decide on practicality, options and implications of solution configurations, considering the user's heterogeneous professional background
-	(K6) Capture and utilize: Build overarching body of knowledge	…the means to include a knowledge base, in order for users to connect, preserve, document and utilize information fragments and their relations, considering the user's preconditions to obtain a balance of a potent solution and manageable effort.

5 Reference Model

We want to motivate a highly strategic and integrated practice in data-driven enhancement projects [EP] in the manufacturing industry domain and to support this mindset by suggesting a framework to guide the efforts. The development of this reference model is driven by the needs identified in industrial practice and numerous research projects and realized by employing well-researched approaches grounded in established theory. We set up a grid-like structure to assign relevant methodologies to the respective analysis project phases and thereby fulfill the design principles formulated in Section 4. We based our approach on three widely established concepts: standard procedure models, the concept of data aggregation and the field of knowledge management. We attempt to provide the means for the effective combination and domain-specific adaption of these concepts while additionally overcoming their shortcomings as described in section 1 and further elaborated in [25] and [3].

We especially want to emphasize the importance of considering the various ***aggregation levels*** as described in table 2 in which information fragments can occur in, calling attention in particular to the intense interaction of all five levels of aggregation implying the necessity to expand awareness to each of them and their interrelations within each step of action. More specifically speaking an integrated consideration and

operationalization is needed throughout all project phases as the strong focus on DM core analysis activities was one of the main hurdles found in the pre-study described in Section 2. The reference model supports practitioners in the inclusion of all aspects, from aggregation level 1, being the least connected state of raw data and the physical system realization and data acquisition up to level 5, comprising the overarching management of highly connected complex information constructs.

Table 2. Aggregation levels

Aggregation level	**Description**
AL 1: Analogous level	Tangible components of the real-life system
AL 2: Representation level	Data objects representing and accompanying the real-life system
AL 3: Transfer level	Measures derived from representing and accompanying data
AL 4: Implementation level	Implementations derived from feature sizes of the transfer level
AL 5: Information level	Highly connected information comprising facts and interrelations, decision support

Data aggregation is often depicted in a form similar to the ***traditional knowledge pyramid***, although revised and refined approaches can be found superseding this strictly hierarchical view. [26] Within the scope of our research we adopt the view that information fragments can exist in various states of aggregation, starting from incrementally small pieces of data like a single binary number, but also forming states of light aggregation as in protocols or logfiles or of higher aggregation like in the form of data sets, tables, charts or reports, where data is set into context and provides declarations exceeding its alpha-numerical value. We therefore deem it valid to speak of information when referring to aggregated data. Data aggregation states then stretch to strongly aggregated forms of where aggregated chunks of information further connect to complex constructs representing relations comprising formal logic thus resembling the processing of insight and thought in the human mind. We therefore argue that the term information is suitable to describe aggregated forms of data and highly aggregated information equals knowledge in the daily use of language. In table 2 we convey this understanding to the manufacturing industry domain introducing an additional level of ***analogous real-life objects*** which the relevant data relates to and originates in.

Relevant objects within AL 1 can be controllers, motors, GPS trackers and sensors or transport systems, accompanied by the respective digital counterparts in AL 2 like output data of controllers, performance data of motors, GPS data and other sensor data. Furthermore AL 2 addresses additional descriptions of the target-system as e.g. conceptual models. Within AL 3 a suitable concept must be chosen to gather, process and contain any relevant information fragments to transfer them to higher levels of aggregation and derive and utilize insight. A suitable concept can be an enterprise-specific analysis framework, an individual adoption of the DM standard processes described in section 3 or domain-specific adoptions like the “DMME: Data mining methodology for engineering applications” as presented in [3]. Within AL 3 and the central analysis project phase of the chosen concept resides the core activity constituting the success of the EP: Proceeding in an intensely iterative character and closely observing the relation to any other grid point highly context-sensitive feature engineering is made possible. Within AL 4 the found facts and interrelations are implemented by integrating the derived insight within physical instantiations, instantiations of digital shadows or digital twins, simulation models or visualizations.

The knowledge base constituting AL 5 can take many forms, from the incorporation by an individual, classical SQL databases or ontologies to intelligent agents. Lastly the successful utilization of the concept will depend on what the respective knowledge base affords. Despite AL 5 constituting the bottleneck of the implementation, the more suitable its chosen way of instantiation is for the occasion the more intense the usage in practice will be. Highly formalized approaches and machine-readable implementations allow for complex and potent operations but require high effort to set up and maintain. Depending on the application

situation the manageable effort of a lightweight solution can advance implementation success. We suggest orienting on existing solutions like for example extensively elaborated for the application of ontologies in the manufacturing domain in [27].

Two more aspects are vital to exploit the full potential of data analytics in the industrial domain: to take into account the ***dimorphic system character*** of the target system consisting of analogous and digital components and to focus on ***context-sensitive engineering of conclusive features*** as this step constitutes the heart of the project and is complemented by the choice and application of fitting tools and methods, only rendered possible by the utilization of aforementioned concepts providing the necessary context. [29]

As pointed out by [29] and further elaborated by [30], the concepts described in Section 3 share the common essence of a stepwise description of the data mining project phases along with similar core principles of the activities performed during the respective steps. Attempting to capture the essence of the various data mining procedure models we derived a generalized version of data mining project phases as can be seen in figure 2. Based on the specification of the analysis project goal in phase 1 (P1) a conceptualization phase follows in phase 2 (P2). The data analysis core activities are performed in phase 3 (P3) and 4 (P4). First data is collected by setting up the necessary physical infrastructure and accumulating all accessible and presumably relevant information fragments, growing and extending the data pool. Then feature engineering, model building and extraction of relations follow, reducing the data build-up to a set of connected information which can then be deployed. Phase 5 (P5) draws on the preceding phases and can and should be conducted in parallel from the start as it preserves and makes available the methodological and meta-information of the data analysis project as well as comprises the supervision of its execution during and after the project.

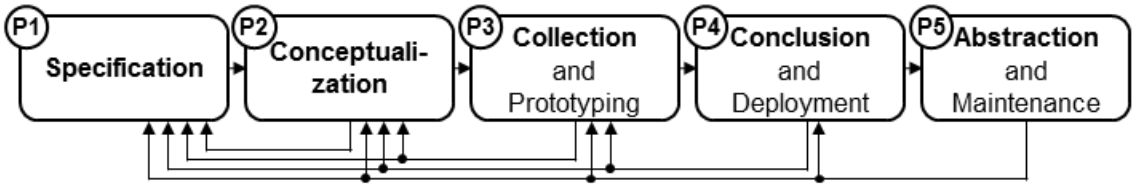


Figure 2. Generalized procedure model

The phases described above provide the reference model with a basic sequence of actions to perform in a data analysis project and can be replaced by any adequate alternative during instantiation, e.g. a standard process or an enterprise specific procedure. Concurrently the necessity to consider various aggregation levels of available and derived information fragments pertains for all project steps. The aggregation level view in combination with the project phases forms a grid as presented in figure 3 to address the methodological repertory of each combination of layer and phase allowing for the mapping of relevant methods accompanied by respective meta-information.

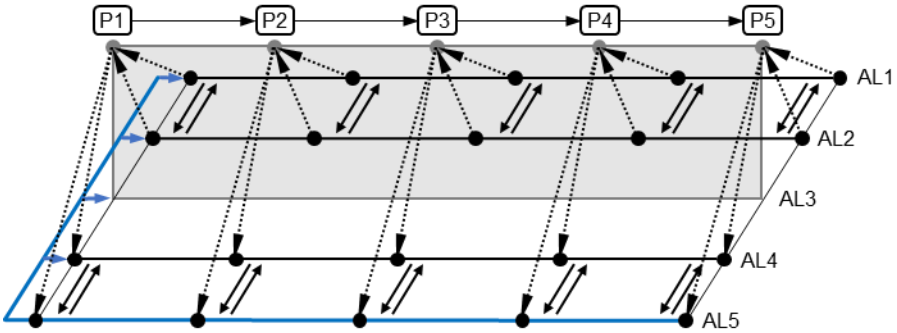


Figure 3. Reference model

At each grid point a template is to be provided to document used methods and their domain-specific application as well as to give an initial information impulse comprising a narrow set of well-established methods along with a continuable list of methods and sufficient search terms. If available, sub-methodologies and detailed sub-selection options are included by grouping them in a hierarchical manner beneath the respective method, providing a template for each hierarchical dimension. The basic or initial selection can be realized by pre-defining a default method for each methodological category as well as by giving a minimum viable implementation strategy.

Within the iterative solution process a ***token of current knowledge*** cycles the defined project phases undergoing permanent revision and thus updating the knowledge base. The active token resembles an assumption about the current state of the targeted artifact, permanently considering the dimorphous character of the target-system. It is the state of the art for nearly any real-life system to be accompanied by a digital counterpart. From our point of view these two sides of reality, the analogous components and digital descriptions and traces mirroring them, form the targeted system and have to be considered continuously to investigate, analyze and enhance this system. For further details on the ***real-life system*** we suggest [30] and [31] on the concepts of ***digital shadow and digital twin*****.** To realize the iterative procedure based on an assumption token it is advisable to orient on existing approaches like the “Conceptual Model of the Learning-Oriented Knowledge Management System” given in [32].

When applying the reference model, a specific model is derived tailored to support the targeted EP. Project phases, components included within the aggregation levels and respective methodological suggestions populating the reference model grid are adapted to their relevance within the given context. To ensure intuitive applicability for practitioners, the reference model and templates should be provided in form of visual content accompanied by textual explanations, preferably by the means of a software application.

6 Discussion and Outlook

In the presented paper we gave an outline towards a framework supporting the systematic data-based creation of insight. The suggested reference model aims at providing the means to accelerate the learning curve within an active data analysis project as well as to build and utilize an overlaying corpus of knowledge exceeding project boundaries. This aim can be addressed by orienting on the sensemaking approach as described by [6] to derive knowledgemaking key activities. To afford the realization of these activities design principles were formulated. Following these principles, we set up a grid-like structure to assign relevant methodologies to the respective analysis project phases while considering the possible aggregation levels information fragments can occur in. The presented reference model offers a guideline for communication, handling and documentation of technological and methodological information thus providing the means for the construction and utilization of an overarching knowledge base.

First application experience in the support of research projects showed the value of the reference model to promote a more integrated method of operation, but also made obvious how providing the means for intuitive applicability is crucial for the successful implementation of the approach. [4] [36]

Future work will be devoted to the demonstration, evaluation and revision of the concept in practice. Additionally, a thorough analysis of existing and common methodological elements will be conducted by the analysis of research publications within leading journals and by assessment of accessible information on their application in practice to develop an appropriate classification and identify any additional elements that should be included. Moreover, having provided the means to document the usage of methods and their specification as well as having examined their classification allows for the construction of a formalized body of knowledge addressing the creation of knowledge itself. Future work will comprise the development of a taxonomy of methodological principles at hand to then be conveyed to an ontology defining logical relations, rules and principles allowing for decision support by typecasting similar EPs and deriving suitable solution approaches. While this paper focused on the motivation and the theoretical grounding of the concept, some consideration should also be given to its compliance with existing standards and tools to accelerate interoperability. The integration with standardized approaches like the *Reference Architecture Model Industrie* 4.0 (RAMI4.0) or with data management aspects like the data lifecycle approach can create synergies and add a helpful dimension to support the organizational implementation of the suggested method within enterprises. [37]

Acknowledgements

Parts of this research were funded by the German Federal Ministry of Education and Research, grant number 03ZZ0212G "Zwanzig20 Agent-3D Verbundvorhaben: QUALIPRO".

Competing interests

The authors declare no competing interests.

References

1. Schäffer, T., Leyh, C.: Master Data Quality in the Era of Digitization - Toward Inter-organizational Master Data Quality in Value Networks: A Problem Identification. In: Piazolo, F., Geist, V., Brehm, L., and Schmidt, R. (eds.) Innovations in Enterprise Information Systems Management and Engineering. pp. 99–113. Springer International Publishing, Cham (2017).
2. Zschech, P., Heinrich, K., Pfitzner, M., Hilbert, A.: Are you up for the challenge? Towards the development of a big data capability assessment model. Proc. 25th Eur. Conf. Inf. Syst. ECIS. 2613–2624 (2017).
3. Huber, S., Wiemer, H., Schneider, D., Ihlenfeldt, S.: DMME: Data Mining Methodology for Engineering Applications – A Holistic Extension to the CRISP-DM Model. In: Procedia CIRP (2018). pp. 403–408., Gulf of Naples, Italy (2018).
4. D. Gliem, C. Laroque., U. Jessen, J. Stolipin, S. Wenzel, W. Kusturica: SimCast – Simulationsgestützte Prognose der Dauer von Logistikprozessen. Abschlussbericht. Universität Kassel, Fachgebiet Produktionsorganisation und Fabrikplanung; Westsächsische Hochschule Zwickau, Fachgebiet Wirtschaftsinformatik (2019).
5. Weick, K.E.: The social psychology of organizing. McGraw-Hill, New York (2006).

6. Seidel, S., Chandra Kruse, L., Székely, N., Gau, M., Stieger, D.: Design principles for sensemaking support systems in environmental sustainability transformations. Eur. J. Inf. Syst. 27, 221–247 (2018). https://doi.org/10.1057/s41303-017-0039-0.
7. Gregor, S., Jones, D.: The anatomy of a design theory. J. Assoc. Inf. Syst. 8, (2007).
8. Zschech, P., Fleißner, V., Baumgärtel, N., Hilbert, A.: Data Science Skills and Enabling Enterprise Systems: Eine Erhebung von Kompetenzanforderungen und Weiterbildungsangeboten. HMD Prax. Wirtsch. 55, 163–181 (2018). https://doi.org/10.1365/s40702-017-0376-4.
9. Debortoli, S., Müller, O., Brocke, J. vom: Vergleich von Kompetenzanforderungen an Business-Intelligence- und Big-Data-Spezialisten: Eine Text-Mining-Studie auf Basis von Stellenausschreibungen. Wirtschaftsinformatik. 56, 315–328 (2014). https://doi.org/10.1007/s11576-014-0432-4.
10. Hevner, A.R., March, S.T., Ram, S., Park, J.: Design Science in Information Systems Research. MIS Q. 28, 75–105 (2004).
11. Hevner, A.R.: A three cycle view of design science research. Scand. J. Inf. Syst. 19, 87–92 (2007).
12. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. J. Manag. Inf. Syst. 24, 45–77 (2007). https://doi.org/10.2753/MIS0742-1222240302.
13. Stachowiak, H.: Allgemeine Modelltheorie. Springer, Wien [u.a.] (1973).
14. Vom Brocke, J., Grob, H.L.: Referenzmodellierung: Gestaltung und Verteilung von Konstruktionsprozessen. Logos Verlag, Berlin (2015).
15. Schlieter, H., Esswein, W.: Reference Modelling in Health Care - State of the Art and Proposal for the Construction of a Reference Model. Enterp. Model. Inf. Syst. Archit. 6, 36–49 (2011). https://doi.org/10.18417/emisa.6.3.3.
16. Kurgan, L.A., Musilek, P.: A survey of Knowledge Discovery and Data Mining process models. Knowl. Eng. Rev. 21, 1 (2006). https://doi.org/10.1017/S0269888906000737.
17. Brachman, R.J., Anand, T.: The Process of Knowledge Discovery in A First Sketch. AAAI Tech. Rep. WS-94-03 (1994).
18. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. 17, 37–37 (1996).
19. Xu, L.D., He, W., Li, S.: Internet of Things in Industries: A Survey. IEEE Trans. Ind. Inform. 10, 2233–2243 (2014). https://doi.org/10.1109/TII.2014.2300753.
20. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.R., Wirth, R.: CRISP-DM 1.0: Step-by-step data mining guide, (2000).
21. IBM Corporation 2016: Analytics Solutions Unified Method. Implementations with Agile principles, ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf, (2016).
22. Weick, K.E., Sutcliffe, K.M., Obstfeld, D.: Organizing and the Process of Sensemaking. Organ. Sci. 16, 409–421 (2005). https://doi.org/10.1287/orsc.1050.0133.
23. Gregor, S., Kruse, L.C., Seidel, S.: The Anatomy of a Design Principle. J. Assoc. Inf. Syst. (2020).
24 Chandra, L., Seidel, S., Gregor, S.: Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions. In: 2015 48th Hawaii International Conference on System Sciences. pp. 4039–4048. IEEE, HI, USA (2015). https://doi.org/10.1109/HICSS.2015.485.

25. Michalczyk, S., Scheu, S.: Designing an analytical Information Systems Engineering method. In: ECIS (2020).

26 Jennex, M.E.: Re-Visiting the Knowledge Pyramid. In: 2009 42nd Hawaii International Conference on System Sciences. pp. 1–7. IEEE, Waikoloa, Hawaii, USA (2009). https://doi.org/10.1109/HICSS.2009.361.

27. Sanfilippo, E.M., Belkadi, F., Bernard, A.: Ontology-based knowledge representation for additive manufacturing. Comput. Ind. 109, 182–194 (2019). https://doi.org/10.1016/j.compind.2019.03.006.

28. Gulcehre, C., Bengio, Y.: Knowledge Matters: Importance of Prior Information for Optimization. J Mach Learn Res. 17, (2013).

29. Mariscal, G., Marbán, Ó., Fernández, C.: A survey of data mining and knowledge discovery process models and methodologies. Knowl. Eng. Rev. 25, 137–166 (2010). https://doi.org/10.1017/S0269888910000032.

30. Ferstl, O.K., Sinz, E.J.: Grundlagen der Wirtschaftsinformatik. Oldenbourg, München (2013).

31. Kritzinger, W., Karner, M., Traar, G., Henjes, J., Sihn, W.: Digital Twin in manufacturing: A categorical literature review and classification. IFAC-Pap. 51, 1016–1022 (2018). https://doi.org/https://doi.org/10.1016/j.ifacol.2018.08.474.

32. Hall, D., Paradice, D., Courtney, J.F.: Building a theoretical foundation for a learning-oriented knowledge management system. J. Inf. Technol. Theory Appl. JITTA. 5, 7 (2003).

33. QualiPro. https://agent3d.de/qualipro. Accessed 10.05.2021.

34. DIN SPEC 91345:2016-04. Deutsches Institut für Normung, DIN SPEC 91345:2016-04. Beuth-Verlag, 2016, https://www.beuth.de/de/technische-regel/din-spec-91345/250940128; Accessed 10.05.2021.

35. Simon, H.A.: The sciences of the artificial. MIT Press, Cambridge, Mass. (2008).

36. Klaus, G., Buhr, M.: Philosophisches Wörterbuch. VEB Enzyklopädie, Leipzig (1971).

37. Angée, S., Lozano-Argel, S.I., Montoya-Munera, E.N., Ospina-Arango, J.-D., Tabares-Betancur, M.S.: Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects. In: Uden, L., Hadzima, B., and Ting, I.-H. (eds.) Knowledge Management in Organizations. pp. 613–624. Springer International Publishing, Cham (2018), doi: https://doi.org/10.1007/978-3-319-95204-8_51

24th International Conference on Business Information Systems (BIS 2021)

Big Data

https://doi.org/10.52825/bis.v1i.60

Published: 02 July 2021

Optimization-based Business Process Model Matching

Merih Seran Uysal[1][https://orcid.org/0000-0003-1115-6601], Dominik Hüser[1], and Wil M.P. van der Aalst[1][https://orcid.org/0000-0002-0955-6940]

[1]Process and Data Science Chair, RWTH Aachen University, Aachen, Germany

{uysal,wvdaalst}@pads.rwth-aachen.de dominik.hueser@rwth-aachen.de

Abstract. The rapid increase in generation of business process models in the industry has raised the demand on the development of process model matching approaches. In this paper, we introduce a novel optimization-based business process model matching approach which can flexibly incorporate both the behavioral and label information of processes for the identifi-cation of correspondences between activities. Given two business process models, we achieve our goal by defining an integer linear program which maximizes the label similarities among process activities and the behavioral similarity between the process models. Our approach en-ables the user to determine the importance of the local label-based similarities and the global behavioral similarity of the models by offering the utilization of a predefined weighting param-eter, allowing for flexibility. Moreover, extensive experimental evaluation performed on three real-world datasets points out the high accuracy of our proposal, outperforming the state of the art.

Keywords: Process Model Matching, Optimization Problem, Integer Linear Programming, Behavioral Similarity

1 Introduction

The ubiquity of advanced capabilities of the digital world enables organizations to generate and store process models which exhibit indispensable activities of their business processes in various domains, e.g., finance, logistics, and production [1, 14, 18]. The resulting increase in uptake of business process model repositories leads to the need for the development of techniques in various fields, e. g. storage of process models, management of repositories, process querying, and process model matching.

Process model matching is the task of finding correspondences between the activities of two given process models. In particular, for very large process model repositories of organizations, it is essential to utilize process model matching techniques in order to determine similar models and merge them, eliminate redundancies, as well as alleviate storage and processing costs, and increase efficiency accordingly.

Most of the existing model matching techniques typically utilize activity labels and process structures to determine process matching in model repositories [12]. However, incorporating the behavior of the underlying process models is indispensable while detecting process match-ing. Unlike label-based and structure-based process matching approaches, behavioral process model matching takes the order of the activities in the models into consideration to attain a more reliable, accurate matching.

In this paper, we introduce a novel business process model matching approach **Opti**mization-based Process **M**odel M**a**tching (*OPTIMA*) which matches the individual components of two

given process models to each other by enabling the incorporation of both the label and behavioral information of the process models. Our proposal exhibits an optimization problem which maximizes the activity label similarities at an individual local level, and simultaneously maximizes the behavioral similarity of the given processes at a global level by utilizing their *relational profiles* [19, 21, 22, 24]. Thanks to the high flexibility of our approach, it is possible for the user to set the importance (i.e. weighting) of the behavioral information to be incorporated, as well as the label information of the process model components. Furthermore, our approach is completely independent of the application of a prior matching of activity labels, exposing a competitive advantage, when compared with some existing approaches. Moreover, our extensive comparative experimental evaluation performed on three real-world datasets points out the competitiveness of our proposal against the existing techniques, in particular outperforming the state of the art in terms of f-score performance.

Our paper is structured as follows: Section 2 gives an overview of the related work regarding business process model matching. Then, Section 3 presents the preliminaries including fundamental information about Petri nets, as well as relational profiles, and similarity functions we define. In Section 4, we introduce our proposal Optimization-based Process Model Matching (*OPTIMA*), followed by Section 5 which presents the extensive experimental results. Our paper is concluded by Section 6 with a conclusion and future work.

2 Related Work

Business process model matching has been a challenging research area where there have been numerous attempts to provide effective and accurate techniques. Process model matching describes the task of finding corresponding transitions in two given process models, whose roots stem from process model similarity [4, 6, 7, 16, 17] and ontology matching [8] relying on structural and label comparison of processes [2, 5]. Researchers have primarily developed label-based matching techniques which assesses the similarity of acitivity labels in process models. Exhibiting a well-known label-based approach, the basic Bag of Words (BoW) matching technique [11] first determines pairwise bag of words similarity among the labels of transitions, and a word similarity function is used, such as Levenshtein [23] or Lin [13], to compute all pairwise similarity scores and find out the highest scores for the matching.

In contrast to ontology and label-based matching, process models exhibit additional behavioral information which cannot be captured by only considering labels or process structures. Based on this fact, researchers have developed further approaches considering the behavioral information of process models. The authors of [12] propose a behavioral model matching approach which considers both label-based similarities and behavioral relations. After determining the semantic similarity of label components, match constraints are derived based on behavioral profiles [22] of the process models. These constraints are utilized towards a matching formalized as an optimization problem and solved by using Markov Logic Network inference. Another further model matching approach is proposed in [3] which is based on the quantitative bisimulation. First, process models are converted into labeled transition systems and then the degree of simulation is computed, followed by solving a linear program, corresponding to the overall bisimulation result. We refer to [16, 17] for a more comprehensive study of model matching approaches.

Since our proposal can incorporate both the information of activity labels and behavior of given two processes regulated by a parameter, it is noteworthy to use the Bag of Words approach, a label-based method, as a baseline for the label-based matching comparison for our evaluations later on.

3 Preliminaries

For investigating process model matching, we first use Petri nets and workflow nets as our formal grounding [1]. Then, we formulate the relational profile exhibiting a compact behavioral representation of a Petri net. Last, we present similarity functions and give the definition of the relation type similarity we require for our proposal later on.

3.1 Petri Nets

Originally introduced by C. Adam Petri [15], Petri nets are the most utilized process modeling language which allows for concurrency modelling, as well as the analysis of process models effectively. Below, we first present the definition of Petri net, labeled Petri net, and workflow net definitions and terms based on [1], serving as fundamentals for our paper.

Definition 1 (Petri net) *A* Petri net *is defined as a triplet* $N = (P, T, F)$ *where* P *is a finite set of* places*,* T *a finite set of* transitions *such that* $P \cap T = \emptyset$*, and* $F \subseteq (P \times T) \cup (T \times P)$ *is the* flow relation *denoting a set of directed arcs. A* marked Petri net *is defined as a pair* (N, M) *where* N *is a Petri net and* $M \in \mathbb{B}(P)$ *is a multi-set over* P *denoting the marking of the net.*

Definition 2 (Labeled Petri net) *Let* $\mathcal{A}$ *denote the* universe of activity labels*. A* Labeled Petri net *is a tuple* $N = (P, T, F, A, \lambda)$ *where* (P, T, F) *is a Petri net,* $A \subseteq \mathcal{A}$ *is the set of* activity labels*, and* $\lambda : T \to A$ *is the* labeling function*.*

For some particular transitions which are not observable, we use the notation τ, i.e. a transition t with $l(t) = \tau$ is unobservable and is referred to as *silent* or *invisible*. Furthermore, elements of $P \cup T$ are referred to as *nodes*. For any $x \in P \cup T$, the *pre-set of* x (a.k.a. input set), denoted $\bullet x$, is the set of nodes with a directed arc to x, i.e. $\bullet x = \{y \mid (y, x) \in F\}$. The *post-set of* x, denoted $x\bullet$, is the set of nodes with a directed arc from x, i.e. $x\bullet = \{y \mid (x, y) \in F\}$.

A marked, labeled Petri net is referred to as *labeled Petri net system*, denoted $S = (N, M_0)$, where $N = (P, T, F, A, \lambda)$ is a labeled Petri net and $M_0 \in \mathbb{B}(P)$ a multi-set over the places P, denoting the *initial marking*. We let $\mathcal{N}$ denote the *universe of marked labeled Petri nets*.

As convention, for any labeled Petri net system $S = (N, M)$ with $N = (P, T, F)$, we let $\mathcal{T}$ denote the *universe of transitions*, and $T_v(S) := \{t \in T \mid \lambda(t) \neq \tau\}$ be the set of non-silent (a.k.a. visible) transitions in S. For sake of simplicity, the notation $T_v(S)$ is replaced by T_S^v in the remainder of the paper, where necessary.

Given a labeled Petri net system (N, M) with $N = (P, T, F, A, \lambda)$, the transition $t \in T$ is *enabled* in marking M, denoted $(N, M)[t\rangle$, iff $\bullet t \leq M$. The *firing rule* $\_[\_\rangle\_ \subseteq \mathcal{N} \times T \times \mathcal{N}$ is the smallest relation satisfying for any $(N, M) \in \mathcal{N}$ and any $t \in T : (N, M)[t\rangle \implies (N, M)[t\rangle(N, M) \setminus \bullet t) \uplus t\bullet$.

For a given labeled Petri net system (N, M_0), a sequence $\sigma = \langle t_1, ..., t_n \rangle \in T^*$, with $n \in \mathbb{N}$, is called *firing sequence* of (N, M_0) iff there exist markings $M_1, ..., M_n$ such that for all i with $0 \leq i < n$, t_{i+1} is enabled in marking M_i, i.e. $(N, M_i)[t_{i+1}\rangle$, and firing t_{i+1} ends up in the marking M_{i+1}, i.e. $(N, M_i)[t_{i+1}\rangle(N, M_{i+1})$.

Workflow nets, a subclass of Petri nets, are highly relevant for business process modeling due to their strength in natural representation of the life-cycle of cases of the underlying process models [1]. The formal definition of workflow net is given below.

Definition 3 (Workflow net) *Given an identifier* $\bar{t} \notin P \cup T$*, a labeled Petri net* $N = (P, T, F, A, \lambda)$ *is called a* workflow net *(*WF-net*) iff (1)* P *contains a source place* $i \in P$ *(a.k.a. input place) such that* $\bullet i = \emptyset$*, (2)* P *contains a sink place* $o \in P$ *(a.k.a. output place) such that* $o\bullet = \emptyset$ *and (3) its short circuit net* $\bar{N} = (P, T \cup \{\bar{t}\}, F \cup \{(o, \bar{t}), (\bar{t}, i)\}, A \cup \{\tau\}, \lambda \cup \{(\bar{t}, \tau)\})$ *is strongly connected, i.e. there is a directed path between any pair of nodes in* $\bar{N}$*.*

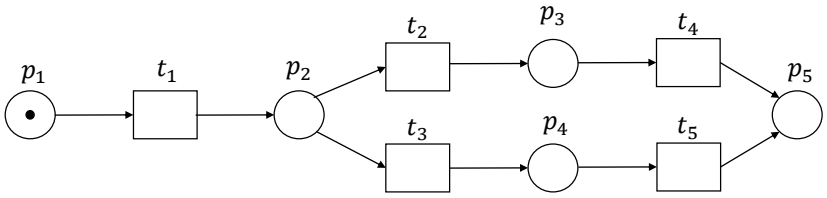


Figure 1. An example workflow net. The notation p_i denotes the i-th place and t_j denotes the j-th transition. The places p_1 and p_5 exhibit the input (aka source) place and output (aka sink) place, respectively.

Since WF-nets can expose processes with errors, such as deadlocks, activities that can never become active, still enabled intermediate transitions in spite of the process termination, etc., we need to define soundness criterion which is commonly used in the literature [20].

A workflow net $N = (P, T, F, A, \lambda)$ with an input place $i \in P$ and an output place $t \in P$ is called *sound* iff (1) $(N, [i])$ is safe, i.e. places cannot hold multiple tokens at the same time (*safeness*), (2) for any marking $M \in [N, [i]\rangle$: $o \in M \Rightarrow M = [o]$ (*proper completion*), (3) for any marking $M \in [N, [i]\rangle : [o] \in [N, M\rangle$ (*option to complete*), (4) for any transition $t \in T$, there is a firing sequence enabling t, i.e. $(N, [i])$ includes no dead transitions (*absence of dead parts*). Furthermore, a Petri net is *free-choice* if any two transitions sharing an input place have identical input sets, i.e. for all transitions $t_1, t_2 \in T, {\bullet}t_1 \cap {\bullet}t_2 \neq \emptyset \Rightarrow {\bullet}t_1 = {\bullet}t_2$. Figure 1 exhibits an example workflow net.

3.2 Relational Profiles

In order to give a compact behavioral representation of a Petri net, an appropriate structure is required which captures the relationships among its transitions. Below, we present the comprehensive definition of the *relational profile*.

Definition 4 (Relational profile) *Let $N = (P, T, F, A, \lambda)$ be a sound free-choice workflow net and $S = (N, M_0)$ the corresponding workflow net system. A* relational profile $\mathcal{R}^S = (\Psi, \Omega)$ *of S is a tuple comprising a set Ψ of* relation types *and an* assignment relation $\Omega \subseteq T \times T \times \Psi$ *which assigns pairs of transitions relation types. A transition $s \in T$ is in a relation $R \in \Psi$ with a transition $t \in T$, denoted sRt, iff $(s, t, R) \in \Omega$. $\mathcal{R}^S$ is called* mutually exclusive relational profile *if for all transitions $s, t \in T$ and all relation types $R_1, R_2 \in \Psi$ with $R_1 \neq R_2 : (s, t, R_1) \in \Omega \Rightarrow (s, t, R_2) \notin \Omega$.*

Since our proposal *OPTIMA* requires that profiles assign at most one relation per pair of transitions, we will consider such relational profiles satisfying the latter via the term *mutually exclusive profiles*, complying with [21].

Example. We consider a relational profile $\mathcal{R}^S = (\Psi, \Omega)$ of the workflow net in Figure 1 and two relation types *eventually-follows relation* $\succ \subseteq T \times T$ and *directly-follows relation* $> \subseteq T \times T$, resulting in $\Psi = \{\succ, >\}$. Note that (t_i, t_j) is in an eventually-follows relation if there exists a firing sequence which fires t_i before t_j. In contrast, (t_i, t_j) is in a directly-follows relation if there exists a firing sequence where t_j is fired after t_i without any visible transition in between. In Figure 1, we realize that $t_1 \succ t_4$ holds but t_1 is not directly-followed by t_4, i.e. $t_1 \not> t_4$, thus, $(t_1, t_4, \succ) \in \Omega$ and $(t_1, t_4, >) \notin \Omega$. In addition, $\mathcal{R}^S$ is not a mutually exclusive relational profile, since two transitions can exhibit more than one relation, e.g. $(t_1, t_2, \succ) \in \Omega$ and $(t_1, t_2, >) \in \Omega$.

3.3 Similarity Functions

After presenting the definition of *relational profile*, we now focus on the similarity computation of two relational profiles of the Petri nets at hand. Since we assume that mutually exclusive profiles are used to represent the behavior of Petri nets, we define a similarity function which

determines the similarity of two given relation types, corresponding to the behavioral similarity of two transitions exposing those relation types, such as directly-follows relation and eventually-follows relation. In this paper, since we consider the relational profiles of the α-relational Profile ($\alpha\mathcal{P}$) [19], the Behavioral Profile ($\mathcal{BP}$) [21, 22], and the BP+ profile ($\mathcal{BPP}$) [24], we define the *relation type similarity* function by using the aforementioned profiles. Please note that this similarity function is not limited to these profiles, and can easily be extended by the further profiles accordingly, where necessary.

Definition 5 (Relation type similarity) *Let S_1 and S_2 be sound and free-choice WF-net systems with relational profiles $\mathcal{R}^{S_1} = (\Psi, \Omega_1)$ and $\mathcal{R}^{S_2} = (\Psi, \Omega_2)$ of type $\mathcal{R} \in \{\mathcal{BP}, \alpha\mathcal{P}, \mathcal{BPP}\}$. The* relation type similarity $sim^{\mathcal{R}} : \Psi \times \Psi \to [0, 1]$ *of two relation types $R_1, R_2 \in \Psi$ is defined by:*

$$sim^{\mathcal{R}}(R_1, R_2) = \begin{cases} \mathbb{1}[R_1 = R_2] & \text{if } \mathcal{R} \in \{\mathcal{BP}, \alpha\mathcal{P}\} \\ w_{R_1,R_2} & \text{if } \mathcal{R} = \mathcal{BPP} \end{cases}$$

where the similarity value w_{R_1,R_2} of BP+ relation types stems from [24](Table 2).

The identification function $\mathbb{1}[\alpha] \in \{0, 1\}$ returns 1 if and only if the statement α is true, i.e. if the relation equivalence holds in the definition above.

Analogously, the *label-based similarity* function $sim^L : \mathcal{T} \times \mathcal{T} \to [0, 1]$ computes the similarity of the given two transitions, which will be utilized in the upcoming section, too.

4 Optimization-based Process Model Matching

In this section, we propose our novel approach **Opti**mization-based Process **M**odel M**a**tching (*OPTIMA*) which takes both local and global information of the underlying process models into consideration. This is achieved by utilizing the label information of the activity labels and the behavior information of both process models.

Our approach is presented as an optimization problem which maximizes the label similarities at an individual local level, and simultaneously maximizes the behavioral similarity of both processes at a global level by using their relational profiles. This is attained by defining an integer linear program which exhibits an optimization problem with a linear objective function, linear constraints, and variables which are defined to be integers [10].

In order to provide flexibility for the user, e.g. process owner, domain expert, etc., we introduce a weighting parameter w which determines how much importance will be attached to label information and behavioral information, aligning with the user intention. Moreover, our proposal is fully independent of the application of a prior matching of transition labels, which constitutes an important competitive advantage in comparison with some existing approaches. For sake of simplicity, the notations $T_v(S_1)$ and $T_v(S_2)$ will be replaced by T_1^v and T_2^v for the remainder of our paper, where required. Below, we first give the formal definition of our novel approach *OPTIMA* and then elaborate on its constraints:

Definition 6 (Optimization-based Model Matching) *Given two sound free-choice WF-net systems $S_1 = (N_1, M_0^{S_1})$, $S_2 = (N_2, M_0^{S_2})$ with $N_1 = (P_1, T_1, F_1, A_1, \lambda_1)$, $N_2 = (P_2, T_2, F_2, A_2, \lambda_2)$, and mutually exclusive relational profiles $\mathcal{R}^{S_1} = (\Psi, \Omega_1)$ of S_1 and $\mathcal{R}^{S_2} = (\Psi, \Omega_2)$ of S_2, let $sim^{\mathcal{R}} : \Psi \times \Psi \to [0, 1]$ be a relation type similarity of the profile type $\mathcal{R}$ and $sim^L : \mathcal{T} \times \mathcal{T} \to [0, 1]$ be a label-based similarity function. The* Optimization-based Model Matching (OPTIMA) $M \subseteq T_1^v \times T_2^v$ *is derived from the optimal solution of the following problem:*

$$\max\ w \sum_{\substack{s_1,s_2 \in T_1^v \\ t_1,t_2 \in T_2^v}} \frac{1}{m^2} y_{s_1,s_2,t_1,t_2} sim^{\mathcal{R}}(\mathcal{R}^{S_1}_{s_1,s_2}, \mathcal{R}^{S_2}_{t_1,t_2}) \ + \ (1-w) \sum_{\substack{s \in T_1^v \\ t \in T_2^v}} \frac{1}{m} x_{s,t} sim^{L}(s,t)$$

$$\text{s.t.} \quad \sum_{s \in T_1^v} x_{s,t} \leq 1 \qquad \forall t \in T_2^v \tag{1}$$

$$\sum_{t \in T_2^v} x_{s,t} \leq 1 \qquad \forall s \in T_1^v \tag{2}$$

$$2y_{s_1,s_2,t_1,t_2} \leq x_{s_1,t_1} + x_{s_2,t_2} \qquad \forall s_1, s_2 \in T_1^v, t_1, t_2 \in T_2^v \tag{3}$$

$$x_{s,t} \in \{0,1\} \qquad \forall s \in T_1^v, t \in T_2^v \tag{4}$$

$$y_{s_1,s_2,t_1,t_2} \in \{0,1\} \qquad \forall s_1, s_2 \in T_1^v, t_1, t_2 \in T_2^v \tag{5}$$

where $w \in [0,1]$ *denotes the weighting parameter, and* $m = \min\{|T_1^v|, |T_2^v|\}$.

The maximum number of simple correspondences of the two nets N_1 and N_2, i.e. the matching of single transitions of Petri nets, is determined by $m := \min\{|T_1^v|, |T_2^v|\}$. According to Constraint (4) above, for transitions $s \in T_1^v$ and $t \in T_2^v$, $x_{s,t} \in \{0,1\}$ indicates if s is matched to t (i.e. $x_{s,t} = 1$) or not (i.e. $x_{s,t} = 0$). Constraints (1) and (2) ensure that every transition of one WF-net is matched to at most one transition of the other WF-net.

The decision variable y is concerned with the aggregation of the information of two x variables: For visible transitions $s_1, s_2 \in T_1^v$ and $t_1, t_2 \in T_2^v$, Constraint (5) indicates if s_1 is matched to t_1 and simultaneously if s_2 is matched to t_2 (i.e. $y_{s_1,s_2,t_1,t_2} = 1$) or not (i.e. $y_{s_1,s_2,t_1,t_2} = 0$). Furthermore, Constraint (3) denotes the relationship between the variables x_{s_1,t_1}, x_{s_2,t_2}, and y_{s_1,s_2,t_1,t_2} which ensures that if $x_{s_1,t_1} = x_{s_2,t_2} = 1$, then the maximization problem results in $y_{s_1,s_2,t_1,t_2} = 1$ due to the nature of the problem definition.

The objective function comprising two summands aims to maximize the average label similarity between matched transitions, and maximize the behavioral similarity of both WF-nets, depending on their relational profiles. Finally, the obtained sum is normalized by the squared number of maximum possible simple correspondences m^2 to provide an objective value in $[0.0, 1.0]$.

5 Experiments

In this section, we first give details about the experimental system setup, datasets, and the process model matching approaches which are used in our evaluations. Then, we will present the extensive evaluation results.

5.1 Experimental Setup

5.1.1 System setup.

The implementation of programs is performed in JAVA 8 and experiments are conducted on 2 $\times$ Intel Xeon Gold 5115 CPUs, each consisting of 10 cores and 20 threads @ 2.40GHz with a total of 512 GB RAM DDR4-2400 and 15 $\times$ 400-AXQU 960 GB SSD with Ubuntu Linux 18.04. In addition, for our proposed *OPTIMA* approach, we utilize Gurobi 8.0.1 [9], and adopt the Petri net and behavioral profile implementation from the jBPT[1] library. The implementation utilized for the evaluation results presented in this paper is available and can be publicly checked out[2].

5.1.2 Datasets.

We use three real-world datasets which arise from the Process Model Matching Contests 2015 [2]. The first dataset is the University Admission Processes (abbreviated by *University*) comprising 36 model pairs derived from 9 models representing the application procedure for students

[1] https://github.com/jbpt/codebase
[2] https://github.com/domhues/ilp-matcher

	Characteristics	*University*	*Birth*	*Asset*
Before conversion to PNML	# model pairs	36*	36	36*
	# transitions (min)	12*	9	1*
	# transitions (max)	45*	25	43*
	# transitions (avg)	24.2*	19.3	18.6*
After conversion to PNML	# model pairs	21	-	17
	# non-silent transitions (min)	16	-	1
	# non-silent transitions (max)	32	-	21
	# non-silent transitions (avg)	24.1	-	6.4
Gold standard	% matched non-silent transitions	25.35%	65.95%	84.86%
	% unmatched non-silent transitions	74.65%	34.05%	15.14%
	% simple correspondences	83.3%	14.0%	22.0%
	% complex correspondences	16.7%	86.0%	78.0%
	% trivial correspondences	33.3%	4.0%	18.3%

Table 1. Characteristic information of the datasets *Birth, University,* and *Asset* ($*$ values are adopted from [2]).

at nine universities in Germany. The second dataset is the Birth Registration Processes (*Birth*) consisting of 36 model pairs that were derived from 9 models representing the birth registration processes of Germany, Russia, South Africa, and the Netherlands. The third dataset is Asset Management Processes (*Asset*) which includes 36 model pairs that were derived from 72 models from an SAP Reference Model Collection covering the fields of finance and accounting. Since the *University* and *Asset* datasets originally include process models of BPMN and EPML formats, respectively, these models are first converted to Petri nets, i.e. PNML format, so that process model matching approaches and our proposal can be evaluated.

It is noteworthy to state that model pairs available in the datasets *Birth, University,* and *Asset* are associated with a *gold standard* indicating the ground truth corresponding to the optimal matching of the process model pairs. The gold standard is derived manually by making use of the human expert knowledge, comprising simple (*1:1*) and complex (*1:n*) correspondences.

Table 1 presents key characteristics about the three aforementioned datasets. As mentioned above, the process models in the datasets *University* and *Asset* are of BPMN and EPML formats, respectively, which are converted to Petri nets, i.e. PNML format. It is visible that the conversion of the models from BPMN and EPML into PNML format affects the number of model pairs which are then used for the matching evaluations. The reason for obtaining less process models after the format conversion is that some transformed models are not free-choice models any more to which relational profiles cannot be applied. Since the *Birth* dataset comprises the process models of the PNML format, there is no need to apply any model conversion, i.e. 36 process model pairs remain for the experimental evaluation. In addition, the *University* dataset indicates the highest number of non-silent transitions (24.1 transitions), while the *Asset* dataset has the lowest number of non-silent transitions (6.4 transitions). Furthermore, the *Asset* dataset includes minimum 1 non-silent transition, while the *University* dataset shows the highest minimum number of non-silent transitions (16).

At the bottom of Table 1, we realize some information regarding the gold standard indicating the percentage of the matched and unmatched transitions, as well as the information of simple and complex correspondences. For the *University* dataset, 25.35% of the non-silent transitions are considered as mapped by the gold standard, while 74.65% remain unmatched. Furthermore, the *University* dataset indicates a high percentage of simple correspondences (83.3%) in the gold standard, while the *Asset* dataset shows a high percentage of complex correspondences (78%) in its gold standard. Please note that a low percentage of simple correspondences corresponds to a high percentage of complex correspondences.

5.1.3 Approaches.

As given above, our proposal *OPTIMA* utilizes both label-based and behavioral information of process models by means of a weighting parameter w. We vary $w \in \{0.0, 0.1, ..., 1.0\}$ and utilize the basic Bag of Words label similarity function with Lin word similarity function. Furthermore, we consider the following individual relational profiles of process models: the α-relational Profile ($\alpha\mathcal{P}$) [19], the Behavioral Profile ($\mathcal{BP}$) [21, 22], and the BP+ profile ($\mathcal{BPP}$) [24]. Please note that the evaluation and comparison of various label similarity functions is not the scope of this work.

In order to provide a fair empirical study, we consider the existing process model matching approaches which in particular take the behavioral information of process models into consideration, too. First, we utilize the Markov Logic Network model matching approach [12] with two different labeling functions, namely **R**efactored label similarity function, as proposed by the authors, and the basic **B**ag of Words label similarity function with Lin word similarity function [13] (abbreviated by *Markov**R*** and *Markov**B***, respectively) so that the results can be compared with those of *OPTIMA*. Then, we use the bisimulation-based model matching approach (*Bisim*) [3] utilizing the basic Bag of Words label similarity function with Lin word similarity function so that the results are comparable with those of our proposal.

It is noteworthy to state that setting $w = 0$ leads to the fact that *OPTIMA* exposes only a label-based process model matching, while setting $w = 1$ provides our proposal to include only behavioral information for process model matching. Hence, regarding the former, we consider a baseline from the class of the label-based process model matching approaches, i.e. the Bag of Words process model matching approach [11] (*BoW*) with Lin word similarity function (with the threshold value of 0.7, as authors suggest).

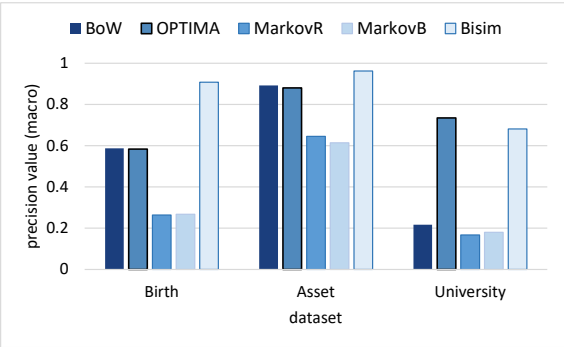
We compare our computed matching results, i.e. found correspondences in our matching, against a gold standard which is generated by authors in [2]. In this way, each found correspondence of the activity pair is determined to be in one of the following classes: true-positive (TP), true-negative (TN), false-positive (FP) or false-negative (FN). Taking this classification into consideration, we then calculate the precision $(TP/(TP + FP))$, the recall $(TP/(TP + FN))$, and the f-score $(2 \times precision \times recall)/(precision + recall)$.

In order to gather comparable, fair evaluation results, we utilize the average of precision, recall, and f-score values, referring to [2]: the *macro* evaluation considers the average of precision, recall, and f-score values over all test cases, while the *micro* evaluation is obtained by first summing up all true/false positives, true/false negatives, and then computing the precision, recall, and f-score values once at the end of the computation. Furthermore, since *Bisim*, *MarkovR, MarkovB*, and our proposal *OPTIMA* consist of parameters which are not necessarily preset by the authors, we present our results obtained by applying the parameters which lead to the highest micro f-score values, as stated in [2]. Due to space limitations, we focus on the precision, recall, and f-score comparison of our proposal and the aforementioned approaches.

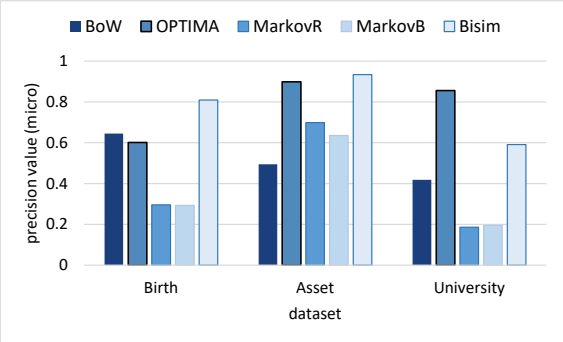
5.2 Experimental Results

After giving details about the experimental setup above, we now present our evaluation results. Based on *micro* and *macro* aggregation of the results, Figure 2 exhibits precision, recall, and f-score results of our proposal *OPTIMA* and four state-of-the-art approaches, respectively. By inspecting the plots in both figures, we note that micro and macro aggregated evaluation results of all approaches expose a similar tendency over utilized real-world datasets (*Birth, Asset, and University*), as anticipated.

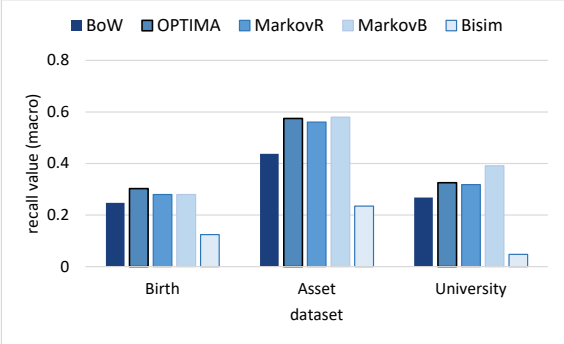
As stated before, we present the results attained by applying the parameters resulting in the highest micro f-score values, as proposed in [2]. The results of our approach *OPTIMA* are obtained by determining the best values attained with the BP+ profile ($\mathcal{BPP}$) with the weighting parameter $w = 0.4$ on the *Birth* dataset, with the BP+ profile ($\mathcal{BPP}$) with the weighting param-



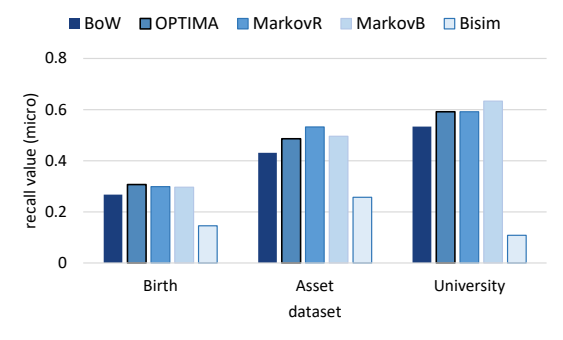
(a) Precision results obtained by *macro* aggregation.



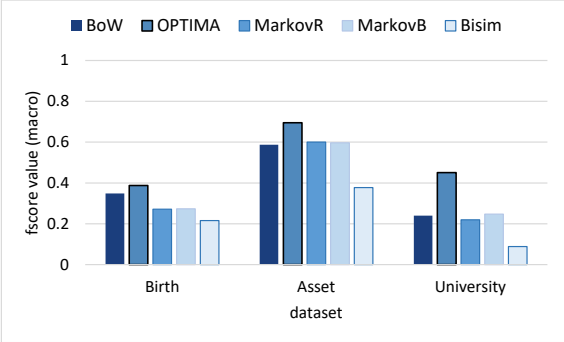
(b) Precision results obtained by *micro* aggregation.



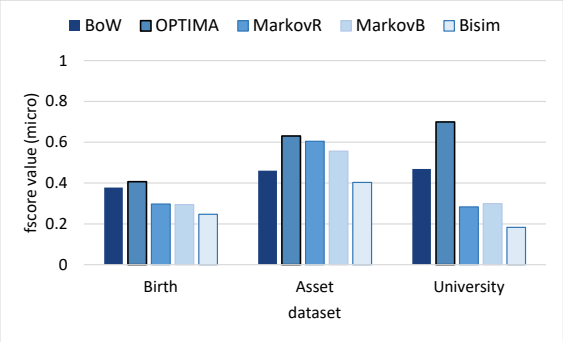
(c) Recall results obtained by *macro* aggregation.



(d) Recall results obtained by *micro* aggregation.



(e) F-score results obtained by *macro* aggregation.



(f) F-score results obtained by *micro* aggregation.

Figure 2. *Micro* and *macro* aggregation results of precision, recall, and f-score measures obtained on three real-world datasets, i.e. Birth Registration Processes (*Birth*), Asset Management Processes (*Asset*), and University Admission Processes (*University*) stemming from Process Model Matching Contest 2015 [2]. For a fair comparison, we utilize (i)Bag of Words process model matching approach [11] (*BoW*) indicating a baseline for a label-based approach (ii) Markov Logic Network model matching approach [12] with two variations *MarkovR* and *MarkovB* utilizing Refactored and Bag of Words label similarity functions (iii) Bisimulation-based model matching approach [3] (*Bisim*). Considering both label and behavior information of process models and being independent of the application of a prior matching of activity labels, our proposal *OPTIMA* outperforms all approaches regarding micro and macro f-score results on all real-world datasets.

eter $w = 0.5$ on the *University* dataset, and with the Behavioral Profile ($\mathcal{BP}$) with the weighting parameter $w = 0.2$ on the *Asset* dataset. Furthermore, for the bisimulation-based approach *Bisim*, a skip-penalty with the value 0.7 for *Birth*, *Asset*, and a skip-penalty with the value 0.9 for *University* lead to the highest values. In addition, the best values for Markov-based approaches *MarkovB*, *MarkovR* are attained by the constraint weight 0.001 for *Birth*, *University*, and the con-

straint weight 0.01 for the *Asset* dataset. Finally, for the Bag of Words process model matching approach (*BoW*) indicating a baseline for a label-based matching model technique, we utilize the threshold of 0.7 which is suggested by the authors in [11].

The results summarized in Figure 2(a)-2(b) report that the *Bisim* approach outperforms other approaches in terms of precision performance on the *Birth* and *Asset* datasets. Furthermore, *OPTIMA* exhibits the highest precision values on the *University* dataset, while the Markov-based approaches show the lowest precision performance. The slightly higher performance of *Bisim* over that of *OPTIMA* can be elucidated by the fact that the quantitative simulation exposes a higher or comparable expressiveness for model matching, when compared with the incorporation of the relational profiles. Furthermore, we observe that the *BoW* approach, indicating only a baseline label-based approach, shows a much higher precision performance than that of *MarkovB* and *MarkovR* on *Birth* and *Asset* comprising complex correspondences in their gold standard, i.e. ground truth. This can be elucidated by the fact that *BoW* can successfully detect complex correspondences, while Markov-based approaches can only find a smaller portion of complex correspondences on these datasets.

The results presented in Figure 2(c)-2(d) provide confirmatory evidence that our proposal *OPTIMA* outperforms existing approaches regarding the recall measure on the *Birth* dataset. In addition, *OPTIMA* indicates a comparable recall performance when compared with *MarkovR* and *MarkovB* on all datasets, which can be explained by that fact that both Markov-based approaches and our proposal can detect simple correspondences successfully. Furthermore, *BoW* shows a comparable recall performance on all datasets, outperforming *Bisim*. An interesting observation is the considerably poor recall performance of *Bisim* on all datasets. This suggests that the applied quantitative simulation technique is eligible for detecting only a small fraction of the relevant results.

A closer examination of macro and micro aggregation indicates that the macro aggregation of the precision, recall, and f-score measures reflects lower values on the *University* dataset than the micro aggregation results. This posits that some model pairs in *University* expose a relatively poor performance in the three measures, directly resulting in lower macro aggregated values since every model matching contributes equally to the computation of macro scores. In contrast, the corresponding micro aggregation results seem to be higher, since the influence of the poor performance of some particular models in the aforementioned dataset does not substantially contribute to the computation of micro aggregation results at all.

As presented in Figure 2(e)-2(f), *OPTIMA* considerably outperforms all state-of-the-art matching approaches regarding both micro and macro aggregation f-score results on all three real-world datasets. The intuition behind this observation lies in the high precision and recall results of our proposed approach, while other approaches exhibit a smaller result either in precision or recall.

6 Conclusion

One of the major challenges and key components in today's organizations is the ever-increasing amounts of business processes, resulting in the need for novel effective process model matching techniques in huge process model repositories. Providing direct insight into process model matching, this paper introduces a novel business process model matching approach *Optimization-based Process Model Matching* (*OPTIMA*) which matches individual components of two given process models to each other by incorporating both label and behavioral information of the process models. We present an optimization problem maximizing the activity label similarities at a local level, and the behavioral similarity of the given processes at a global level by leveraging relational profiles. Being fully independent of any prior matching of activity labels, our proposal shows high competitiveness against existing techniques, in particular outperforming the state

of the art in terms of accuracy performance.

An interesting direction for future work concerns the analysis of the complex correspondences which can potentially shed on light on the further matching strategies. Furthermore, we intend to conduct research into the evaluations on various real-world datasets in order to gain more insights, as well as examine the execution time of our approach to evaluate its performance. In addition, examining the efficiency of the proposal, as well as reducing the number of variables in the maximization problem to attain higher efficiency can be dedicated for future examination.

Acknowledgments

We would like to thank the Alexander von Humboldt (AvH) Stiftung for supporting our research.

References

[1] van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer, Heidelberg, 2 edn. (2016)

[2] Antunes, G., Bakhshandeh, M., Borbinha, J., Cardoso, J., Dadashnia, S., Francescomarino, C.D., Dragoni, M., Fettke, P., Gal, A., Ghidini, C., Hake, P., Khiat, A., Klinkmüller, C., Kuss, E., Leopold, H., Loos, P., Meilicke, C., Niesen, T., Pesquita, C., Peus, T., Schoknecht, A., Sheetrit, E., Sonntag, A., Stuckenschmidt, H., Thaler, T., Weber, I., Weidlich, M.: The Process Model Matching Contest 2015. In: EMISA'15: International Workshop on Enterprise Modelling and Information Systems Architecture. pp. 127–155. GI, Innsbruck, Austria (Sep 2015)

[3] Becker, J., Breuker, D., Delfmann, P., Dietrich, H.A., Steinhorst, M.: Identifying Business Process Activity Mappings by Optimizing Behavioral Similarity. In: AMCIS. vol. 1, p. Paper 21 (01 2012)

[4] Becker, M., Laue, R.: A Comparative Survey of Business Process Similarity Measures. Computers in Industry **63**(2), 148 – 167 (2012)

[5] Cayoglu, U., Dijkman, R., Dumas, M., Fettke, P., García-Bañuelos, L., Hake, P., Klinkmüller, C., Leopold, H., Ludwig, A., Loos, P., Mendling, J., Oberweis, A., Schoknecht, A., Sheetrit, E., Thaler, T., Ullrich, M., Weber, I., Weidlich, M.: Report: The Process Model Matching Contest 2013. In: Lohmann, N., Song, M., Wohed, P. (eds.) Business Process Management Workshops. pp. 442–463. Springer International Publishing, Cham (2014)

[6] Dijkman, R.M., Dumas, M., van Dongen, B.F., Käärik, R., Mendling, J.: Similarity of Business Process Models: Metrics and Evaluation. Inf. Syst. **36**(2), 498–516 (2011)

[7] Dumas, M., García-Bañuelos, L., Dijkman, R.M.: Similarity Search of Business Process Models. IEEE Data Eng. Bull. **32**(3), 23–28 (2009)

[8] Euzenat, J., Shvaiko, P.: Ontology Matching. Springer Publishing Company, Incorporated, 2nd edn. (2013)

[9] Gurobi Optimization LLC: Gurobi Optimizer Reference Manual (2019), `http://www.gurobi.com`

[10] Hillier, F., Lieberman, G.: Introduction to Linear Programming. McGraw-Hill (1990)

[11] Klinkmüller, C., Weber, I., Mendling, J., Leopold, H., Ludwig, A.: Increasing Recall of Process Model Matching by Improved Activity Label Matching. In: Business Process Management. pp. 211–218. Springer Berlin Heidelberg (2013)

[12] Leopold, H., Niepert, M., Weidlich, M., Mendling, J., Dijkman, R., Stuckenschmidt, H.: Probabilistic Optimization of Semantic Process Model Matching. In: Business Process Management. pp. 319–334. Springer Berlin Heidelberg (2012)

[13] Lin, D.: An Information-theoretic Definition of Similarity. In: Proc. of the 15th International Conference on Machine Learning. vol. 98, pp. 296–304. Morgan Kaufmann (1998)

[14] Pegoraro, M., Uysal, M.S., van der Aalst, W.M.P.: Discovering Process Models from Uncertain Event Data. In: Business Process Management Workshops. pp. 238–249. Springer International Publishing, Cham (2019)

[15] Petri, C.A.: Kommunikation mit Automaten. Schriften des Rheinisch-Westfälischen Institutes für Instrumentelle Mathematik an der Universität Bonn, Technische Hochschule, Darmstadt. (1962), https://books.google.de/books?id=NCZMvAEACAAJ

[16] Schoknecht, A., Thaler, T., Fettke, P., Oberweis, A., Laue, R.: Similarity of Business Process Models – A State-of-the-Art Analysis. ACM Comput. Surv. **50**(4), 52:1–52:33 (Aug 2017)

[17] Thaler, T., Schoknecht, A., Fettke, P., Oberweis, A., Laue, R.: A Comparative Analysis of Business Process Model Similarity Measures. In: Business Process Management Workshops. pp. 310–322. Springer International Publishing, Cham (2017)

[18] Uysal, M.S., van Zelst, S.J., Brockhoff, T., Ghahfarokhi, A.F., Pourbafrani, M., Schumacher, R., Junglas, S., Schuh, G., van der Aalst, W.M.: Process Mining for Production Processes in the Automotive Industry. In: Industry Forum at BPM 2020 co-located with 18th International Conference on Business Process Management (BPM 2020), Sevilla, Spain (2020)

[19] van der Aalst, W., Weijters, T., Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering **16**(9), 1128–1142 (Sep 2004)

[20] van der Aalst, W.: The Application of Petri-nets to Workflow Management. Journal of Circuits, Systems and Computers **8**(1), 21–66 (1998)

[21] Weidlich, M., Mendling, J., Weske, M.: Efficient Consistency Measurement Based on Behavioral Profiles of Process Models. IEEE Transactions on Software Engineering **37**(3), 410–429 (May 2011)

[22] Weidlich, M., Mendling, J., Weske, M.: Computation of Behavioural Profiles of Process Models. Business Process Technology, Hasso Plattner Institute for IT-Systems Engineering. Potsdam (2009)

[23] Weigel, A., Fein, F.: Normalizing the Weighted Edit Distance. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing. vol. 2, pp. 399–402 vol.2 (Oct 1994)

[24] Wen, L., Song, J., Wang, J., Kumar, A.: BP+: An Improved Behavioral Profile Metric for Process Models. https://www.researchgate.net/publication/286932844 (2015), accessed: 01.02.2021

24th International Conference on Business Information Systems (BIS 2021)

Big Data

https://doi.org/10.52825/bis.v1i.66

Published: 02 July 2021

Database-less Extraction of Event Logs from Redo Logs

Dorina Bano[1], Tom Lichtenstein[1], Finn Klessascheck[1], and Mathias Weske[1]

[1]Hasso Plattner Institute, University of Potsdam, Germany

Abstract. Process mining is widely adopted in organizations to gain deep insights about running business processes. This can be achieved by applying different process mining techniques like discovery, conformance checking, and performance analysis. These techniques are applied on event logs, which need to be extracted from the organization's databases beforehand. This not only implies access to databases, but also detailed knowledge about the database schema, which is often not available. In many real-world scenarios, however, process execution data is available as redo logs. Such logs are used to bring a database into a consistent state in case of a system failure. This paper proposes a semi-automatic approach to extract an event log from redo logs alone. It does not require access to the database or knowledge of the database schema. The feasibility of the proposed approach is evaluated on two synthetic redo logs.

Keywords: Event Log, Redo Log, Database Schema, Log Extraction, Process Mining

1 Introduction

An event log is an intrinsic ingredient of process mining that is comprised of process execution data [1]. In traditional process mining, event logs are extracted from the databases of a given organization. Typical event log extraction approaches require extensive access to the database tables and detailed knowledge about the database schema [2]. If this information is not available, other sources of information have to be tapped to extract event logs, primarily redo logs. Redo logs are an attractive choice, since these logs not only hold values stored in the database, but also a temporal ordering of the modifications that were applied to these data.

Several Data Base Management Systems (DBMSs), like Oracle RDBMS [1], provide redo logs to store conducted data operations and use them to bring the database into a consistent state in case of system failure. Existing approaches propose solutions on how to derive event logs from redo logs. These approaches rely on information about the database and its schema [3]–[5]. In this paper we argue that it is possible to extract an event log by just considering redo logs as a single source of information.

Therefore, this paper proposes database-less extraction of event logs from redo logs. We are following a two-step semi-automatic approach where in the first step the database schema is automatically inferred from the redo log. The inferred schema is evaluated by a domain expert and it is used to extract an event log based on selected case notion. In the second step, the database schema is used to correlate the redo log events into the event log traces. The feasibility of our approach is tested on two synthetic redo logs.

[1]https://www.oracle.com/database/technologies/

The reminder of this paper is organized as follows. Section 2 briefly discusses the basic notions needed to understand the rest of the paper. The approach for extracting an event log from redo log is described in Section 3. The related work are briefly discussed in section 4. Consecutively, section 5 provides an evaluation of the proposed approach whiles section 6 concludes the paper.

2 Preliminaries

This section introduces the basic notions and concepts regarding the event log and redo log, which we refer to throughout this paper.

The starting point of any process mining techniques is an event log, which is defined as a ordered collection of events. Each event is represented by three mandatory attributes: the case identifier, which identifies the process execution instance; the activity name, which represents a well-defined step of the process execution; and, the timestamp, which represent the time occurrence of the event. In addition, an event log can store several optional attributes such as resources and organizations, showing the business unit where the process is executed. All events pertaining to the same case identifier establish the case.

One can make use of a plethora of event log extraction approaches to construct an event log around a specific case notion from a database of a given organization [2], [6]. Therefore, it is possible to extract several event logs, each pertaining to a different case notion, from the same source of information. To achieve that goal, the extraction procedure requires access to the database and deep knowledge of the schema. However, this access and knowledge cannot always be secured. Instead, usually another kind of log is more ready available—*Redo logs*.

A Redo log stores all changes of the database as they occur. Each entry in a redo log corresponds to an transaction executed by the database system. Database systems like Oracle RDBMS enable redo logs to store the historic view on what has happened in the system. In a real-life scenario these logs are used to restore the database to the consistent state in case of system failures.

Each transaction in the redo log is represented as an SQL statement (called redo entry) and consist of (1) the *operation* made upon a certain database table i.e., insert, delete, and update; (2) the affected *attributes* and the corresponding values, (3) the *row id* on which the statement must be applied, (4) the *timestamp* of the statement occurrence.

Some examples of the redo entries are illustrated in Listing 1, which shows several redo entries in a medical database system. Suppose that a patient is admitted to a hospital and afterwards diagnosed by a doctor. Each statement represents a redo entry. In the first redo entry two attribute value are inserted into the *Patients* table. While in the second second one an update operation is made upon the *Diagnoses* table. The third redo entry includes a delete operation over the *Admissions* table.

```
1 insert into "SYSTEM"."PATIENTS"("ID","GENDER") values ('86','M') AAAT
    ; 08-JAN-2021 12:46:15
2 update "SYSTEM"."DIAGNOSES" set "ID" = '584' where "ADMISSION_ID" = '
    101' and ICD_CODE='P599' and ROWID = 'BADR'; 08-JAN-2021 12:45:33
3 delete from "SYSTEM"."ADMISSIONS" where "ID" = '32' and "PATIENT_ID"
    = '34' and ROWID = 'SABD'; 08-JAN-2021 12:46:27
```

Listing 1. Redo log fragment: each statement corresponds to an transaction made upon the database and is called redo entry

In this paper we provide a method for extracting an event log from a redo log without the knowledge of the underlying database schema. Below a detailed explanation of each step is

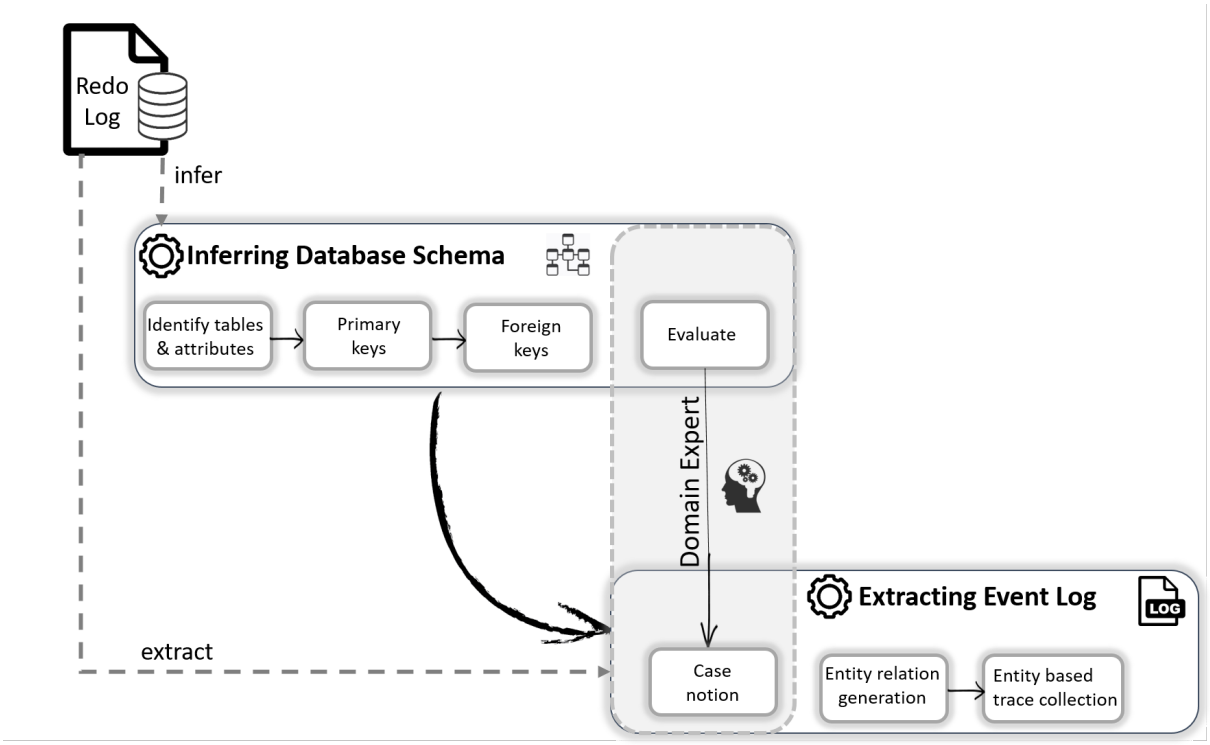


Figure 1. Overview of the approach followed to discover an event log from a redo log

given.

3 Extraction of event logs from redo logs

As mentioned in the previous section, the redo logs store all the changes of the database as they occur in the form of SQL statements. Our aim is to extract an event log by only considering the redo log as a single source of information as shown in Figure 1.

To correlate the redo entries into the event log traces a database schema is, nevertheless, needed. Therefore, the first step of our approach is to infer the database schema from a redo log without having access to the actual database. Afterwards, the domain experts comes into play to evaluate the inferred database schema and decide about the case notion by having in mind the goal of the process model. The last step of our approach is is to extract the event log from the redo log in conjunction with the database schema. The resulting event log is defined as a set of traces and each trace contains a set of events.

The assumption is that the redo log contains sufficient information to be able to extract the database schema and, in consequence, the event log. The bigger the redo log is and the more variation of the software system transactions it contains, the more accurate the inferred database schema becomes.

3.1 Inferring the database schema from redo logs

An important step that has to be taken before extracting an event log is to discover the database schema, which is used to correlate the data taken from the redo log into the event log traces. The database schema inferring process consists of identifying (1) the *database tables with their attributes*; (2) the *primary key* of each table; (3) the *foreign key* relations needed to determine the relation between tables.

Database schema tables and attributes detection To discover the database schema tables we look at each redo entry and filter out the table names. For each operation made upon the certain table in the redo entry a new database table is constructed. If the table name already exists in the database schema then we move forward to the next redo entry. Since each redo entry contains only one operation made upon one table then this step requires only one iteration over all redo log entries.

Once the database tables are constructed we have to identify for each table the affected attributes. Following the same idea, we go through all the redo entries and extract for each table the affected attribute names. If a new attribute is detected for a table, it is added to the corresponding table. Otherwise, the attribute is ignored as it is already in the schema.

The next step of the schema extraction is to define the relation between tables based on the attribute values involved in each redo entry. This implies the need to discover the primary keys and foreign keys. Since the redo logs do not contain explicit information about these types of keys we argue that this information can be extracted from the data pertaining to each attribute.

Primary key detection Primary keys are important constraints in the database indicating the attributes relations that hold in a database. By definition a primary key attribute contains only unique values [7]. This implies the need of checking the values of each attribute discovered from the previous step. If duplicate values appear for a certain attribute then this attribute is not a primary key candidate.

However, checking for unique values is necessary but not sufficient to detect the primary key. It might happen that a certain attribute contains unique values and can easily be misinterpreted as a primary key (e.g., an attribute which stores the value of the account balance). To overcome this issue, we further check if the attribute values appear in ascending order throughout the redo log. If this is not a case, the attribute is not considered as a primary key candidate even if its values are unique.

To increase the accuracy in the primary key identification step, we are also considering the attribute name suffix. In order to make the database more readable and easy to maintain, it is a common practice in reality for the primary keys to contain a suffix like: 'key', 'id', 'nr' [8].

Foreign key detection The foreign keys are used to identify the relations between the previously created tables. One difference between the primary keys and foreign keys is that we can have several foreign keys but only one primary key for each table. Another difference is that foreign keys require the existence of the primary key since they are used to reference a primary key.

To identify the foreign keys between the constructed tables and the predefined primary keys we rely on the *inclusion dependency* relation, which means that all the values of the referencing attribute must be contained in the referenced attribute. For example, if we have two attributes called X and Y respectively, each pertaining to two different tables in the database, we can say that all attribute values of foreign key Y must be present in the attribute values of the primary key X. This condition needs to be satisfied only on one direction, which implies that the primary key attribute might have additional values.

Nevertheless, the inclusion dependency is required but it does not suffice to detect the foreign keys. It might happen that there are, for example, 100 entries in the primary key attribute, 50 of which appear in another attribute of a different table. This means that this attribute is a candidate for foreign key. To increase the chances for discovering the correct foreign key in the same way as we did with the primary keys we will consider also the suffix attribute name.

Before using the inferred database schema as input for the event log extraction step we leave to the domain expert to review it.

3.2 Event log extraction

The event log extraction takes as input the redo log and the inferred database schema. It is the database schema that helps us to correlate the redo entries into the event log traces. The domain expert supports this transformation by picking the case notion while having in mind the desired process view. Specifically, the selected case notion denotes which database schema table will be used to determine the event log traces.

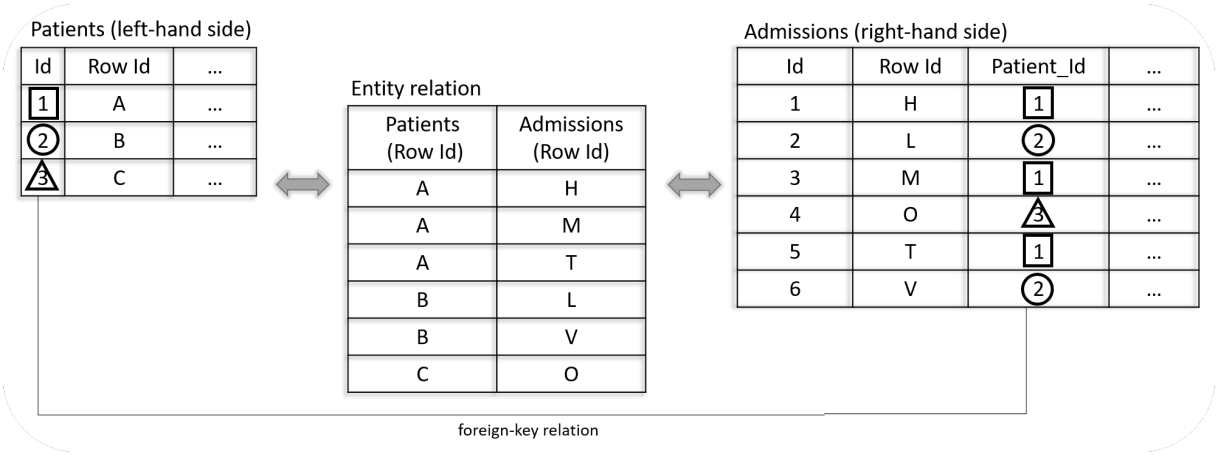


Figure 2. Example of an entity relation table creation. *Prescription* table is in relation with the *Pharmacy* table via the foreign key between *User Id* and *Id* attribute

Entity Relation Generation For each pair of tables in the database schema that are in relation via the discovered foreign keys the following method is applied: For each entity on the left-hand side of the referenced table it is checked whether one or more entities exist on the right-hand side of the depended table (i.e., the one that contains the foreign key attribute). If this is the case, the *Row Id* of the depended table together with the *Row Id* of the referenced table are added to an entity relation table.

This is done until all entities of the left-hand side table have been considered. The method is repeated for all tables that are in a foreign key relation outputting an entity relation table for each pair. This requires the *Row Id* to be unique. However, it might happen that several rows in the database might have the same row id. This occurs if a delete operation has happened on that row before. If a certain row is deleted from the database, its row id can be reassigned to another row via the insert operation. Therefore, a pre-processing step is needed to overcome this limitation.

Before constructing the entity relation table the redo log is parsed and whenever a redo entry that contains a delete operation occurs it's row id is tracked. If another redo entry is inserted and it occupies one of the tracked row id, then a unique suffix is appended to that row id. This is repeated as soon as the next delete statement for that row id is encountered. In this way, we make sure that each row id of the parsed redo log belongs exactly to one redo entry.

Figure 2 provides an example how the entity relation table can be created between two tables (called *Patients* and *Admissions*) that are in relation through the foreign key. Suppose that from the previously inferred database schema we get the information that each table has an primary attribute called *Id*. In addition, the *Patient_Id* attribute of the *Admissions* table (right-hand side) is assigned as a foreign key of the *Patients* table (left-hand side). For the entity with *Id* equal to *1* in the *Patients* table we will check all entities in the *Admissions* table that have the same value to the foreign key attribute (i.e., *Patient_Id*). As a result there are three entities that satisfy this conditions (highlighted with rectangle). In consequence, the corresponding *Row Id* of both tables are added to the entity relation table (first three entities of the entity relation table).

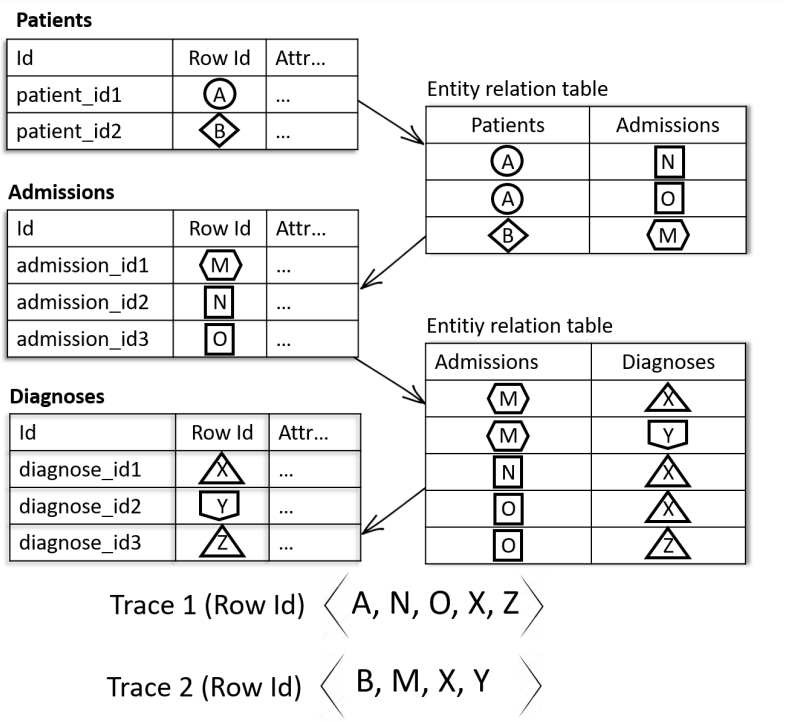


Figure 3. An example of event log extraction after applying the entity-based trace collection step on the inferred database schema. The Patients table is picked as root class/case notion by a domain expert

The same logic is followed for the entity with *Id* equal to 2 (highlighted with circle) and 3 (highlighted with triangle) in the *Patients* table. The resulted entity relation table has six relations in total. By default, there cannot be duplicates due to the uniqueness of the primary key. Once the relation is constructed for each pair of the tables that are in a foreign key relation, the next step is to define the event log traces.

Entity-based Trace Collection The goal of the entity-based trace collection step is to define the event log traces by considering the entity relation table and the redo entries. In process mining each trace is defined as as a collection of events. Each event is determined by a case id, activity name, timestamp and list of other optional attributes. To discover an event log trace we are considering for each row id of the root class (from now on let us call it RCID), picked by the domain expert, all row ids (from other tables) that are in relation with the RCID. This implies that for each RCID an event log trace is constructed.

To discover all row ids that are in relation with one RCID, the entity relation table comes into play. Starting from the the first RCID, we check the entity relation table for the entities that are in relation with this RCID. Let us call all entities that are in relation with the given RCID the *target entities relation*. Each target entity relation corresponds to one or more *Row Ids* of another table. Therefore the same step is followed to find all *Row Ids* that are in relation with the second table. In the same way we iterate through all tables for which a predefined entity relation table exists. For each RCID a trace is created as a list of all discovered Row Ids relating to it. Each event in this trace has a case id equal to the RCID entity. The event's activity name and timestamp are extracted from the redo log entries based on the respective Row Id event. Depending on the scope of the to-be-discovered business process model, several optional attributes can be defined via the attributes lists considered in each redo log entry. All the traces defined through this method constitute the event log.

Figure 3 illustrates an example of the entity-based trace collection based on the scenario

presented in Section 2. The patient was admitted to a hospital and afterwards diagnosed by the doctor. Suppose that the inferred schema for this scenario has three tables *Patients*, *Admissions* and *Diagnoses*. The Admissions table is in a foreign key relation with the Patients table, and Diagnoses is in a foreign key relation with Admissions. For each pair of tables the corresponding entity relation table is illustrated in Figure 3. Let us assume that the domain expert has selected the Patients table as a root class (case notion). For RCID equal to A the entity relation table is checked to identify the target entities relation. There are two target entities N and O in the entity relation table which at the same time defines the Row Id of the related table. In the same way we go further and check the related Row Id of N and O by considering the entity relation between Admissions and Diagnoses.

We iterate through all RCID of the root class and identify a list of Row Ids which are in relation with A and B. Therefore, from the example illustrated in Figure 3 two traces are discovered: $\langle A, N, O, X, Z \rangle$ and $\langle B, M, X, Y \rangle$. As the last step, for each *Row Id* in the predefined trace the redo log is queried to retrieve the activity name and time. The case id of the first trace is equal to the Row Id of RCID, which is equal to A in the Patients table while the case id of the second trace is likewise defined as B.

4 Related work

Discovering an event log from a redo log is recently subject to research work. Murillas et al. in [4] propose a three-step approach (scope, bind and classify) to extract an event log from the changes captured by the underlying database. The author assumes that the class model and event model, which capture the changes of the underlying database, are known upfront. In contrast, we are extracting an event log by considering only the redo log as a single source of information.

Another approach for extracting an event log from the redo log is presented in [5]. The event log extraction step requires three objects as input: the redo log, the data model and the trace id pattern. The latter is extracted from the database and it is used to determine the case notion needed to produce the event log traces. More specifically, the trace id pattern is used to find a common set of attributes between different classes while considering the primary key and foreign key relations. Instead, our approach assumes the lack of the database and such trace id pattern can be derived from the redo log entries.

In [3], the authors give an comparison between the traditional process mining and redo log process mining based on the event log extraction phase. They have applied both approaches on real-life data and come to the conclusion that the event log extracted from the redo log are richer in terms of number of events. In contrast to our approach the extraction of the event logs requires as input the redo log and the data model which is accessed via the connected Oracle database.

To the best of our knowledge, this is the first work on transforming database transactions into an event log without requiring live access to the database.

5 Evaluation

Our approach is evaluated based on two synthetic redo logs. The first one is extracted from the MIMC_III (Medical Information Mart for Intensive Care III) real-life dataset [9] which contains electronic health records (EHRs) related to patients admitted to the critical care unit (CCU) at the Beth Israel Deaconess Medical Centre (BIDMC), in Boston, USA. A script is written to simulate a redo log based on the provided MIMIC dataset. The approach presented in this paper is implemented as a Scala-based CLI tool and can be found on GitHub [2]. The Oracle

[2] https://github.com/fyndalf/redo-log-parser

Figure 4. Inferred database schema from the simulated MIMC redo log

LogMiner [3] is used to archive and export the redo logs.

The approach presented in Section 3.1 is applied to the simulated redo log and the inferred database schema is illustrated in Figure 4. After comparing it with the original schema, we come to the conclusion that they are similar.

As the next step, we chose two case notions: *Patients* and *Admissions*. As the last step of our approach the event log was extracted based on the selected case notion. The process model was then discovered from the extracted event log via the Inductive Visual Miner algorithm [10], which is a ProM plug-in [11]. The discovered process model is illustrated in Figure 5.

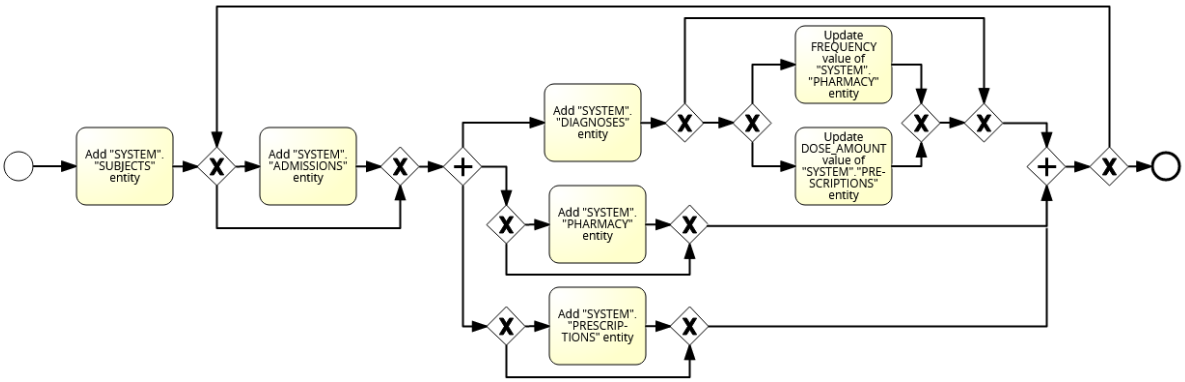


Figure 5. MIMC discovered process model for the Patient. The Inductive Visual Miner algorithm is applied with 70% threshold

From the discovered process we observe that the process starts with the insertion in the *Patient* table, which is happening because the *Patients* is selected as a root class. Afterwards, the insertion in the *Admissions* table can optionally happen. That is followed by the insertion in the *Diagnoses* table. In consequence the insertion of the *Prescriptions* or *Pharmacy* table can take place. The creation of all tables is followed by either updating the *Frequency* attribute in the *Pharmacy* table or updating the *DOSE_AMOUNT* attribute in the *Prescription* table which concludes the process.

Our approach is applied also to the *Ticket selling* dataset which used in [5]. This dataset

[3] https://docs.oracle.com/en/database/oracle/oracle-database/18/sutil/oracle-logminer-utility.html

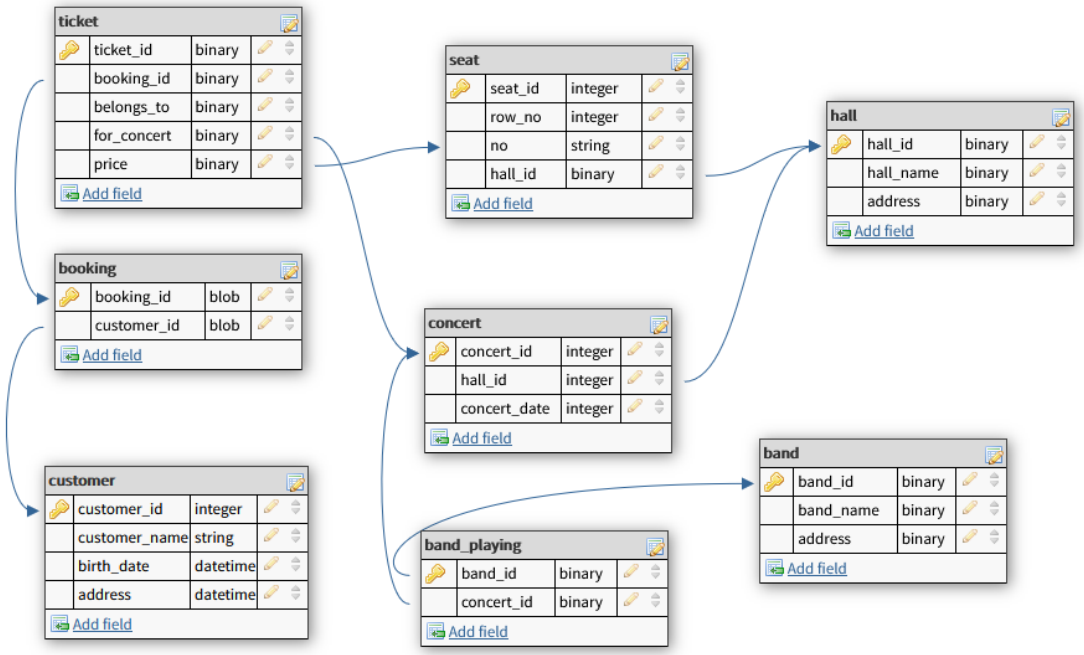


Figure 6. Inferred database schema from the Ticket selling redo log

contains events regarding a portal for selling concert tickets. The inferred database schema is shown in Figure 6 and it is compared with the original schema presented in [5]. There is only one extra relation inferred between *price attribute* in *Ticket table* and *no attribute* (seat number) in *Seat table* that is not manifested in the original schema. We checked the data and concluded that this seems to be an inconsistency in the original data: every ticket in the redo log has a price equal to 35 and there is one seat with the seat number equal to 35. We are assuming that the data are simulated and non-constant prices would not lead to this result.

6 Conclusion

In this paper we propose an approach to extract an event log from the redo log in the absence of a database. We show that the access to whole database and extensive knowledge about the database schema is not always necessary to extract an event log. To realize this, we propose a semi automatic two-step approach to output the desired event log. First, it is shown how to infer a database schema solely from the redo log. Then, the event log extraction is achieved by considering both the inferred database schema and the redo log. The assistance of the domain expert is required to select a case notion by having in mind the desired process model goal.

The feasibility of the approach is proven by developing a prototype which is applied on two synthetic redo logs. The inferred database schema it is proven to be similar to the original one proving the effectiveness of our approach. The discovered event log conforms to the XES [4] standard and can be used by the the process mining experts to apply different techniques like discovery, conformance checking and performance analysis.

As future work, we are aiming to further exploit redo logs as a rich source of information to enrich the discovered process model with additional information such as data objects. In addition we would like to further improve the performance and usability of our implementation tool.

[4]IEEE Task Force on Process Mining. XES Standard Definition. www.xes-standard.org

References

[1] W. M. P. van der Aalst, *Process Mining - Data Science in Action, Second Edition*. Springer, 2016, ISBN: 978-3-662-49850-7. DOI: 10.1007/978-3-662-49851-4. [Online]. Available: https://doi.org/10.1007/978-3-662-49851-4.

[2] S. Remy, L. Pufahl, J.-P. Sachs, E. P. Böttinger, and M. Weske, "Event log generation in a health system: A case study," in *Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13-18, 2020, Proceedings*, D. Fahland, C. Ghidini, J. Becker, and M. Dumas, Eds., ser. Lecture Notes in Computer Science, vol. 12168, Springer, 2020, pp. 505–522. DOI: 10.1007/978-3-030-58666-9_29. [Online]. Available: https://doi.org/10.1007/978-3-030-58666-9%5C_29.

[3] E. G. L. de Murillas, G. E. Hoogendoorn, and H. A. Reijers, "Redo log process mining in real life: Data challenges & opportunities," in *Business Process Management Workshops - BPM 2017 International Workshops, Barcelona, Spain, September 10-11, 2017, Revised Papers*, ser. Lecture Notes in Business Information Processing, vol. 308, Springer, 2017, pp. 573–587. DOI: 10.1007/978-3-319-74030-0_45. [Online]. Available: https://doi.org/10.1007/978-3-319-74030-0%5C_45.

[4] W. M. P. van der Aalst, "Extracting event data from databases to unleash process mining," in *BPM - Driving Innovation in a Digital World*, J. vom Brocke and T. Schmiedel, Eds., Springer, 2015, pp. 105–128. DOI: 10.1007/978-3-319-14430-6_8. [Online]. Available: https://doi.org/10.1007/978-3-319-14430-6%5C_8.

[5] E. González-López de Murillas, W. van der Aalst, and H. Reijers, "Process mining on databases: Unearthing historical data from redo logs," English, *Lecture Notes in Computer Science*, no. 9253, pp. 367–385, 2015, ISSN: 0302-9743. DOI: 10.1007/978-3-319-23063-4_25.

[6] K. Diba, K. Batoulis, M. Weidlich, and M. Weske, "Extraction, correlation, and abstraction of event data for process mining," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, e1346, 2020. DOI: 10.1002/widm.1346. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1346. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1346.

[7] L. Jiang and F. Naumann, "Holistic primary key and foreign key detection," *J. Intell. Inf. Syst.*, vol. 54, no. 3, pp. 439–461, 2020. DOI: 10.1007/s10844-019-00562-z. [Online]. Available: https://doi.org/10.1007/s10844-019-00562-z.

[8] A. Rostin, O. Albrecht, J. Bauckmann, F. Naumann, and U. Leser, "A machine learning approach to foreign key discovery," 2009.

[9] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160 035, 2016.

[10] S. J. J. Leemans, D. Fahland, and W. van der Aalst, "Process and deviation exploration with inductive visual miner," in *Proceedings of the BPM Demo Sessions Co-located with the 12th International Conference on Business Process Management (BPM)*, ser. CEUR Workshop Proceedings, vol. 1295, CEUR-WS.org, 2014, p. 46.

[11] B. F. van Dongen, A. K. A. de Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters, and W. van der Aalst, "The ProM Framework: A New Era in Process Mining Tool Support," in *Applications and Theory of Petri Nets 2005, 26th International Conference, ICATPN*, ser. Lecture Notes in Computer Science, vol. 3536, Springer, 2005, pp. 444–454. [Online]. Available: https://doi.org/10.1007/11494744%5C_25.

24th International Conference on Business Information Systems (BIS 2021)
Big Data
https://doi.org/10.52825/bis.v1i.67

Published: 02 July 2021

Towards a Concept for Building a Big Data Architecture with Microservices

Aamir Shakir, Daniel Staegemann, Matthias Volk, Naoum Jamous and Klaus Turowski

Otto-von-Guericke University Magdeburg

Abstract. Microservices and Big Data are renowned hot topics in computer science that have gained a lot of hype. While the use of microservices is an approach that is used in modern software development to increase flexibility, Big Data allows organizations to turn today's information deluge into valuable insights. Many of those Big Data architectures have rather monolithic elements. However, a new trend arises in which monolithic architectures are replaced with more modularized ones, such as microservices. This transformation provides the benefits from microservices such as modularity, evolutionary design and extensibility while maintaining the old monolithic product's functionality. This is also valid for Big Data architectures. To facilitate the success of this transformation, there are certain beneficial factors. In this paper, those aspects will be presented and the transformation of an exemplary Big Data architecture with somewhat monolithic elements into a microservice favoured one is outlined.

Keywords: Big Data, Microservice, Success Factors, Software Design, Software Architecture

1 Introduction

A continuous growth of the overall amount of data that is created, captured and analyzed heavily affects today's infrastructures as a whole [1]. With these advancements, an increase in processing speed is desirable. Consequently, the performance of conventional methods and technologies are growing insufficient, thus motivating the invention of new techniques and the need of modern data architectures, as they are denoted by the terms Big Data and Big Data architecture [2]. Organizations that implement aforementioned technologies can experience a significant increase in performance [3] due to more efficient decision making, reduced expenses and a more result-oriented portfolio of services while improving customer relations [4]. However, implementing the respective applications is still a challenging task [4]. Contributing to this problem, the newly implemented algorithms and solutions might evolve over time, thus rendering the effort pointless, if the Big Data architecture is not adapted accordingly. This results in a rising demand for designs, which consist of building blocks that can be modified and replaced independently from each other, contrasting isolated implementations and applications [5]. Implementing microservices as the building blocks of Big Data architecture appears to be a feasible solution due to high degree of modularization [6]. Therefore, the research question is:

What Factors should be considered to build a microservice favored Big Data architecture?

To provide an answer, at first, the concepts of microservices and Big Data are described. This is followed by a look at what research has already done in the area of microservices in Big Data. This shall be a foundation to derive success factors for a successful microservice based

Big Data architecture. Afterwards, an exemplary Big Data architecture with isolated elements will be transformed into a microservice favored Big Data architecture. he advantages and disadvantages are discussed and compared with the specified success factors. Finally, a conclusion is given, also highlighting potentially beneficial directions for future research endeavors.

2 Background

Big Data and microservices are introduced in the following. Their mode of operation and important properties are described. Also monolithic architectures as alternatives to microservice based systems are described and a state of the art of microservices in Big Data is given.

2.1 Big Data

While initially being referred to as a synonym for large amounts of data that cannot be easily handled by relational databases and technologies of that time, today Big Data covers a variety of advanced data characteristics, technologies, paradigms and methods [1]. During this time, the concept has undergone significant changes that dramatically changed the term from a hype topic [7] to the foundation of most of the data-driven and data-intensive projects known today [4]. Despite that long lasting maturation and a highly active research community [8], no distinct and universally applied definition was found that precisely describes the nature and elements of that term [1]. Notwithstanding that, according to one of the most widely used definitions, Big Data "consists of extensive data sets -primarily in the characteristics of volume, variety, velocity, and/or variability -that require a scalable architecture for efficient storage, manipulation, and analysis" [9]. Another definition which can be used is that Big Data is a field that systematically deals, extracts information of, or finds ways to analyze data volumes that are for example too complex, too fast-moving, too large or too weakly structured to be analyzed using manual and conventional data processing methods [10]. Similar to the pure definition itself, many differences about the description of the data exist. While some of the data characteristics are observed as core characteristics, namely volume, variety and velocity, others are treated unequally [11]. Volume refers to the size and number of elements to be processed, the variety focuses on the structure of data, which can be either unstructured, semi-structured or structured. Furthermore, the velocity describes the speed the data is incoming and processed with [12]. Apart from the pure lack of experts and qualified staff [13], the comprehensive planning, engineering and integration of architectures represents a cumbersome task [1]. Many practitioners and researchers noted this problem and attempted to reduce the prevailing complexity through the design and developments of promising solutions, such as reference architectures [14], decision support systems [4], automation approaches [15] or the application of new technologies [6]. Especially in times at which highly decentralized or loosely coupled environments are sought more than ever, as in the case of very large business application scenarios, the use of Big Data in combination with the latter remains desirable.

2.2 Monolithic Architectures

Monolithic architectures follow a traditional model for software in which the structure is a single and indivisible unit [16]. A monolith has one code base with multiple modules, like the described Big Data architecture in Section 5. Those modules can consist of one or multiple layers [17]. Figure 1 shows an example with three layers [18]:

I Front-end user interface that runs on user devices
II Logic components that run on a server
III Back-end database in which the application data is stored

Therefore, the modularization of such systems is limited by the resources they share (i.e, the database) and the components cannot be executed independently [19]. The central component and control can lead to a large code base, which makes the code difficult to understand and modify as well as less expandable [20]. As a result, the development slows down. Another effect

is that continuous deployment can become difficult [21], since a small change to a part of the application requires the entire monolith to be rebuilt and deployed [21]. Big Data technologies are often very specific, and so is the connection between them. Special adapters, interfaces or services must be provided so that they are compatible. A ”loosening” of the system and flexible exchange is only possible to a very limited extent, if at all. Accordingly, such a system can appear monolithic to the user, whereas the conceptual design and implementation, on closer inspection, appear rather isolated and less flexible.

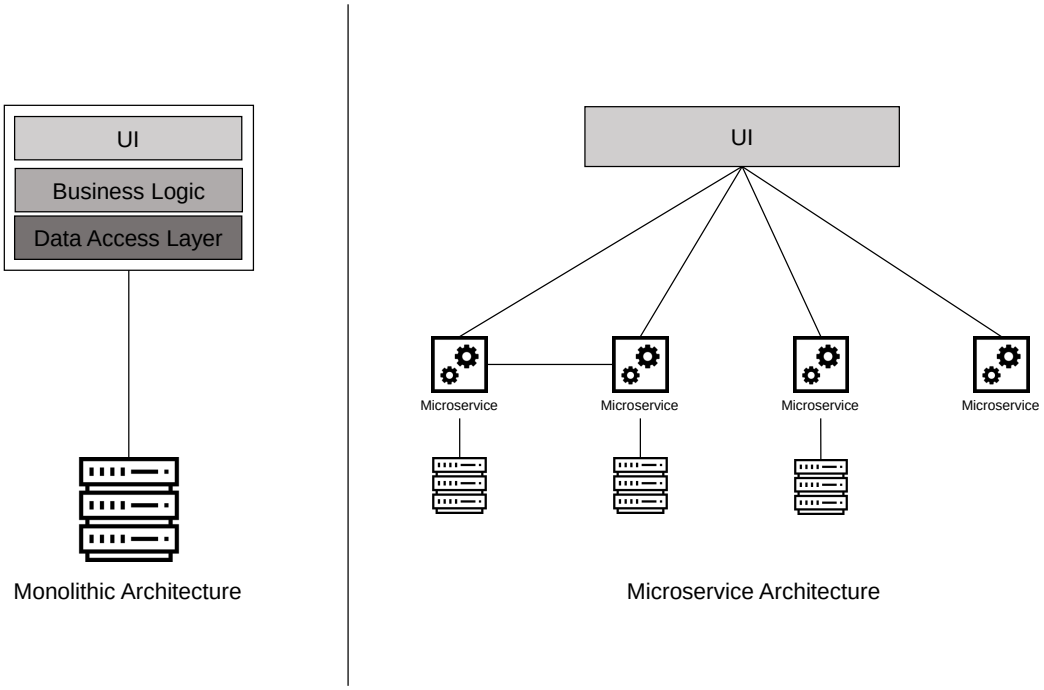


Figure 1. On the left side there is a monolithic architecture which contains all features and services versus on the right side the microservices architecture with decomposed services that work together to get the same functionality.

2.3 Microservices

Microservices are an opposite approach to monolithic architectures [17]. While there is no precise definition for the term microservice [22], they can be described as an approach to develop a single application by decomposing it into a suite of small services [23]. Each of them runs in their own process and communicates with lightweight mechanisms. These services, because being so independent, only need a minimum of central management [24]. They are based on business functions and can be deployed separately by continuous deployment tools. Due to their independence, they can be written in different programming languages and can be based on different technologies, for example for the data storage. Common characteristics for microservices are described in the following. The first one is *”Organized around Business Capabilities”*. The microservice approach leads to structuring teams around business capabilities instead of traditionally building teams based on the technology layer. Consequently, the teams are cross-functional and encompass the full range of skills required for development and prevent the “logic everywhere” siloed architectures [22]. This comes with the constraint that the team realizing this concept can’t be based on strict hierarchical communication [25]. The second one is *”Componentization, or Modularity”*. Componentization is known to be a generally good practice in software engineering. Achieving a high degree of modularity is often considered as difficult [26]. Because systems are broken down into services that are independently deployable, componentization is achieved with the microservice architecture by design. For small internal changes, only the affected service has to be redeployed. The third one is *”Decentralized Data Management”*. Every microservice has its own storage and its own technology

to manage its store. This leads to a decentralized data storage which is isolated from other services. Different services can therefore have different conceptual models, e.g. they operate on different attributes of the same data entities. The fourth one is *"Evolutionary Design"*. The evolutionary design is a typical feature of a microservice architecture, where services decomposition is used as a driving force to enable frequent and controlled changes in the system [27]. The microservices are specialized, so minor changes on the feature request can lead to implementing new services which can be easily added to the existing application. However, in case of only small changes, those can usually be based on their predecessor. The fifth one is *"Smart endpoints and dumb pipes"* [22]. The services run isolated from each other. They are decoupled, cohesive and have their own domain logic. They use lightweight protocols for communication, which are often used in a REST-ish manner and as basic as possible. The sixth one is *"Products not Projects"*. The aim of most application developments is to deliver a piece of software which is considered to be complete. While developing a service, the team preferably develops its service as a standalone product. The seventh one is *"Decentralized Governance"*. Because of the decentralized architecture, which relies on independently deployable components, the centralized governance can be relaxed [28]. Every service in the system can be build based on its own technology that is most suitable for the task. This leads to high flexibility in the choice of tools and implementation technology. Additionally it allows us to adopt to new technologies on a smaller scale first [23]. The properties are summarized in Table 1.

Table 1. An overview of the typical properties of microservices.

Property	**Description**
Organized around Business Capabilities	Developer teams are structured around Business Capabilities instead of the technology layer.
Componentization or Modularity	The system is broken down into services that are independently deployable.
Decentralized Data Management	Every service has its own storage and its own technology to manage its store.
Evolutionary Design	For each service the best fitting technology (e.g. programming language, framework) can be used.
Smart Endpoints and Dumb Pipes	Components need to be able to access required data and other components through pipelines [22].
Products not Projects	The microservice is supposed to be a finished standalone product.
Decentralized Governance	Every service in the system can be build based on its own technology that is most suitable for the task.

3 Microservices in Big Data

There are several contributions for example [27]–[29], which explore to what extent Big Data can be used in microservice architectures. This is generally considered in the context of IoT. Furthermore, some approaches also use microservices to build a Big Data architecture. Zhelev and Rozeva showed that microservices can be used to build an event-driven Big Data architecture that relies heavily on the implementation of microservices [30]. Also Miao et al. implemented a microservice based Big Data Analysis Platform for Online Educational Applications [31]. Both describe how they build their architecture and they also described that the main challenge is to keep the data integrity. Freymann et al. described how modular services such as microservices can be used to tackle the six fundamental challenges: *Volume*, *Velocity*, *Variety*, *Complexity*, *Veracity* and *Value* [6]. They found that the modular architecture should have the properties described in Table 2. Like Zhelev and Miao, they outlined that data integrity is a challenge but really important for a successful Big Data architecture. Furthermore, the use of microservices

for the application of test-driven development to the Big Data domain is proposed in the work of Staegemann et al. [32].

Table 2. In the table the properties of the modular architecture are described, which was used to build a Big Data architecture [6].

Property	Description
Modularity	The architecture divides and structures a system into software and hardware modules realized by microservices.
Adaptability	This is the ability of the architecture to modify and extend a system [33].
Scalability	Scalability supports the expansion of a solution horizontally and vertically by its hardware and software components [6].
Data Handling	A proper deal handling is important to organize large amount of data.
Distributed System	Distributed Systems ”enable load balancing, distribution of computational power, data storage and efficient parallel” [6].
Infrastructure Management	An overall management of the system to get transparency [34].

4 Success Factors of Microservices in Big Data

Based on the findings from Section 3 and the definitions and characteristics of Big Data and microservices, certain factors can be identified that facilitate a successful implementation of microservice favored Big Data architectures. The projects described in Section 3 have taken data integrity as a challenge in implementing their architecture. In many Big Data architectures, a structured and segregated system is created, to prevent unplanned modifications during the data processing, transformation or persistence. With the modular structure, there is no such instance. Every process can hold its own data and change it. In turn, this change can be relevant for other processes. If these changes are not communicated further, this can lead to serious errors and failure. Therefore, the first success factor for microservice favored Big Data architectures is maintaining data integrity. Many projects using microservices for Big Data systems have tried to apply best practices. But since their systems often require a high degree of flexibility and specialization, the implementation was massively facilitated by loosening them a bit. This effect has also been described by Krylovskiy et al. for building large-scale Smart City IoT platforms [27]. A microservice architecture with Big Data is implemented successfully by deviating from best practices for the architectures. This shows that it isn't mandatory to always follow strict rules and that it is also possible to deviate from them if it makes the implementation easier. This is the second success factor. Based on the findings described in Section 3 and the characteristics of microservices described in Section 2.3, further important factors can be extracted. One of them is that microservices benefit greatly from their extensibility. This is also highlighted in the work of Freymann et al. [6]. New components can be added without further ado. This makes it possible to swap out components with relative ease. Furthermore, it is easy to scale the system horizontally and vertically [28]. Another success factor it the technological independence of the individual components. This also allows to use the most suitable technology for each component, instead of being forced to adhere to predefined standards. For example, [35] describes a fraud detection system that uses different databases - one for the user's past activity, another for a blacklist, a third one for a white list of activities etc. Each of these services can follow a different architecture and a different internal mechanism. Lastly, the

individual independent components must communicate in a way among themselves that fits the architecture they are following. When dealing with microservices, this is done via pipelines, which follow the ”smart endpoints and dumb pipes” approach. That means, the components should not aim at forming fine-grained and complex communication structures with other components. Instead, they should try to be ”as decoupled and as cohesive as possible” [22]. Those factors are also summarized in Table 3.

Table 3. The success factors for building a microservice favored Big Data architecture are described in this table.

Success Factor	Description
Data Integrity	The data has to be consistent through every process and every change.
Don't be a Slave of Rules	Best practices can be relaxed if it is deemed beneficial to build a successful architecture.
Extensibility	To refine or update an existing architecture, new components are to be added easily [22].
Technological Independence	For each service the best fitting technology (e.g. programming language, framework) can be used.
Pipelines	Components need to be able to access required data and other components through pipelines [22].

5 Example Transformation

In the previous section, the success factors were explained. In the following, a Big Data architecture is transformed into a microservice favored architecture. There are many different Big Data architectures, generic ones such as the Lambda architecture [36], the Kappa architecture [37] as well as domain-specific ones as in the example of the IoT Big Data Architecture proposed in [38]. In the following, a widely used initial architecture is used for the transformation [38]. This can be used as an example to show how one should also proceed with other systems. Due to its already modular structure, there are few hurdles to overcome. As will be shown in the following, this architecture benefits greatly from the transformation. The architecture is presented first, followed by an explanation of the transformation. The advantages and challenges are then discussed and the success factors presented in Section 4 are examined.

5.1 Exemplary Big Data Architecture

The fundamental building blocks of Big Data architectures are *data sources*, *batch processing*, *real time message ingestion*, *analytical data storage*, *analysis and reporting* and *orchestration*, as they are visualized in Figure 2.

To start with Big Data applications, one or multiple sources of data are required. The specific kind of sources may vary and may, inter alia, involve application data from databases, files produced by the application itself (e.g. log files) or real-time data from sensors or other IoT devices. Before an analysis can be performed, the data has to be processed first. The initial filtering, aggregation and further preparation for analysis is usually conducted via *batch processing* [38]. This process of reading and writing source files to a new destination after filtering them takes a long time, depending on the quantity of data. If the data set consists of real-time messages, the application has to provide a way to handle them. Such real-time messages are received in the form of a stream, from which the data can for example be saved by simply dropping the messages into a folder or analyzed in real-time, with only the results being stored. To support scale-out processing, reliable delivery and other queueing semantics, an ingestion store as a buffer may be needed. Once preparations are finished, Big Data applications usually store the data in a structured format, which in turn can be efficiently used by analysis tools. To

further the aspect of efficiency, the components mentioned above have to be connected and applied repeatedly on the data. Technologies that aid the automation of these workflows are called *orchestration technologies* [38].

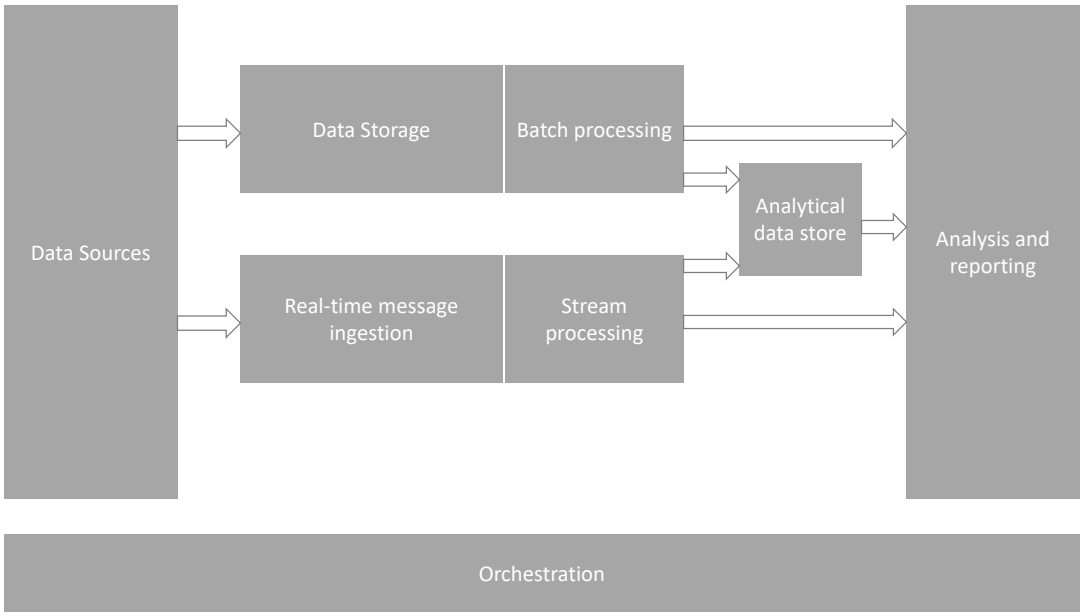


Figure 2. The structure of a Big Data architecture. It contains the components described above and also the dependencies between them. (As shown in [38].)

5.2 Transformation

The central control of the routines in the orchestration is closely interwoven in the presented architecture. Therefore, slightly simplified, it can be abstracted as one element. In the first step, this module is removed. For the repeating routines, an extra module can be written in each case, which accesses the required components through interfaces. Thus the functionality is maintained, but at the same time one moves away from closely intertwined constructs. Since the described architecture already possesses a modular structure, the individual components can be transferred into their own services. The following must be considered with the transfer:

- Dependencies: Required services are addressed via pipelines.
- Interfaces: Other services can use the results, which must be provided by interfaces.

If each service ensures these points, a coherent transformation can take place. Corresponding services can address other services via pipelines and interfaces. This means, that the dependencies can be transferred from the old model without further ado. So the previously thought-out processing steps are retained, but at the same time the rigid structure is broken. By its very nature, a Big Data architecture is usually dealing with very large amounts of data. The *data source* is a separate service. However, it must be ensured in any case that, if a service wants to make constant changes to a certain part of the dataset, this is also communicated to the *data service*. This is necessary to ensure the consistency of the data, one of the success factors mentioned in Section 4. Likewise, from a performance point of view, it makes a lot of sense for certain services to keep a part of the data with them as well. At this point, the best practice of microservices, which states that each service owns its own data, is relaxed. This is in accordance with the second success factor (”Don’t be a slave of rules”). However, this is not possible with the Big Data architecture, since the data cannot always be clearly assigned to each service. It therefore makes sense for a service to cache certain data, but for the entire dataset to be managed by an extra service. In further research it should be investigated in

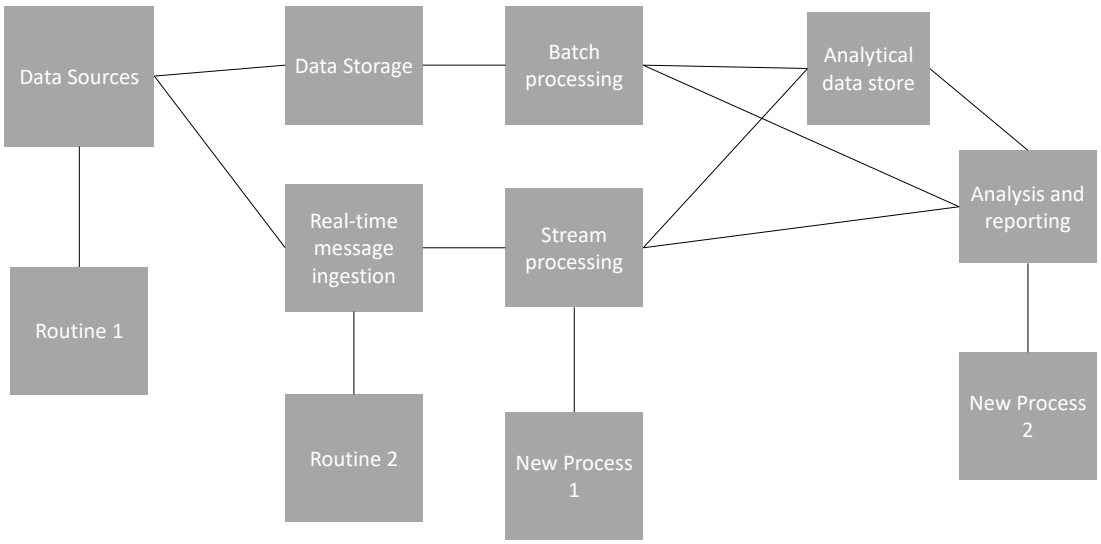


Figure 3. The result of the transformation. The whole structure is significantly more modular, but still contains all components except for the orchestration. In addition, two services have been added as examples, each of which shows a routine that would otherwise find a place in the orchestration.

which dimension the overhead and the performance change it. Up to this point no clear statement can be made. Thanks to the use of interfaces, we can now extend our structure easily by new components as are exemplary shown in Figure 3. By this, our architecture fulfills another of the success factors from Section 4. Figure 3 also shows the other components the now microservice-based architecture resulting from the transformation that has been undertaken.

5.3 Advantages and Challenges

The architecture shown combines a lot of positive aspects from both worlds. That means, the advantages of the individual mother architectures can also be transferred here. By decomposing the individual components into services, the ideal technology (e.g., specific algorithms or programming languages) can be used for each service. This architecture is also extensible. New services can be added without further ado. Via pipelines and appropriate interfaces they can request the processing necessary data of the respective services. With extensibility comes scalability. Big Data architectures support horizontal scaling [39]. Since the functionality was split into microservices, with the dependencies and components being adopted, not only the architecture is horizontally and vertically scalable, but also distinct parts of it can be adjusted to the respective circumstances and needs [40]. This makes the microservice favored architecture just as powerful as the reference architecture but at the same time brings the advantages of microservice architectures: Technological Independence, Evolutionary Design and Extensibility [41]. Big Data architectures are already complex from scratch. Although the transfer to a microservice architecture removes some of the complexity, new complexity is added from another side. This is because organizing the interfaces is complex during the initial transition. Another problem is the overhead of maintaining data consistency.

5.4 Evaluation of the Transformation in View of the Success Factors

In the following, we will compare the success factors presented in Section 4 with the architecture presented. During the transformation, effort is taken to set pipelines in such a way that, if a service wants to make consistent changes to the data, it must communicate this to the service

managing the data, data sources, and this assumes the change. Likewise, a service compares its data with the data of the data service. This fulfills the point of data consistency and thus the first success factor. The communication between the individual services is very much based on pipelines, which are only available for data transfer. The data is in turn made available by well thought-out interfaces. This principle of smart endpoints and dumb pipes is fulfilled and thus also the second success factor, *Pipelines*. As already shown, new services can be added without further ado, so the third success factor - *Extensibility* - is also fulfilled. By decoupling them into individual services, they can be implemented in technologies that are best suited to them. That in turn is the next success factor, *Technology Independence*. During the transformation it was shown that the best practices were relaxed to ensure the most efficient and functional transfer possible. Therefore, the last success factor *Don't be a slave of rules* has also been fulfilled.

6 Conclusion

This paper showed which factors should be fulfilled in order to successfully build a microservice favored Big Data achitecture. Furthermore, an exemplary Big Data architecture was transformed step by step into a microservice based architecture on the conceptual level and examined with regard to the success factors. Thus, it was shown that this approach offers great potential in theory. In future research, the practical implementation should be examined in more detail. From a theoretical point of view, questions of data management and synchronization are still interesting. This would also be a question to be investigated in future research. In general, microservices offer a great perspective in Big Data and should be focused on more in the future.

References

[1] M. Volk, D. Staegemann, and K. Turowski, "Big Data," in *Handbuch Digitale Wirtschaft*, ser. Springer Reference Wirtschaft, T. Kollmann, Ed., Wiesbaden: Springer Fachmedien, 2020, pp. 1–18, ISBN: 978-3-658-17345-6. DOI: 10.1007/978-3-658-17345-6_71-1.

[2] F. X. Diebold, "On the Origin(s) and Development of the Term 'Big Data'," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2152421, Sep. 2012. DOI: 10.2139/ssrn.2152421.

[3] O. Müller, M. Fay, and J. vom Brocke, "The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics," *Journal of Management Information Systems*, vol. 35, no. 2, pp. 488–509, Apr. 2018, ISSN: 0742-1222, 1557-928X. DOI: 10.1080/07421222.2018.1451955.

[4] M. Volk., D. Staegemann., S. Bosse., R. Häusler., and K. Turowski., "Approaching the (big) data science engineering process," in *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security - Volume 1: IoTBDS,*, INSTICC, SciTePress, 2020, pp. 428–435. DOI: 10.5220/0009569804280435.

[5] D. Staegemann, M. Volk, C. Daase, and K. Turowski, "Discussing relations between dynamic business environments and big data analytics," *Complex Syst. Informatics Model. Q.*, vol. 23, pp. 58–82, 2020. DOI: 10.7250/csimq.2020-23.05. [Online]. Available: https://doi.org/10.7250/csimq.2020-23.05.

[6] A. Freymann, F. Maier, K. Schaefer, and T. Böhnel, "Tackling the Six Fundamental Challenges of Big Data in Research Projects by Utilizing a Scalable and Modular Architecture," in *5th International Conference on Internet of Things, Big Data and Security, IoTBDS 2020. Proceedings*, 2020, pp. 249–256, ISBN: 978-989-758-426-8.

[7] L. Hung and F. Jackie, *Hype Cycle for Emerging Technologies, 2012*. [Online]. Available: https://www.gartner.com/en/documents/2100915/hype-cycle-for-emerging-technologies-2012 (visited on 02/14/2021).

[8] A. Parlina, K. Ramli, and H. Murfi, "Theme Mapping and Bibliometrics Analysis of One Decade of Big Data Research in the Scopus Database," *Information*, vol. 11, no. 2, p. 69, Feb. 2020, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. DOI: 10.3390/info11020069.

[9] W. L. Chang and N. Grady, "NIST Big Data Interoperability Framework: Volume 1, Definitions," Oct. 2019, Last Modified: 2020-01-07. [Online]. Available: https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-definitions.

[10] T. Breur, "Statistical Power Analysis and the contemporary "crisis" in social sciences," *Journal of Marketing Analytics*, vol. 4, no. 2, pp. 61–65, Jul. 2016, ISSN: 2050-3326. DOI: 10.1057/s41270-016-0001-3.

[11] N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. A. Abbasi, and S. Salehian, "The 10 Vs, Issues and Challenges of Big Data," in *Proceedings of the 2018 International Conference on Big Data and Education*, ser. ICBDE '18, New York, NY, USA: Association for Computing Machinery, Mar. 2018, pp. 52–56, ISBN: 978-1-4503-6358-7. DOI: 10.1145/3206157.3206166.

[12] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, Apr. 2015, ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2014.10.007.

[13] A. Gardiner, C. Aasheim, P. Rutner, and S. Williams, "Skill Requirements in Big Data: A Content Analysis of Job Advertisements," *Journal of Computer Information Systems*, vol. 58, no. 4, pp. 374–384, Oct. 2018, ISSN: 0887-4417. DOI: 10.1080/08874417.2017.1289354.

[14] C. Avci, B. Tekinerdogan, and I. Athanasiadis, "Software architectures for big data: A systematic literature review," *Big Data Analytics*, vol. 5, Aug. 2020. DOI: 10.1186/s41044-020-00045-1.

[15] A. M. Fernández, D. Gutiérrez-Avilés, A. Troncoso, and F. Martínez–Álvarez, "Automated Deployment of a Spark Cluster with Machine Learning Algorithm Integration," *Big Data Research*, vol. 19-20, p. 100 135, Mar. 2020, ISSN: 2214-5796. DOI: 10.1016/j.bdr.2020.100135.

[16] A. Tiwana, "Chapter 5 - Platform Architecture," in *Platform Ecosystems*, A. Tiwana, Ed., Boston: Morgan Kaufmann, Jan. 2014, pp. 73–116. DOI: 10.1016/B978-0-12-408066-9.00005-9.

[17] D. Namiot and M. Sneps-Sneppe, "On micro-services architecture," *International Journal of Open Information Technologies*, vol. 2, p. 4, 2014.

[18] C. Fan and S. Ma, "Migrating monolithic mobile application to microservice architecture: An experiment report," in *2017 IEEE International Conference on AI Mobile Services (AIMS)*, Jun. 2017, pp. 109–112. DOI: 10.1109/AIMS.2017.23.

[19] A. Bucchiarone, N. Dragoni, S. Dustdar, S. Larsen, and M. Mazzara, "From monolithic to microservices: An experience report from the banking domain," *IEEE Software*, vol. 35, pp. 50–55, May 2018. DOI: 10.1109/MS.2018.2141026.

[20] P. Karwatka, *Monolithic architecture vs microservices*, Jan. 2020. [Online]. Available: https://divante.com/blog/monolithic-architecture-vs-microservices/ (visited on 02/14/2021).

[21] F. Ponce Mella, G. Márquez, and H. Astudillo, "Migrating from monolithic architecture to microservices: A rapid review," in *Proceedings of the 38th International Conference of the Chilean Computer Science Society*, Sep. 2019.

[22] F. Martin and L. James, *Microservices - a definition of this new architectural term*, Mar. 2014. [Online]. Available: https://martinfowler.com/articles/microservices.html (visited on 02/14/2021).

[23] M. Amundsen and M. Mclarty, *Microservice Architecture: Aligning Principles, Practices, and Culture*. Sebastopol, CA: O'Reilly Media, Inc, USA, Aug. 2016, ISBN: 978-1-4919-5625-0.

[24] P. Drews, I. Schirmer, B. Horlach, and C. Tekaat, "Bimodal Enterprise Architecture Management - The Emergence of a New EAM Function for a BizDevOps-based fast IT," Oct. 2017. DOI: 10.1109/EDOCW.2017.18.

[25] M. E. Conway, "How do committees invent," *design organization criteria*, 1968.

[26] D. Faitelson, R. Heinrich, and S. Tyszberowicz, "Functional Decomposition for Software Architecture Evolution," in, Jul. 2018, pp. 377–400. DOI: 10.1007/978-3-319-94764-8_16.

[27] A. Krylovskiy, M. Jahn, and E. Patti, "Designing a Smart City Internet of Things Platform with Microservice Architecture," in *2015 3rd International Conference on Future Internet of Things and Cloud*, Aug. 2015, pp. 25–30. DOI: 10.1109/FiCloud.2015.55.

[28] L. Sun, Y. Li, and R. A. Memon, "An open IoT framework based on microservices architecture," *China Communications*, vol. 14, no. 2, pp. 154–162, Feb. 2017, Conference Name: China Communications, ISSN: 1673-5447. DOI: 10.1109/CC.2017.7868163.

[29] R. K. Naha, S. Garg, D. Georgakopoulos, P. P. Jayaraman, L. Gao, Y. Xiang, and R. Ranjan, "Fog computing: Survey of trends, architectures, requirements, and research directions," *IEEE Access*, vol. 6, pp. 47 980–48 009, 2018. DOI: 10.1109/ACCESS.2018.2866491.

[30] S. Zhelev and A. Rozeva, "Using microservices and event driven architecture for big data stream processing," in *AIP Conference Proceedings*, vol. 2172, Nov. 2019, p. 090 010. DOI: 10.1063/1.5133587.

[31] K. Miao, J. Li, W. Hong, and M. Chen, "A Microservice-Based Big Data Analysis Platform for Online Educational Applications," *Scientific Programming*, Jun. 2020, ISSN: 1058-9244 Pages: e6929750 Publisher: Hindawi Volume: 2020. DOI: 10.1155/2020/6929750.

[32] D. Staegemann, M. Volk, N. Jamous, and K. Turoski, "Exploring the applicability of test driven development in the big data domain," in *Proceedings of the 2020 ACIS*, Dec. 2020.

[33] K. Lehmann and A. Freymann, "Demo Abstract: Smart Urban Services Platform a Flexible Solution for Smart Cities," in *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, Apr. 2018, pp. 306–307. DOI: 10.1109/IoTDI.2018.00052.

[34] R. Peinl, F. Holzschuher, and F. Pfitzer, "Docker Cluster Management for the Cloud - Survey Results and Own Solution," *Journal of Grid Computing*, vol. 14, no. 2, pp. 265–282, Jun. 2016, ISSN: 1572-9184. DOI: 10.1007/s10723-016-9366-y.

[35] J. Scott, *Using microservices to evolve beyond the data lake*. [Online]. Available: https://www.oreilly.com/content/using-microservices-to-evolve-beyond-the-data-lake (visited on 02/21/2021).

[36] M. Kiran, P. Murphy, I. Monga, J. Dugan, and S. S. Baveja, "Lambda architecture for cost-effective batch and speed big data processing," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct. 2015, pp. 2785–2792. DOI: 10.1109/BigData.2015.7364082.

[37] T. Zschörnig, R. Wehlitz, and B. Franczyk, "A Personal Analytics Platform for the Internet of Things - Implementing Kappa Architecture with Microservice-based Stream Processing," in *Proceedings of the 19th International Conference on Enterprise Information Systems*, Jan. 2017, pp. 733–738. DOI: 10.5220/0006355407330738.

[38] Z. Tejada, *Big data architectures - Azure Architecture Center*. [Online]. Available: https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data (visited on 02/14/2021).

[39] A. Ali and M. Abdullah, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics," *International Journal of Integrated Engineering*, vol. 11, Sep. 2019. DOI: 10.30880/ijie.2019.11.06.015.

[40] S. J. Fowler, *Production-Ready Microservices: Building Standardized Systems Across an Engineering Organization*, 1st edition. Sebastopol, CA: O'Reilly Media, Dec. 2016, 4. Scalability and Performance - Production-Ready Microservices, ISBN: 9781491965979.

[41] D. Taibi, V. Lenarduzzi, and C. Pahl, "Processes, Motivations, and Issues for Migrating to Microservices Architectures: An Empirical Investigation," *IEEE Cloud Computing*, vol. 4, no. 5, pp. 22–32, Sep. 2017, Conference Name: IEEE Cloud Computing, ISSN: 2325-6095. DOI: 10.1109/MCC.2017.4250931.

24th International Conference on Business Information Systems (BIS 2021)
Smart Infrastructures
https://doi.org/10.52825/bis.v1i.51

Published: 02 July 2021

Execution of Multi-Perspective Declarative Process Models using Complex Event Processing

Niklas Ruhkamp[1], Stefan Schönig[1][https://orcid.org/0000-0002-7666-4482]

[1] Institute for Management Information Systems, University Regensburg, Germany

Abstract. The Internet of Things (IoT) enables continuous monitoring of phenomena based on sensing devices as well as analytics opportunities in smart environments. Complex Event Processing (CEP) comprises a set of techniques for making sense of the behavior of a monitored system by deriving higher level knowledge from lower level system events. Business Process Management (BPM) attempts to model processes and ensures that executed processes conform with a predefined sequence. In IoT scenarios frequently a large number of events has to be analyzed in real-time to allow an instant response. While BPM reaches its limits in such situations, CEP is able to analyze and process high volume streams of data in real-time. The evaluation and execution of rules and models of both paradigms are currently based on separate formalisms and are frequently implemented in heterogeneous systems. The presented paper integrates both domains by proposing an execution approach for multi-perspective declarative process process models completely based on CEP. The efficiency of the combined paradigms is validated in an implemented demonstration with simulated and real-life sensor data.

Keywords: Process Execution, Complex Event Processing, MP-Declare, Event-driven systems

The world is increasingly linked through a large number of connected devices, typically embedded in electrical/electronical components and equipped with sensors and actuators, that enable sensing, (re-)acting, collecting and exchanging data via various communication networks including the Internet of Things (IoT). As such, it enables continuous monitoring of phenomena based on sensing devices (wearable devices, beacons, smartphones, machine sensors, etc.) as well as analytics opportunities in smart environments (smart homes, connected cars, smart logistics, Industry 4.0, etc.) [1]. Event processing focuses on capturing and processing events in real-time, for detecting changes or trends indicating opportunities or problems. *Complex Event Processing* (CEP) comprises a set of techniques for making sense of the behavior of a monitored system by deriving higher level knowledge from lower level system events. CEP and *Business Process Management* (BPM) have traditionally been considered very separate from each other [2], [3]. BPM attempts to model processes and ensures that executed processes conform with a predefined sequence. In addition, BPM offers methods to analyze business processes and to search for potential improvements [4]. In complex situations a large number of events has to be analyzed in real-time to allow an instant response. While BPM reaches its limits in such situations, CEP is able to analyze and process high volume streams of data in real-time. To implement novel scenarios in the area of the IoT, e.g., Industry 4.0 and Smart Home scenarios, a combination of these two domains is becoming an active field of research under the term event-driven BPM [5]. Both domains provide their own advantages, which can complement each other [2]. Consider, e.g. an Industry 4.0 environment that produces a large amount

of raw data. Using CEP, this data can be processed efficiently. In the context of predictive maintenance the failure of a machine can be predicted long time in advance. The combination of CEP and BPM would not only allow to detect such a pattern using CEP but also to define how human operators have to react to such a failure prediction based on activities defined in the underlying process model.

The evaluation and execution of rules and models of both paradigms, however, are currently based on separate formalisms and are frequently implemented in heterogeneous systems. First attempts to combine both paradigms have already been done and are mentioned in related work. An integrated execution approach combining both worlds in one engine is still missing. We fill this research gap and integrate both domains by proposing an execution approach for business process models completely based on CEP. For this purpose, the multi-perspective extension of the declarative process modeling language Declare [6], MP-Declare [7] is used for mapping predefined constraints to CEP-queries, which are then executed by a CEP-engine. We implemented[1] our approach based on the *Esper*[2] CEP-engine and the corresponding Esper Query Language (EQL). Additionally, screencasts and videos demonstrate the real-life application of the approach using sensor data provided via MQTT[3].

The remainder of this paper is structured as follows: Section 1 describes fundamentals and related work. In Section 2.1 we introduce our approach to execute MP-Declare constraints using CEP queries. Section 2.2 describes the implementation of the integrated exeuction engine. The approach is validated with simulated and real data in Section 3 and Section 4 concludes the paper.

1 Background and Related Work

Next, we describe event-driven systems, basics of multi-perspective declarative process modelling and give an overview of related approaches.

1.1 Event-Driven Systems

Processes in our everyday life and business related procedures are influenced and triggered by various events. The processing, interpretation and reaction to such events is therefore an important part of how companies work. The three basic steps of event-driven systems are: *(i) Sense:* The starting point is the recognition of relevant information or facts by sources like sensors. This information is interpreted as events and reflects a relevant part of the state of reality. For event-driven systems the events must be recognized immediately at the time of their occurrence to guarantee real-time processing. Otherwise, the value of the information decreases. *(ii) Process:* In this step, the analysis of the detected events is performed. Events are aggregated, correlated, abstracted, classified, or if necessary discarded. Here, we seek for patterns in and between the event streams, which express certain relationships and dependencies between the events. *(iii) Respond:* If a pattern is recognized, the system can react with a corresponding action, e.g., warnings or the triggering of a business activity. The generation of new events is also a possible reaction, as for instance the generation of complex events on a higher level of abstraction.

As soon as data from event sources like sensors, network data, or news tickers arrive, they are processed by a CEP engine using predefined rules to detect patterns and derive complex events. This process can be repeated on several levels of abstraction. Subsequently, predefined actions are triggered or the obtained information is transferred to other systems, e.g., databases, message channels, or information systems. This way, processes are not only monitored but additionally automatic actions can be triggered [5].

[1] Available at `https://github.com/NiklasRuhkamp/MP-Declare-To-CEP`

[2] EsperTech - Esper Documentation: `https://www.espertech.com/esper/&sc=SUR`

[3] `https://vimeo.com/512049878`

1.2 Multi-Perspective Declarative Process Modelling

Declarative process modelling approaches like Declare [6] have proven to be suitable means to capture activities and agent interactions within flexible environments additionally involving real world objects [1]. A central shortcoming of the process modeling language Declare is the fact that constraints only apply to activities while other perspectives such as time and data perspectives are ignored. In real world scenarios like IoT applications these are important factors that must be considered to model realistic processes. To include these perspectives, Declare has been extended by a multi-perspective version called *MP-Declare* [7].

Semantics of Declare are extended by further conditions, which refer to the payload of events and must be *fulfilled* to satisfy a constraint namely the *activation condition*, the *correlation condition* and the *target condition*. These additional conditions will be demonstrated using the example of template *Response (A, B)* in the context of the IoT. With standard Declare, the constraint would be formulated as *Response (machine failure, order maintenance)*. Machine failure is the *activation*, while order maintenance is the *target*. This constraint defines that if a machine is down, the maintenance department must be informed at some point in the future to initiate the maintenance. With MP-Declare semantics additional conditions can now be added. The *activation condition* φ_a, in this case, is the fact that the machine is a production-critical one, whose failure would cause a standstill of production within minutes. If this condition does not occur while machine failure does, the constraint is not activated. This is formally written as $A \wedge \varphi_a(x)$ which means that when action *A* occurs, the condition φ_a must be true for *x*. The *correlation condition* φ_c is already part of the *target* and addresses the payload of both, event *A* and event *B*. Therefore, φ_c must be valid when the *target* arrives. It is formulated as $B \wedge \varphi_a(x,y)$. In this context, an example would be that if a machine from certain vendor is down, the maintenance department of the same vendor must be informed and not the one of another vendor. Additionally, there is the *target condition* φ_t which only refers to the payload of event *B* and is written as $\varphi_t(y)$. Depending on the use case a time period can also be defined in which the rule must be fulfilled [8]. The complete constraint would therefore be described as: if a machine essential for ongoing production is down, maintenance has to be initiated by the same vendor, which also has to be available for maintenance within the defined time frame. This process would ensure that the machine is repaired by the corresponding maintenance right after a problem is detected with minimal consequences for the production process.

1.3 Related Work

The use of a CEP engine to execute multi-perspective declarative process models is yet not well studied. The paper by Soffer et al. [2] is one of the most important sources for the combination of CEP and BPM in general, and specifically for concepts of integrating BPM and CEP. The paper provides a state-of-the-art review of current research of the symbiosis of CEP and BPM and describes challenges and opportunities. Janiesch et al. [9] are dealing with the combination of IoT and BPM in a research and practitioner's point of view. A general framework for event-driven BPM is presented in [10]. Li et al. [11] provide a translation of the block-structured part of Business Process Execution Language (BPEL) into events in order to be able to execute them using CEP. Although BPEL is not a modeling language, but rather an execution language, there is a mapping between BPMN and BPEL [12]. Cicekli and Cicekli [13] use an imperative process modeling language called control-flow-graph, which works with basic control flow patterns. To execute them and increase their expressiveness, they provide corresponding event rules. Another attempt to realize event-driven systems is done by RuleML [14] using rule detection to trigger processes. Hens et al. [15] use imperative languages such as BPMN and YAWL and divide them into small chunks. The start and completion of these are then processed by a CEP-engine. However, this cannot be seen as a complete execution on the CEP-engine since the individual chunks are still handled by the process engine. In the work of Daum et al. [16], they investigate the integration of BPM and CEP. However, they investigate how process models can be supported or extended through CEP, but not the execution of these models on a

CEP engine. Another approach of integrating external events into business models is done by Mandal et al. [17] combining a process engine and a event engine in a heterogeneous system. There exist also first approaches for the execution of declarative rules using CEP. Jergler et al. [18] propose a version of the Guard-Stage-Milestone model (GSM) based on CEP to specify life cycle processes of business artifacts. This model contains Event-Condition-Action rules (ECA-rules), which can be executed by a CEP engine. Soffer [19] also suggests an ECA-based execution of declarative models. Approaches to detect declarative process models from event-logs were devised, such as in [20]. According to [21], some works use process stream mining to do so. In [22], Schönig et al. introduce that SQL can be used to derive multi-perspective declarative models from logs. In fact, these approaches are using static event logs and do not process the data as a stream. Therefore, the latency between occurrence of a constraint and its detection is not acceptable for time critical cases. Another problem could be the storage requirements, if processing the data in real-time is not possible. Burattin et al. [23] propose a framework for the discovery of declarative process models. They combine algorithms for stream mining and algorithms for the online discovery of Declare models to get a real-time representation of the process. Another attempt in this direction is proposed by [24] using Hoeffding trees. A similar approach is presented in [25] by examining event streams to process models using prefix-alignments.

In the context of Big Data traditional approaches may reach their limits due to the high data volume. Therefore, a mapping to CEP, which is geared towards Big Data, could perform more efficiently. Wu et al. [26] are using SASE [27] over streams of RFID-events in order to detect matching patterns and to feed an external monitoring application. Another work dealing with monitoring in combination with CEP is introduced by [28]. In this approach CEP is used in combination with Business Activity Monitoring (BAM) to monitor cloud BPM.

None of these approaches deals with the execution of multi-perspective declarative models itself. A first attempt was made in [29]. In this work, MP-Declare constraints are transformed into the modelling language Alloy and executed afterwards. In summary, the execution of complete multi-perspective declarative process models via CEP is still unexplored. The motivation of the work at hand is to study and implement a solution which integrates MP-Declare process models entirely into CEP and thereby bridging the gap between these two paradigms.

2 Execution of MP-Declare Models using CEP

This chapter explains how multi-perspective declarative MP-Declare models can be executed on top of a CEP-engine. Therefore, we introduce the concept of how MP-Declare models can be mapped to CEP queries.

2.1 Concept

MP-Declare constraints essentially consist of two components. The *activation* of the constraint is on the left-hand-side of the constraint. As soon as it occurs, the constraint is *activated*. In addition to this, the *activation condition* must also be *fulfilled*. On the right-hand-side is the *target* including the *correlation condition* and the *target condition*, which must also occur to *fulfill* the constraint. A time period within the *target* must occur can be specified additionally. Therefore, a concept is needed for applying CEP rules to examine a stream of data for the *fulfillment* or *violation* of MP-Declare constraints and thus to execute entire MP-Declare process models by means of a CEP-engine.

CEP is able to recognize patterns in an event stream and react to them. Furthermore, incoming low-level event streams can be filtered by CEP and, if necessary, depending on the application, a new stream of events can be generated including the relevant data. This feature is now used to apply MP-Declare constraints to incoming event streams. The basic idea here is to detect the left-hand-side of the constraint - the *activation* - using CEP. As soon as event of

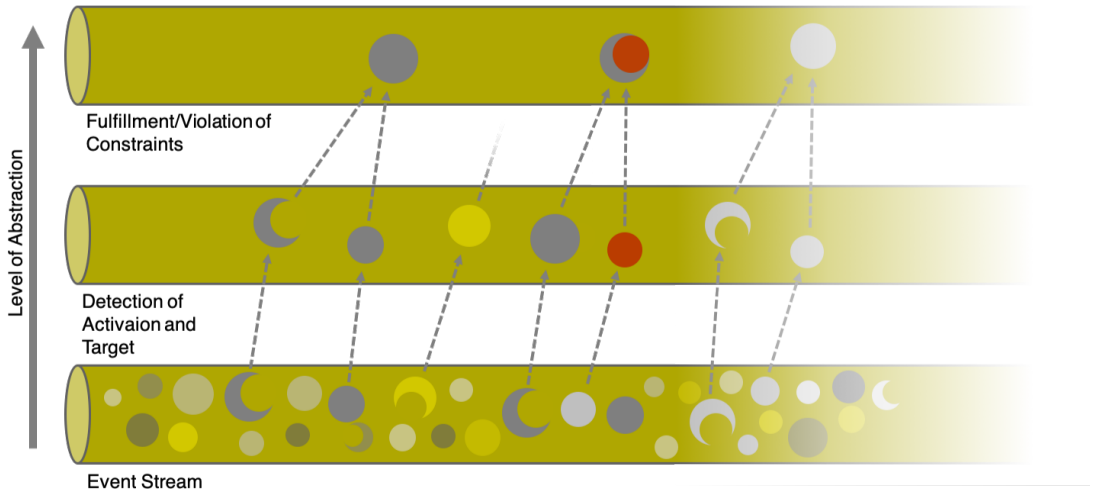


Figure 1. Recognition of Activation and Target in Event Streams of different Levels of Abstraction

a new process instance occur in the event stream and an *activation* of a constraint has been detected, the right-hand-side of the constraint - the *target* - must arrive. Thus, CEP must be able to store all events of process instances where the *activation* of a constraint occurred, and then to process the event stream to see if the *target* occurs as well. In this case, this instance *fulfills* the constraint. If the *target* is not found in the event stream, this process instance *violates* the rule or can not *fulfill* the rule so far.

This procedure is illustrated in Figure 1. On the lowest level of abstraction the incoming event streams can be seen. These are completely unordered, unfiltered and from different data sources. The CEP-engine accesses this stream and examines it for incoming *activations*. The engine has to store all *activated* constraints until they are *fulfilled* or *violated*. To implement this concept, streams of different abstraction levels are used as shown in Figure 1. *Activations* are stored on an intermediate level of abstraction. On this level also *targets* and *violations* are represented. Once an *activation* occurs, the CEP-engine searches for the corresponding patterns in the event stream to find the *target*. As soon as it arrives, the *target* is also added to the higher-level stream. The intermediate level stream represents all available *activations* and *targets*. Finally, at the highest abstraction level stream, the CEP-engine checks whether the constraints have been *fulfilled* or *violated*. For this purpose, it examines all *activations* and checks if the intermediate stream contains the matching *targets* or a *violation*. Additionally, the CEP engine examines the payload of the events to determine whether required conditions are met. As soon as a constraint can be detected as *fulfilled* or *violated*, it is added to the highest-level stream. As shown as the dark grey event in Figure 1, both the *activation* and the *target* are in the intermediate stream and hence the constraint is *fulfilled*. As shown as the green event, only the *activation* occurs without the matching *target*. The constraint is not yet *fulfilled* but might be in the future. This enables CEP to examine the event stream for CEP constraints and thus execute the entire MP-Declare models in form of a set of constraints using the CEP engine.

Besides the time perspective (whether the *target* has to occur before or after the *activation*), the constraints can be divided into three categories. The first group is constraints that can only be *fulfilled* or are not yet *fulfilled* but not directly *violated*. These consist of an *activation* and a *target*. The time period in which the *target* has to appear is potentially infinite.

The second group of constraints are those which can be *fulfilled* or can be *violated* directly. The *target B* of *alternate response*, for example, must occur without other instances of event *B* in between. Therefore, the constraint is *violated*, as soon as another *B* occurs between *A* and *B*.

The negating constraints, which are marked by the prefix *"not"* belong to the third category. These are *violated* as soon as the *activation* of a constraint is followed by the corresponding *target*, which according to the constraint must not occur. As soon as both sides of a *"not"* constraint occur, this rule is *violated*.

For constraints of the first category, however, the constraint is *fulfilled* if both sides occur and are not yet *fulfilled* when the *activation* occurs but not (yet) the *target*. In brief, the engine waits for the *fulfillment* of activated constraints. The procedure for the second category is different. The CEP engine does not have to wait to see whether the constraint is *fulfilled* at some point in the future. The engine does not only search for the *fulfillment* but also for *violations* of activated constraints by looking for occurring of the opposite of the constraints.

To highlight this behavior, Figure 1 illustrates the detection of a constraint of the second category, using the example of *Chain Response* defined as: if *A* occurs, *B* must occur next, which in turn means no other events must occur in between. In the event stream the matching *target* occurs after *activation*, but not immediately after its *activation*. The next event after the *activation* is the event marked in red. As a result the CEP-engine can mark the rule as *violated*. Using multiple streams at different levels of abstraction, MP-Declare models can entirely be mapped to CEP rules and executed using CEP-engine.

2.2 Implementation

For the execution of MP-Declare models by means of CEP, the CEP-Engine Esper is used in this work. The CEP-engine works like an inverted database. In a conventional database, the data sets are fixed, and the data is accessed dynamically using a query language. However, with a CEP-engine the query rules are known from the beginning, and the data is then loaded in the form of an event stream and checked for the rules in real time. Incoming event streams can be read via input adapters and connectors. These streams are characterized by very high volume, fast emergence of new data and occurrence in real-time. Esper also provides access to historical data in memory to provide querying possibility on historical events. The engine manages the registered statements, examines the streams for these statements and stores the results in the form of Plain Old Java Objects (POJO). These are then available for downstream systems via the output adapter.

2.2.1 Transforming MP-Declare to EQL

Esper provides its own SQL-like Event Processing Language (EPL) ccalled Event Query Language (EQL). Here, queries essentially consist of SELECT-, FROM- and WHERE-constructs. EQL-queries are used to examine the event-stream for incoming *activations*, *targets* and *violations*. The EQL-query to detect, for instance, the target of a *Response* constraint in the context of machine failures and the order of a maintenance looks as follows:

```
SELECT a.id, b.company FROM PATTERN[
every a = MachineFailureEvent(productionCritical = true)
-> b= OrderMaintenanceEvent(available = true)]
WHERE a.manufacturer = b.company
```

The *PATTERN* defines that all cases, in which a production critical machine fails, *followed by* a maintenance order for which the company must be available at that time, are to be considered. The term *"every"* defines that the query should not be made only once, but all instances which *fulfill* this pattern should be returned. The WHERE clause checks the *correlation condition*. In case of backward-looking constraints like *Precedence* the EQL-query looks similar to the query of *Response*. The difference is that the *target* is on the left-hand-side and the *activation* is on the right-hand-side. Therefore, the engine is storing all potential *targets* of a backward-looking constraint as long as the corresponding *activation* follows and the constraint can finally be detected as *fulfilled*. As depicted in Figure 1 the engine has to search for *violations* in some

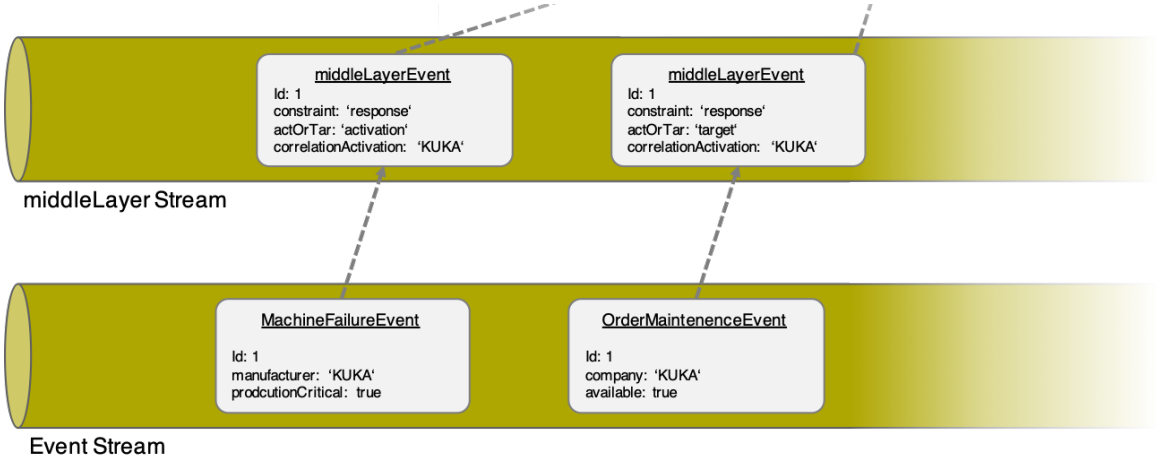


Figure 2. Example of events in the middleLayer-stream to fulfill Response

cases, too.

2.2.2 Constraint Builder

The most important component is the *ConstraintBuilder*. In order to execute MP-Declare rules using CEP, these must be translated into EQL-queries for the engine to be able to process them. The *ConstraintsBuilder* is composed of a for-loop, which iterates over all added constraints in the *constraintAndConditionList*, handed over by the *ConstraintScreen*. After all required contents are assigned, an if-else-statement is used to search for the suitable constraint type since each constraint type has different requirements.

2.2.3 middleLayer-Stream

As mentioned before (cf. Section 2.1) and shown in Figure 1 an additional stream is used, called *"middleLayer"* and is necessary to handle *activations*, *targets* and *violations* on a higher level of abstraction. This stream has four properties: The first one (integer *"id"*) is used to store the identifier of each event in the *middleLayer*-stream. This enables the engine to handle the different instances of processes. If the process does not have multiple instances, the identifier remains at its default value null. The second property (string *"constraint"*) is storing the type of the constraint. As soon as the user defines a specific name for a constraint, this name will be used instead. This allows handling several constraints of the same type. For example: if two different constraints of the type *Response* were defined, it might happen, that a *target* of the second constraint is assigned to the *activation* of the first one. In this case the first *Response* would mistakenly be *fulfilled*. To prevent this from happening, unique names need to be set. To recognize whether an event of the *middleLayer*-stream is an *activation*, a *target* or a *violation*, the third property (string *"actOrTar"*) is used. The last property (string *"correlationActivation"* solves a similar problem as the second one and is only used if the constraint includes a *correlation condition*. It ensures that, for example, the activating *MachineFailureEvent* of a "KUKA"-machine, can only be followed by an *OrderMaintenanceEvent* for which the company is also "KUKA" to solve the *Response* constraint. An illustration of the *middleLayer-Stream* is shown in Figure 2.

2.2.4 Assignment of Process Instances

It is sometimes important to process events context-dependent. For example, if a machine fails and needs to be repaired, it is important to check the event stream of the maintenance orders for an order that was placed for that particular machine and not for another one, to ensure the maintenance of the correct machine. By using the *CREATE CONTEXT* clause, Esper divides events into context partitions.

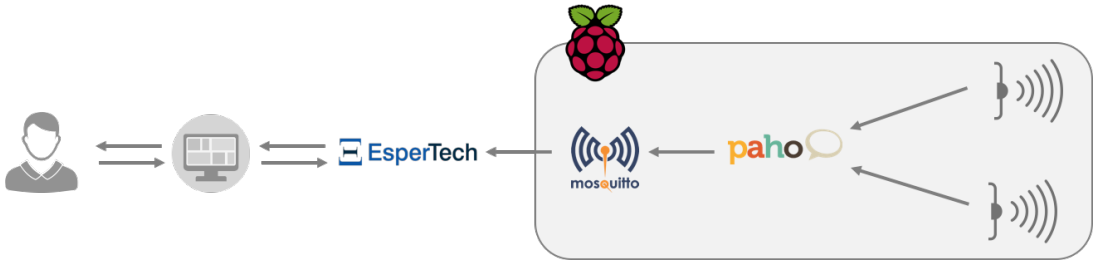


Figure 3. Environment to simulate a MQTT-Stream

3 Evaluation

We extensively evaluated our approach in terms of functionality and performance. In order to validate the presented approach events were first sent to the CEP-engine from the Java-environment in a predefined sequence. The simulation of the sample events is implemented on a new thread. During the validation, 84 different iterations were made. Each constraint was tested in seven different variations. The tool was able to detect all *activations*, *violations* and *fulfillments* of each constraint prooving the funtionality and correctness. In addition, the latency between the occurrence of an event and the assignment of a *violation* or *fulfillment* was measured, to check the performance in terms of efficiency of the tool. The gap between the occurrence of an event and the detection of the constraint, is used as a measure of latency. According to the result of this measurement, the latency was only 1 msec. If the event does not only affect one but several different *activations*, the latency increases. The maximum latency in this measurement was 4 msec. On average, the latency was 1.91 msec. A longer period of time is to be expected for constraints like *Precedence*, where the *violation* can not be detected directly, which is therefore not included in the latency validation. The engine is waiting for 1 second, if the *activation* is followed by the detection of a *target* within this time. If not, the constraint is *violated*. Therefore, the highest possible latency for detection is 1 second. The results of the validation show that our tool is efficient at the detection of MP-Declare constraints while achieving a 100 percent success rate. Additionally, we evaluated our approach with a real-life approach. Here, an IoT setup was simulated as illustrated in Figure 3. An ultrasonic distance sensor and a push-button matrix are connected to a Raspberry Pi 4B. The sensor data was sent via MQTT (Message Queuing Telemetry Transport) to the tool, running on another machine. MQTT is a lightweight, publish/subscribe messaging transport protocol, often used in the context of IoT. The sensors are connected to the GPIO-board of the Raspberry Pi. The Raspberry Pi is also used as an MQTT message broker, using Eclipse Mosquitto. The data of the sensors are read with Python and published to the MQTT broker, using Paho. A screencast and video demonstrating the application and functionality are available online.[4]. It has proven that constraints involving real-life sensor data can be sent to the engine via MQTT and are processed correctly by the engine.

4 Conclusion and Future Work

Our approach presents an efficient, scalable and reliable approach to examine a stream for MP-Declare constraints in real-time, using complex event processing. The CEP-engine Esper has been implemented and adapted such that it is able to detect *activations*, *fulfillments* and *violations*. For this purpose, besides the low-level event streams, different streams of a higher level of abstraction are used to store *activations* and to listen for following *targets* or *violations*. This way, MP-Declare constraints are executable even for unstructured high-volume streams of events. The graphical user interface offers an intuitive and easy way to formulate MP-Declare constraints. The user has full flexibility in adding new constraints. A screencast

[4] https://vimeo.com/512049878

available online[5] demonstrates the functionality of the tool. After the process has been started, the engine is giving real-time feedback to the user. This work provides an implementation of the discussed approach, to successfully and quickly execute MP-Declare process models. The tool can also easily be extended by predefined actions, which should apply as soon as a constraint is *violated*. Therefore, the system can easily be implemented to automatically detect whether predefined restrictions have been *violated* and to initiate necessary reactions. The validation has proven that this approach is highly reliable while achieving low latency. As future work, another validation in a more realistic environment is recommended. Since CEP is designed for Big Data and implemented to process huge amounts of data in real-time, it is expected that such integration is going to be successful. The integration into the context of IoT-environments like Industry 4.0 should demonstrate, that even multiple large streams can successfully be integrated and examined.

References

[1] G. Meroni, C. D. Ciccio, and J. Mendling, "An artifact-driven approach to monitor business processes through real-world objects," in *ICSOC*, vol. 10601, 2017, pp. 297–313.

[2] P. Soffer *et al.*, "From event streams to process models and back: Challenges and opportunities," Information Systems, vol. 81.2019, pp. 181–200, 2019.

[3] S. Schönig, L. Ackermann, S. Jablonski, and A. Ermer, "Iot meets BPM: a bidirectional communication architecture for iot-aware process execution," *Softw. Syst. Model.*, vol. 19, no. 6, pp. 1443–1459, 2020.

[4] D. M. Goetz, "Integration of business process management and complex event processing," 2010.

[5] R. Bruns and J. Dunkel, *Complex Event Processing*. 2015, ISBN: 978-3-658-09898-8.

[6] W. M. P. van der Aalst, M. Pesic, and H. Schonenberg, "Declarative workflows: Balancing between flexibility and support," *Comput. Sci. Res. Dev.*, vol. 23, no. 2, pp. 99–113, 2009.

[7] A. Burattin *et al.*, "Conformance checking based on multi-perspective declarative process models," *Expert Syst. Appl.*, vol. 65, pp. 194–211, 2016.

[8] H. van der Aa, K. J. Balder, F. M. Maggi, and A. Nolte, "Say it in your own words: Defining declarative process models using speech recognition," in *BPM Forum*, vol. 392, 2020, pp. 51–67.

[9] C. Janiesch *et al.*, "The internet of things meets business process management: A manifesto," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 4, pp. 34–44, 2020. DOI: 10.1109/MSMC.2020.3003135.

[10] R. von Ammon, C. Emmersberger, T. Greiner, F. Springer, and C. Wolff, "Event-driven business process management," in *Fast Abstract, Second International Conference on Distributed Event-Based Systems, DEBS 2008, Rom, Juli 2008*, 2008. [Online]. Available: https://epub.uni-regensburg.de/6829/.

[11] G. Li, V. Muthusamy, and H.-A. Jacobsen, "A distributed service-oriented architecture for business process execution," *ACM Trans. Web*, vol. 4, no. 1, 2010.

[12] M. Weidlich, G. Decker, A. Großkopf, and M. Weske, "BPEL to BPMN: the myth of a straight-forward mapping," in *OTM Confederated International Conferences*, vol. 5331, 2008, pp. 265–282.

[13] "Formalizing the specification and execution of workflows using the event calculus," *Information Sciences*, vol. 176, no. 15, pp. 2227–2267, 2006.

[5] https://vimeo.com/512048601

[14] A. Paschke, "The reaction ruleml classification of the event / action / state processing and reasoning space," *CoRR*, vol. abs/cs/0611047, 2006. arXiv: cs/0611047. [Online]. Available: http://arxiv.org/abs/cs/0611047.

[15] P. Hens, M. Snoeck, G. Poels, and M. D. Backer, "Process fragmentation, distribution and execution using an event-based interaction scheme," *J. Syst. Softw.*, vol. 89, pp. 170–192, 2014.

[16] M. Daum, M. Götz, and J. Domaschka, "Integrating cep and bpm: How cep realizes functional requirements of bpm applications (industry article)," in *International Conference on Distributed Event-Based Systems*, 2012, pp. 157–166.

[17] S. Mandal, M. Hewelt, and M. Weske, "A framework for integrating real-world events and business processes in an iot environment," in *OTM 2017 Conferences*, 2017, pp. 194–212.

[18] M. Jergler, H.-A. Jacobsen, M. Sadoghi, R. Hull, and R. Vaculın, "Safe distribution and parallel execution of data-centric workflows over the publish/subscribe abstraction," in *ICDE*, 2016, pp. 1498–1499.

[19] P. Soffer, "A state-based intention driven declarative process model," *International Journal of Information System Modeling and Design*, vol. 4, pp. 44–64, 2013.

[20] F. M. Maggi, M. Montali, and U. Bhat, "Compliance monitoring of multi-perspective declarative process models," in *EDOC*, 2019, pp. 151–160.

[21] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. 2011, ISBN: 3642193447.

[22] S. Schönig, C. D. Ciccio, F. M. Maggi, and J. Mendling, "Discovery of multi-perspective declarative process models," in *ICSOC*, vol. 9936, 2016, pp. 87–103.

[23] A. Burattin, M. Cimitile, F. M. Maggi, and A. Sperduti, "Online discovery of declarative process models from event streams," *Transactions on Services Computing*, vol. 8, no. 6, pp. 833–846, 2015.

[24] N. Navarin, M. Cambiaso, A. Burattin, F. M. Maggi, L. Oneto, and A. Sperduti, "Towards online discovery of data-aware declarative process models from event streams," in *IJCNN*, 2020, pp. 1–8.

[25] S. J. van Zelst, A. Bolt, M. Hassani, B. F. van Dongen, and W. M. P. van der Aalst, "Online conformance checking: Relating event streams to process models using prefix-alignments," *Int. J. Data Sci. Anal.*, vol. 8, no. 3, pp. 269–284, 2019.

[26] E. Wu, Y. Diao, and S. Rizvi, "High-performance complex event processing over streams," in *International Conference on Management of Data6*, 2006, pp. 407–418.

[27] University of Massachusetts, *SASE Home*, http://sase.cs.umass.edu&sc=SUR, Online; accessed 20 December 2020.

[28] J. N. Martínez Garro, P. Bazán, and F. J. Díaz, "Using bam and cep for process monitoring in cloud bpm," *Computer Science and Technology*, vol. 16, no. 01, 2016.

[29] L. Ackermann, S. Schönig, S. Petter, N. Schützenmeier, and S. Jablonski, "Execution of multi-perspective declarative process models," in *OTM Conferences*, vol. 11230, 2018, pp. 154–172.

24th International Conference on Business Information Systems (BIS 2021)
Smart Infrastructures
https://doi.org/10.52825/bis.v1i.68

Published: 02 July 2021

Exploring Potential Impacts of Self-Sovereign Identity on Smart Service Systems

An Analysis of Electric Vehicle Charging Services

Daniel Richter[1] [https://orcid.org/0000-0003-1549-5467], and Jürgen Anke[1] [https://orcid.org/0000-0002-9324-9387]

[1] Digital Service Systems Group, HTW Dresden, Germany

Abstract. Self-sovereign identity (SSI) is a new paradigm, which puts users back in control of their own digital identity. This does not only strengthen the position of the users but implies new interaction schemes that may improve interoperability and usability. Smart services systems enable the integration of resources and activities and use smart products as boundary objects. As such systems typically involve digital interactions between multiple actors, it can be assumed that utilising SSI has a positive impact on them. To investigate how these potential improvements manifest themselves, we investigate electric vehicle charging as example of a smart service system. At the core of our conceptual analysis is the service process, which we extract from a reference model. Based on a SWOT analysis, we identify areas for transformation and derive an SSI-enabled interaction model for an electric vehicle charging service. The evaluation of the new process shows that SSI can reduce complexity of integration with partners and can provide a better customer experience through simplified registration and authentication. Moreover, SSI might even lead to the disintermediation of actors in the service system. Although SSI is still emerging, our findings underline its relevance as a mechanism to establish trust in smart service systems through the seamless and standardised integration of digital identities for humans, organisations, and things.

Keywords: Self-Sovereign Identity, E-Mobility, Service Design, Smart Service Systems, Service Process

1 Introduction

1.1 Motivation

Digital services require trust between the involved parties. Therefore, participants in transactions need to prove their identity in some way. As the internet has no built-in identity protocol [1], various approaches have been developed in the past. In the isolated model, online services are at the centre of the identity ecosystem, as each one requires users to register a user account with them, which then can be used together with a password to log in [2]. This leads to many user accounts (“logins”), which are spread among various service providers. In the federated model, dedicated identity providers are utilised, where the user registers once. Afterwards, their identity can be verified at online services that support the respective identity provider [2]. Popular examples include so-called social logins, such as the ones offered by Google, Facebook, and Microsoft. Their major drawback is the dependency of both the user and the service on identity providers [3] as well as the potential tracking of the usage behaviour [4] to create detailed user profiles. These might become part of data-driven business models such as targeted advertising [5]. Additionally, they are attractive targets for identity theft [4].

Self-sovereign identity (SSI) is the most recent approach and enables users to manage their digital identities on their own [2], [6]. In addition to strengthening users, SSI could also provide benefits from an interoperability and process perspective [4]. The SSI approach has mainly been discussed within the field of security, privacy, and distributed system but barely within information systems research. At present, there are only few academic papers on the application of SSI in business scenarios. They include the application of SSI in "know-your-customer" processes in banking [7], access to public health services [8], remote management of industrial equipment [3], payback programmes in retail [9], student exchange [10], e-petitions [11], assigning medical information to persons without regular identity, e.g. to fight Covid-19 [12]. Most of these scenarios represent typical business process or digital services.

In this paper, we aim to identify potential impacts of SSI in smart service systems, which can be defined as "service systems in which smart products are boundary-objects that integrate resources and activities of the involved actors for mutual benefit" [13, p. 12]. Examples for smart services include car sharing, pay-per-use models for industrial equipment, and remote diagnostics for household appliances [13], [14]. Digital identities are relevant to smart service systems as they need mechanisms for establishing trust between the involved actors. Typically, this requires the development of proprietary apps, individual customer registration processes, and legally compliant management of personally identifiable data (PII). Moreover, these measures need to be reconciled between all actors in the service ecosystem, which causes interoperability challenges. At the same time, users are confronted with complicated tasks like registration, managing passwords, setting up multi-factor authentication that impede the adoption of such services [15]. SSI promises lower usage barriers through simplified verification of identities and thus fewer steps for registration and authentication. Both privacy as well as ease of use have been found to have a positive impact on customer experience of smart services [14].

For our study, we have chosen the smart service system of contractual electric vehicle (EV) charging with roaming capabilities. In this domain, there are multiple actors involved in both the operation of charging points and the billing of the charging. Roaming providers act as intermediaries to simplify access to charging points in different regions. Existing research on electric mobility highlights the problem of interoperability and complexity [16], e.g., through a model-based framework [17]. Therefore, we have posed following research questions:

- RQ1: How can SSI improve the design of an EV charging service (EVCS) system with roaming capabilities?
- RQ2: What are implications of these changes for actors involved in the service system?

Our expected result are (1) a set of areas in which smart service systems can benefit from SSI, and (2) implications for involved actors, if SSI is used in such systems. With that, we aim to create an initial understanding of the potential that SSI might provide for the design of smart service systems. Furthermore, we strive to uncover future research opportunities in this field.

1.2 Methodology

Our *research objective* is to understand the implications of applying SSI to smart service systems. As SSI has not been widely used yet, we intended to explore the potential impacts of SSI on the specific use case of roaming in EVCS using a conceptual analysis of a reference process. For that, we have chosen the following *research approach:*

1. Identify characteristics of SSI from literature that are relevant for service system design,
2. Extract reference model of service system from roaming standards as a basis for a qualitative comparison,
3. Perform a SWOT analysis to identify potential areas of improvement of the service system through the exploitation of potential benefits created by SSI,
4. Derive an improved model of SSI-enabled EV charging with roaming capabilities,
5. Discuss the implications for the involved actors resulting from these changes.

2 Conceptual Foundation

2.1 Application of Self-Sovereign Identities in Business Scenarios

The paradigm of SSI puts the user at the centre of the identity ecosystem and back in control of his or her identity data. Identities for various subjects can be registered, resolved, updated or revoked without a central authority [4]. Identities are rooted in immutable data registries, which are also not necessarily controlled by a central authority like a government or a private company [4], [6]. Such registries are typically implemented using distributed ledger technologies (DLT) [18]. Using a software client called “wallet”, the user can collect cryptographically secured claims (credentials) about various attributes of his or her identity, which are provided by trustworthy institutions (issuers). The user (holder) can then present a subset of these attributes to service providers (verifiers) upon request [4], [6] in a peer-to-peer (P2P) communication, i.e., without an intermediary. As the credential is cryptographically verifiable, the issuer does not need to be contacted, which prevents tracking and correlation of user activities. Finally, digital identities are not only applicable to persons, but also to organisations and objects, e.g., technical products [3].

There is an ongoing effort to increase the interoperability of SSI-based solutions through standards. Most notably are the W3C standards ”Decentralized Identifiers” (DID) [19] and ”Verifiable Credential” (VC) [20]. A DID is an identifier according to the Uniform Resource Identifier specification, which uniquely identifies a DID subject. A DID refers to a DID document that defines service endpoints, which can be used to interact with the DID subject. It is common practice to create a new DID for each new business relationship, which prevents correlation of user data and thus ensures a high level of data protection [11]. VCs allow for complex interactions between participating entities, which are identified by their DIDs. The authenticity of a VC’s contents is verifiable by cryptographic methods. Furthermore, the integrity of a VC can be proven by persisting a hash in an immutable data registry serving as a fingerprint.

In the existing literature on SSI in business scenarios, the following benefits are mentioned:

- User data does not need to be stored at the service provider [7]
- Identity issuer is not involved in identity access, verification, and resolution [7], [8], [10]
- Identity data integrity and availability is ensured [7]
- Privacy of users is ensured [7]–[12]
- Users can trade their data for rewards [9]
- Users can link different identity attributes from various issuers on their own [10]
- Anonymous participation in transactions [11], [12]

Besides these, also critical issues were raised that limit the adoption of SSI:

- Lack of effective key management [7]
- Insufficient knowledge on alignment of technical, institutional, and societal aspects [12]
- Frameworks are not production-ready and prone to feature changes [3]
- SSI is in early stage and lacks of widespread adoption [3]
- Incomplete standardisation hampers interoperability [3]

2.2 Electric Vehicle Charging Services

EVCS are embedded in a larger ecosystem of smart energy and mobility services. The diverse actors within this ecosystem are characterised by different goals and business models. These actors and their relationships are described by the E-Mobility Systems Architecture (EMSA) [17] as well as the industry standard Open Charge Point Interface (OCPI) [21]. We used these as a theoretical foundation to derive an exemplary model of an EVCS with roaming capabilities. Roaming can be understood as a special feature of such services allowing a customer to charge at a multitude of charging stations of operators they have no direct contractual relationship with [22]. Services without explicit contractual relationships such as ad-hoc charging were out of scope of our research.

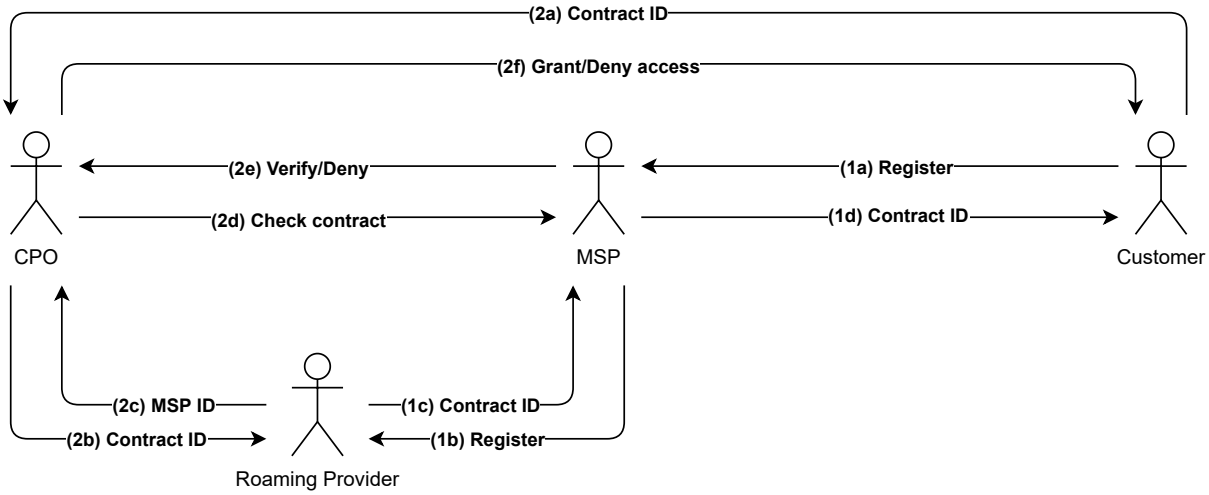


Figure 1. Interactions during registration and authentication of customers in an EVCS

2.2.1 Common Roles in Roaming Scenarios

There are four main roles in EVCS systems as described in EMSA [17] and OCPI 2.2 [21]. The descriptions given below are based on the assumption of a hub-based service topology for the exchange of contractual data [21], to provide an EVCS with roaming capabilities.

The **Charging Point Operator (CPO)** manages and operates charging points. CPOs enable potential customers to recharge their EVs. However, CPOs are geographically limited in acquiring new customers and compete locally with other CPOs. As a result, CPOs depend on MSPs as brokers to market their charging services and reach customers outside their main region of activity, in order to gain a competitive advantage.

The **Mobility Service Provider (MSP)** provides customers with services to enhance personal mobility, including access to charging points of several CPOs under a common contract. This simplifies billing for customers since the MSP handles the payment processing with involved partners. Furthermore, by aggregating demand, MSPs may offer lower prices to customers. The role of MSP may also be assumed by a CPO that is active in multiple regions.

The **Roaming Provider** acts as intermediary between MSPs and CPOs and manages a business technology platform for the uniform exchange of charging authorisations. Roaming providers form regional ecosystems for technological cooperation and creation of common standards.

The **Customer** initiates the charging session, confirms its completion, and pays for the process. Furthermore, a customer enters contractual relationships with an MSP to take advantage of mobility options or alternative billing models. Specifically in roaming scenarios, a customer only indirectly pays for the charging provided by the CPO via the MSP. In the domain of electric mobility, the customers experience regarding the relationship to service providers was described as negative, especially mentioning a lack of integration [14].

2.2.2 Interactions during Roaming Scenario

Using the specifications from OCPI 2.2 [21], we examined the following three processes to describe a roaming scenario in an EVCS system:

1. Registration of CPO and MSP to the roaming platform
2. Registration of customer to an MSP
3. Authentication of customer at a charging point

For clarity, we modelled the interactions within these processes in figure 1 without the initial registration of CPO and MSP to the roaming platform.

During registration, the customer provides the selected MSP with PII data that is relevant to invoicing. Once the customer has decided on a tariff, the contract information is transmitted to the roaming platform. It generates an identifier, which allows the mapping of the customer to possible CPOs according to the selected tariff. Finally, the customer is provided with the identifier, which can be presented at the charging points as stipulated in the service agreements. The contract identifier is further passed on by the corresponding CPO to the roaming platform, which locates the associated MSP. The MSP is then contacted by the CPO regarding the validity of the contract corresponding to the presented identifier. If the authentication is successful, a confirmation message is sent that leads to the release of the charging point.

3 Results

3.1 Identification of Options for Applying SSI in EV Charging Services

To explore the expected impacts of SSI on EVCS systems, we conduct a SWOT analysis. The internal perspective is based on the analysis of the presented exemplary EVCS system. The external perspective is represented by the opportunities and threats associated with SSI as described in section 2.1. Figure 2 shows an overview of the analysis inputs.

First, roaming strengths as identified by analysis of the industry standard documents should be highlighted. Charging services with clearly defined and uniform interfaces for customer interactions have a high degree of maturity. If the service processes can be handled via an application issued by the MSP, a streamlined activity of registration can be assumed, as no physical documents have to be exchanged to establish the contract. At the service providers, the lack of a requirement to manage physical contract documents reduces to redundancies in master data. While agreeing on a common roaming platform simplifies the processes between MSPs and CPOs to a certain extent and responsibility can be handed over, weaknesses still arise from this intermediation scenario.

The quasi-standardisation by roaming platforms leads to an accumulation of market power. MSPs and CPOs are dependent on the roaming providers to effectively manage their services, as they can only easily enter into business relationships with partners who are on the same platform. Being active on several platforms is either not possible due to regional monopolies or is economically unreasonable. In a competitive scenario, this leads to a limitation of the number of charge points that can be used by a customer under a single contract and thus negatively impacts the added value provided by roaming roaming capabilities.

The opportunities for potential applications of SSI approaches can be taken directly and indirectly from the so-called ten principles of SSI defined by Allen [23]. The most important opportunity is also the eponymous property of the SSI paradigm: Users should gain sovereignty over the data they own and thus enable them to decide which data is shared with whom. In conjunction with zero-knowledge proofs, it is also possible to prove the correctness of data without disclosing its contents [7]. The use of zero-knowledge proofs could therefore contribute to a lower utilisation of networks and databases. This could also reduce risks associated with data theft and improve data security in general. The use of DLT to implement SSI solutions could also increase transaction security, as it records business activities immutably in a decentralised ledger. This immutability provides the basis for the verifiability of the integrity of data used for business purposes and thus enables the issuance of verifiable documents without involving third parties. Initial efforts to standardise data formats and processes related to the aforementioned opportunities potentially enable the production of a uniform protocol for identity data exchange that is independent of specialised providers.

Despite the many opportunities for customers and service providers, the SSI concept is still in its infancy. While key components of the SSI architecture have been clarified and standards such as DID and VC emerge, there are still uncertainties in the details related to specific use

SSI Opportunities	SSI Threats
• Service providers do not have to maintain user data • Identity issuer not involved in identity access, verification and resolution • Integrity and availability of identity data is ensured • Users can consciously employ data in transactions • Users can link different identity attributes independently • Anonymous transaction participation	• Lack of effective key management • Insufficient knowledge on alignment of technical, institutional and societal aspects • Low number of production-ready frameworks lead to investment uncertainties • Low rate of adoption among users and service providers • Incomplete standardisation hampers interoperability
EV Charging Services Strengths	**EV Charging Services Weaknesses**
• Defined authentication method per service provider • Unique identifiers for authentication of end customers • Coordinated business partners due to role model • (Proprietary) standards for IT handling of business processes • Potentially low number of customer interactions	• Dependencies between business partners • Complex, interdependent processes • Intransparent business network • Interoperability across standards not guaranteed • Cumbersome administration of tariffs • Closed ecosystems

Figure 2. SWOT analysis inputs

cases. For example, the schema of VCs must be defined in advance to ensure interoperability between business partners in an industry. Additionally, for each industry a set of authorised institutions must be maintained that are widely accepted as trustworthy issuers. This complicates the design of such use cases with SSI approaches, as assumptions have to be made under uncertainty. On the other hand, this flexibility, which is still present at the beginning, enables the introduction of well-founded proposals for improving the SSI ecosystem.

From the mutual consideration of SSI opportunities and threats as well as strengths and weaknesses associated with EVCS, we can derive four strategic options for the application of SSI in these kinds of service systems.

- **Option 1:** Integrating SSI approaches into the contract creation process could simplify registration and subsequent authentication processes.
- **Option 2:** Using DIDs and VCs, MSP and CPO could emancipate themselves from the roaming provider as a trusted source of business identities.
- **Option 3:** The use of existing ecosystems and protocols from the domain of EVCS could relieve uncertainties related to the SSI standards under development.
- **Option 4:** The gradual implementation of components and the exchange with industry partners could promote interoperability between providers and a smooth transition to the SSI paradigm.

Based on options 1 and 2, highlighting the opportunities of SSI, the main direction of an EVCS system utilising SSI can be derived. With SSI approaches dedicated providers of identities and associated data objects are no longer necessary. Thus the role of roaming provider becomes obsolete. Instead of using a central platform for the exchange of contract data, SSI assumes an actor-centric perspective. SSI techniques can be applied at two points in the scenario of roaming. First, customers, MSP, and CPO can create and manage their own business identities utilising DIDs. Second, the business relationships between customer and MSP and between MSP and CPO as well as the associated rights and obligations can be represented in the form of VCs.

Cooperation in a complex business network requires a high degree of coordination, which is illustrated in particular by options 3 and 4. Since the removal of a central business partner from the service ecosystem represents a drastic change, the distribution of roles and tasks of the actors also changes in order to convey the same value to a customer. In the following sections we focus on the interactions that are relevant to the service provisioning at its core. As such, the financial clearing of the consumed services will be out of scope.

3.2 An SSI-enabled EV Charging Service System

Applying SSI technologies to an exemplary roaming service system following the presented options leads to changes in the interaction structure of said system. With the elimination of the roaming platform provider three actors remain in the EVCS system leading to three possible interactions which we have conceptually examined. An overview of the issuance and presentation of the credentials involved in the SSI-based service is provided in figure 3. The details of the actors' interactions are explained in the subsequent sections.

3.2.1 Interactions between CPO and MSP

The CPO provides the central service to a customer in EV charging. Thus, the CPO commissions an MSP to mediate their charging points under certain conditions. With VCs, it is possible to map the relationship between customers and MSP in a verifiable and persistent way. Therefore, the CPO can create a VC with the authorisation for these mediation activities and sign it with their private key. This "Broker VC" may be secured against misuse by two measures. First, the DID of the MSP could be included in the VC schema. Only the specified MSP and third parties authorised by this MSP could then prove that the MSP is in control. Second, a hash function could be used to record a fingerprint of the VC in an immutable registry so that its issuance could be proven. Furthermore, this "Broker VC" could contain an automatic expiration date following the contractual agreements between MSP and CPO. After the VC issuance, the associated MSP can store it in a self-selected location. With this VC, the MSP can verifiably broker the CPOs charging points at their own pricing, which makes up the core of an EVCS with roaming capabilities. After this, the mediation activities of the MSP is be electronically authorised by the "Broker VC", which is used as proof in case the services are contested by any party.

3.2.2 Interactions between MSP and Customer

When customers decides to enter a contract for the roaming service, their software agents (wallet) create a DID with a corresponding DID document specifically for the business relationship with the MSP. Within the DID document, service endpoints are be specified which are necessary to interact with the customer within this specific business relationship. The mapping of the contractual relationship could be handled by using another VC. On a basic level, the VC issued by the MSP to the customer only need to contain a statement of being a customer of the MSP. The type of customer and therefore the type of contract would need to be defined by the VC type, which should be based on a common schema. Also, this "Service VC" could contain tariff regulations and the duration of the contract. Including the contract data in the same VC could lead to the sharing of data at the charging point, which is not necessary for plain authorisation. To enable selective disclosure of individual components of the "Service VC", the chosen signature procedure would need to include each individual attribute.

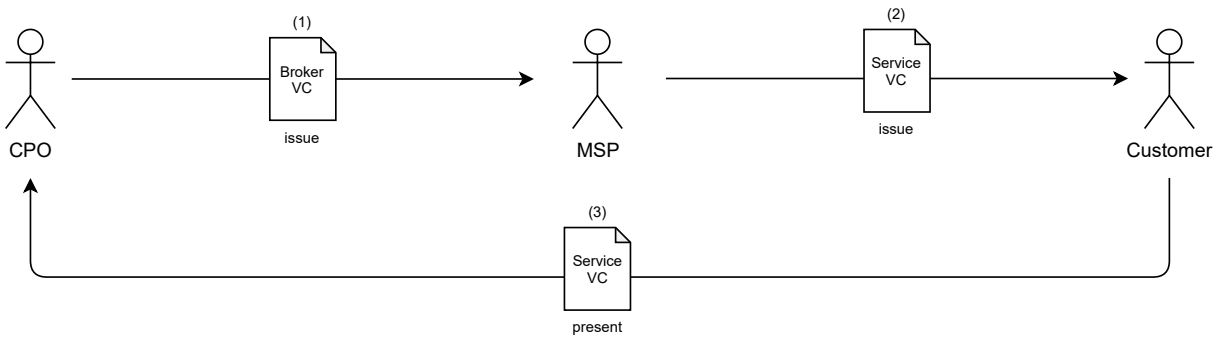


Figure 3. Issuance and presentations of broker and service credentials in SSI-enabled EV charging scenario

3.2.3 Interactions between Customer and CPO

Once the service VC is be issued, customers can authenticate themselves at the charging points of the partner CPOs mediated by the MSP. After physically connecting the vehicle to a charging point, the customer would be asked for authentication to start the charging process. For that, the charging station management sends a request to the CPO's backend to create a so-called challenge. This challenge may be displayed in the form of a QR code at the graphical user interface of the charging point. A scan of this code forwards the customers to their wallet application, wherein they can accept the request to present proof of a valid roaming contract. After selecting the corresponding attributes of the "Service VC" issued by the MSP, these are sent to the CPO backend using the smartphone's mobile data connection. The correct destination could either be stored directly with the challenge or could be determined by resolving a DID and accessing a dedicated service endpoint.

A total of three checks would have to be performed on the presented VC, in order to ensure a sufficient level of certainty about the correctness of the transmitted claims:

- **Integrity** This can be implemented by using the proof algorithms of VC and its presentation. A positive outcome of the presentation proves that no changes were made to the data. Since the included VC consists of derived attributes from the Service VC, a proof would include both the correctness of the stated claim of charging authorisation and the integrity of the underlying Service VC.
- **Validity** This would include a review of the expiration date to determine if the underlying contract is still effective. Furthermore, it could be checked whether a valid MSP can be found for the issuing DID.
- **Schematic correctness** The schema used by the VC would need to be compared with the schema definition agreed upon between MSP and CPO.

If these requirements are met, the customer is successfully authenticated and the requested charging process will be authorised. Otherwise, using an endpoint for messages from the customer's DID document or via the interfaces of the charging point, an error report should be sent, including suggestions for an alternative method for starting the charging process.

4 Discussion

Based on the results presented in the previous section, we discuss their impact for the design of smart service systems, as well as implications for the actors involved in service systems.

4.1 Impact of SSI on Smart Service System Design

With regards to the service system design, we can distinguish the impact on the service experience and customer experience. On the service experience, the major difference is the reduction of complexity through SSI as the standardised method for P2P credential exchange. It replaces commonly used registration processes, which result in additional processing steps for data verification and more complex customer data to be maintained. Another benefit of SSI is the improved interoperability of IT systems between different actors in a service system. Instead of expensive direct point-to-point communication between backend systems, the data exchange takes place via credential exchange via the customers' wallets.

This also has implications for the customer experience, as customers can directly authenticate themselves with IT systems of service providers using credentials from a trusted issuer, such as banks or municipalities. Avoiding complex registration processes makes services more accessible, as fewer steps are required for establishing a trusted relationship between the actors in the service system. Additionally, as such wallet apps are used in a variety of occasions, a growing number of users become familiar with their use. This is an improvement over today's situation in which the use of a new service requires the installation and setup of a service-

specific app. Overall, SSI contributes towards a seamless customer experience with fewer apps, less effort for learning the handling of unfamiliar apps, and simpler authentication of the user.

4.2 Implications of SSI for Involved Actors

From the changes in the smart service system outlined above, we can derive implications for the involved actors, which can be distinguished into customer, service provider, and broker (such as roaming providers).

Customer For customers, SSI-enabled smart service systems could offer a more comfortable enrolment and authentication process. If the customer already uses a digital wallet, he or she can provide a proof of his identity without a complex registration process. Once a connection to the service provider is established, authentication can be performed automatically and without passwords. This streamlines the overall customer experience. and satisfies customers needs for independence and easier integration [14]. Additionally, an SSI-enabled smart service might be preferred by customers as the exchange of data between customer and service provider is transparent and under control of the customer. This improves privacy, which is one of the most frequently mentioned benefits expected for business scenarios (see section 2.1).

Service Provider For service providers SSI can potentially lower the entry barrier for customer on-boarding and simplify the authentication for service usage. Besides, it makes the service more attractive through higher levels of privacy. However, this assumes that the customer already has an digital wallet and VCs for his or her identity from a trusted issuer in place.

Besides authentication, VCs can also be used for various purposes dealing with proving status or permissions to other parties. This offers an opportunity for businesses that strive for efficient coordination between companies to reduce their cost. The specific requirements of a use case are determined in the business practice itself and by external regulations. Considering the regulatory requirements for the official calibration of charging points' measurement equipment, SSI approaches could be used not only to identify customers, business partners, and contracts but also to ensure the integrity of charging data. A VC could be created that verifiably persists the relevant attributes, including duration and type of the charging process, a process identifier, as well as the power consumed in kilowatt-hours.

On the operational level, the cost for customer data processing might be reduced. As customers can provide verified and up-to-date identity data on demand, the cost on GDPR-compliant processing of customer data at the service provider is lowered. Furthermore, expensive procedures to ensure and maintain data quality of customer data are not required anymore. On the other hand, initial investments for enabling existing systems for SSI need to be considered.

Broker We can also observe changes to the overall service ecosystem. SSI provides the transformative capabilities to map and enrich different commercial structures. As SSI puts persons and organisations into control of their identity data, they can interact with other actors through P2P credential exchange. This weakens the position of platforms that act as broker between actors in a service ecosystem, which can eventually lead to their disintermediation.

Our EVCS example shows the emancipation of MSP and CPO from the roaming provider, which is therefore not needed anymore. This is mainly enabled by the secure P2P credential exchange between various actors. To offer a roaming service to a customer, MSP and CPO must agree on aspects such as pricing, internal billing, and the charging methods offered. Furthermore, a common schema for the contractual VCs must be defined and stored in a suitable location with verifiable integrity. These schemas form the basis for the data structures of the VCs used in the service process and decide to a large extent on the interoperability of the service offerings of actors in a service ecosystem. For each roaming service offered, a common

VC schema is required to verifiably map the authorisations and roles in the business scenario. To make this approach accessible to providers of other scenarios and thus to bring about an integration of the business network, coordination at the industry level is necessary.

4.3 Limitations

The insights gained in this paper are based on a single domain, which is commercial charging processes and the service architecture of roaming Therefore, the described disintermediation effects might not be transferable to other service systems. Another limitation is the characteristic of this work being a conceptual analysis. As the validation of the proposed concepts within a real roaming ecosystem was out of scope, our insights are based on a theoretical analysis. The actual impact of the application of SSI will be highly dependent on the concrete implementation and service design of real organisations and service ecosystems. Therefore, an empirical analysis is required to verify the effects in a real-world implementation of the presented concept.

4.4 Future research

Based on our study, we identified potentials for future research along four research topics:

Customer experience of smart services: There is a need to better understand the effects of SSI for the enrolment process, which essentially creates a ”bring your own identity” model. Here, interaction designs and customer experience should be systematically investigated to create design knowledge for smart service systems. Usability research has dealt with different enrolment processes, however they did not cover SSI yet [15]. Existing research on customer experience of smart service has yet to consider the effects of SSI, particularly on privacy, ease of use, accessibility, and controllability [14].

Interoperability: SSI changes the mechanism on how data is transferred between systems. Instead of direct communication, the user (identity holder) transfers data between systems via his or her wallet. This reduces the need for complex integration but requires consensus about the semantics of data in VCs. This can be achieved through credential schemata. Open questions in this area are related to the schema definition process, the selection of actors to issue schemata, as well as to their storing, resolving, and versioning.

Drivers and barriers for adoption: The introduction of SSI in smart service systems can reduce cost for proprietary apps, enrolment procedures, and GDPR-compliant handling of PII data. Enabling of existing applications for SSI on the other hand might be expensive due to a lack of reusable, production-ready components such as integration tools and software libraries. With regard to benefits, SSI can lower the entry barrier for new users and improve privacy. Therefore, cost-benefit analysis should be conducted to evaluate effects in real world cases to develop evaluation models which help practitioners to make informed decisions with respect to both, costs arising from the technical implementation as well as changing business models. Regarding the latter, SSI as a paradigm leads to a shift in the availability of data from service providers to customers. This enables selective disclosure and conscious transactions using this data from the customers perspective. It is yet to be examined from the service providers perspective whether potentially lowered costs in the handling of PII data outweighs the impacts on data utilisation as part of their business models.

Service ecosystems: The disruptive potential of SSI with regard to disintermediation of actors in service ecosystems needs to be studied from an economic as well as strategic management perspective. While SSI might make the service of certain actors less relevant, it can be assumed the it will also create the need for new ones, e.g., trusted issuers, quality assurers, operators of DLT infrastructure, wallet providers, and integration services. While roles in multi-actor smart service innovation have been a topic of recent research [24], it does neither consider the specifics of SSI nor the potential strategic options for actors threatened by disintermediation.

5 Conclusions

Establishing trust between actors in smart service system is a task that leads to various challenges with existing approaches. The inherent characteristics of self-sovereign identity promise to simplify the establishing of trust in services processes but also influence the overall design of the service system. Using the example of an EVCS, we gained an initial understanding of the impacts that SSI can have on smart service systems. Our findings underline the potential of SSI, which should therefore be considered in the design of such systems.

As an emerging concept, it is not surprising that academic works on the application of SSI in business scenarios is scarce, let alone in the context of smart service systems. With the ongoing proliferation of SSI in various industries and application domains, empirical research can be conducted to better understand how the expected benefits of SSI will manifest themselves.

[1] K. Cameron, "The laws of identity," 2005. [Online]. Available: https://www.identityblog.com/stories/2005/05/13/TheLawsOfIdentity.pdf.

[2] O. Avellaneda, A. Bachmann, A. Barbir, J. Brenan, P. Dingle, K. H. Duffy, E. Maler, D. Reed, and M. Sporny, "Decentralized identity: Where did it come from and where is it going?" *IEEE Communications Standards Magazine*, vol. 3, no. 4, pp. 10–13, 2019, ISSN: 2471-2825. DOI: 10.1109/MCOMSTD.2019.9031542.

[3] P. C. Bartolomeu, E. Vieira, S. M. Hosseini, and J. Ferreira, "Self-sovereign identity: Use-cases, technologies, and challenges for industrial iot," in *Proceedings, 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Piscataway, NJ: IEEE, 2019, pp. 1173–1180, ISBN: 978-1-7281-0303-7. DOI: 10.1109/ETFA.2019.8869262.

[4] K. C. Toth and A. Anderson-Priddy, "Self-sovereign digital identity: A paradigm shift for identity," *IEEE Security & Privacy*, vol. 17, no. 3, pp. 17–27, 2019, ISSN: 1540-7993. DOI: 10.1109/MSEC.2018.2888782.

[5] B. Cyphers and G. Gebhart, *Behind the one-way mirror: A deep dive into the technology of corporate surveillance*, Electronic Frontier Foundation, Ed., 2019. [Online]. Available: https://www.eff.org/wp/behind-the-one-way-mirror.

[6] A. Mühle and A. Grüner, "A survey on essential components of a self-sovereign identity," 2018. [Online]. Available: https://www.researchgate.net/publication/326459642_A_Survey_on_Essential_Components_of_a_Self-Sovereign_Identity.

[7] R. Soltani, U. Trang Nguyen, and A. An, "A new approach to client onboarding using self-sovereign identity and distributed ledger," in *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, IEEE, 7/30/2018 - 8/3/2018, pp. 1129–1136, ISBN: 978-1-5386-7975-3. DOI: 10.1109/Cybermatics{\textunderscore}2018.2018.00205.

[8] D. W. Chadwick, R. Laborde, A. Oglaza, R. Venant, S. Wazan, and M. Nijjar, "Improved identity management with verifiable credentials and fido," *IEEE Communications Standards Magazine*, vol. 3, no. 4, pp. 14–20, 2019, ISSN: 2471-2825. DOI: 10.1109/MCOMSTD.001.1900020.

[9] K. Wittek, L. Lazzati, D. Bothe, A.-J. Sinnaeve, and N. Pohlmann, "An ssi based system for incentivized and self-determined customer-to-business data sharing in a local economy context," in *2020 IEEE European Technology and Engineering Management Summit (E-TEMS)*, IEEE, 2020, pp. 1–5, ISBN: 978-1-7281-0903-9. DOI: 10.1109/E-TEMS46250.2020.9111805.

[10] A. Stasis, N. Triantafyllou, P. Georgakopoulos, R. L. Armitt, and P. Kavassalis, "Designing an academic electronic identity management system for student mobility using eidas eid and self-sovereign identity technologies," in *26th Annual EUNIS Congress*, 2020.

[11] R. Karatas and I. Sertkaya, "Self sovereign identity based e-petition scheme," *International Journal of Information Security Science*, vol. 9, no. 4, pp. 213–229, 2020.

[12] R. B. Gans, J. Ubacht, and M. Janssen, "Self-sovereign identities for fighting the impact of covid-19 pandemic," *Digital Government: Research and Practice*, vol. 2, no. 2, pp. 1–4, 2021, ISSN: 2691-199X. DOI: 10.1145/3429629.

[13] D. Beverungen, O. Müller, M. Matzner, J. Mendling, and J. Vom Brocke, "Conceptualizing smart service systems," *Electronic Markets*, vol. 29, no. 1, pp. 7–18, 2019, ISSN: 1019-6781. DOI: 10.1007/s12525-017-0270-5.

[14] L. Gonçalves, L. Patrício, J. Grenha Teixeira, and N. V. Wünderlich, "Understanding the customer experience with smart services," *Journal of Service Management*, vol. 31, no. 4, pp. 723–744, 2020, ISSN: 1757-5818. DOI: 10.1108/JOSM-11-2019-0349.

[15] C. Porter, "Design shortcomings in e-service enrolment processes," *International Journal of E-Services and Mobile Applications*, vol. 10, no. 3, pp. 1–18, 2018, ISSN: 1941-627X. DOI: 10.4018/IJESMA.2018070101.

[16] N. Masuch, E. Eryilmaz, T. Küster, U. Pletat, J. Fähndrich, T. Theodoropoulos, M. Koukovini, N. S. Hadjidimitriou, and N. Dellas, "Decentralized service platform for interoperable electro-mobility services throughout europe," in *Towards User-Centric Transport in Europe 2: Enablers of Inclusive, Seamless and Sustainable Mobility*, B. Müller and G. Meyer, Eds., Cham: Springer International Publishing, 2020, pp. 184–199, ISBN: 978-3-030-38028-1. DOI: 10.1007/978-3-030-38028-1{\textunderscore}13.

[17] B. Kirpes, P. Danner, R. Basmadjian, H. d. Meer, and C. Becker, "E-mobility systems architecture: A model-based framework for managing complexity and interoperability," *Energy Informatics*, vol. 2, no. 1, p. 28, 2019. DOI: 10.1186/s42162-019-0072-4.

[18] D. van Bokkem, R. Hageman, G. Koning, L. Nguyen, and N. Zarin, "Self-sovereign identity solutions: The necessity of blockchain technology," *arXiv e-prints*, arXiv:1904.12816, 2019. [Online]. Available: https://arxiv.org/pdf/1904.12816v1.pdf.

[19] M. Sporny, D. Longley, and D. Chadwick, *Verifiable credentials data model 1.0: Expressing verifiable information on the web*, 2020. [Online]. Available: https://www.w3.org/TR/vc-data-model/.

[20] D. Reed, M. Sporny, D. Longley, C. Allen, R. Grant, M. Sabadello, and J. Holt, *Decentralized identifiers (dids) v1.0: Core architecture, data model, and representations*, 2020. [Online]. Available: https://www.w3.org/TR/did-core/.

[21] EVRoaming Foundation, *Ocpi 2.2: Open charge point interface*, 12.06.2020. [Online]. Available: https://evroaming.org/app/uploads/2020/06/OCPI-2.2-d2.pdf.

[22] J. Ratej, B. Mehle, and M. Kocbek, "Global service provider for electric vehicle roaming," in *2013 World Electric Vehicle Symposium and Exhibition (EVS27)*, Nov. 2013, pp. 1–11. DOI: 10.1109/EVS.2013.6914941.

[23] C. Allen, *The path to self-sovereign identity*, 2016. [Online]. Available: http://www.lifewithalacrity.com/2016/04/the-path-to-self-soverereign-identity.html.

[24] J. Anke, J. Pöppelbuß, and R. Alt, "It takes more than two to tango: Inter-organizational collaboration in smart service systems engineering," *Schmalenbach Business Review*, vol. 72, no. 4, pp. 599–634, 2020. DOI: 10.1007/s41464-020-00101-2.

24th International Conference on Business Information Systems (BIS 2021)

Knowledge Graphs

https://doi.org/10.52825/bis.v1i.39

Published: 02 July 2021

Domain-specific Event Abstraction

Finn Klessascheck[1], Tom Lichtenstein[1], Martin Meier[1], Simon Remy[1], Jan Philipp Sachs[2,4], Luise Pufahl[3], Riccardo Miotto[4,5], Erwin Böttinger[2,4], and Mathias Weske[1]

[1]Hasso Plattner Institute (HPI), University of Potsdam, Potsdam, Germany

[2]Digital Health Center, HPI, University of Potsdam, Potsdam, Germany

[3]Software & Business Engineering, Technische Universitaet Berlin, Berlin, Germany

[4]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, USA

[5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA

Abstract. Process mining aims at deriving process knowledge from event logs, which contain data recorded during process executions. Typically, event logs need to be generated from process execution data, stored in different kinds of information systems. In complex domains like healthcare, data is available only at different levels of granularity. Event abstraction techniques allow the transformation of events to a common level of granularity, which enables effective process mining. Existing event abstraction techniques do not sufficiently take into account domain knowledge and, as a result, fail to deliver suitable event logs in complex application domains. This paper presents an event abstraction method based on domain ontologies. We show that the method introduced generates semantically meaningful high-level events, suitable for process mining; it is evaluated on real-world patient treatment data of a large U.S. health system.

Keywords: Process mining, Event abstraction, Domain knowledge, Healthcare

1 Introduction

Many organizations have an inherent interest to monitor and understand their processes. For example, analyzing and adopting processes can improve their overall efficiency, e nsure that legal requirements are met, and maintain a desired quality level. To this end, process mining provides techniques to analyze processes based on event data recorded during their execution. However, such event data is not always available in the necessary format and often differs in its granularity in complex settings. This also applies to treatment processes in hospitals, which are typically highly heterogeneous, complex, multidisciplinary, ad-hoc, and susceptible to change [1]. In the past, process mining has been proven as a technique well-suited to derive an understanding of medical processes, like patient-flows, and to improve them accordingly [2].

When extracting event data for process mining of electronic health records (EHRs), multiple data sources have to be tapped into, including hospital information systems. This variety of data sources and differences in data granularity leads to a mismatch in the level of abstraction between different events. Moreover, the fact that many events are documented manually by physicians or other medical personnel typically leads to varying degrees of detail in the recorded events. Using the resulting event logs would result in complex process models [3]–[5]. In order to generate event logs with events of comparable granularity and to elicit useful process models, event abstraction is needed. To this end, a rich set of research works on event abstraction

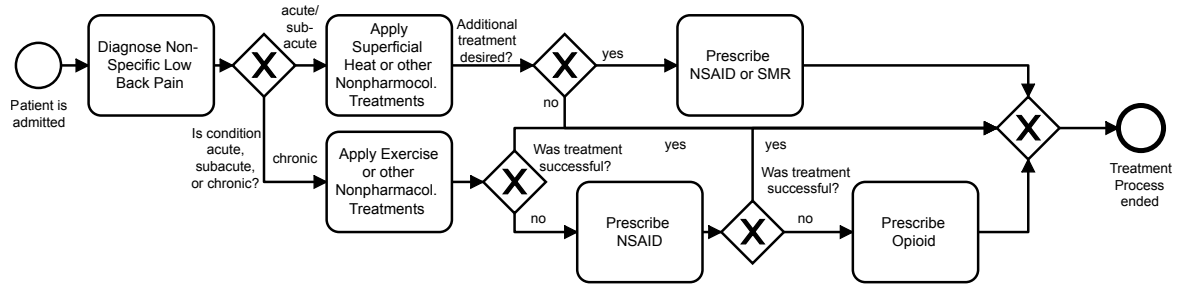


Figure 1. BPMN process model describing the treatment of LBP as per [7]. The events appearing in the event log of List. 1 belong to the same process but do not align with the activities of the process model, and make it difficult to put them into relation.

methods and techniques exists, as shown in [3], [6]. But fewer research works utilize existing domain knowledge. Those approaches mostly follow a bottom-up principle by starting from the observed event data. While the resulting process models are less complex and already improve the process analysis, bottom-up approaches do not guarantee a conceptually sound abstraction. There is a high prevalence of standards and ontologies in the medical field, containing domain-specific information such as medications, procedures, and diagnoses. This knowledge can be used to enrich abstraction mechanisms by actively selecting a use case-specific level of abstraction.

This paper presents an event abstraction method based on domain knowledge for process analysts to generate event logs for process mining. The method is domain-agnostic but was developed and tested for the healthcare domain. In the remainder of this paper, a motivating example is given in Sect. 2 followed by an overview of related work in Sect. 3. Afterwards, the domain-specific abstraction method, is presented in Sect. 4. Our approach is applied to the back pain treatment process data of a large health system in the U.S. in Sect. 5. The work is concluded and future research discussed in Sect. 6.

2 Motivating Example

In this section, we discuss the treatment process of non-specific low back pain (LBP), which is one of the most frequent complaints, as a motivating example. The evidence-based framework for diagnosis and treatment of LBP in the U.S. context is defined by a guideline from the American College of Physicians, which outlines under which circumstances and in what order certain interventions are to be conducted or medications to be prescribed [7], as shown in Fig. 1. As such, the guideline encompasses recommendations with regards to the prescription of a multitude of different drug classes, e.g., opioids (e.g., Oxycodone), non-steroidal anti-inflammatory drugs (NSAID; e.g., Ibuprofen), or skeletal muscle relaxants (SMR; e.g., Diazepam). Based on this, it could be of particular interest for a health system to which extent current treatment processes comply with the given clinical guideline.

```
<event>
    <date key=''time:timestamp'' value=''2018-08-04T13:04''/>
    <string key=''concept:name'' value=''NAPROXEN 500 MG TABLET''/>
    <string key=''event:context'' value=''EPIC MEDICATION''/>
    <string key=''event:code'' value=''19918''/>
</event>
<event>
    <date key=''time:timestamp'' value=''2018-09-01T13:59''/>
    <string key=''concept:name'' value=''OXYCODONE 5 MG TABLET''/>
    <string key=''event:context'' value=''EPIC MEDICATION''/>
    <string key=''event:code'' value=''20627''/>
</event>
```

Listing 1. Excerpt of a real-world event log containing low-level events.

The event data recorded in the hospital information system (HIS) is stored in a detailed manner due to medical necessity and billing purposes, including precise information about, e.g., the dosage of medications. Listing 1, extracted out of an EHR database, shows the event names that result from this data. While the event log contains fine-grained events, the treatment guideline modeled in Fig. 1 is concerned with more abstract, high-level activities (e.g., *Prescribe NSAID*)[1]. This makes it difficult to compare the guideline with the recorded treatment event data. Due to the fine-grained event data, discovering a process model without processing it beforehand would result in a highly complex model that would be hard to interpret. It would include, for example, numerous activities, splits and branches for each drug and its potential dosage. Thus, event abstraction is needed to enable the alignment of low-level events with the high-level activities of a clinical guideline and to discover more human-readable process models.

An easy approach would be to abstract all events that refer to the administration of drugs into the high-level event *Administer Drug*. On one side, this would reduce the complexity. On the other side, valuable information, like the administered drug, would be discarded, such that this abstraction would not be useful and diverge from the level of detail given by the clinical guideline. It would, for example, not be possible to differentiate between non-pharmacological treatments and pharmacological treatments or non-opioid and opioid treatments, as per the guideline. However, this categorization cannot be simply made with the available event information from the data warehouse (cf. List. 1), so that additional information, e.g., about the drug class or the effect mechanisms, is necessary. Standards and ontologies that contain information about medical concepts such as medications, procedures, and diagnoses, already exist in healthcare. Thus, in this work, we explore the application of such medical knowledge resources for exploiting the medical context of events and determining the right level of abstraction.

3 Related Work

In the following, we discuss related work on event abstraction in general and with a particular focus on healthcare. Diba et al. [6] identifies event abstraction as one out of three major tasks for event log generation. Similarly, van Zelst et al. [3] provides a taxonomy for event abstraction techniques and categorizations. In general, abstractions techniques can be distinguished based on their information richness, which again depends on the required input [6]. The first category comprises clustering and unsupervised learning approaches like [8], [9], which require no additional input. In addition, Richetti et al. [9] applies natural language processing to discover activities and objects from the event labels. Based on their semantic relations, similar activities get clustered and substituted. In contrast, Tax et al. [10] applies supervised learning to map low-level events to activity executions that requires labelled training data to learn the probabilistic mappings. Leemans et al. [11] utilizes event attributes to derive hierarchical models encapsulating low-level events in sub-processes. The approach lacks of a mechanism to guarantee semantically sound abstraction groups.

Another category comprises approaches using behavioural patterns to transfer low-level events into higher-level activities. As an example, Mannhardt et al. [12] provides a supervised event abstraction technique. With the help of behavioural activity patterns, domain knowledge is captured, based on which events are matched. Additionally, the approach by Baier et al. [13] utilizes enriched process models. They describe the issue of *n:m* mappings from events to activities and use an annotated process model to identify mappings between events and activities in which they additionally encoded domain knowledge. With the help of a gamified crowdsourcing approach, Sadeghianasl et al. [4] requires a group of domain-experts to participate in the game to equalize activity labels. Lastly, the authors of [14] present a knowledge-based

[1] Note that, while the example of this section only consists of medication-related events, the overall problem exists for diagnostic-/treatment-related events as well.

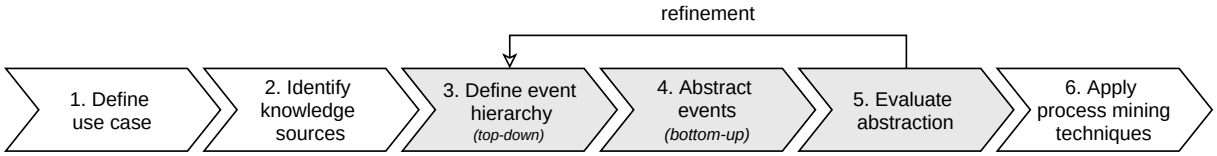


Figure 2. Method for domain-specific event abstraction. After defining the use case and identification of relevant knowledge sources, an adequate abstraction level is derived in an iterative manner.

abstraction mechanism based on domain-specific ontologies. Depending on a set of rules, events are mapped to the fundamental concepts of the ontologies; multiple events matching the same ground term within a given time window get merged into macro-activities. However, this approach heavily relies on predefined mapping rules.

For the healthcare domain, Mans et al. [15] illustrates the problem of different levels of data granularity by providing an overview of the spectrum of available data in an HIS. The authors classify different source systems based on the abstraction level and the data accuracy. Those findings are confirmed in a case study [16] on event log generation in healthcare in which event abstraction challenges and the need for suitable techniques, like abstraction tables, are discussed. In order to investigate treatment processes, the authors of [5] use, similarly to [11], sub-processes to encapsulate low-level events of treatment processes in sub-processes. The gold standard of the approach relies on the encoding of domain-knowledge within the activity labels, which might be added manually.

Most of the presented approaches are data-driven only and rely on structural features like control-flow pattern, while only a few also consider additional contextual information, e.g., [12], [14]. We will refer to those approaches as bottom-up. Still, they are missing mechanisms to generate a consistent level of abstraction, which is also medically sound.

4 Domain-specific Abstraction Method

In order to overcome the limitations of existing approaches, we introduce a method for domain-specific event abstraction. While the domain for the abstraction is not limited in any sense, one concrete abstraction instance can only be applied within one specific domain, and is not reusable across multiple domains. In the context of this work, the term abstraction means the unification of event identifiers for related events, so that several events are not combined to one event, but assigned to the same event class. The presented method combines two basic concepts of event abstraction, namely *bottom-up* and *top-down*. In the following, both concepts are presented, as well as the concrete abstraction procedure.

4.1 Basic Concepts of Event Abstraction

The first concept, which will subsequently be referred to as *bottom-up*, uses the low-level event log as its only input. Most of the abstraction approaches described in Sect. 3 can be considered to follow this concept. The advantage of a bottom-up abstraction is that little or no additional domain-specific information is required. However, such approaches produce results which might include high-level events at different levels of abstraction. Depending on the use case, it may be desirable to maintain a certain level of abstraction across the output event log as well as to produce high-level sound events for a domain. An abstraction that relies exclusively on the data level is therefore not sufficient in this case.

One way to overcome the lack of domain-specific information is to use an abstraction procedure following the concept of *top-down* abstraction. Top-down approaches first define the goal of the abstraction for a use case based on domain-specific information and map the low-level events to the targeted high-level events. In comparison to the bottom-up concept, the patterns

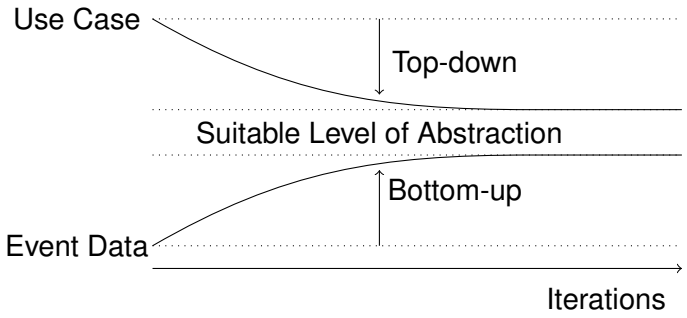


Figure 3. Conceptual depiction of the abstraction method. The use case and event data are alternately investigated from the top-down and bottom-up perspective to further restrict the abstraction scope.

and rules are not detected in the log, but defined by analysts in advance. The domain-specific patterns and rules are used to group related low-level events together using an *assignment algorithm* [12], [14].

The *top-down* approach allows to explicitly define the desired level of abstraction and can thus significantly improve the analysis of the process. However, an accurate abstraction of low-level events requires a deep understanding of the underlying data. Additionally, the abstraction must be conducted in such a way that it fits to the present data, which can be very complex and detailed, and is hard for the analyst to know beforehand. This can make exploratory analyses particularly difficult.

4.2 Alternating Event Abstraction

In order to find a suitable level of abstraction considering the available low-level event data as well as the abstraction goal of a use case, abstraction concepts of bottom-up and top-down are combined. Figure 2 depicts the method, which consists of six steps. In the first step, a use case needs to be defined due to the nature of the involved top-down concept. The use case may be guided by a domain-specific question, e.g., relating a treatment process to clinical guidelines or regulatory aspects. Next, knowledge sources relevant for the use case are collected and used as input for the subsequent steps. Before actual process mining techniques can be applied in step 6, an adequate abstraction level needs to be derived from steps 3 to 5 in an iterative manner. In these steps, the data and use case are alternately investigated from the top-down and the bottom-up perspective to further restrict the abstraction scope from each perspective with every iteration, as illustrated in Fig. 3. In order to achieve this, this paper proposes a hierarchy-based abstraction procedure.

Event Hierarchy. Since the desired granularity of an abstraction strongly depends on the use case, an *event hierarchy* is created as a basis for the abstraction of low-level event identifiers, which describes groups of related terms that are essential for the analysis. The event hierarchy is designed from a top-down perspective. It consists of several *abstraction sets*. Each abstraction set consists of an *abstraction term* and a *list of subordinate terms*. The abstraction term serves as the high-level event identifier for each low-level event that can be assigned to the abstraction set with the help of the corresponding subordinate terms. The general structure of an event hierarchy, as well as an exemplary abstraction set in the medical context, is illustrated in Fig. 4.

A relevant abstraction term in the medical context could be, for example, a class of drugs, such as NSAID. The low-level events are then mapped to the subordinate terms contained in the lists of the abstraction sets and abstracted to the corresponding abstraction term. The mapping is done by checking for each event whether its name contains one or multiple terms from the lists of the subordinate terms. Depending on the quality of the provided abstraction

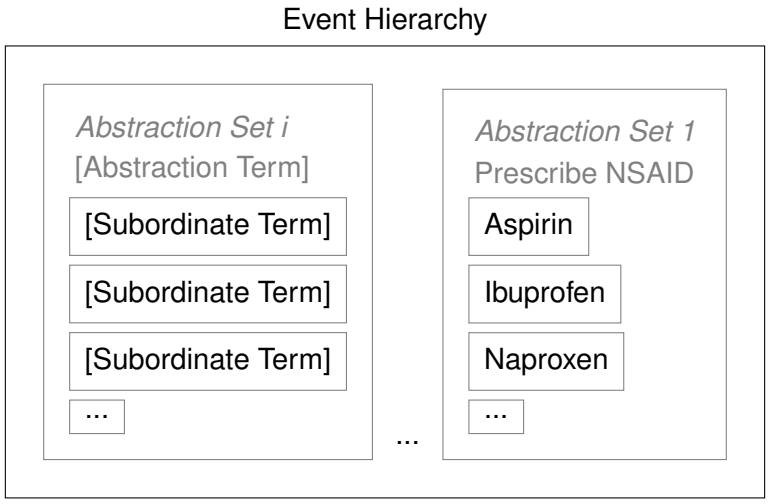


Figure 4. General structure of an event hierarchy with an example abstraction set.

sets, a single event name can contain different terms from different abstraction sets. In this case, either a domain expert can decide on the assignment, or the Levenshtein distance [17] or other kinds of distance measures could be used to determine the term most closely related to the event identifier. In order to avoid misleading abstractions, the lists of subordinate terms of different abstraction sets should not overlap. Having distinct abstraction sets allows a precise view of the data, which is directly tailored to the use case. After finding a matching term, the event identifier is then changed to the abstract term of the abstraction set. To preserve as much information as possible, the original name is added as an attribute to the event. If an event cannot be assigned to an abstraction set, the event keeps its original name. An abstraction set can additionally be marked as a filter set. If an event is assigned to an abstraction set that is marked as a filter, it is not included in the resulting event log.

Iterative Approximation of the Abstraction Level. Each iteration starts from the top-down perspective with the creation or refinement of the event hierarchy. After the abstraction, the resulting high-level events are analyzed from the bottom-up perspective. More concrete, it can be determined whether more abstraction sets should be added, or whether additional subordinate terms derived from low-level events that were not previously covered by any abstraction set should be added to one. Special attention should be paid to events that have not been abstracted to high-level events and to abstraction sets that abstract large parts of the event log. This evaluation can be done by either directly examining the resulting high-level events or by evaluating a process model, mined from the resulting event log. With that, the resulting event hierarchy is progressively refined towards the selected use case. Because of this, it cannot be used arbitrarily for other use cases.

5 Application to a Real-World Use Case

This section presents an exemplary application of the abstraction method to the real-world use case of Sect. 2. While the method itself is domain-agnostic in its application, a concrete example from the healthcare domain is suited to illustrate the concrete steps and their usefulness. As the first two steps (i.e., the use case definition and the collection of additional knowledge sources) are described in Sect. 2, this section evaluates and illustrates steps three to five.

5.1 Experimental Setup and Data Preparation

In this work we used EHR data from the Mount Sinai Health System, a large health system located in New York, NY, which generates a high volume of structured, semi-structured and unstructured data from inpatient, outpatient, and emergency room visits. Patients in the system

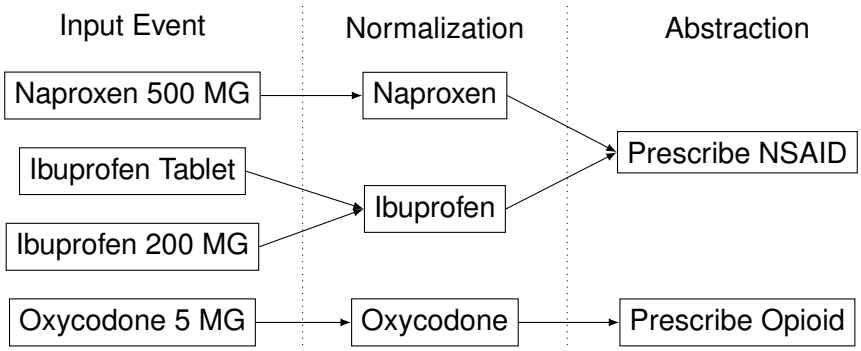


Figure 5. Example event abstraction process to illustrate the normalization and abstraction from the event log excerpt in List. 1 to the event log in List. 2.

can have up to 15 years of follow-up data. We accessed a de-identified dataset containing approximately 8.1 million patients from eight hospitals within the system, spanning the years from 2003 to 2018. This research was approved by the Institutional Review Board (IRB)[2] of the institution fully compliant with the HIPAA[3] regulations. In particular, in this study we included diagnosis codes (ICD-9/10), medications and procedures.

In order to be able to apply the method to the available data, preprocessing steps are necessary. Those steps include the normalization of event descriptors. This is necessary to not encode dosage information and other factors, which would otherwise lead to highly complex event hierarchies, without providing any additional value for the process analysis. The normalization is part of our prototypical implementation and is based on manually crafted rules derived by a medical expert. The normalization of the event log excerpt given in List. 1 is illustrated in Fig. 5, while the normalized event descriptors can be found in List. 2 attached as attributes. After the normalization, the iterative abstraction is applied to the low-level event log as described in Sect. 4.

```
<event>
    <date key=''time:timestamp'' value=''2018-08-04T13:04''/>
    <string key=''concept:name'' value=''Prescribe NSAID''/>
    <string key=''event:description'' value=''NAPROXEN 500 MG TABLET''/>
    <string key=''event:normalized'' value=''Naproxen''/>
    <string key=''event:context'' value=''EPIC MEDICATION''/>
    <string key=''event:code'' value=''19918''/>
</event>
<event>
    <date key=''time:timestamp'' value=''2018-09-01T13:59''/>
    <string key=''concept:name'' value=''Prescribe Opioid''/>
    <string key=''event:description'' value=''OXYCODONE 5 MG TABLET''/>
    <string key=''event:normalized'' value=''Oxycodone''/>
    <string key=''event:context'' value=''EPIC MEDICATION''/>
    <string key=''event:code'' value=''20627''/>
</event>
```

Listing 2. Example event log extract after the application of the abstraction method.

5.2 Evaluation

To demonstrate the usefulness of the proposed method, we created and analyzed two event logs using a prototypical implementation[4]. A detailed description of the event log generation pipeline can be found in [16]. Each log comprises 2,000 patients diagnosed with low back pain from the Mount Sinai Health System with a focus on the treatment processes. When creating the first event log, no abstraction was applied, and all events were included with the granularity

[2] IRB-19-02369

[3] Health Insurance Portability and Accountability Act

[4] https://github.com/bptlab/fiber2xes

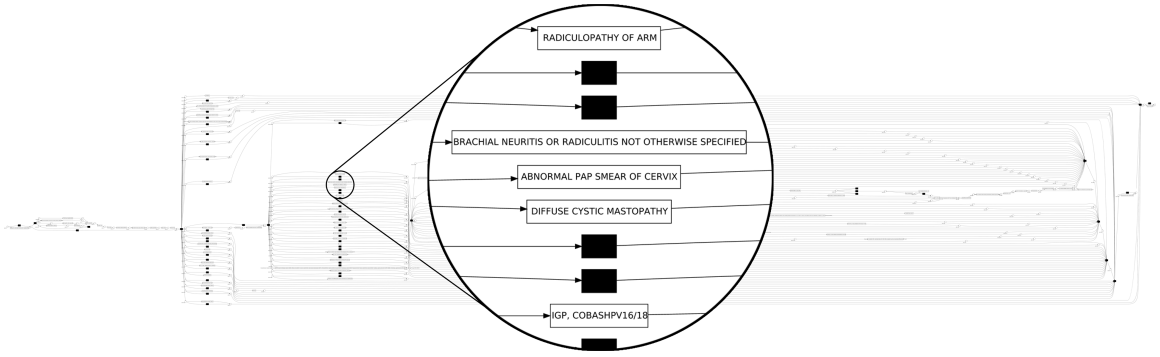


Figure 6. Discovered process model based on only three randomly selected traces from the event log without any abstraction and 60% noise filter.

as they were recorded in the data warehouse. To mine process models form the different event logs, we used the inductive miner plug-in [18] of ProM[5].

Fig. 6 shows the resulting process model, involving only three randomly selected traces of the event log without any abstraction. A visual examination of the diagram is very complex because of the large number of nodes and edges. A comparison to the model representing the clinical guideline in Fig. 1 is hardly possible. This complexity also complicates a variant analysis. The event log without any abstraction contains 1,998 trace variants and 28,540 different event classes. In contrast, Fig. 7 depicts the process model discovered from an event log, where our method was applied during its generation. Apparently, this model is much more compact. The abstracted event log contains 1,300 variants and the seven event classes that were previously defined in the event hierarchy. With this event log, the activities can be directly related to those present in the clinical guideline in Fig. 1. However, the abstracted event log still contains a high number of trace variants, which reflects the flexible nature of treatment processes. In summary, the event abstraction has resulted in a simpler process model that facilitates both interpretation and alignment with the clinical guideline.

5.3 Discussion

The combination of top-down and bottom-up concepts takes advantage of existing structured knowledge and outlines a clear procedure to produce sound abstractions with regard to the underlying domain that lead to semantically meaningful events. In other words, it enables the domain-specific event abstraction of event data. While this method requires manual work and input of a domain expert, the resulting abstractions have been shown capable of increasing the understandability and interpretability of the event logs. Abstractions produced by applying the presented method are comprehensible - there is no “black box” leading to abstracted events, and domain experts are always able to verify the validity of the abstraction. Furthermore, the method makes it possible to easily vary the level of abstraction when investigating a given use case. Also, resulting abstractions are highly flexible and configurable. In contrast to Leonardi et al. [14], no custom ontology needs to be defined, since the abstraction sets are directly derived from existing knowledge sources, like existing standards. This also increases their reusability in other event hierarchies as they have a certain general validity. The presented method could also be combined with existing abstraction techniques discussed in Sect. 3, such as the work of Mannhardt et al. [12], in order to add further perspectives to the resulting event log. They could be used, for example, to add a temporal perspective to the abstraction, where subsequent steps of a larger treatment process, are abstracted into a single event. Furthermore, this method could be improved by adding support for automating the creation of the event hierarchy, or evaluation of the abstraction results based on different metrics.

[5] http://www.promtools.org/

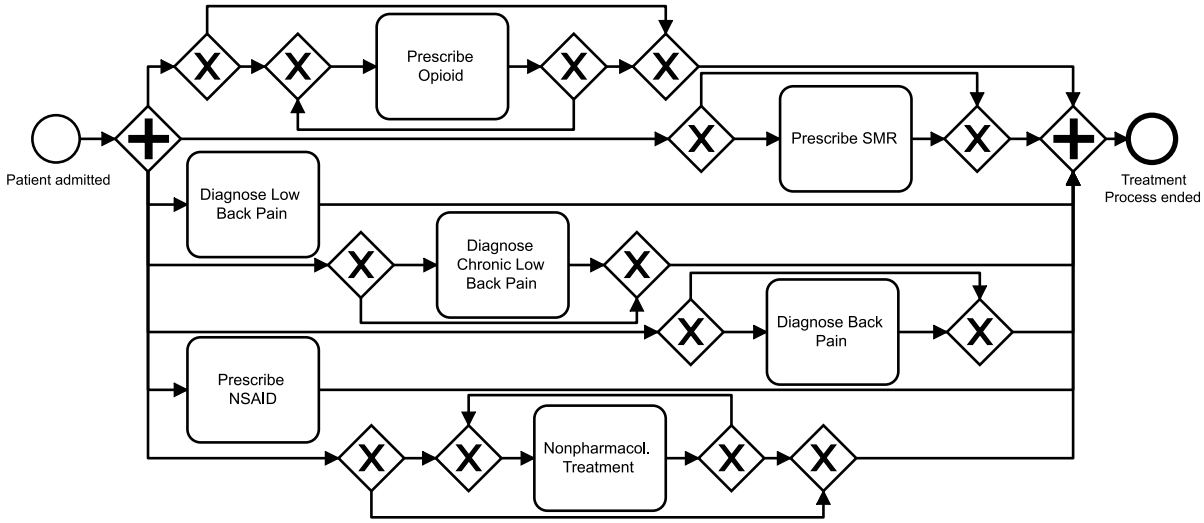


Figure 7. Discovered process model, including all 2,000 patients based on the event log after event abstraction, with 40% noise filter.

6 Conclusion

Different levels of granularity in event data hamper the effective use of process mining techniques. Therefore we proposed a structured method to achieve a semantically meaningful event abstraction. By combining top-down and bottom-up approaches, conceptually sound abstractions from low-level events are reached. Applying the method to a real-world event log from a U.S. health system shows the usefulness of the approach. Currently, we have developed and tested the domain-specific method exclusively for healthcare. Still, the steps are general and might also be useful for other domains, which need to be further studied. For future work, different metrics to evaluate the quality of generated event hierarchies and abstraction sets are worth investigating, as they could help guiding the application of the method through the different iterations. Moreover, interfaces for integrating existing ontologies such as RxNorm or SNOMED CT, as well as a more semi-automated approach for creation of event hierarchies, could reduce the manual effort.

Acknowledgments. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award numbers S10OD026880. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

References

[1] P. Homayounfar, “Process mining challenges in hospital information systems,” in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, 2012, pp. 1135–1140.

[2] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, “Process mining in healthcare: A literature review,” *Journal of Biomedical Informatics*, vol. 61, pp. 224–236, 2016.

[3] S. J. van Zelst, F. Mannhardt, M. de Leoni, and A. Koschmider, “Event abstraction in process mining: Literature review and taxonomy,” *Granular Computing*, pp. 1–18, 2020.

[4] S. Sadeghianasl, A. H. M. ter Hofstede, S. Suriadi, and S. Turkay, “Collaborative and interactive detection and repair of activity labels in process event logs,” in *2nd International Conference on Process Mining, ICPM*, 2020, pp. 41–48.

[5] X. Lu, A. Gal, and H. A. Reijers, “Discovering hierarchical processes using flexible activity trees for event abstraction,” in *2nd International Conference on Process Mining, ICPM*, 2020, pp. 145–152.

[6] K. Diba, K. Batoulis, M. Weidlich, and M. Weske, “Extraction, correlation, and abstraction of event data for process mining,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020.

[7] A. Qaseem, T. J. Wilt, R. M. McLean, and M. A. Forciea, “Noninvasive treatments for acute, subacute, and chronic low back pain: A clinical practice guideline from the american college of physicians,” *Annals of Internal Medicine*, vol. 166, no. 7, pp. 514–530, 2017.

[8] F. Folino, M. Guarascio, and L. Pontieri, “Mining predictive process models out of low-level multidimensional logs,” in *Advanced Information Systems Engineering (CAiSE)*, ser. LNCS, vol. 8484, Springer, 2014, pp. 533–547.

[9] P. H. P. Richetti, F. A. Baião, and F. M. Santoro, “Declarative process mining: Reducing discovered models complexity by pre-processing event logs,” in *Business Process Management BPM*, ser. LNCS, vol. 8659, Springer, 2014, pp. 400–407.

[10] N. Tax, N. Sidorova, R. Haakma, and W. M. P. van der Aalst, “Event abstraction for process mining using supervised learning techniques,” in *Proceedings of SAI Intelligent Systems Conference (IntelliSys)*, ser. LNNS, vol. 15, Springer, 2016.

[11] S. J. J. Leemans, K. Goel, and S. J. van Zelst, “Using multi-level information in hierarchical process mining: Balancing behavioural quality and model complexity,” in *2nd International Conference on Process Mining, ICPM*, 2020, pp. 137–144.

[12] F. Mannhardt, M. de Leoni, H. A. Reijers, W. M. P. van der Aalst, and P. J. Toussaint, “Guided process discovery - A pattern-based approach,” *Information Systems*, vol. 76, pp. 1–18, 2018.

[13] T. Baier, J. Mendling, and M. Weske, “Bridging abstraction layers in process mining,” *Information Systems*, vol. 46, pp. 123–139, 2014.

[14] G. Leonardi, M. Striani, S. Quaglini, A. Cavallini, and S. Montani, “Towards semantic process mining through knowledge-based trace abstraction,” in *International Symposium on Data-Driven Process Discovery and Analysis*, ser. LNBIP, vol. 340, Springer, 2017, pp. 45–64.

[15] R. Mans, W. M. P. van der Aalst, R. J. B. Vanwersch, and A. J. Moleman, “Process mining in healthcare: Data challenges when answering frequently posed questions,” in *Process Support and Knowledge Representation in Health Care - BPM, Revised Selected Papers*, ser. LNCS, vol. 7738, Springer, 2012, pp. 140–153.

[16] S. Remy, L. Pufahl, J.-P. Sachs, E. P. Böttinger, and M. Weske, “Event log generation in a health system: A case study,” in *International Conference on Business Process Management*, ser. LNCS, Springer, 2020.

[17] V. I. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions and Reversals,” *Soviet Physics Doklady*, vol. 10, p. 707, 1966.

[18] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, “Discovering block-structured process models from event logs containing infrequent behaviour,” in *Business Process Management Workshops*, ser. LNBIP, vol. 171, Springer, 2013, pp. 66–78.

24th International Conference on Business Information Systems (BIS 2021)

Knowledge Graphs

https://doi.org/10.52825/bis.v1i.63

Published: 02 July 2021

Ontological Modeling of the State Economic Development Policy for Cultural Industries

Kostiantyn Tkachenko[1][https://orcid.org/0000-0003-0549-3396], Olha Tkachenko[1] [https://orcid.org/0000-0003-1800-618X], Oleksandr Tkachenko[2][https://orcid.org/0000-0001-6911-2770], Mariia Proskurina[3] [https://orcid.org/0000-0002-7701-9784], Iryna Parkhomenko[3][https://orcid.org/0000-0002-8328-6774].

[1] State University of Infrastructure and Technology, Ukraine

[2] National Aviation University, Ukraine

[3] Kyiv University of Culture, Ukraine

Abstract. The article discusses an ontological approach to solving the problem of forming state policy of economic development of cultural and creative industries and the corresponding intellectual-information management systems. The purpose of this article is to develop an effective toolkit (based on ontologies) for making optimal decisions in the field of state regulation of the cultural and creative industries, taking into account the dynamic factors of the external environment. The ontological approach considered in the article assumes the presence of three levels of models: meta-ontology, models of subject areas of cultural and creative industries and models of making appropriate management decisions on the formation of economic development policy of cultural and creative industries. The novelty of the proposed approach lies in the purposeful nature of ontological modeling of such complex system as the state policy of economic development of cultural and creative industries. The system under consideration has certain goals, tasks, resources, processes, factors of influence, risks and other characteristics. These characteristics include, in particular, the structure of the model, the ability to highlight the essential objects of real relations of the considered subject areas, the ability to represent knowledge for the joint work of specialists in computer modeling, the processing of expert knowledge and the generation of management decisions within the framework of the corresponding intellectual-information systems.

Keywords: Ontological modeling, domain, information system, scenario design, cultural and creative industries

Introduction

One of the most important tasks of supporting decision-making (managerial and political) in the field of state policy for the development of cultural industries is the task of determining the set of acceptable alternatives for the implementation of this policy, followed by the possibility of choosing the best solution according to the specified optimality criteria.

There are a lot of models (analytical, optimization) of control processes for the functioning and development of complex systems [1], but for solving the assigned tasks (forming a state policy for the development of cultural industries), their use is ineffective due, first of all, to the scale and complexity of the subject area (domain) under consideration.

The use of the concept "industry" in the context of culture determines the orientation of public policy measures primarily on the economic component of the sector. It is also

advisable to consider cultural and creative industries (CCI) together as a single sector of socio-economic activity. This sector is characterized as multisystem and dynamism of its development, as well as the complexity of interdisciplinary connections.

The state policy for the development of cultural industries involves the management of institutions and the formation of information and cultural space. Mandatory elements of CCI public administration are [2]:

- setting policy priorities;
- sources of financing for the development of CCI;
- management principles (management codex);
- building an optimal and efficient system for managing subindustries, considering the regional specifics.

Among the main directions of the state policy implementation for the development of CCI, it is possible to single out the following:

• creation of an internal market, which forms the information and cultural space of the country;
• development of the creative class (this requires reforms in the education system and labor legislation);
• creation of an appropriate business climate for the development of creative entrepreneurship (creation of optimal conditions for the commercialization of talent and ideas, favorable tax regime, protection of the producer of a national cultural product, etc.).

The economic development of the sector of cultural and creative industries, being an object of state policy, requires the creation of an adequate model for determining the strategy and tactics of subject area development, making appropriate management decisions and carrying out forecast calculations. Such a model should take into account the complexity, multidimensionality, many goals of economic development of the subject area under consideration, factors of influence, etc.

Semantic or ontological models should be used to display the whole complex of CCI problems and their economic development.

The purpose of this study is to develop an effective toolkit (based on ontologies) for making optimal decisions in the field of state regulation of CCI, considering the dynamic factors of the external environment.

The existing restrictive conditions impose rather strict requirements on the generated set of alternatives (options) of management decisions. Because of this, the range of admissible alternatives either deliberately narrows or completely excludes the possibility of making an optimal decision (we can only talk about a partially optimal solution).

In practice, this leads to errors of two types: investing in ineffective projects within the framework of the state policy to support the industry, or refusal from more strategically profitable decisions due to an incorrect planning model.

Even if an optimization model is developed and an optimal solution or a set of feasible solutions is obtained for the problem under consideration, then it should be borne in mind that the conditions under which this model and this solution are obtained can change quite quickly.

That is why planning (management) should be considered as a continuous process that allows to adapt public policy (public administration project) to external and internal changing conditions.

At the same time, it should be considered that not every model and not every solution has a sufficient margin of stability so that they can continue to be used without significant adjustment throughout the entire life cycle of the project or its individual stages.

Management decisions aimed at the economic development of the CCI should be formed for:

- strategies for the economic development of the CCI taking into account external influence;
- strategies for economic development of the CCI taking into account internal conditions;
- creation and promotion of a national cultural product (cultural projects) in domestic and foreign markets;
- resource management (material, financial, legal, personnel, etc.);
- technological support of economic development;
- development of the CCI infrastructure.

For the subject area under consideration, it is proposed that the generation of managerial decisions is a complex informal process.

To form possible options for management decisions, expert knowledge is required on:

- subject area;
- the current and predicted state of its external environment (factors of influence: external and internal, the hierarchy of the goals of hierarchical development);
- the specifics of the management of CCIs, their subsectors, and their economic development.

The problem of finding an expert even for traditional industrial projects is often difficult to solve, and for innovative creative projects, the risks of which are especially high and there is no experience in implementing similar projects, it can become unsolvable.

Experts are not always able to determine all possible scenarios for the development of CCI, and multivariate planning is either not always justified at a sufficient analytical level, or does not have a strategy.

The technology of state policy formation based on the ontological approach can be an addition to the expert knowledge about subject area (domain), the features and possible options for the economic development of subject area [3], [4].

Features of Ontological Modeling of The State CCI Economic Development Policy

The implementation of the state policy of CCI economic development presupposes the presence of a large number of alternative solutions that take into account both the organizational structure of the subject area (the organizational and institutional structure of the state management of culture and CCI) and the scenario of its functioning.

Scenario planning is the most common approach to solving problems at the structural level when planning the economic development of CCI, including in public administration [5].

The scenario for the implementation of public policy in the subject area under consideration is a combination of conditions (external and internal) that lead to certain results, to the efficiency and financial feasibility of specific activities and projects or to the creation of a certain creative product.

Before deciding on the feasibility of implementing certain government measures and defining a strategy for public policy in general, possible scenarios should be investigated. This will make it possible to determine the area of sustainability of the subject area to dynamically changing environmental factors and the prospects for the implementation of the policy of support and development of the CCI.

To describe the subject area, it is necessary to define such components as goals, tasks, activities, results and resources.

Goals are a set (aggregate) of measures of state regulation of the external and internal environment of the CCI (I, intent).

Tasks are a set of actions (procedures, works and activities) that must be carried out in order to achieve the goals of public policy in the CCI (T, task).

Works (operations) are a set of processes aimed at solving problems and obtaining results, requiring the necessary investment of time and resources (O, operation).

Results are a set of political and socio-economic events that embody the goals of politics - these are decisions, creation of institutional structures, implementation of targeted programs, adoption of laws (Rt, result).

Resources are the set of objects needed to get work done (Res). The necessary resources are allocated for the execution of each work.

Thus, the considered domain (D), without taking into account the influence factors, can be represented as:

$$D = (I, T, O, Rt, Res) \quad (1)$$

Since the state policy is implemented in a certain environment, called the CCI environment, in addition to describing the domain (see (1)), the scenario includes a description of the external conditions under which the generation of managerial decisions and their implementation is carried out.

The external conditions of implementation (factors of influence) can change quite dynamically throughout the entire life cycle of a complex system, which is the state policy in general, and the policy in the field of CCI, in particular. These factors are sources of possible risks of the implementation of an effective policy to support the economic development of the CCI and have a direct impact on the process of forming possible options for the corresponding management decisions.

Thus, the considered domain (D) taking into account the external factor of influence (Fe) and internal factor of influence (Fi) can be represented as:

$$D = (I, T, O, Rt, Res, Fe, Fi) \quad (2)$$

When determining the subject area development scenario, all of the above information must be taken into account. As a rule, such information is an unformalized or weakly formalized description, and the complexity of scenario compilation is associated with the need to consider the development of the subject area internal processes and their not always obvious connections with the subject area environment factors and among themselves.

The use of an ontological approach will make it possible to automate the generation of public policy strategies that allow achieving goals taking into account the most important and probable risks. This contributes to a significant increase in objectivity and optimization of the management decision-making process.

Before proceeding with the description of the ontology, it is necessary to determine for what purpose it is being created and what tasks it will solve.

In this work, ontological modeling is used to develop a variety of scenarios for the economic development of the CCI. The proposed ontology should ensure the completeness of coverage of the feasible solutions area to achieve the goals of state policy in the field of culture.

The implementation of the state policy of CCI economic development is a complex process, which is influenced by many factors; therefore, the formalization of the structural level of multivariate calculations will improve the quality of effective management decisions.

Decisions made at the planning level should be formed with taking into account the possible impact of negative and positive risks of various nature and varying degrees of controllability.

Implementation of the optimal state policy in the CCI will minimize the consequences or avoid potential risks for the economic entities of the sector. Among such risks, one should, in particular, highlight:

- geopolitical (threat to the internal information space, cultural intervention, convergence);
- social (creation of semi-public goods, polarization of public opinion, discrimination, lack of qualified personnel);
- technical and technological (lack of modern equipment for the production and consumption of a cultural product, lack of infrastructure, etc.);
- economic (growth in costs for the production of a cultural product, economic instability, inflation, a decrease in consumer solvency, unpredictability of consumer behavior, export-import strategy, tax burden, etc.);
- market (falling product prices, piracy, decrease in sales volumes, increased competition, shadow market, etc.);
- financial (high costs for the production of prototypes, lack of funding for innovative projects, lack of credit programs for startups, high level of financial investment for some industries in the sector, etc.);
- production (quality of a cultural product, creation of a standardized cultural product, copyright protection, etc.);
- political (making ineffective political decisions, lobbying interests, a threat to the decentralization process, lack of specialized legislation in the sector, insufficient level of ensuring the cultural rights of citizens, etc.).

The typification of risks by their levels of manageability (fully manageable, partially managed and unmanaged) contributes to:

- setting priorities in risk management;
- reducing risks in the development of production and financial strategies for the implementation of state policy in the field of CCI;
- the use of information technology and intelligent information systems for making management decisions.

Thus, the toolkit for generating scenarios for the implementation of state policy in the field of culture should have knowledge about:

- available implementation strategies not only at the level of a predefined set of actions;
- the effectiveness of the strategy implementation in accordance with reality;
- reactions to unknown situations (risks, challenges, opportunities) that may arise during the life cycle of the implementation of public policy.

When using the ontological approach, it is important to separate the ontology of the domain from the ontology of problems solved in this domain [6], [7], [8].

This makes it possible to more conveniently describe the dynamic processes of the problems being solved on the basis of static data and knowledge of the domain. Therefore, to ensure the subsequent intellectualization of the generation of options (alternatives) of management decisions, it is advisable to use:

- ontological model of the domain;
- ontological model of the process of constructing an appropriate scenario for the implementation of state policy in the field of culture;
- an ontological model of the process of constructing an appropriate scenario for the economic development of the cultural sphere.

The ontological model proposed by the authors can be represented as $M = \langle V, C, K, L, A_l \rangle$, where:

V – a set of nodes (primary elements, terms of domain);

C – a set of connecting elements of ontological model (of the corresponding ontograph), each of which defines a certain fragment of ontological model (of the ontograph);

K is the set of key vertices of the ontograph, each of which defines a certain class of equivalent (of the same type) elements of the ontograph ($K \subset V$);

L is a set of labels of elements (alphabet of elements) of the ontograph, each of which specifies a certain base class of equivalent elements of the ontograph.

Such classes of elements of ontological model include following classes:

- nodes of the ontograph,
- connecting elements of the ontograph,
- key vertices of the ontograph,
- ontograph labels,
- incidence relations defined on the set of elements of the ontograph.

A_I is a set of incidence relations that are defined on a set of ontograph elements. All incidence relationships are binary oriented relationships.

To implement each ontological model, it is necessary to select directly the entities and the relationships between them (knowledge base), as well as the interpretation functions defined for this knowledge base.

The interpretation functions will make it possible to use the created knowledge base for solving various problems of forming the state policy of economic development of the CCI, presented in metaontology and ontologies of a lower level.

The complex of ontological models used to generate management decisions in the formation of the state policy of economic development of the CCI is presented in the form of a semantic network of ontologies (ontograph) [9].

The ontograph reflects the relationship between the individual components of the cultural industry (its structural and organizational elements, etc.) and is intended for:

- choice of management decisions;
- establishing links between components at the same level and at different levels.

The vertices in the ontograph are various models of the subject area CCI, and the edges are the relations of detailing and decomposition of these models (Figure 1).

The complex of models includes:

- metaontology, as a model of the upper level of generalization;
- ontological model of the domain (cultural and creative industries);
- an applied ontology of the formation procedure (generation) and adoption of managerial decisions (strategic and tactical) in the management systems of CCI efficiency.

Metaontology is used as a tool for integrating various models of cultural and creative industries and the most general description of the state policy of their economic development. The domain ontological model details some of the top-level metaontology concepts.

The applied ontology of the procedure for the formation and adoption of managerial decisions is intended for the design of appropriate information systems based on ontologies [10]. The complex of models can be extended by various models of the subject area. The model of the subject area - cultural and creative industries - is shown in Figure 1.

This model displays the main directions of the implementation of state policy in the context of the development of the creative economy.

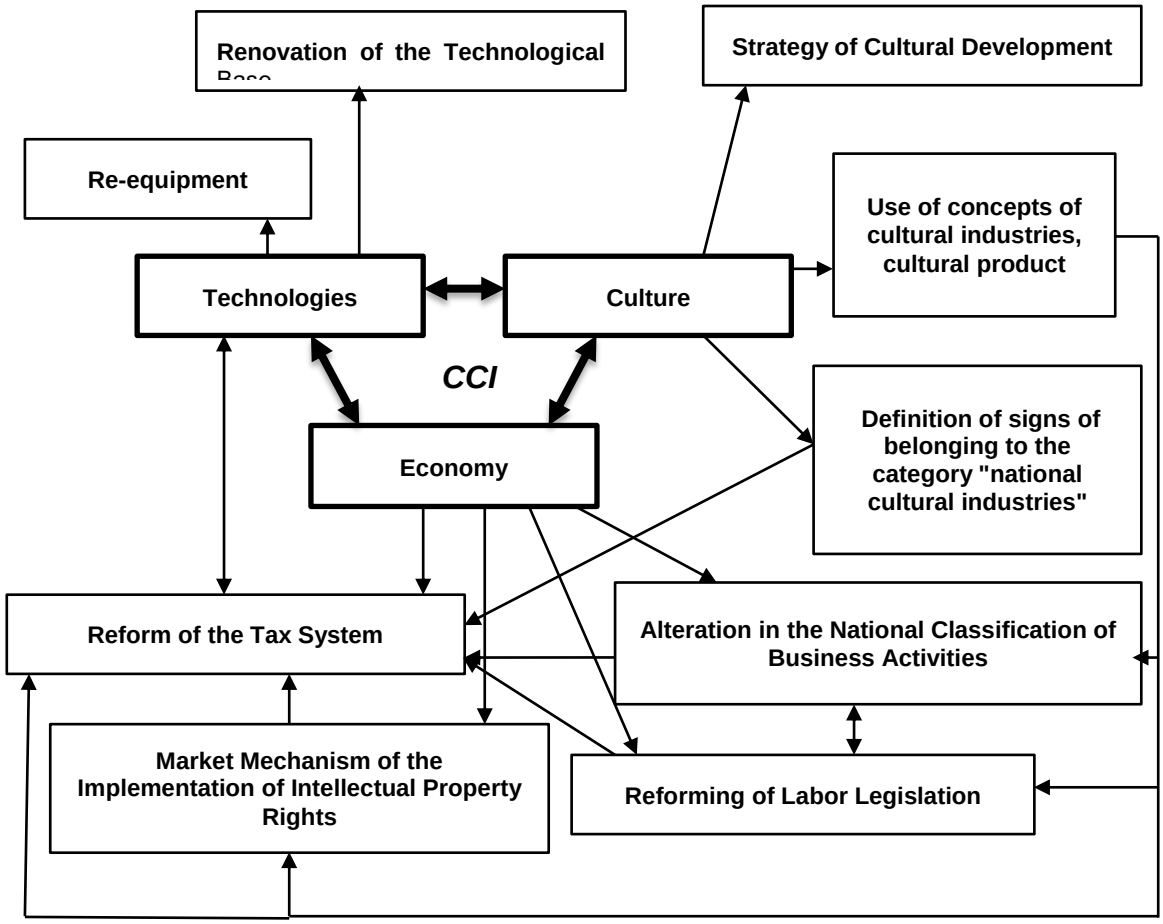


Figure 1. Model of CCI

The ontological model of domain is a formalized description of knowledge about the goals, tasks, works, results and resources of all their cultural projects.

The ontological model is the semantic basis for creating information support for the processes of forming the state policy of economic development of the CCI.

The purpose of the developed ontological model of domain in the context of solving the problem of forming scenarios of state policy in the field of CCI is to describe the events of the scenario, drawing up a plan for the economic development of CCI, then when constructing an ontological model, the concepts are highlighted based on their influence on the formation of positive and negative financial flows generated by a cultural project.

The questionnaire survey of the Ukrainian CCI stakeholders (more than 50 representatives of SSH (show business, cinema and theater figures, heads of avent agencies, creative directors of advertising companies, etc.)) made it possible to form a list of significant factors for the implementation of professional activities in the sector and the development of CCI state policy.

A fragment of the ontograph of the public policy generalized model for the development of the CCI with an indication of the main concepts for the classes Management, Development and Policy_factures is built in the Protege environment [11], [12] and is shown in Figure 2.

Figure 2. Ontograph for the State_policy class.

A more detailed description of the ontograph for the Development class is shown in Figure 3.

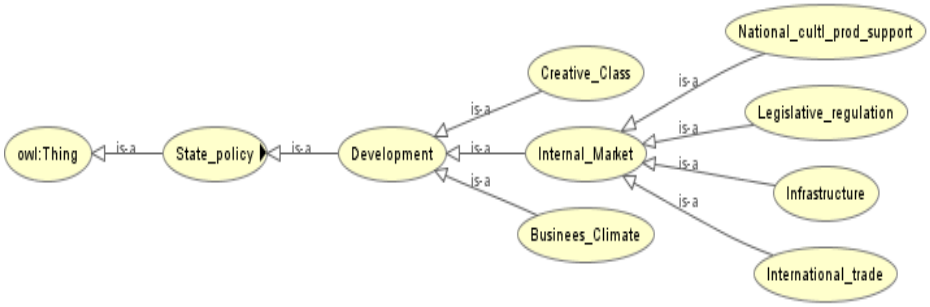


Figure 3. Ontograph for the Development class.

The detailed description of the ontograph for the Culture_Management class is shown in Figure 4.

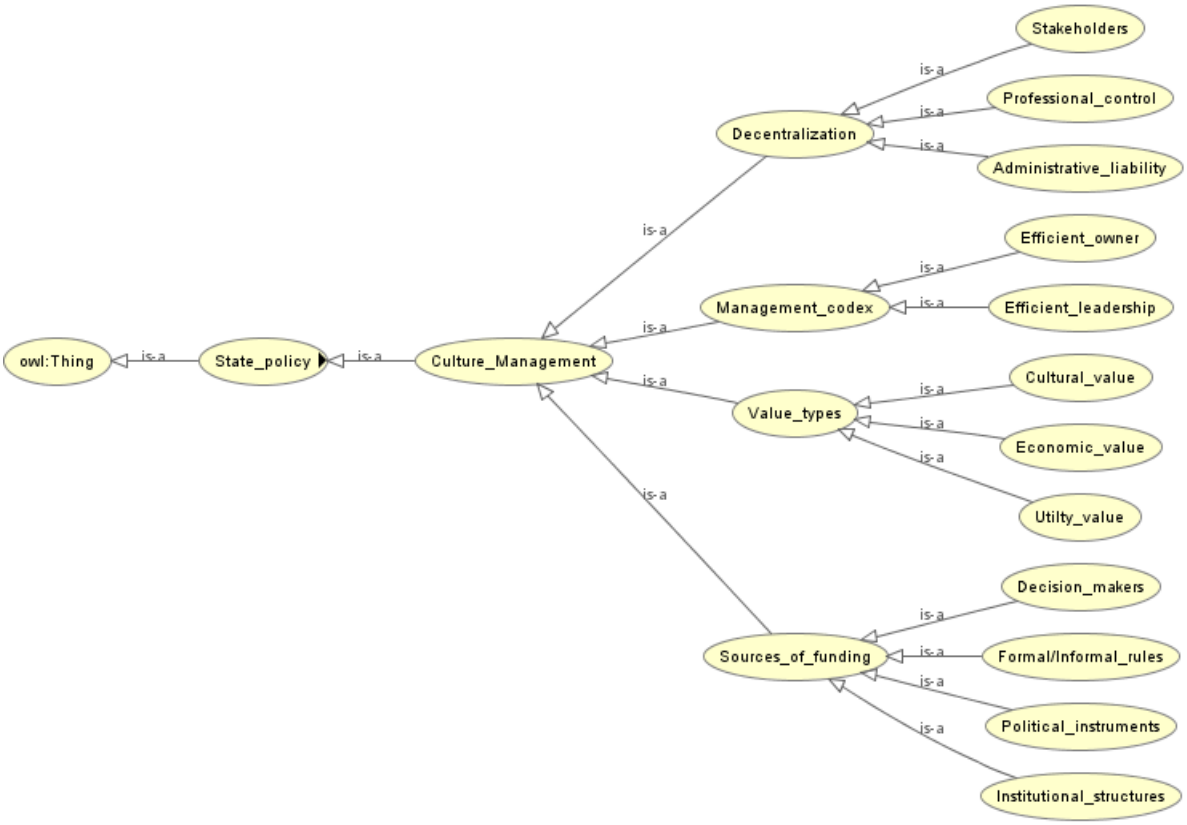


Figure 4. Ontograph for the Culture_Management class.

The structure of the knowledge model based on the proposed ontological approach to modeling should ensure the relationship between:

- alternative strategies for the formation and implementation of the state policy of economic development of the CCI;
- the most possible and probable risks, the negative consequences of which can be minimized through the implementation of a particular strategy;
- characteristics and features of the formation (generation) of control decisions at the current stage of the life cycle of the CCI project.

The strategic scenario of the state policy for the economic development of the CCI can be represented by a complex of optimal management decisions in the production, marketing and financial areas of the CCI project, taking into account the acceptable level of risk, the desired results, the degree of attainability of goals, and the necessary resources.

Formation of Scenarios of the CCI Development

Creation of scenarios for the formation of the state economic policy of the CCI development takes into account the planning goals and requires the replenishment of the knowledge base with the necessary information to generate appropriate management decisions.

The degree of scenario detailing directly affects the detail of the subsequent financial calculations for the CCI project and the choice of tools for quantitative substantiation of management decisions.

When formulating scenarios for the state policy of CCI economic development at the initial stages of the life cycle, the following is considered:

- the general direction of the formation (planning) of the CCI development;
- duration of the CCI project investment stage;
- the duration of the practical implementation of management decisions in the CCI;

- the term for obtaining results from the implementation of the adopted decisions.

Generation of management decisions that lead to adaptive transformation of scenarios uses:

- a formalized description of scenarios for the formation of the state policy of economic development of the CCI;
- knowledge about the subject area and its environment, accumulated in the knowledge base of the corresponding intelligent information system;
- reasonable forecasts of the influence of potential risk factors (external and internal).

The proposed approach to the formation of the state policy of the CCI economic development on the basis of ontological modeling is built on the following principles.

- formation of scenarios (their decomposition and synthesis) is performed with taking into account the following:
  - the logic of the implementation processes of state support for the development of CCI;
  - political, managerial and other aspects of the CCI project based on the corresponding ontological model;
- variety of scenarios should be based on observed and / or predicted cultural trends,
- content of the scenarios varies depending on the stage of implementation of the state policy for the development of the CCI.

The proposed technology, on the one hand, forms a framework for full discrete coverage of the area of permissible management decisions, and on the other hand it allows the scenarios to be flexibly filled with events with taking into account the required planning detail, as well as taking into account the assumptions of the CCI project authors and the decision maker (in their role can be customers and the state).

The use of ontological modeling is effective when negotiating between stakeholders and participants in the CCI project, for example:

- project manager - resource providers (in the formation of managerial decisions and conditions for cooperation to ensure strategies for the state policy of economic development of the CCI);
- project manager - potential producers of a cultural product, information agencies (in the formation of management decisions and conditions for cooperation to ensure information strategies);
- project manager - potential investors, patrons, sponsors (when forming managerial decisions and terms of cooperation to ensure project financing).

The work did not provide for the use of Wikidata, which may not always fully reflect some specific features of CCI.

Conclusion

The unified approach to the generation of management decisions and the corresponding intelligent information system has been developed on the basis of ontological modeling of the processes that form the state policy of CCI economic development.

Using ontological modeling and an appropriate knowledge base, it is possible to generate management decisions at any stage of the life cycle of a cultural project.

Ontographs of various levels of generalization are formed in ontological modeling - metaontology, domain ontology and applied ontology for generating the corresponding management decisions.

Metaontology is based on classes of concepts, a system of goals and strategies, organizational and functional structures, a system of criteria for the effectiveness of state policy of economic development of the CCI.

The ontological model of the subject area can be applied at such stages as the development and formalization of the state policy strategy of the CCI economic development, the formation of the corresponding scenarios for the CCI functioning and development.

References

1. M.R. Kogalovsky, L.A. Kalinichenko, ”Conceptual and ontological modeling in information systems”, *Programming and Computer Software*, vol. 35, pp. 241–256, 2009, doi: https://doi.org/10.1134/S0361768809050016
2. A. C. Pratt, T. Hutton, ”Reconceptualising the relationship between the creative economy and the recession: learning from the financial crisis. Cities: Asian-European Perspectives”, Heidelberg: Springer. 2013, pp. 86–95, doi: https://doi.org/10.1016/j.cities.2012.05.008.
3. F. Neuhaus, Otto von-Guericke, S. Ray, Ram D. Sriram, ”Toward Ontology Evaluation across the Life Cycle”. nvlpubs.nist.gov. https://nvlpubs.nist.gov/nistpubs/ir/2014/NIST.IR.8008.pdf (accessed Feb.1, 2021). doi: 10.6028/NIST.IR.8008
4. E.M. Sanfilippo, ”Feature-based product modelling: an ontological approach”, *International Journal of Computer Integrated Manufacturing*, vol. 31(11), pp. 1097–1110, 2018, doi: https://doi.org/10.1080/0951192X.2018.1497814.
5. C. List, ”Levels: descriptive, explanatory, and ontological”. lse.ac.uk. http://eprints.lse.ac.uk/87591/1/List_Levels%20descriptive_2018.pdf (accessed Feb.1, 2021), doi: https://doi.org/10.1111/nous.12241
6. R. Ali, D. Luther, “Scenario Planning: Strategy, Steps and Practical Examples”. netsuite.com. https://www.netsuite.com/portal/business-benchmark-brainyard/industries/articles/cfo-central/scenario-planning.shtml (accessed Feb.1, 2021).
7. K. Munir, “The use of ontologies for effective knowledge modelling and information retrieval“. *Applied Computing and Informatics*, 14(2), pp. 116-126, 2018. https://doi.org/10.1016/j.aci.2017.07.003.
8. O. Tkachenko, A. Tkachenko, K. Tkachenko, ”Ontological Modeling of Situational Management”, *Digital platform: information technology in the sociocultural area*, vol. 3. № 1, pp. 22-32, 2020, doi: 10.31866/2617-796x.3.1.2020.206096.
9. A. Gelfert, “The Ontology of Models”, in Magnani L., Bertolotti T. (eds*) Springer Handbook of Model-Based Science*, Springer Handbooks, Springer, Cham, 2017.
10. S.E. Greger, S.V. Porshnev, “Building an ontology of information system architecture”. *Fundamental Research*, № 10, pp. 2405-2409, 2013.
11. ”Protégé 5.5”. protege.stanford.edu. https://protege.stanford.edu/products.php accessed Feb.1, 2021).
12. M.A. Musen. “The Protégé project: A look back and a look forward AI Matters”. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), 2015, 10.1145/2757001.2757003.

24th International Conference on Business Information Systems (BIS 2021)
Knowledge Graphs
https://doi.org/10.52825/bis.v1i.64

Published: 02 July 2021

Semantic Representation of Domain Knowledge for Professional VR Training

Jakub Flotyński[1], Paweł Sobociński[1], Sergiusz Strykowski[1], Dominik Strugała[1], Paweł Buń[2], Filip Górski[2], Krzysztof Walczak[1]

[1]Poznań University of Economics and Business
Niepodległości 10, 61-875 Poznań, Poland
e-mail: flotynski@kti.ue.poznan.pl
[2]Poznań University of Technology
Piotrowo 3, 60-965 Poznań, Poland

Abstract. Domain-specific knowledge representation is an essential element of efficient management of professional training. Formal and powerful knowledge representation for training systems can be built upon the semantic web standards, which enable reasoning and complex queries against the content. Virtual reality training is currently used in multiple domains, in particular, if the activities are potentially dangerous for the trainees or require advanced skills or expensive equipment. However, the available methods and tools for creating VR training systems do not use knowledge representation. Therefore, creation, modification and management of training scenarios is problematic for domain experts without expertise in programming and computer graphics. In this paper, we propose an approach to creating semantic virtual training scenarios, in which users' activities, mistakes as well as equipment and its possible errors are represented using domain knowledge understandable to domain experts. We have verified the approach by developing a user-friendly editor of VR training scenarios for electrical operators of high-voltage installations.

Keywords: semantic web, ontologies, virtual reality, training, scenarios

1 Introduction

Progress in the quality and performance of graphics hardware and software observed in recent years makes realistic interactive presentation of complex virtual spaces and objects possible even on commodity hardware. The availability of diverse inexpensive presentation and interaction devices, such as glasses, headsets, haptic interfaces, motion tracking and capture systems, further contributes to the increasing applicability of virtual (VR) and augmented reality (AR) technologies. VR/AR applications have become popular in various application domains, such as e-commerce, tourism, education and training. Especially in training, VR offers significant advantages by making the training process more efficient and flexible, reducing the costs, liberating users from acquiring specialized equipment, and eliminating risks associated with training in a physical environment.

Training staff in virtual reality is becoming widespread in various industrial sectors, such as production, mining, gas and energy. However, building useful VR training environments requires competencies in both programming and 3D modeling, as well as domain knowledge, which is necessary to prepare practical applications in a given domain. Therefore, this process typically

involves IT specialists and domain specialists, whose knowledge and skills in programming and 3D modeling are usually low. Particularly challenging is the design of training scenarios, as it typically requires advanced programming skills, and the level of code reuse in this process is low. High-level componentization approaches commonly used in today's content creation tools are insufficient because the required generality and versatility of these tools inevitably leads to high complexity of the content design process. Therefore, the availability of user-friendly tools for domain experts to design VR training scenarios using domain knowledge becomes essential to reduce the required time and effort, and consequently promote the use of VR in training.

A number of solutions enabling efficient modeling of VR content using techniques for domain knowledge representation have been proposed in previous works. In particular, the semantic web provides standardized mechanisms to describe the meaning of any content in a way understandable to both users and software. The semantic web is based on description logics, which permit formal representation of concepts, roles and individuals. Such representations can be subject to reasoning, which leads to the inference of implicit knowledge based on explicit knowledge, as well as queries including arbitrarily complex conditions. These are significant advantages for the creation and management of content by users in different domains. However, usage of the semantic web requires skills in knowledge engineering, which is not acceptable in the practical preparation of VR training. Thus, the challenge is to elaborate a method of creating and managing semantic VR scenarios, which could be employed by users who do not have advanced knowledge and skills in programming, 3D modeling and knowledge engineering.

In this paper, we propose a new method of building and managing VR training scenarios based on semantic modeling techniques with a user-friendly editor. The editor enables domain experts to design scenarios in an intuitive visual way using domain knowledge described by ontologies. Our approach takes advantage of the fact that in concrete training scenes and typical training scenarios, the variety of 3D objects and actions is limited. Therefore, it becomes possible to use ontologies to describe available training objects and actions, and configure them into complex scenarios based on domain knowledge.

The work described in this paper has been performed within a project aiming at the development of a flexible VR training system for electrical operators. All examples, therefore, relate to this application domain. However, the developed method and tools can be similarly applied to other domains, provided that relevant 3D objects and actions can be identified and semantically described.

The remainder of this paper is structured as follows. Section 2 provides an overview of the current state of the art in VR training applications and a review of approaches to semantic modeling of VR content. Section 3 describes an ontology of training scenarios. The proposed method of modeling training scenarios is described in Section 4. An example of a VR training scenario along with a discussion of the results is presented in Section 5. Finally, Section 6 concludes the paper and indicates possible future research.

2 Related Work

2.1 Training in VR

VR training systems enable achieving a new quality in employee training. With the use of VR, it becomes possible to digitally recreate real working conditions with a high level of fidelity. Currently available systems can be categorized into three main groups: desktop systems, semi-immersive systems and fully immersive systems. Desktop systems use mainly traditional presentation and interaction devices, such as a monitor, mouse and keyboard. Semi-immersive systems use advanced VR/AR devices for presentation, e.g., head-mounted displays (HMD), and interaction, e.g., motion tracking. Immersive systems use advanced VR/AR devices for

both presentation and interaction. Below, examples of VR training systems within all of the three categories are presented.

The ALEn3D system is a desktop system developed for the energy sector [1]. The system enables interaction with 3D content displayed on a 2D monitor screen, using a mouse and a keyboard. Scenarios implemented in the system mainly focus on training the operation of power lines and consist of actions performed by line electricians. The system includes two modules: a VR environment and a course manager. The VR environment can operate in three modes: virtual catalog, learning and evaluation. The course manager is a browser application that allows trainers to create courses, register students, create theoretical tests and monitor learning progress.

An example of a semi-immersive system is the IMA-VR system [2]. It enables specialized training in a virtual environment aimed at transferring motor and cognitive skills related to the assembly and maintenance of industrial equipment. The specially designed IMA-VR hardware platform is used to work with the system. The platform consists of a screen and a haptic device. This device allows a trainee to interact and manipulate virtual training scenes. The system records accomplished tasks and statistics, e.g., time, required assistance, errors made and correct steps.

An example of a fully immersive AR system is the training system for repairing electrical switchboards developed by Schneider Electric in cooperation with MW PowerLab [3]. The system is used for training in operation on electrical switchboards and replacement of their parts. The system uses the Microsoft HoloLens HMD. After a user puts on the HMD, the system scans the surroundings for an electrical switchboard. The system can work in two ways: providing tips on a specific problem to be solved or providing general tips on operating or repairing the switchboard.

2.2 Semantic modeling of VR content

A number of works have been devoted to ontology-based representation of 3D content, including a variety of geometrical, structural, spatial and presentational elements. A comprehensive review of the approaches has been presented in [4]. Existing methods are summarized in Table 1. Five of the methods address the low (graphics-specific) abstraction level, while six methods address a high (general or domain-specific) abstraction level. Three of those methods are general—may be used with different domain ontologies. For the methods that address a high abstraction level in specific application domains, the domains are indicated.

Table 1. Comparison of semantic 3D content modeling methods

Approach	**Level of Abstraction**	
	Low (3D graphics)	High (application domain)
De Troyer et al. [5]–[9]	✓	general
Gutiérrez et al. [10], [11]	✓	humanoids
Kalogerakis et al. [12]	✓	-
Spagnuolo et al. [13]–[15]	-	humanoids
Floriani et al. [16], [17]	✓	-
Kapahnke et al. [18]	-	general
Albrecht et al. [19]	-	interior design
Latoschik et al. [20]–[22]	-	general
Drap et al. [23]	-	archaeology
Trellet et al. [24], [25]	-	molecules
Perez-Gallardo et al. [26]	✓	-

The presented review indicates that there is a lack of a generic method that could be used for creating interactive VR training scenarios in different application domains. The existing

ontologies are either 3D-specific (with focus on static 3D content properties) or domain-specific (with focus on a single application domain). They lack domain-independent conceptualization of actions and interactions, which could be used by non-technical users in different domains to generate VR applications with limited help from graphics designers and programmers. In turn, the solutions focused on 3D content behavior, such as [27], [28], do not provide concepts and roles for representation of training scenarios.

3 Ontological Representation of VR Training Scenarios

A *scenario ontology* has been designed to enable semantic representation of VR training scenarios. The scenario ontology consists of a TBox and an RBox. The TBox is a specification of classes (concepts) used to describe training scenarios. The RBox is a specification of properties (roles) of instances (individuals) of the classes. A particular training scenario is an ABox including instances of TBox classes described by RBox properties. The scenario ontology and particular training scenarios are separate documents implemented using the RDF, RDFS and OWL standards. RDF is the data model for the ontology and scenarios. In turn, RDFS and OWL provide vocabularies, which enable expression of such relations as concept and role inclusion and equivalence, role disjointedness, individual equality and inequality, and negated role membership.

The entities specified in the scenario ontology as well as the relations between them are depicted in Fig. 1. The entities encompass classes (rectangles) and properties (arrows) that fall into three categories describing: the workflow of training scenarios, objects and elements of the infrastructure, and equipment necessary to execute actions on the infrastructure.

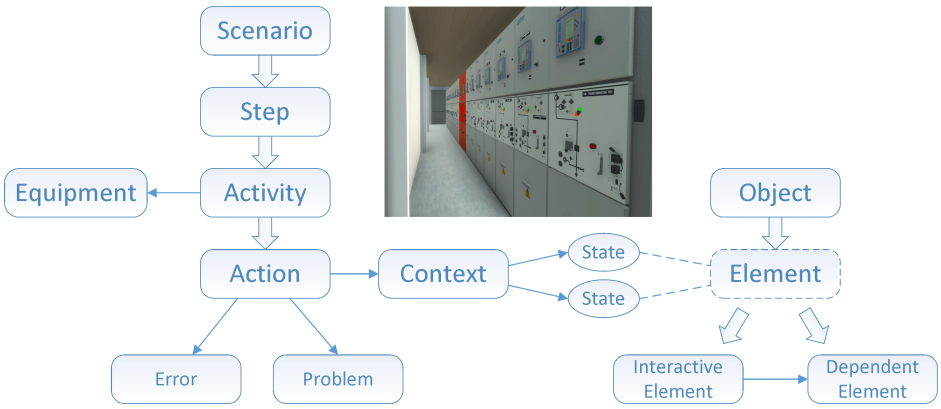


Figure 1. Ontology of VR training scenarios

Every *scenario* is represented by an individual of the *Scenario* class. A scenario consists of at least one *Step*, which is the basic element of the workflow, which consists of at least one *Activity*. Steps and activities correspond to two levels of generalization of the tasks to be completed by training participants. Activities specify equipment required when performing the works. In the VR training environment, it can be presented as a toolkit, from which the user can select the necessary tools. Steps and activities may also specify protective equipment. *Actions*, which are grouped into activities, specify particular indivisible tasks completed using the equipment specified for the activity. Actions are executed on infrastructural components of two categories: *Objects* and *Elements*, which form two-level hierarchies. A technician, who executes an action, changes the *State* of an object's element (called *Interactive Element*), which may affect elements of this or other objects (called *Dependent Elements*). For example, a control panel of a dashboard is used to switch on and off a transformer, which is announced on the panel and influences the infrastructure. N-ary relations between different entities in a scenario are represented by individuals of the *Context* class, e.g., associated actions, elements and states. Non-typical situations in the workflow are modeled using *Errors* and *Problems*.

While errors are due to the user, e.g., a skipped action on a controller, problems are due to the infrastructure, e.g., a controller's failure.

4 Designing VR Training Scenarios

The concept of the method of modeling VR training scenarios is depicted in Fig. 2. The method consists of two main stages, which are accomplished using two modules of the editor we have developed. At the first stage, electricians who directly train new specialists provide primary information about scenarios using the Scenario Editor tool. At the second stage, the information collected from the first stage is used by the managers of technical teams to refine, manage and provide scenarios in their final form using the Semantic Scenario Manager. Next, the final scenarios are used to train specialists with the VR application.

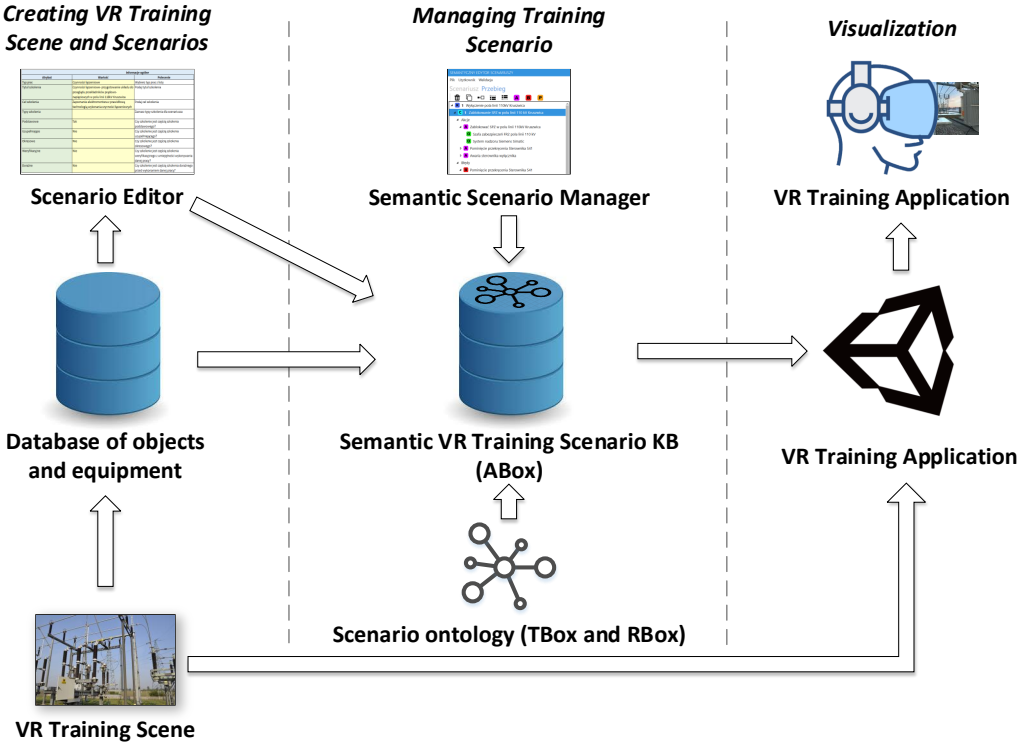


Figure 2. Knowledge base driven design of VR training scenarios

4.1 Scenario Editor

The Scenario Editor is a visual tool based on MS Excel. Its main goal is to enable efficient and user-friendly collection of data about training scenarios by electricians who directly work with trainees and the high-voltage installations.

Scenarios are stored as Excel files based on a specific scenario template. A single scenario is represented by several worksheets, each worksheet contains numerous rows with data. Data in a row is organized in a pair <attribute, value>. Rows containing data relating to the same topic are grouped into sections, where each section is identified by a header. The Scenario Exporter has been implemented as an Excel extension using C# programming language. Its class diagram is presented in Fig. 3.

The *OntologyStore* class is responsible for managing mappings between scenario content (scenario sections and rows within the sections) and elements of the scenario ontology (classes and properties). The mappings are stored in a template file—the same file which is used by the Scenario Editor. While instantiating, the *OntologyStore* class parses the template file and builds in-memory object-oriented representation of the mappings.

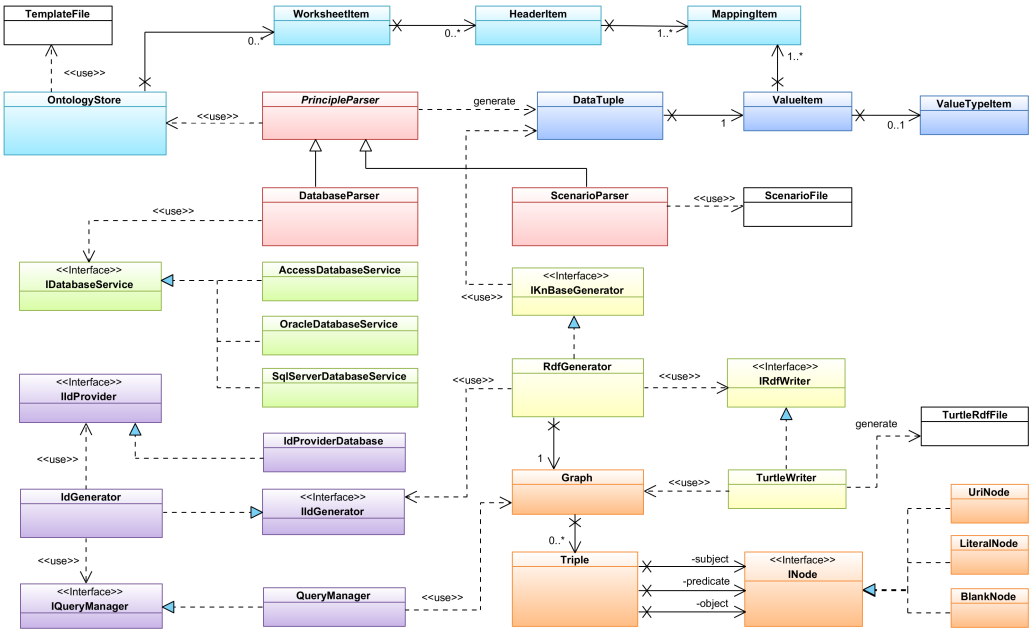


Figure 3. The Scenario Exporter class diagram

Each row in the template file is described by the corresponding mapping unit(s). A single mapping unit consists of three entities: Class, Property and Range. The *Class* entity defines a class which will be assigned to a domain individual introduced in the row of scenario content. Examples of such domain individuals are Scenario Step, Step Activity, and Activity Action. The *Property* entity defines an object property or a data property. The domain of that property is a class specified inline or above a row the given property is associated with. If it is a data property, the Range entity must be void; in this case, while exporting scenario content, the value inserted in a given scenario row is used as the object of the serialized triple. If it is an object property, the Range entity must be set to a class the object property refers to with optional name of a data property specified. While exporting scenario content, when no name of the data property is specified, the last seen individual of that class is used as the object of the serialized triple. Otherwise, when the name of the data property is specified, the last seen individual having the specific property value is used.

The mapping units can be aggregated, i.e., more than a single mapping unit can be specified for a single scenario row. In this case, while exporting scenario content for a single row, more than one RDF triple will be generated.

The resulting knowledge base includes data from two sources: the Excel file containing scenario content and a database of scene objects and equipment. The classes responsible for parsing those data sources are the *ScenarioParser* and the *DatabaseParser* respectively, both inheriting from the parent abstract class *PrincipleParser*. The parser classes generate instances of the *DataTuple* class, which represents data in an agnostic manner, i.e., independently of its origin. While conducting a parse, the parser classes use the *OntologyStore* class to obtain references to the appropriate mappings; the references are stored in instances of the *DataTuple* class together with the data value. To gain independence from the physical storage of data in various databases, the *DatabaseParser* class uses implementations of the *IDatabaseService* interface.

The *RdfGenerator* class represents an implementation of the *IKnBaseGenerator* interface for generating a semantic knowledge base in a form of RDF triples. The generating process performs as follows. First, the generator is fed with instances of the *DataTuple* class containing data values together with corresponding mappings to ontology elements. Then, the generator iterates through all data tuples and transforms them to appropriate RDF triples according to

mappings. Because, in general, a data tuple can have several mapping units assigned, each data tuple can result in more than one RDF triple generated.

The generated RDF triples are stored in a form of a semantic graph represented by the *Graph* class. An RDF triple is represented by the *Triple* class and consists of three entities: subject, predicate and object. These entities are included within the graph as its nodes and are represented by various classes being implementations of the *INode* interface:

- the *UriNode* class: a node with a full identifier (a name), used to uniquely represent an RDF triple entity within the whole graph,
- the *LiteralNode* class: a node with a literal text value, enriched with optional metadata: data type and language, used to store single data values of scenario content,
- the *BlankNode* class: an anonymous node (without a public identifier), used to group a set of other nodes into a subgraph.

The *IIdGenerator* interface defines a method for generating RDF triples with domain-specific identifiers for individuals of objects, elements and states included in a knowledge base. The *IdGenerator* class, which implements this interface, first uses the *IQueryManager* implemented as the *QueryManager* class to query the semantic graph for all mentioned above individuals. Next, it uses the *IIdProvider* implemented as the *IdProviderDatabase* class to retrieve the appropriate identifiers from the database of objects and equipment. Finally, RDF triples with the identifiers are generated and asserted into a semantic graph implemented through the *Graph* class.

A semantic graph can be serialized to a text file or saved to a remote triple store. The *TurtleWriter* class is used to serialize a graph to a text file compliant with Turtle syntax.

4.2 Semantic Scenario Manager

The Semantic Scenario Manager is an intuitive visual tool based on Windows Presentation Foundation, which is used by the managers of electricians' teams. Its main goal is to enable refinement and management of the particular training scenarios on the basis of data provided by the electricians using the Scenario Editor.

The Semantic Scenario Manager presents a user with a number of simple and intuitive forms enabling modification of scenario elements. The forms include the names of attributes as well as textboxes or drop-down lists, where the user can provide the necessary information (Fig. 4). The values presented in the drop-down lists are acquired from the scenario ontology. The user needs to provide general information, such as the type of work and a scenario title. Also, the scenario must be classified as elementary, complementary, regular, or verifying. Next, based on the type of work, the user gives information about the works: their category, symbol, technology used and workstation number. The last step is to provide which elements of protective equipment are necessary to complete the training. The user can choose the equipment from a list.

After completing the general information about the scenario, the manager can review and modify the particular steps, activities and actions that trainees need to perform in this scenario. In each scenario, at least one step with at least one activity with at least one action must be specified (cf. Section 3). Actions are associated with interactive and dependent objects' elements as well as possible problems and errors that may occur during the action.

The manager can refine and manage the details of the scenario by editing its tree view, which is a widespread and intuitive form of presentation of hierarchical data (Fig. 5). The hierarchy encompasses the scenario steps, activities, actions, problems, errors and objects, which are distinguished by different icons. The user can expand and collapse the list of sub-items for every item in the tree. The user can also visually add, modify and delete the items in

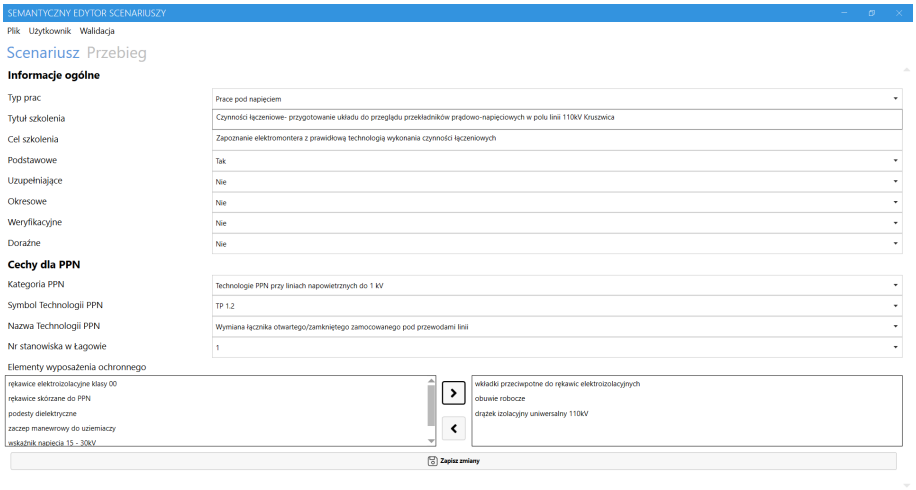


Figure 4. Basic scenario data in the Semantic Scenario Manager

the tree using the toolbar and the context menu. The order of the steps, activities and actions can be altered by dragging and dropping.

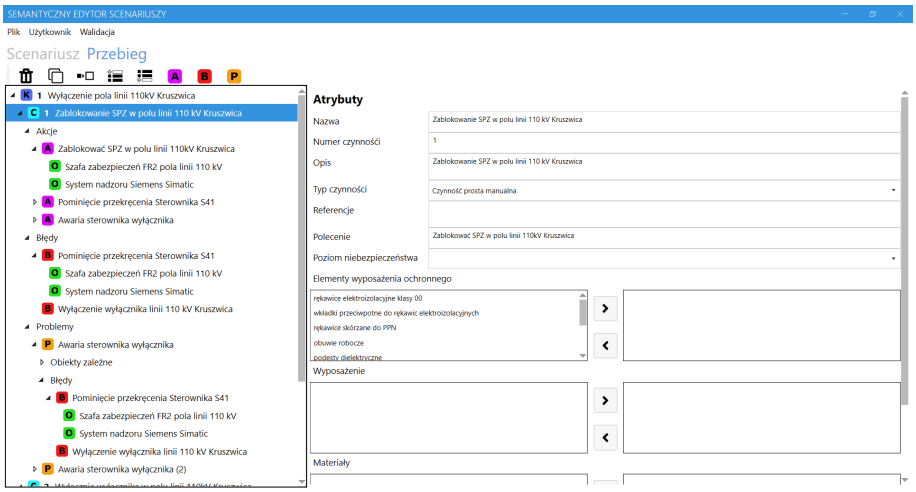


Figure 5. Scenario details tab with the tree view and a form for an activity

During the scenario design, the manager can potentially make a mistake leading to unexpected results in the VR training scene. For that reason, the Semantic Scenario Manager validates the entire scenario against the Scenario Ontology (cf. Section 3) to check whether the scenario is correct. The validation is the consistency checking process on the Scenario Ontology combined with the ABox describing the scenario. It verifies multiple elements of the scenario, including mandatory fields and permitted values, the number of steps, activities and actions, as well as relations between individual instances of classes. The Semantic Scenario Manager highlights the incorrect attributes and the encompassing tree items.

5 Demonstration and Discussion

Training of employees in practical industrial environments requires designing new and modifying existing training scenarios efficiently. In practice, the number of scenarios is by far larger than the number of training scenes. One of the possible applications of our approach is the representation of the training of operators of high-voltage installations. In this case, typically, one 3D model of an electrical substation is associated with at least a dozen different scenarios. These scenarios include learning daily maintenance operations, reactions to various problems that may occur in the installation as well as reactions to infrastructure malfunction.

In the presented approach, all scenarios are knowledge bases structured according to the generic scenario ontology. The scenario ontology consists of 343 axioms, 18 classes, 34 object properties and 47 datatype properties, which can be used in different scenarios. A scenario knowledge base is an ABox specifying a concrete training scenario consisting of steps, activities and actions, along with its elements and infrastructure objects, which are described by classes and properties specified in the scenario ontology (Fig. 6). Scenario knowledge bases are encoded in OWL/Turtle.

To perform training, a scenario knowledge base is imported into the VR Training Application by an importer module, which – based on the scenario KB – generates the equivalent object model of the scenario. An example view of a user executing the ”Karczyn” VR training scenario action is presented in Fig. 7.

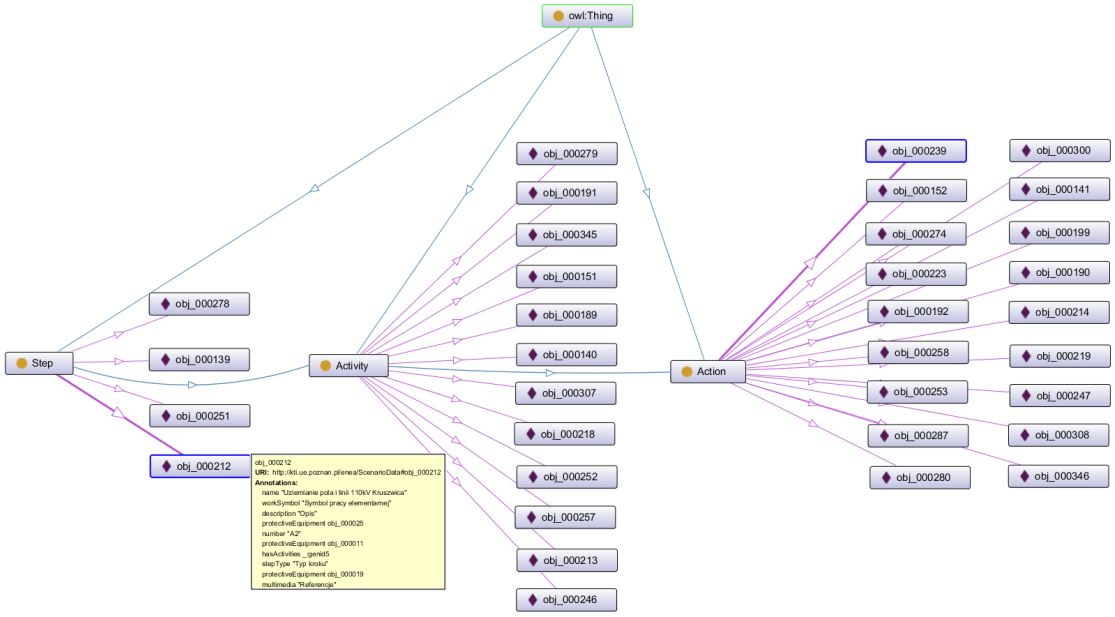


Figure 6. VR training scenario represented as a semantic knowledge base (fragment)

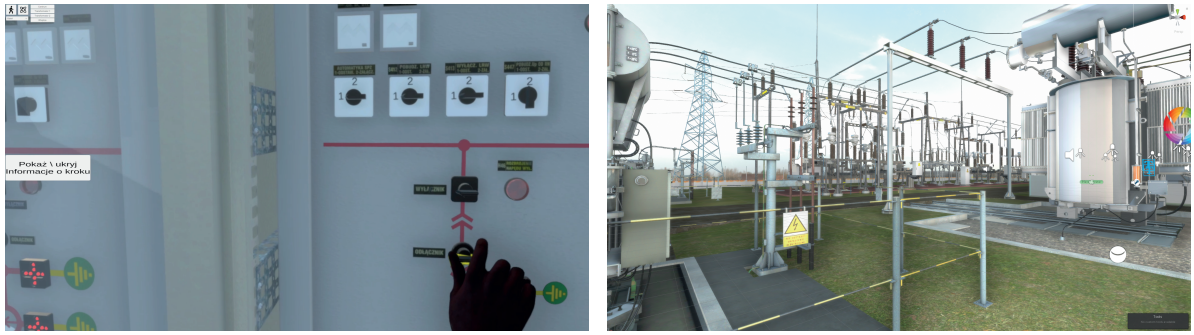


Figure 7. VR training scenario – control room view (left), outside view (right)

The example scenario ”Karczyn” covers the preparation of a trainee for specific maintenance work and consists of 4 steps, 11 activities and 17 actions. For each action, there are dependent objects (44 in case of this scenario). For each step, activity, action and object, the scenario provides specific attributes (9-10 for each item). For each attribute, the name, value, command and comment are provided. In total, the specification of the course of the scenario consists of 945 rows in Excel. In addition, there are 69 rows of specification of errors and 146 rows of specification of problems. The scenario also covers protective equipment, specific work equipment and others.

The generic scenario ontology (TBox) encoded in OWL takes 1,505 lines of code and 55,320

bytes in total. The ”Karczyn” scenario saved in Turtle (which is a more efficient way of encoding ontologies and knowledge bases) has 2,930 lines of code and 209,139 bytes in total.

Implementation of the ”Karczyn” scenario directly as a set of Unity 3D C# scripts would lead to very complex code, difficult to verify and maintain even by a highly-proficient programmer. The design of such a scenario is clearly beyond the capabilities of most domain experts dealing with the everyday training of electrical workers.

An important aspect to consider is the size of the scenario representations. The total size of the ”Karczyn” Unity 3D project is 58 GB, while the size of the executable version is only 1.8 GB. Storing 20 scenarios in editable form as Unity projects would require 1.16 TB of disk space. Storing 20 scenarios in the form of semantic knowledge bases requires only 4MB of storage space (plus the size of the executable application).

The use of semantic knowledge bases with a formal ontology described in this paper enables the concise representation of training scenarios and provides means of editing and verifying scenarios correctness with user-friendly and familiar tools.

6 Conclusions and Future Works

The approach proposed in this paper enables the semantic representation of training scenarios, which is independent of particular application domains. The representation can be used in various domains when accompanied by domain-specific knowledge bases and 3D models of objects. In this regard, it differs from the approaches summarized in Table 1, which are not related to training, even if they permit representation of 3D content behavior.

The approach enables flexible modeling of scenarios at a high level of abstraction using concepts specific to training instead of forcing the designer to use low-level programming with techniques specific to computer graphics. The presented editor, in turn, enables efficient and intuitive creation and modification of the scenarios by domain experts. Hence, the method and the tool make the development of VR applications, which generally is a highly technical task, attainable to non-technical users allowing them to use the terminology of their domains of interest in the design process.

Future works include several elements. First, the environment will be extended to support collaborative creation of scenarios by distributed users. Second, we plan to extend the training application to support not only the training mode, but also the verification mode of operation with appropriate scoring based on user's performance. Finally, we plan to extend the scenario ontology with concepts of parallel sequences of activities, which can be desirable for multi-user training, e.g., in firefighting.

Acknowledgments

The research work presented in this paper has been supported by the European Union from the European Regional Development Fund within the Smart Growth Operational Programme 2020-2024. The project is executed within the priority axis ”Support for R&D Activity of Enterprises” of the National Centre for Research and Development under the contract POIR.01.01.01-00-0463/18.

References

[1] R. Patoni. (). “Alen 3D,” [Online]. Available: https://www.scribd.com/document/245796121/ALEN-3D.

[2] N. Gavish, T. Gutiérrez, S. Webel, J. Rodríguez, M. Peveri, U. Bockholt, and F. Tecchia, "Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks," *Interactive Learning Environments*, vol. 23, no. 6, pp. 778–798, 2015. DOI: 10.1080/10494820.2013.815221. eprint: https://doi.org/10.1080/10494820.2013.815221.

[3] Schneider Electric. (2017). "HoloLens Application on Premset," [Online]. Available: https://www.youtube.com/watch?v=RpXyagutoZg.

[4] J. Flotyński and K. Walczak, "Ontology-Based Representation and Modelling of Synthetic 3D Content: A State-of-the-Art Review," *Computer Graphics Forum*, vol. 35, pp. 329–353, 2017. DOI: 10.1111/cgf.13083.

[5] W. Bille, O. De Troyer, B. Pellens, and F. Kleinermann, "Conceptual modeling of articulated bodies in virtual environments," in *Proceedings of the 11th International Conference on Virtual Systems and Multimedia (VSMM)*, H. Thwaites, Ed., Archaeolingua, Ghent, Belgium: Archaeolingua, 2005, pp. 17–26.

[6] F. Kleinermann, O. De Troyer, H. Mansouri, R. Romero, B. Pellens, and W. Bille, "Designing semantic virtual reality applications," in *In Proceedings of the 2nd INTUITION International Workshop, Senlis*, 2005, pp. 5–10.

[7] B. Pellens, O. De Troyer, W. Bille, F. Kleinermann, and R. Romero, "An ontology-driven approach for modeling behavior in virtual environments," in *Proceedings of On the Move to Meaningful Internet Systems 2005: Ontology Mining and Engineering and its Use for Virtual Reality (WOMEUVR 2005) Workshop*, R. Meersman, Z. Tari, and P. Herrero, Eds., Springer-Verlag, Agia Napa, Cyprus: Springer-Verlag, 2005, pp. 1215–1224.

[8] O. De Troyer, F. Kleinermann, B. Pellens, and W. Bille, "Conceptual modeling for virtual reality," in *Tutorials, posters, panels and industrial contributions at the 26th Int. Conference on Conceptual Modeling - ER 2007*, J. Grundy, S. Hartmann, A. H. F. Laender, L. Maciaszek, and J. F. Roddick, Eds., ser. CRPIT, vol. 83, Auckland, New Zealand: ACS, 2007, pp. 3–18.

[9] O. De Troyer, F. Kleinermann, H. Mansouri, B. Pellens, W. Bille, and V. Fomenko, "Developing semantic VR-shops for e-Commerce.," *Virtual Reality*, vol. 11, no. 2-3, pp. 89–106, 2007.

[10] M. Gutiérrez, "Semantic virtual environments, EPFL," 2005.

[11] M. Gutiérrez, D. Thalmann, and F. Vexo, "Semantic virtual environments with adaptive multimodal interfaces.," in *MMM*, Y.-P. P. Chen, Ed., IEEE Computer Society, Jan. 26, 2005, pp. 277–283, ISBN: 0-7695-2164-9.

[12] E. Kalogerakis, S. Christodoulakis, and N. Moumoutzis, "Coupling ontologies with graphics content for knowledge driven visualization," in *VR '06 Proc. of the IEEE conference on Virtual Reality*, Alexandria, Virginia, USA, Mar. 2006, pp. 43–50.

[13] M. Attene, F. Robbiano, M. Spagnuolo, and B. Falcidieno, "Semantic Annotation of 3D Surface Meshes Based on Feature Characterization," in *Proceedings of the Semantic and Digital Media Technologies 2nd International Conference on Semantic Multimedia*, ser. SAMT'07, Genoa, Italy: Springer-Verlag, 2007, pp. 126–139, ISBN: 978-3-540-77033-6.

[14] F. Robbiano, M. Attene, M. Spagnuolo, and B. Falcidieno, "Part-Based Annotation of Virtual 3D Shapes," *2013 International Conf. on Cyberworlds*, vol. 0, pp. 427–436, 2007.

[15] M. Attene, F. Robbiano, M. Spagnuolo, and B. Falcidieno, "Characterization of 3D Shape Parts for Semantic Annotation," *Comput. Aided Des.*, vol. 41, no. 10, pp. 756–763, Oct. 2009, ISSN: 0010-4485. DOI: 10.1016/j.cad.2009.01.003.

[16] L. De Floriani, A. Hui, L. Papaleo, M. Huang, and J. Hendler, "A semantic web environment for digital shapes understanding," in *Semantic Multimedia*, Springer, 2007, pp. 226–239.

[17] L. Papaleo, L. De Floriani, J. Hendler, and A. Hui, "Towards a semantic web system for understanding real world representations," in *Proceedings of the Tenth International Conference on Computer Graphics and Artificial Intelligence*, 2007.

[18] P. Kapahnke, P. Liedtke, S. Nesbigall, S. Warwas, and M. Klusch, "ISReal: An Open Platform for Semantic-Based 3D Simulations in the 3D Internet," in *International Semantic Web Conference (2)*, 2010, pp. 161–176.

[19] S. Albrecht, T. Wiemann, M. Günther, and J. Hertzberg, "Matching CAD object models in semantic mapping," in *Proceedings ICRA 2011 Workshop: Semantic Perception, Mapping and Exploration, SPME*, 2011.

[20] M. Fischbach, D. Wiebusch, A. Giebler-Schubert, M. E. Latoschik, S. Rehfeld, and H. Tramberend, "SiXton's curse - Simulator X demonstration," in *Virtual Reality Conference (VR), 2011 IEEE*, M. Hirose, B. Lok, A. Majumder, and D. Schmalstieg, Eds., 2011, pp. 255–256. [Online]. Available: http://dx.doi.org/10.1109/VR.2011.5759495.

[21] M. E. Latoschik and H. Tramberend, "Simulator X: A Scalable and Concurrent Software Platform for Intelligent Realtime Interactive Systems," in *Proceedings of the IEEE VR 2011*, 2011.

[22] D. Wiebusch and M. E. Latoschik, "Enhanced Decoupling of Components in Intelligent Realtime Interactive Systems using Ontologies," in *Software Engineering and Architectures for Realtime Interactive Systems (SEARIS), proceedings of the IEEE Virtual Reality 2012 workshop*, 2012, pp. 43–51.

[23] P. Drap, O. Papini, J.-C. Sourisseau, and T. Gambin, "Ontology-based photogrammetric survey in underwater archaeology," in *European Semantic Web Conference*, Springer, 2017, pp. 3–6.

[24] M. Trellet, N. Ferey, M. Baaden, and P. Bourdot, "Interactive visual analytics of molecular data in immersive environments via a semantic definition of the content and the context," in *Immersive Analytics (IA), 2016 Workshop on*, IEEE, 2016, pp. 48–53.

[25] Trellet, M., Férey, N., Flotyński, J., Baaden, M., Bourdot, P., "Semantics for an integrative and immersive pipeline combining visualization and analysis of molecular data," *Journal of Integrative Bioinformatics*, vol. 15 (2), pp. 1–19, 2018.

[26] Y. Perez-Gallardo, J. L. L. Cuadrado, Á. G. Crespo, and C. G. de Jesús, "GEODIM: A Semantic Model-Based System for 3D Recognition of Industrial Scenes," in *Current Trends on Knowledge-Based Systems*, Springer, 2017, pp. 137–159.

[27] J. Flotyński, M. Krzyszkowski, and K. Walczak, "Semantic Composition of 3D Content Behavior for Explorable Virtual Reality Applications," in *Proc. of EuroVR 2017, LNCS*, J. Barbic, M. D'Cruz, M. E. Latoschik, M. Slater, and P. Bourdot, Eds., Springer, 2017, pp. 3–23. DOI: 10.1007/978-3-319-72323-5_1.

[28] J. Flotyński and K. Walczak, "Knowledge-based representation of 3D content behavior in a service-oriented virtual environment," in *Proceedings of the 22nd International Conference on Web3D Technology, Brisbane (Australia), June 5-7, 2017,*, ACM, New York, 2017, Article No 14, ISBN: 978-1-4503-4955-0. DOI: 10.1145/3055624.3075959.

24th International Conference on Business Information Systems (BIS 2021)
Knowledge Graphs
https://doi.org/10.52825/bis.v1i.65

Published: 02 July 2021

Mapping of ImageNet and Wikidata for Knowledge Graphs Enabled Computer Vision

Dominik Filipiak[1][https://orcid.org/0000-0002-4927-9992], Anna Fensel[1,2][https://orcid.org/0000-0002-1391-7104], and Agata Filipowska[3][https://orcid.org/0000-0002-8425-1872]

[1]Semantic Technology Institute (STI) Innsbruck, Department of Computer Science, University of Innsbruck, Austria
[2]Wageningen University & Research, The Netherlands
[3]Department of Information Systems, Poznań University of Economics and Business, Poland

Abstract. Knowledge graphs are used as a source of prior knowledge in numerous computer vision tasks. However, such an approach requires to have a mapping between ground truth data labels and the target knowledge graph. We linked the ILSVRC 2012 dataset (often simply referred to as ImageNet) labels to Wikidata entities. This enables using rich knowledge graph structure and contextual information for several computer vision tasks, traditionally benchmarked with ImageNet and its variations. For instance, in few-shot learning classification scenarios with neural networks, this mapping can be leveraged for weight initialisation, which can improve the final performance metrics value. We mapped all 1 000 ImageNet labels – 461 were already directly linked with the *exact match* property (P2888), 467 have exact match candidates, and 72 cannot be matched directly. For these 72 labels, we discuss different problem categories stemming from the inability of finding an exact match. Semantically close non-exact match candidates are presented as well. The mapping is publicly available at https://github.com/DominikFilipiak/imagenet-to-wikidata-mapping.

Keywords: ImageNet, Wikidata, mapping, computer vision, knowledge graphs

Introduction

Thanks to deep learning and convolutional neural networks, the field of computer vision experienced rapid development in recent years. ImageNet (ILSVRC 2012) is one of the most popular datasets used for training and benchmarking models in the classification task for computer vision. Nowadays, an intense effort can be observed in the domain of few- [17] or zero-shot learning [16] which copes with various machine learning tasks, for which training data is very scarce or even non-available. More formally, N-way K-shot learning considers a setting, in which there are N categories with K samples to learn from (typically $K \leq 20$ in few-shot learning). This is substantially harder from standard settings, as deep learning models usually rely on a large number of samples provided. One of the approaches to few-shot learning considers relying on some prior knowledge, such as the class label. This can be leveraged to improve the performance of the task. For instance, Chen et al. [4] presented Knowledge Graph Transfer Network, which uses the adjacency matrix built from knowledge graph correlations in order to create class prototypes in a few-shot learning classification. More generally, knowledge-embedded machine learning systems can use knowledge graphs as a source of information for improving performance metrics for a given task. One of these knowledge graphs is Wikidata [15], a popular collaborative knowledge graph.

Our main research goal concentrates on facilitating general-purpose knowledge graphs enabled computer vision methods, such as the aforementioned knowledge graph transfer network. In this paper, we provide a mapping between ImageNet classes and Wikidata entities, as this is the first step to achieve this goal. Our paper is inspired by and built on top of the work of Nielsen [12] – he first explored the possibility of linking ImageNet WordNet synsets with Wikidata. We also aim at providing detailed explanations for our choices and compare the results with these provided by Nielsen. Our publicly available mapping links WordNet synset used as ImageNet labels with Wikidata entities. It will be useful for the aforementioned computer vision tasks. Practical usage scenarios consider situations in which labelling data is a costly process and the considered classes can be linked to a given graph (that is, for few- or zero-shot learning tasks). However, simpler tasks, such as classification, can also use context knowledge stemming from rich knowledge graph structure (in prototype learning [18], for instance).

The remainder of this paper is structured as follows. In the next section, we briefly discuss related work. Then, in the third section, we provide detailed explanations about the mapping process, which is focused on the entities which do not have a perfect match candidate. The next section provides some analytics describing the mapping, as well as a comparison with automated matching using a NERD tool – *entity-fishing* [7]. The paper is concluded with a summary. Most importantly, the mapping is publicly available[1].

Background and Related Work

To provide a mapping between ILSVRC 2012 and Wikidata, it is necessary to define some related terms first. This requires introducing a few additional topics, such as WordNet, since some concepts (such as structure) in ILSVRC 2012 are based on the former. This section provides a comprehensive overview of these concepts. We also enlist the existing literature on the same problem of finding this specific mapping. To the best of our knowledge, there were only two attempts to achieve this goal – both are described below.

WordNet is a large lexical database of numerous (primarily) English words [11]. Nouns and verbs have a hierarchical structure and they are grouped altogether as *synsets* (sets of synonyms) in WordNet. Historically, this database paid a significant role in various pre-deep learning era artificial intelligence applications (it is still used nowadays, though). ImageNet [5] is a large image database, which inherited its hierarchical structure from WordNet. It contains 14,197,122 images and 21841 WordNet-based synsets at the time of writing, which makes it an important source of ground-truth data for computer vision. ImageNet Large Scale Visual Recognition Challenge (abbreviated as ILSVRC) [13] was an annual competition for computer vision researchers. The datasets released each year (subsets of original ImageNet) form a popular benchmark for various tasks to this day. The one released at ILSVRC 2012 is particularly popular and commonly called ImageNet[2] up to this date. It gained scientific attention due to the winning architecture AlexNet [9], which greatly helped to popularise deep learning. ImageNet is an extremely versatile dataset – architectures coping with it usually have been successful with different datasets as well [2]. Models trained on ImageNet are widely used for transfer learning purposes [8].

Launched in 2012, Wikidata [15] is a collaborative knowledge graph hosted by Wikimedia Foundation. It provides a convenient SPARQL endpoint. To this date, it is an active project and it is an important source of information for e.g. Wikipedia articles. Due to its popularity, size, and ubiquity, Wikidata can be considered as one of the most popular and successful knowledge graph instances along with DBpedia [1] and Freebase-powered [2] Google Knowledge graph. Given the recent interest in the ability to leverage external knowledge in computer vision tasks [4], it would be therefore beneficial to map ImageNet classes to the correspond-

[1] https://github.com/DominikFilipiak/imagenet-to-wikidata-mapping

[2] From now on, we will refer to ILSVRC 2012 dataset as simply ImageNet, unless directly stated otherwise.

ing Wikidata entities. The idea itself is not new, though the full mapping was not available to this date. To the best of our knowledge, Nielsen [12] was the first to tackle this problem. He summarised the encountered problems during preparing the mapping and classified them into few categories. These categories include missing synsets on Wikidata, matching with a disambiguation page, discrepancies between ImageNet and WordNet, differences between the WordNet and the Wikidata concepts with similar names, and multiple semantically similar items in WordNet and Wikidata. Nielsen described his effort in detail, though the full mapping was not published. Independently, Edwards [6] tried to map DBpedia and Wikidata to ImageNet (in a larger sense, not ILSVRC 2012) using various pre-existing mappings and knowledge graph embeddings methods, such as TransE [3], though the results of such mapping have not been published as well. Contrary to these papers, we publish our mapping.

Mapping

This section is devoted to the mapping between the ImageNet dataset and Wikidata. First, we explain our approach in order to provide such mapping. Then, we identify and group key issues, which occurred in the mapping process. We also provide more detailed explanations for the most problematic entities.

To prepare the mapping, we follow the approach and convention presented by Nielsen. Namely, we use synset names from WordNet 3.0 (as opposed to, for example, WordNet 3.1). That is, we first check the `skos:exactMatch` (`P2888`) property in terms of an existing mapping between Wikidata entities and WordNet synsets. This has to be done per every ImageNet class. For example, for the ImageNet synset `n02480855` we search for `P2888` equal to `http://wordnet-rdf.princeton.edu/wn30/02480855-n` using Wikidata SPARQL endpoint. Listing 1 provides a SPARQL query for this example.

```
SELECT *
WHERE {
  ?item wdt:P2888 ?uri.
  FILTER STRSTARTS(STR(?uri),
  ↪ "http://wordnet-rdf.princeton.edu/wn30/02480855-n").
}
```

Listing 1. Matching WordNet with Wikidata entities using SPARQL.

As of November 2020, there are 461 already linked synsets out of 1000 in ImageNet using `wdt:P2888` property. For the rest, the mapping has to be provided. Unlike Edwards [6], we do not rely on automated methods, since the remaining 539 entities can be checked by hand (although we test one of them in the next section). Using manual search on Google Search, we found good `skos:exactMatch` candidates for the next 467 ImageNet classes. These matches can be considered to be added directly to Wikidata, as they directly reflect the same concept. For the vast majority of the cases, a simple heuristics was enough – one has to type the synset name in the search engine, check the first result on Wikipedia, evaluate its fitness and then use its Wikidata item link. Using this method, one can link 928 classes in total (with 467 entities matched by hand).

Sometimes, two similar concepts were yielded. Such synsets were a subject of qualitative analysis, which aimed at providing the best match. Similarly, sometimes there is no good match at all. At this stage, 72 out of 1000 classes remain unmatched. Here, we enlist our propositions for them. We encountered problems similar to Nielsen [12], though we propose a different summary of common issues. We categorised these to the following category problems:

Table 1. Mapping – hyponymy.

WordNet 3.0 synset		Wikidata Entity	
n03706229	(*magnetic compass*)	Q34735	(*compass*)
n02276258	(*admiral*)	Q311218	(*vanessa*)
n03538406	(*horse cart*)	Q235356	(*carriage*)
n03976467	(*Polaroid camera, Polaroid Land camera*)	Q313695	(*instant camera*)
n03775071	(*mitten*)	Q169031	(*glove*)
n02123159	(*tiger cat*)	Q1474329	(*tabby cat*)
n03796401	(*moving van*)	Q193468	(*van*)
n04579145	(*whisky jug*)	Q2413314	(*jug*)
n09332890	(*lakeshore*)	Q468756	(*shore*)
n01685808	(*whiptail, whiptail lizard*)	Q1004429	(*Cnemidophorus*)
n03223299	(*doormat*)	Q1136834	(*mat*)
n12768682	(*buckeye, horse chestnut, conker*)	Q11009	(*nut*)
n03134739	(*croquet ball*)	Q18545	(*ball*)
n09193705	(*alp*)	Q8502	(*mountain*)
n03891251	(*park bench*)	Q204776	(*bench*)
n02276258	(*admiral*)	Q311218	(*vanessa*)

hyponymy, *animals and their size, age, and sex*, *ambiguous synsets*, and *non-exact match*. Each of these represents a different form of a trade-off made in order to provide the full mapping. This is not a classification in a strict sense, as some of the cases could be placed in several of the aforementioned groups.

Hyponymy. This is the situation in which the level of granularity of WordNet synsets did not match the one from Wikidata. As a consequence, some terms did not have a dedicated entity. Therefore, we performed semantic inclusion, in which we searched for a more general "parent" entity, which contained this specific case. Examples include *magnetic compass* (extended to *compass*), *mitten* (*glove*), or *whisky jug* (*jug*). The cases falling to this category are presented in Table 1.

Animals and their size, age, and sex. This set of patterns is actually a subcategory of the hyponymy, but these animal-related nouns provided several problems worth distinguishing. The first one considers a situation in which a WordNet synset describes the particular sex of a given animal. This information is often missing on Wikidata, which means that the broader semantic meaning has to be used. For example, *drake* was mapped to *duck*, whereas *ram, tup* to *sheep*. However, while *hen* was linked to *chicken*, for *cock, rooster* (n01514668) there exist an exact match (Q2216236). Another pattern considers distinguishing animals of different age and size. For instance, *lorikeet* in WordNet is defined as "any of various small lories". As this definition is a bit imprecise, we decided to use *loriini*. In another example *eft* (juvenile newt) was linked to *newt*. Similarly, there is eastern and western *green mamba*, but WordNet defines it as "the green phase of the black mamba". The breed of poodle has three varieties (*toy*, *miniature*, and *standard poodle*), but Wikidata does not distinguish the difference between them – all were therefore linked to *poodle* (Q38904). These mappings are summarised in Table 2.

Ambiguous synsets. This is a situation in which a set of synonyms does not necessarily consist of synonyms (at least in terms of Wikidata entities). That is, for a synset containing at least two synonyms, there is at least one possible Wikidata entity. At the same time, the broader term for a semantic extension does not necessarily exist, since these concepts can be mutually exclusive. For instance, for the synset *African chameleon, Chamaeleo chamaeleon* there exist two candidates on Wikidata, *Chamaeleo chamaeleon* and *Chamaelo africanus*. We choose the first one due to the WordNet definition – "a chameleon found in Africa". Another synset,

Table 2. Mapping – animals and their size, age, and sex.

WordNet 3.0 synset		Wikidata Entity	
n01847000	(*drake*)	Q3736439	(*duck*)
n01514859	(*hen*)	Q780	(*chicken*)
n01806143	(*peacock*)	Q201251	(*peafowl*)
n02412080	(*ram, tup*)	Q7368	(*sheep*)
n01820546	(*lorikeet*)	Q15274050	(*loriini*)
n01631663	(*eft*)	Q10980893	(*newt*)
n01749939	(*green mamba*)	Q194425	(*mamba*)
n02113624	(*toy poodle*)	Q38904	(*poodle*)
n02113712	(*miniature poodle*)	Q38904	(*poodle*)
n02113799	(*standard poodle*)	Q38904	(*poodle*)
n02087046	(*toy terrier*)	Q37782	(*English Toy Terrier*)

academic gown, academic robe, judge's robe contains at least two quite different notions – we have chosen *academic dress*, as this meaning seems to be dominant in the ImageNet. *Harvester, reaper* is an imprecise category in ImageNet since it offers a variety of agricultural tools, not only these suggested by the synset name. *Bonnet, poke bonnet* has a match at Wikidata's *bonnet*, though it is worth noticing that ImageNet is focused on babies wearing this specific headwear. The mapping of this category can be found in Table 3.

Non-exact match. Sometimes, however, there is no good exact match for a given synset among Wikidata entities. At the same time, the broader term might be too broad. This leads to unavoidable inaccuracies. For example, for *nipple* we have chosen its meronym, *baby bottle*. Interestingly, *nipple* exists in Polish Wikidata, though it does not have any properties, which makes it useless in further applications. Other examples involve somewhat similar meaning – *tile roof* was mapped to *roof tile*, or *steel arch bridge* to *through arch bridge*. *Plate rack* was linked to *dish drying cabinet*, though it is not entirely accurate, as the ImageNet contains pictures of things not designated to drying, but sometimes for dish representation. In other example, we map *baseball player* to *Category:baseball players*. ImageNet contains photos of different kinds of *stemware*, not only *goblet*. *Cassette* was linked to a more fine-grained synset (*audio cassette*) as the images present audio cassettes in different settings. Table 4 summarises the mappings falling into this category.

ImageNet itself is not free from errors, since it is biased towards certain skin colour, gender, or age. This is a great concern for ethical artificial intelligence scientists since models trained on ImageNet are ubiquitous. There are some ongoing efforts to fix it with a more balanced set of images, though [19]. Beyer et al. [2] enlisted numerous problems with ImageNet, such as single pair per image, restrictive annotation process, or practically duplicate classes. They proposed a set of new, more realistic labels (ReaL) and argued that models trained in such a setting achieve better performance. Even given these drawbacks, ImageNet is still ubiquitous. Naturally, the presented mapping inherits problems presented in ImageNet, such as these in which images roughly do not present what the synset name suggests. This problem was previously reported by Nielsen [12] – he described it as a discrepancy between ImageNet and WordNet. As for some examples, this might include *radiator*, which in ImageNet represents home radiator, whereas the definition on Wikidata for the same name describes a bit more broad notion (for instance, it also includes car radiators). *Monitor* is a similar example since it might be any display device, though in ImageNet it is connected mostly to a computer display. *Sunscreen, sunblock, sun blocker* represent different photos of products and their appliance on the human body, which look completely different and might be split into two distinct classes.

Table 3. Mapping – ambiguities.

WordNet 3.0 synset		Wikidata Entity	
n01694178	(*African chameleon, Chamaeleo chamaeleon*)	Q810152	(*Chamaeleo africanus*)
n02669723	(*academic gown, academic robe, judge's robe*)	Q1349227	(*academic dress*)
n02894605	(*breakwater, groin, groyne, mole, bulwark, seawall, jetty*)	Q215635	(*breakwater*)
n01755581	(*diamondback, diamondback rattlesnake, Crotalus adamanteus*)	Q744532	(*eastern diamondback rattlesnake*)
n03110669	(*cornet, horn, trumpet, trump*)	Q202027	(*cornet*)
n03899768	(*patio, terrace*)	Q737988	(*patio*)
n04258138	(*solar dish, solar collector, solar furnace*)	Q837515	(*solar collector*)
n03016953	(*chiffonier, commode*)	Q2746233	(*chiffonier*)
n02114548	(*white wolf, Arctic wolf, Canis lupus tundrarum*)	Q216441	(*Arctic wolf*)
n13133613	(*Ear, spike, capitulum*)	Q587369	(*Pseudanthium*)
n04509417	(*unicycle, monocycle*)	Q223924	(*unicycle*)
n01729322	(*hognose snake, puff adder, sand viper*)	Q5877356	(*hognose*)
n01735189	(*garter snake, grass snake*)	Q1149509	(*garter snake*)
n02017213	(*European gallinule, Porphyrio porphyrio*)	Q187902	(*Porphyrio porphyrio*)
n02013706	(*limpkin, aramus pictus*)	Q725276	(*limpkin*)
n04008634	(*projectile, missile*)	Q49393	(*projectile*)
n09399592	(*promontory, headland, head, foreland*)	Q1245089	(*promontory*)
n01644900	(*tailed frog, bell toad, ribbed toad, tailed toad, Ascaphus trui*)	Q2925426	(*tailed frog*)
n02395406	(*hog, pig, grunter, squealer, Sus scrofa*)	Q787	(*pig*)
n02443114	(*polecat, fitch, foulmart, foumart, Mustela putorius*)	Q26582	(*Mustela putorius*)
n03017168	(*chime, bell, gong*)	Q101401	(*bell*)
n02088466	(*bloodhound, sleuthhound*)	Q21098	(*bloodhound*)
n03595614	(*jersey, t-shirt*)	Q131151	(*t-shirt*)
n03065424	(*coil, spiral, volute, whorl, helix*)	Q189114	(*spiral*)
n03594945	(*jeep, land rover*)	Q946596	(*off-road vehicle*)
n01753488	(*horned viper, cerastes, sand viper, horned asp, Cerastes cornutus*)	Q1476343	(*Cerastes cerastes*)
n03496892	(*harvester, reaper*)	Q1367947	(*reaper*)
n02869837	(*bonnet, poke bonnet*)	Q1149531	(*bonnet*)

Table 4. Mapping – non-exact matches.

WordNet 3.0 synset		Wikidata Entity	
n04311004	(*steel arch bridge*)	Q5592057	(*through arch bridge*)
n01737021	(*water snake*)	Q2163958	(*common water snake*)
n07714571	(*head cabbage*)	Q35051	(*white cabbage*)
n01871265	(*tusker*)	Q7378	(*elephant*)
n04344873	(*studio coach, day bed*)	Q19953097	(*sofa bed*)
n07714571	(*head cabbage*)	Q35051	(*(white) cabbage*)
n03961711	(*plate rack*)	Q1469010	(*dish drying cabinet*)
n04505470	(*typewriter keyboard*)	Q46335	(*typewriter*)
n03825788	(*nipple*)	Q797906	(*baby bottle*)
n04435653	(*tile roof*)	Q268547	(*roof tile*)
n09835506	(*baseball player*)	Q7217606	(*Category:baseball players*)
n02966687	(*carpenter's kit, tool kit*)	Q1501161	(*toolbox*)
n02860847	(*bobsleigh* – sleigh)	Q177275	(*bobsleigh* – sport)
n04493381	(*tub, vat*)	Q152095	(*bathtub*)
n03443371	(*goblet*)	Q14920412	(*stemware*)
n02978881	(*cassette*)	Q149757	(*audio cassette*)

Analytics

We also check to what extent the process can be automated, as it might be useful for larger subsets of ImageNet (in a broad sense). In this section, we present the results of such an investigation. We also provide a concise analysis of the number of direct properties, which is a crucial feature in spite of the future usage of the mapping in various computer vision settings.

Foppiano and Romary developed *entity-fishing* [7], a tool for named entity recognition and disambiguation (abbreviated as NERD). This tool can be employed in order to provide an automatic mapping between ImageNet and Wikidata. We used indexed data built from the Wikidata and Wikipedia dumps from 20.05.2020. For this experiment, each synset is split on commas. For example, a synset *barn spider, Araneus cavaticus* (n01773549) is split into two synset elements: *barn spider* and *Araneus cavaticus*. For each of these elements, the term lookup service from *entity-fishing* is called, which searches the knowledge base for given terms in order to provide match candidates. Since this service provides a list of entities ranked by its conditional probability, we choose the one with the highest value.

We start with the 461 already linked instances, which can be perceived as ground-truth data for this experiment. Among them, for 387 (84%) synset elements there was at least one correct suggestion (for example, for a previously mentioned synset *barn spider* and *Araneus cavaticus* at least one was matched to Q1306991). In particular, 286 (62%) synsets were correctly matched for all its elements (for example, for a previously mentioned synset *barn spider* and *Araneus cavaticus* were both matched to Q1306991). While these results show that NERD tools can speed up the process of linking by narrowing down the number of entities to be searched for in some cases, it does not replace manual mapping completely – especially in more complicated and ambiguous cases, which were mentioned in the previous section. Nevertheless, for the remaining 539 synsets which were manually linked, an identical NERD experiment has been performed, which resulted in similar figures. For 448 (83%) synsets, *entity-fishing* provided the same match for at least one synset element. Similarly, for 342 synsets (63%) the tool yielded the same match for all elements. Albeit these figures can be considered as relatively not low, they prove that the mapping obtained in such a way might consider some discrepancies and justify the process presented in the previous section.

Table 5. Most popular properties in the mapping (occurrences of the same properties of a given entity was counted as one).

property	label	count
P646	(*Freebase ID*)	932
P373	(*Commons category*)	927
P18	(*image*)	911
P8408	(*KBpedia* ID)	687
P910	(*topic's main category*)	681
P279	(*subclass of*)	659
P1417	(*Encyclopædia Britannica Online ID*)	618
P8814	(*WordNet 3.1 Synset ID*)	551
P31	(*instance of*)	524
P1014	(*Art & Architecture Thesaurus ID*)	482

Similarly to Nielsen, we also count the number of the direct properties available in Wikidata. This is a crucial feature since it enables to leverage knowledge graph structure. Listing 2 shows the query used for obtaining the number of properties for `Q29022`. The query was repeated for each mapped entity. Figure 1 depicts a histogram of direct properties for the 1000 mapped classes. This histogram presents the right-skewed distribution (normal-like after taking the natural logarithm) with the mean at 28.28 (σ = 22.77). Only one entity has zero properties (*wall clock*).

In total, there are 992 Wikidata entities used in the mapping, as some of them were used several times, like the mentioned *poodle*. These entities displayed 626 unique properties in total. The most popular ones are listed in Table 5. In the optics of computer vision, an important subset of these categories includes `P373` (*Commons category*) `P910` (*topic's main category*), and `P279` (`subclass of`), as they directly reflect hyponymy and hypernymy with the knowledge graph. Such information can be later leveraged in the process of detecting (dis-)similar nodes in a direct way. For example, using e.g. graph path distance in SPARQL for entities sharing common ancestors considering a given property. However, SPARQL does not allow to count the number of arbitrary properties between two given entities. Using graph embedding is a potential workaround for this issue. For example, one can calculate the distances from 200-dimensional pre-trained embeddings provided by the PyTorch-BigGraph library [10]. Another possible direction considers leveraging other linked knowledge graphs, such as Freebase (*P646*), which is linked to the majority of considered instances.

```
SELECT (COUNT (?property) AS ?count)
WHERE {
    wd:Q29022 ?property [] .
    FILTER STRSTARTS(STR(?property),
    ↪ "http://www.wikidata.ord/prop/direct/")
}
```

Listing 2. Counting direct properties for a single mapped entity. Based on: Nielsen [12].

Summary

In this paper, we presented a complete mapping of ILSVRC 2012 synsets and Wikidata. For 461 classes, such a mapping already existed in Wikidata with `skos:exactMatch`. For other 467

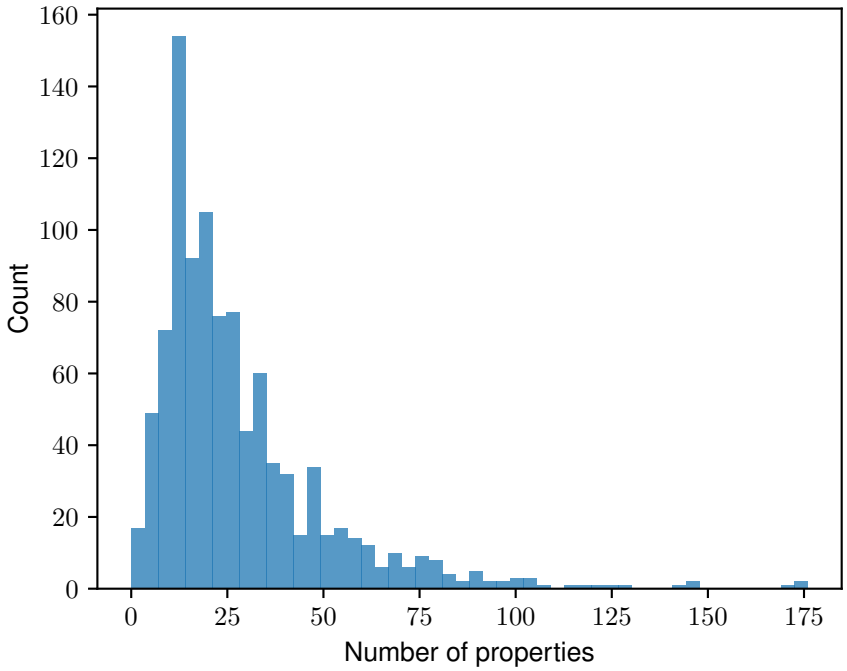


Figure 1. A histogram of direct properties.

classes, we found candidates, which match their corresponding synset. Since 72 classes do not have a direct match, we proposed a detailed justification of our choices. We also compared our mapping with the one obtained from an automated process. To the best of our knowledge, we are the first to publish the mapping ImageNet and Wikidata. The mapping is publicly available for use and validation in various computer vision scenarios.

Future work should focus on empirically testing the mapping. Our results are intended to be beneficial for general-purpose computer vision research since the graphs can be leveraged as a source of contextual information for various tasks, as our analysis showed that the vast majority of the linked entities have a certain number of direct properties. This fact can be utilised according to the given computer vision task. For example, it may be used to generate low-level entity (label) embeddings and calculate distances between them in order to create a correlation matrix used in Knowledge Graph Transfer Network [4] in the task of few-shot image classification. This architecture leverages prior knowledge regarding the semantic similarity of considered labels (called correlation matrix in the paper), which are used for creating class prototypes. These prototypes are used to help the classifier learn novel categories with only few samples available. Correlations might be calculated using simple graph path distance, as well as using more sophisticated low-dimensional knowledge graph embeddings and some distance metrics between each instance. In this case, this will result in a 1000×1000 matrix, as there are 1000 labels in ImageNet. Embeddings from pre-trained models might be used for this task (such as the aforementioned PyTorch-BigGraph embeddings).

Future work might also consider extending the mapping in a way that allows considering larger subsets of ImageNet (in a broad sense), such as ImageNet-6K [4], the dataset, which consists of 6000 ImageNet categories. Preparation of such a large mapping might require a more systematic and collaboratively-oriented approach, which can help to create, verify and reuse the results [20]. The presented approach can also be used for providing mappings with other knowledge graphs and ImageNet. Another possible application might consider further mapping to the actions, which might be particularly interesting for applications in robotics, where robots would be deciding which actions to take based on such mappings [14].

CRediT – Contributor Roles Taxonomy

Dominik Filipiak: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, writing – original draft. **Anna Fensel**: conceptualization, funding acquisition, project administration, writing – review & editing, validation, resources. **Agata Filipowska**: writing – review & editing, validation, resources.

Acknowledgements

This research was co-funded by Interreg Österreich-Bayern 2014-2020 programme project *KI-Net: Bausteine für KI-basierte Optimierungen in der industriellen Fertigung* (grant agreement: AB 292).

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

[2] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26:2787–2795, 2013.

[4] R. Chen, T. Chen, X. Hui, H. Wu, G. Li, and L. Lin. Knowledge graph transfer network for few-shot recognition. In *AAAI*, pages 10575–10582, 2020.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] C. Edwards. Linking knowledge graphs and images using embeddings. https://cnedwards.com/files/studyabroad_report.pdf, 2018.

[7] L. Foppiano and L. Romary. entity-fishing: a dariah entity recognition and disambiguation service. *Journal of the Japanese Association for Digital Humanities*, 5(1):22–60, 2020.

[8] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[10] A. Lerer, L. Wu, J. Shen, T. Lacroix, L. Wehrstedt, A. Bose, and A. Peysakhovich. PyTorch-BigGraph: A Large-scale Graph Embedding System. In *Proceedings of the 2nd SysML Conference*, Palo Alto, CA, USA, 2019.

[11] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[12] F. Å. Nielsen. Linking imagenet wordnet synsets with wikidata. In *Companion Proceedings of the The Web Conference 2018*, pages 1809–1814, 2018.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[14] I. Stavrakantonakis, A. Fensel, and D. Fensel. Matching web entities with potential actions. In *SEMANTICS (Posters & Demos)*, pages 35–38. Citeseer, 2014.

[15] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[16] W. Wang, V. W. Zheng, H. Yu, and C. Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[17] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[18] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu. Robust classification with convolutional prototype learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3482, 2018.

[19] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

[20] A. V. Zhdanova and P. Shvaiko. Community-driven ontology matching. In *European Semantic Web Conference*, pages 34–49. Springer, 2006.

24th International Conference on Business Information Systems (BIS 2021)
Artificial Intelligence
https://doi.org/10.52825/bis.v1i.38

Published: 02 July 2021

Contextual Personality-aware Recommender System Versus Big Data Recommender System

Marcin Szmydt[1][https://orcid.org/0000-0003-4392-0205]

[1]Poznań University of Economics and Business, Poland

Abstract. Many personality theories suggest that personality influences customer shopping preference. Thus, this research analyses the potential ability to improve the accuracy of the collaborative filtering recommender system by incorporating the Five-Factor Model personality traits data obtained from customer text reviews. The study uses a large Amazon dataset with customer reviews and information about verified customer product purchases. However, evaluation results show that the model leveraging big data by using the whole Amazon dataset provides better recommendations than the recommender systems trained in the contexts of the customer personality traits.

Keywords: Recommender system, Predictive modelling, Big five, Personality traits, Big data analytics, Amazon dataset

Introduction

Recommender systems (RSs) are nowadays a very important element that is influencing customer digital experience in electronic services. Many major companies such as Amazon, Netflix, or Spotify are successfully employing effective RSs in their businesses and are seeking to improve their algorithms even further. Therefore, research in this domain seems justified.

There are three main types of recommender systems: content-based (CB), collaborative filtering (CF), and hybrid recommender systems [1]. CB uses similarities among items, e.g., recommending movies of the same genre or news articles on the same topic. A slightly different technique is used in CF. It exploits similarity and relationships among users to provide recommendations [2]. RSs algorithms exploit a different kind of information about the user or items to provide the most accurate recommendations [3].

Exploiting customer personality data seems very appealing to many researchers since it was explored in many studies related to RSs [4]–[6]. Personality theories researchers claim that human personality traits have a significant influence on customer preferences and subsequently on behavior [7], [8]. Therefore, they seem to be a promising predictor of customer behavior. It is especially important in digital markets where customer personality characteristics can be inferred from their digital footprints [9], [10].

1 Research Background

1.1 Customer Personality Traits Identification

In the existing literature, there are many different personality models and personality descriptions [11]–[13]. However, the most commonly used personality model is the Five-Factor Model

(FFM), also known as the Big Five model, proposed by [14] and extended by the work of [15]. According to this approach, there are five basic dimensions of personality: extraversion, neuroticism, openness to experience, agreeableness, and conscientiousness. The Big Five model has been verified in a significant number of empirical studies and has been subjected to psychometric verification on many occasions [16], [17].

Considering the above, personality traits can be successfully used in many different research applications and business scenarios. However, before personality traits can be used, they must be identified in the first place. The most obvious and usually the most reliable approach for identifying FFM personality traits is through psychological questionaries. There were developed many different questionaries for this purpose [18], [19]. However, those questionaries require the user a considerable take time to complete and it is not an easy task to persuade users to donate their time to complete them. Therefore, collecting such data using this technique can be is very expensive and large-scale datasets with personality traits data collected from questionaries are extremely rare. For this reason, researchers and practitioners are trying to infer customer or user personality traits from other data sources such as social media [20], [21], multiple types of digital footprints [22], user-written texts [23]–[26], or speech and video (e.g., face detection and analysis)[25].

1.2 Personality-based Recommender Systems

Through the years, there have been many attempts to incorporate personality traits into RSs. Several publications by [27]–[30] propose and examine an interesting application of personality-based RS (TWIN) in online tourism domain. Their RS produces recommendations based on the user personality model retrieved from the plain text. For their study, they have collected 14,000 text reviews of 1,030 people. To evaluate the performance of the TWIN system they applied their RS to suggest hotels by filtering out reviews produced by people with like-minded views to those of the user. Unfortunately, most of their work is focused on extracting the right personality type from the text, and little is said about the efficiency of the recommendations provided by the RS.

The study carried out by [31] introduced a novel Active-Learning (AL) technique for addressing the cold-start problem in RSs. Their proposed technique uses the FFM as its basis to provide a user with personalized rating requests, without completely relying on explicit feedback (e.g. ratings) or implicit feedback (e.g. item views or purchases) which is usually not available in cold-start situations. Their study claims that their AL method leads to a higher increase in the number of acquired user ratings in comparison to a state-of-the-art rating elicitation strategy. The downside of this study is undoubtedly a small evaluation dataset that covered only 108 participants (required to fill a personality questionnaire).

Very extensive state-of-the-art research related to the application of personality data in RS was presented by [32]. The paper describes different personality models (with the main focus on FFM), a correlation between personality and user preferences, personality identification techniques, an overview of the publicly available datasets for RS, different applications of personality data in RS (cold-start problem, diversity cross-domain recommendations, group recommendations), and open issues and challenges related to the usage of personality in RS.

An interesting example of how to incorporate user personality profile acquires through analysis of the written reviews to RS domain is presented by [5]. Their goal in this study was to incorporate user personality traits into RS and find out whether it would allow improving the accuracy of predicted ratings. The technique used for rating prediction was Kernelized Probabilistic Matrix Factorization (KPMF). The evaluation of their study was based on the experiment which was conducted on the (crawled) IMDB dataset of 2,087 users and 3,500 movies. The ability to identify the personality traits was based on a supervised model trained on the publicly available MyPersonality dataset (social media dataset of 250 users with their personality traits). They have trained six different models and calculated RMSEs based on the test dataset. The

results suggest that the worst score was achieved by the non-optimized Matrix Factorization model, and the most effective model uses a combination of the textual features and the predicted personality scores. Unfortunately, KPMF does not seem to be easily scalable for big datasets.

The six month's study on 1,800 users described by [33] also suggests that it is possible to improve user satisfaction when we integrate users' personality traits into the process of generating recommendations. A recent line of research keeps investigating possible applications of personality data in the RS, especially, in the context of user digital footprints such as text reviews.

The study by [34] identifies the lifestyle of a customer by analyzing text reviews published on Amazon and predicts consumers' purchasing preferences. The interesting results of their experiment conducted on Amazon Review Dataset show that online lifestyles significantly improve recommendation performance and outperform the widely used FFM personality traits as a whole.

A similar study on Amazon Review Dataset was conducted by [35]. The paper suggests that movie preferences correlate with specific product purchase preferences. This finding seems to be in line with the lifestyle preferences correlation.

Another FFM personality-based RS based on text reviews was proposed by [36]. However, the authors added to the model user's level of knowledge about various domains. Their results claim that the proposed model performs better in both MAE and RMSE metrics compared to the other two models (CTR and TWIN).

Finally, RS for e-clothing store based on personality traits, demographics, and behavior of customers in time context was presented by [37]. Their proposed method was compared with different baselines (matrix factorization and ensemble). The results revealed that the proposed method led to a significant improvement in traditional CF performance, and with a significant difference (more than 40%), performed better than all baselines.

1.3 Literature Review Conclusions

Summing up this literature review, the most common model of personality is the FFM, which is composed of the factors openness, conscientiousness, extraversion, agreeableness, and neuroticism. It is suitable for RS since it can be quantified with feature vectors that describe the degree to which each factor is expressed in a user. There are different ways of acquisition of personality traits factors. Generally, those techniques can be grouped into explicit techniques (e.g., questionnaires) and implicit techniques (e.g., identification based on social media, text, or other electronic behavior). While explicit techniques provide relatively accurate assessments of the personalities they are intrusive and time consuming for potential users. However, predicting personality from online texts is a growing trend for researchers. Moreover, FFM traits can be incorporated to RS using pre-filtering [38], KPMF [39], Convex collective matrix factorization [40], or Consistent collective matrix [41]. However, all the approaches besides pre-filtering are not easily scalable and implementable in big data environments. Most researchers working in the area of RS agree that user personality data can improve the quality of recommendations. However, there are still open issues and challenges that need to be addressed to improve the adoption of personality in RS. First of all, most of the studies were based only on a small number of participants (very often ranging from 50 to about 100 participants). Therefore, there is a significant research gap of studies leveraging Big Data for personality-based RS. Moreover, many of the state-of-the-art methods are not easily scalable for large datasets and Big Data technologies.

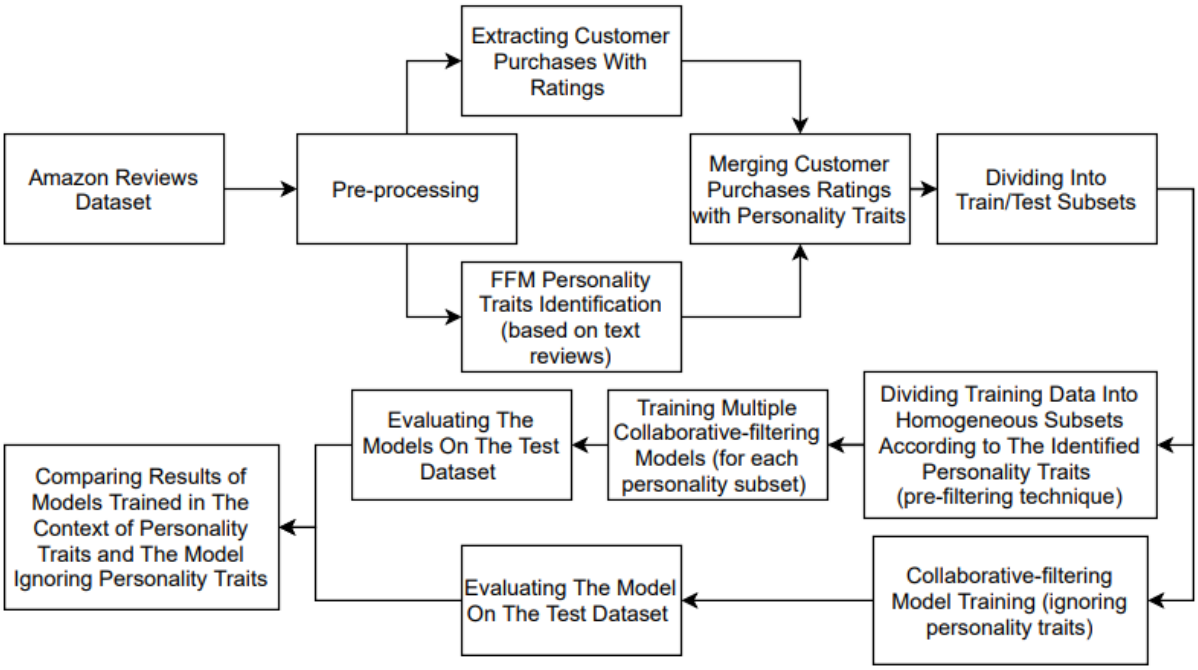


Figure 1. Research Framework

2 Experimental Design

2.1 Research Framework

The main goal of the experiment conducted in this study was to integrate the information contained in the users' text reviews into a RS, and in particular, investigate whether FFM personality traits, as reflected in the text generated by users, would allow improving the Root Mean Square Error (RMSE) of predicted ratings. To incorporate FFM personality traits into the Collaborative-filtering model it was decided to use the contextual pre-filtering technique since it is easily scalable and the easiest to implement in the first place. Figure 1 presents a research framework for the designed experiment. The first step of the experiment was preprocessing the Amazon Reviews. Then, based on the text-reviews, FFM personality traits of Amazon users were identified. For the given group of users product purchases with ratings were extracted and merged with personality data. The next step involved creating a RS model that incorporated personality traits (using pre-filtering) and RS without taking into account personality traits. Finally, both RS models were evaluated and compared.

2.2 Dataset

The analysis was carried out using a subset of Amazon Reviews Dataset collected by [42] and publicly available[1]. The initial dataset covers 233.1 million Amazon reviews between 1998 and 2018. However, to capture the latest trends of customers' behavior and to limit computational power required to process the data, the selected subset used for this study covered the last two years available in the dataset (from the 1st of October 2016 to the 1st of October 2018). Moreover, only reviews of the users with at least five text reviews were selected. Additional filtering was applied to remove empty reviews, errors, and those that did not have a verified purchase status. The final dataset used in this study covered 34,467,155 reviews of 2,968,635 users. The dataset size applied in this study is a significant advantage since there are very few studies of personality-based recommender systems that leverage big data. Different subsets of the same Amazon Reviews Dataset were used in different research scenarios by [34], [36], [43]–[45] and many more scholars. For the purpose of machine learning algorithms, this

[1] http://jmcauley.ucsd.edu/data/amazon/

dataset was divided into the training dataset and testing dataset in proportion 80:20. Therefore, the training dataset covered 27,573,175 customer records and the testing dataset covered 6,893,980 customer records.

2.3 Personality Prediction Engine

To identify FFM personality traits from the text reviews there was used a pre-trained model based on the research with open source code published by [46]. The author of the code was inspired by the work of [24]. Publicly available pre-trained model[2], according to the author, was trained on four different datasets: Stream-of-consciousness Essays, The NRC Emotion Lexicon, Myers-Briggs Personality Type Dataset, and the Scraped Data From Reddit. Stream-of-consciousness Essays dataset is a publicly available dataset of 2,468 anonymous essays tagged with the authors' FFM personality traits. It is the gold standard from psychology since the data was collected in a controlled environment [47]. The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). It is also publicly available[3] and covers about 14,000 words. Myers-Briggs Personality Type Dataset can be found freely on Kaggle[4]. This dataset was collected through the PersonalityCafe forum and provides 8,600 rows of data on peoples' personality type, as well as what they have written. Finally, the Scraped Data From Reddit is the only dataset that is not publicly available. This dataset was used in the research by [48] and was provided by the author of the paper. It covers scraped data from personality subreddits, where people show their personality types in the forum and therefore provide labeled text comments and posts.

The author of the pre-trained model combined all those different sources into one mutual dataset, extracted features from text to vectorize the data with bags of words and GloVe approach, and tested several supervised classification learning algorithms (SVM, Decision Tree, Naive Bayes, Logistic Regression, and Random Forest). The best models for predicting specific FFM personality traits were selected for the final model. The evaluation of the model presented by the author achieves the following accuracy 77.18% (Extraversion), 61.74% (Neuroticism), 75.51% (Agreeableness), 70.34% (Conscientiousness), and 80.39% (Openness). Those results are within the range of the state-of-the-art papers analyzed in the literature review. Therefore, the usage of this pre-trained model seems justified. Similar approaches were presented in papers by [5] or [36].

2.4 Product Recommender Engine

For the purpose of the research, the CF algorithm was chosen for the RS engine since it is relatively accessible in the implementation across different domains and the Amazon Review Dataset contains the (essential for the CF) product ratings. Specifically, Alternating Least Squares (ALS) matrix factorization technique available in the *spark.ml* library was implemented in PySpark according to the Spark documentation[5]. The experiment based on two approaches was designed. The first approach aimed to construct a RS based on one large dataset ignoring the personality traits, while the second approach involved a pre-filtering technique to incorporate personality traits into the RS.

2.4.1 Big Data RS.

The first approach was based on one large training dataset containing solely user ratings for products they purchased. To perform hyperparameter tuning based on 5-fold cross-validation, the sub dataset containing 567,917 records (user ratings) was selected (using random sampling). It allowed to significantly reduce computing power required to train and cross-validate

[2]https://github.com/jkwieser/personality-detection-text

[3]https://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[4]https://www.kaggle.com/datasnaek/mbti-type

[5]https://spark.apache.org/docs/latest/ml-collaborative-filtering.html

models with a different set of hyperparameters. The set of paremeters for the model tuning was based on the experience with other similar projects and the literature. The following code snipped represents a parameter grid used in hyperparameter tuning:

```
param_grid = ParamGridBuilder() \
.addGrid(als.rank, [10, 50, 100, 150]) \
.addGrid(als.regParam, [.01, .05, .1, .15]) \
.build()
```

Then, having selected the best hyperparameters the model was trained on the whole training dataset (without sampling) and evaluated on the test dataset using RMSE score.

2.4.2 Contextual Personality-aware RS.

The second approach aimed to investigate whether FFM personality traits, as reflected in the text generated by users, would allow improving the Root Mean Square Error (RMSE) of predicted ratings. The pre-filtering technique was used to divide the training dataset into homogenous datasets according to the identified personality traits of the users. The threshold for personality traits was set to 0.5 since the evaluation of this predictive model (described by the author) also used the same value. It means that if a given user was assigned by the model value 1 then the probability of a given FFM trait for him/her was more than 0.5, otherwise, 0 was assigned. The subsets of the training dataset were created using two filters: selecting user data according to every FFM combination and selecting user data according to the particular personality traits (selecting users with a given FFM trait, ignoring other traits).

3 Results

3.1 Evaluation Criteria

Recommender systems are popularly evaluated through two main measures: Root Mean Squared Error (RMSE) and Mean Absolute Error(MAE) [49], [50]. However, most cost functions in Machine Learning avoid using MAE and rather use a sum of squared errors or Root Means Squared Error. Moreover, the famous Netflix Prize competition also selected RMSE score as the evaluation criteria [51]. Therefore, for this study RMSE is used as an evaluation metric. The smaller RMSE, the better the RS. In this case, it allowed comparing RS without personality traits and RS with incorporated personality traits.

3.2 Evaluation Results

Regarding the Big Data RS, the results of the hyperparameter tuning of the ALS model revealed that the best performing model (according to RMSE) consisted of the following parameters: als.rank=150 and als.regParam=0.15. Then, the ALS model with those hyperparameters was trained on the whole training dataset. Evaluation conducted on the test dataset achieved the RMSE score of 1.1498. It seems to be a satisfactory result for a RS. Evaluation of the Contextual Personality-aware RS was also conducted on the same test dataset. 37 subsets were selected that correspond to the combinations of the personality data traits. Then, for each personality-homogenous group, there were trained a RS model with hyperparameter tuning (the same parameter grid as used in the RS without personality traits). Each RS was evaluated on part of the test dataset which covered users with corresponding personality traits. Detailed evaluation results with comparison are presented in the Table 1 and Table 2.

The above results indicate that the RS trained on personality-homogenous groups achieves worse average RMSE scores than the RS trained on one diversified big dataset. It may indicate that applying big data is more efficient than using smaller homogenous personality-based groups.

EXT	NEU	AGR	CON	OPN	TRAINING SIZE	TESTING SIZE	CPRS RMSE	BDRS RMSE
0	0	0	0	0	12086	2972	2.3469	1.2508
0	0	0	0	1	230965	57500	1.7227	1.2107
0	0	0	1	0	6949	1770	2.1362	1.1954
0	0	0	1	1	81455	20173	1.9830	1.2034
0	0	1	0	0	13023	3307	2.2307	1.2298
0	0	1	0	1	223708	55574	1.7468	1.2038
0	0	1	1	0	8005	1986	2.2390	1.1879
0	0	1	1	1	91404	22726	1.9603	1.1813
0	1	0	0	0	36605	9061	2.1506	1.2466
0	1	0	0	1	1339587	334993	1.4730	1.2122
0	1	0	1	0	20568	5299	2.2398	1.2062
0	1	0	1	1	406794	101676	1.6215	1.2011
0	1	1	0	0	39359	9559	2.1474	1.2106
0	1	1	0	1	1265592	316421	1.4581	1.1979
0	1	1	1	0	21121	5177	2.2114	1.1703
0	1	1	1	1	504280	126517	1.5560	1.1829
1	0	0	0	0	18692	4652	2.3970	1.1517
1	0	0	0	1	997850	249192	1.4310	1.1634
1	0	0	1	0	10976	2761	2.2642	1.1487
1	0	0	1	1	262347	65518	1.6328	1.1506
1	0	1	0	0	49798	12323	1.9741	1.0615
1	0	1	0	1	2159161	539039	1.3084	1.1383
1	0	1	1	0	14245	3600	2.3315	1.0873
1	0	1	1	1	328325	82385	1.5692	1.1450
1	1	0	0	0	52366	13097	2.0576	1.1777
1	1	0	0	1	6239635	1561631	1.2565	1.1612
1	1	0	1	0	31665	8003	2.1272	1.1244
1	1	0	1	1	1737094	435300	1.3494	1.1470
1	1	1	0	0	138069	34554	1.6726	1.0941
1	1	1	0	1	9269617	2316122	1.1722	1.1246
1	1	1	1	0	36538	9136	2.1399	1.1100
1	1	1	1	1	1925296	481956	1.3220	1.1334
						WEIGHTED AVERAGE RMSE	**1.3134**	**1.1498**

Table 1. RMSE Scores For Contextual Personality-aware Recommender Systems (CPRS) versus the Big Data Recommender Systems (BDRS) - Disjoint Datasets

EXT	NEU	AGR	CON	OPN	TRAINING SIZE	TESTING SIZE	CPRS RMSE	BDRS RMSE
1	*	*	*	*	23 271 674	5 819 269	1.1438	1.1403
*	1	*	*	*	23 064 186	5 768 502	1.1581	1.1494
*	*	1	*	*	16 087 541	4 020 382	1.1567	1.1371
*	*	*	1	*	5 490 239	1 370 806	1.2534	1.1511
*	*	*	*	1	27 064 012	6 765 821	1.1513	1.1503

Table 2. RMSE Scores For Contextual Personality-aware Recommender Systems (CPRS) versus the Big Data Recommender Systems (BDRS) - Overlapping Datasets

4 Discussion

Based on the evaluation results, the paper contributes in pointing out to other researchers that, even though personality traits are indeed very important in RS, incorporating personality traits using contextual pre-filtering is not as efficient as leveraging the whole dataset. Since those findings are based on Amazon.com's large dataset covering many different product domains, it allows expecting that the results can be generalized to other e-commerce platforms as well.

5 Limitations

Every study has limitations and this research is no exception. First of all, this experiment was based only on verified reviews. Hence, the people who purchased the product without reviewing it are not considered in this analysis. Secondly, to identify FFM personality traits from the text reviews there was used a pre-trained model trained on different texts. Otherwise, the study would require a huge number of Amazon users to fill in the FFM personality traits questionnaire which would be a difficult task to accomplish. However, as mentioned before, similar approaches were used by other researchers in this domain as well. Finally, the analysis was based on user accounts that might be shared with others (e.g. members of the family).

6 Future Work

First of all, future research should investigate further fragmentation of personality trait levels rather than having only two states (0 and 1). Exploiting different levels of personality traits (e.g., low, medium, and high levels of extraversion) may improve the accuracy of RS. Moreover, other techniques than pre-filtering can be explored to incorporate personality traits into RS. Finally, exploring similarities in the way users write text reviews (different than FFM personality traits) by applying NLP techniques may be also a good direction to extend this study.

References

[1] P. B. Thorat, R. Goudar, and S. Barve, "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," *International Journal of Computer Applications*, vol. 110, no. 4, pp. 31–36, 2015.

[2] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 145–186, ISBN: 978-0-387-85820-3.

[3] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[4] M. Onori, A. Micarelli, and G. Sansonetti, "A comparative analysis of personality-based music recommender systems.," in *Empire@ RecSys*, 2016, pp. 55–59.

[5] P. Potash and A. Rumshisky, "Recommender system incorporating user personality profile through analysis of written reviews.," in *EMPIRE@ RecSys*, 2016, pp. 60–66.

[6] F. A. Paiva, J. A. Costa, and C. R. Silva, "A personality-based recommender system for semantic searches in vehicles sales portals," in *International Conference on Hybrid Artificial Intelligence Systems*, Springer, 2017, pp. 600–612.

[7] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: A meta-analysis," *Personnel psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[8] S. V. Paunonen and M. C. Ashton, "Big five factors and facets and the prediction of behavior.," *Journal of personality and social psychology*, vol. 81, no. 3, p. 524, 2001.

[9] D. H. Kluemper, P. A. Rosen, and K. W. Mossholder, "Social networking websites, personality ratings, and the organizational context: More than meets the eye? 1," *Journal of Applied Social Psychology*, vol. 42, no. 5, pp. 1143–1172, 2012.

[10] D. Azucar, D. Marengo, and M. Settanni, "Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis," *Personality and individual differences*, vol. 124, pp. 150–159, 2018.

[11] C. S. Hall, G. Lindzey, and J. B. Campbell, "Theories of personality," Wiley New York, Tech. Rep., 1957.

[12] M. Zuckerman, D. M. Kuhlman, J. Joireman, P. Teta, and M. Kraft, "A comparison of three structural models for personality: The big three, the big five, and the alternative five.," *Journal of personality and social psychology*, vol. 65, no. 4, p. 757, 1993.

[13] D. Cervone and L. A. Pervin, *Personality: Theory and research*. John Wiley & Sons, 2015.

[14] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[15] L. R. Goldberg, "The structure of phenotypic personality traits.," *American psychologist*, vol. 48, no. 1, p. 26, 1993.

[16] O. P. John, S. Srivastava, *et al.*, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[17] B. d. E. Raad and M. E. Perugini, *Big five factor assessment: Introduction.* Hogrefe & Huber Publishers, 2002.

[18] G. V. Caprara, C. Barbaranelli, L. Borgogni, and M. Perugini, "The "big five questionnaire": A new questionnaire to assess the five factor model," *Personality and individual Differences*, vol. 15, no. 3, pp. 281–288, 1993.

[19] C. Barbaranelli, G. V. Caprara, A. Rabasca, and C. Pastorelli, "A questionnaire for measuring the big five in late childhood," *Personality and individual differences*, vol. 34, no. 4, pp. 645–664, 2003.

[20] D. Blackwell, C. Leaman, R. Tramposch, C. Osborne, and M. Liss, "Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction," *Personality and Individual Differences*, vol. 116, pp. 69–72, 2017.

[21] D. J. Kuss and M. D. Griffiths, "Online social networking and addiction—a review of the psychological literature," *International journal of environmental research and public health*, vol. 8, no. 9, pp. 3528–3552, 2011.

[22] D. Azucar, D. Marengo, and M. Settanni, "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis," *Personality and Individual Differences*, vol. 124, no. September 2017, pp. 150–159, 2018, ISSN: 01918869. DOI: 10.1016/j.paid.2017.12.018. [Online]. Available: https://doi.org/10.1016/j.paid.2017.12.018.

[23] B. Y. Pratama and R. Sarno, "Personality classification based on twitter text using naive bayes, knn and svm," in *2015 International Conference on Data and Software Engineering (ICoDSE)*, IEEE, 2015, pp. 170–174.

[24] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[25] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, pp. 1–27, 2019.

[26] H. Ahmad, M. Z. Asghar, A. S. Khan, and A. Habib, “A systematic literature review of personality trait classification from textual content,” *Open Computer Science*, vol. 10, no. 1, pp. 175–193, 2020.

[27] A. Roshchina, J. Cardiff, and P. Rosso, “User profile construction in the twin personality-based recommender system,” in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, 2011, pp. 73–79.

[28] ——, “Evaluating the similarity estimator component of the twin personality-based recommender system.,” in *LREC*, 2012, pp. 4098–4102.

[29] A. Roshchina, “TWIN : Personality-based Recommender System,” 2012.

[30] A. Roshchina, J. Cardiff, and P. Rosso, “A comparative evaluation of personality estimation algorithms for the twin recommender system,” in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 2011, pp. 11–18.

[31] M. Elahi, M. Braunhofer, F. Ricci, and M. Tkalcic, “Personality-based active learning for collaborative filtering recommender systems,” in *Congress of the Italian Association for Artificial Intelligence*, Springer, 2013, pp. 360–371.

[32] M. Tkalcic and L. Chen, “Personality and recommender systems,” in *Recommender systems handbook*, Springer, 2015, pp. 715–739.

[33] T. T. Nguyen, F. M. Harper, L. Terveen, and J. A. Konstan, “User personality and user satisfaction with recommender systems,” *Information Systems Frontiers*, vol. 20, no. 6, pp. 1173–1189, 2018.

[34] Y. Huang, H. Liu, W. Li, Z. Wang, X. Hu, and W. Wang, “Lifestyles in amazon: Evidence from online reviews enhanced recommender system,” *International Journal of Market Research*, vol. 62, no. 6, pp. 689–706, 2020.

[35] M. Szmydt, “How do movie preferences correlate with e-commerce purchases? an empirical study on amazon,” in *International Conference on Business Information Systems*, Springer, 2020, pp. 184–196.

[36] S. Yakhchi, A. Beheshti, S. M. Ghafari, and M. Orgun, “Enabling the analysis of personality aspects in recommender systems,” *arXiv preprint arXiv:2001.04825*, 2020.

[37] S. Khodabandehlou, S. A. H. Golpayegani, and M. Z. Rahman, “An effective recommender system based on personality traits, demographics and behavior of customers in time context,” *Data Technologies and Applications*, 2020.

[38] G. Adomavicius and A. Tuzhilin, “Context-aware recommender systems,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA: Springer US, 2011, pp. 217–253, ISBN: 978-0-387-85820-3.

[39] T. Zhou, H. Shan, A. Banerjee, and G. Sapiro, “Kernelized probabilistic matrix factorization: Exploiting graphs and side information,” *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012*, pp. 403–414, 2012. DOI: 10.1137/1.9781611972825.35.

[40] G. Bouchard, D. Yin, and S. Guo, “Convex collective matrix factorization,” in *Artificial intelligence and statistics*, PMLR, 2013, pp. 144–152.

[41] S. Gunasekar, M. Yamada, D. Yin, and Y. Chang, “Consistent collective matrix completion under joint low rank structure,” in *Artificial Intelligence and Statistics*, PMLR, 2015, pp. 306–314.

[42] J. Ni, J. Li, and J. McAuley, “Justifying recommendations using distantly-labeled reviews and fine-grained aspects,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 188–197.

[43] A. I. Hariadi and D. Nurjanah, “Hybrid attribute and personality based recommender system for book recommendation,” in *2017 International Conference on Data and Software Engineering (ICoDSE)*, IEEE, 2017, pp. 1–5.

[44] T. U. Haque, N. N. Saber, and F. M. Shah, “Sentiment analysis on large scale amazon product reviews,” in *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, IEEE, 2018, pp. 1–6.

[45] I. Dematis, E. Karapistoli, and A. Vakali, “Fake review detection via exploitation of spam indicators and reviewer behavior characteristics,” in *International Conference on Current Trends in Theory and Practice of Informatics*, Springer, 2018, pp. 581–595.

[46] J. Wieser, *Personality prediction from text*, `https://github.com/jkwieser/personality-detection-text`, 2020.

[47] J. W. Pennebaker and L. A. King, “Linguistic styles: Language use as an individual difference.,” *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[48] M. Gjurković and J. Šnajder, “Reddit: A gold mine for personality prediction,” in *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, 2018, pp. 87–97.

[49] R. Katarya and O. P. Verma, “An effective collaborative movie recommender system with cuckoo search,” *Egyptian Informatics Journal*, vol. 18, no. 2, pp. 105–112, 2017.

[50] T. Mohammadpour, A. M. Bidgoli, R. Enayatifar, and H. H. S. Javadi, “Efficient clustering in collaborative filtering recommender system: Hybrid method based on genetic algorithm and gravitational emulation local search algorithm,” *Genomics*, vol. 111, no. 6, pp. 1902–1912, 2019.

[51] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

24th International Conference on Business Information Systems (BIS 2021)
Artificial Intelligence
https://doi.org/10.52825/bis.v1i.40

Published: 02 July 2021

Generating a Condensed Representation for Positive and Negative Association Rules

A Condensed Representation for Association Rules

Bemarisika Parfait[1][bemarisika7@yahoo.fr], and Totohasina André[1][andre.totohasina@gmail.com]

[1]ENSET, Université d'Antsiranana, Madagascar

Abstract. Given a large collection of transactions containing items, a basic common association rules problem is the huge size of the extracted rule set. Pruning uninteresting and redundant association rules is a promising approach to solve this problem. In this paper, we propose a Condensed Representation for Positive and Negative Association Rules representing non-redundant rules for both exact and approximate association rules based on the sets of frequent generator itemsets, frequent closed itemsets, maximal frequent itemsets, and minimal infrequent itemsets in database $\mathcal{B}$. Experiments on dense (highly-correlated) databases show a significant reduction of the size of extracted association rule set in database $\mathcal{B}$.

Keywords: Association rules, Generator itemsets, Closed itemsets, Maximal itemsets, Minimal infrequent itemsets.

1 Introduction and Motivations

Positive and negative association rules (PNAR) mining have been studied extensively in Data mining problem. Let X and Y be two disjoints itemsets, an association rule $X \rightarrow Y$ states that a significant proportion in database $\mathcal{B}$ containing items in the premise (or antecedent) X also contain items in the consequent (or conclusion) Y. This rule can indicate the positive relations between different items, is called positive association rule (PAR) in database $\mathcal{B}$. the association rule at other three forms $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$ and $\overline{X} \rightarrow \overline{Y}$, which can indicate the negative relations between items in database $\mathcal{B}$, are called negative association rules (NAR) in database $\mathcal{B}$.

A basic common association rules problem is the huge number of association rules generated many of which are uninteresting (Definition 1) and redundant (Definition 2). Many approaches [13], [14], [16], based on traditional measure confidence [1], has been developed for reducing the size of the extracted rule set. However, no method to prune uninteresting association rules (UAR) has been found in the literature. Indeed, this classic measure confidence is not efficient to prune uninteresting rules. In addition, these approaches are insufficient, because they consider only the positive association rules, and this, with less selective pair support-confidence [1]. Therefore, discovering NAR, which can be interest to several domains [4], [6], [11], [15] such as Artificial Intelligence, Machine Learning, Data Mining, Big Data, Visualization, Marketing, Web mining, etc, is much more less developed than PAR due to the significant problem complexity caused by high computational cost and huge search space in calculating NAR candidates.

In this paper, we propose a Condensed Representation representing non-redundant positive and negative association rules based on *generator itemsets*, *closed itemsets*, *maximal*

itemsets and *minimal infrequent itemsets*. The main contributions are summarized as follows. 1) We propose GC2M algorithm for mining simultaneously all frequent generators, all frequent closeds, all maximal frequent itemsets, and all minimal infrequent itemsets. GC2M is an abbreviation of *Generator itemsets, Closed itemsets, Maximal itemsets, and Minimal infrequent itemsets*. 2) We introduce a formal definition for uninteresting association rules (UAR), then propose an efficient strategy for pruning UAR using M_{GK} measure [7]. 3) We propose an efficient strategy for search space pruning. 4) We propose three new efficient bases based on M_{GK} measure : Concise Basis for Positive Approximate Rules (CBA), Concise Basis for Negative Exact Rules (CBE^{-}), and Concise Basis for Negative Approximate Rules (CBA^{-}). We prove that these concise bases are a lossless representation of non-redundant rules since all valid rules can be derived from these (cf. Theorems 2, 3, 4 and 5). 7) Based on these formalizations, we develop an efficient algorithm, called CONCISE, to discover non-redundant rules.

This paper is organized as follows. Section 2 discusses the related works. Section 3 gives the basic concepts. A Condensed Representation for PNARs is detailed in Section 4. Section 5 presents the experimental results. Conclusion and future work are given in Section 6.

2 Related works

The approaches of association rules mining can be roughly divided into two categories: i Bases of positive association rules, and ii Bases of negative association rules.

In positive basis, we present Duquenne-Guigues basis [10]. Without going into the details of its calculation, this approach is not informative. Bastide's approach [2] adapts Duquenne-Guigues basis. However, it inherits the same flaws as Guigues's approach [10]. In [13], the authors define two bases: *Exat Min-Max Association Rules* and *Approximate Min-Max Association Rules*. Despite their indisputable interests, these two bases contain UARs, and not complete (i.e. they do not generate the negative association rules). In [14], Pasquier defines two bases: *Generic Base for Exact Rules* and *Generic Base for Approximate Rules*. However, this approach is still incomplete and not optimal: it extracts only the positive association rules, many of which are UARs due to the confidence. Xu's approach [16] also extends Pasquier's approach [13], and defines two bases: *Reliable Approximate Basis* and *Reliable Exact Basis*, using CF (Certainty Factor). Similar to Pasquier's approach [14], Xu's approach [16] is also incomplete, it only considers positive rules, don't consider negative association rules.

In negative basis, it is important to mention that the extraction of negative rules is less developed compared to that of positive rules. Note that it emerges from the bibliographic study conducted so far that Feno's approach [7] is the first approach to have studied the problem of bases for negative rules. It extends the Pasquier's approach [13], and defines four bases: *Basis for Exact Positive rules (BPE), Basis for Approximate Positive rules (BPA), Basis for Exact Negative Rules (BNE)* and *Basis for Approximate Negative Rules (BNA)*. However, this approach is not informative, because it selects the premises from a positive borders [12] (or pseudo-closed [13]) which intuitively returns the maximal elements, not in accordance with the notion of minimal premise. It is not very selective due to the use of critical value (cf. Equation (4)) when selecting valid rules. In addition, its formulation of negative exact rules is not appropriate which can present a high memory for searching space. Recently, Dong et al. [5] propose an efficient method for pruning redundant negative and positive rules, using Confidence and Correlation coefficient. Similar to Pasquier's approach, no methods to prune UARs has been found. In particular, Dong's approach does not consider a concept of bases for non-redundant rules, then its configuration semantics is not comparable to our approach.

From this quick literature, mining informative association rules is still a major challenge, for several reasons. On the one hand, the majority of existing approaches are limited on positive association rules which are not sufficient to guarantee the interest of knowledge extraction. On

the other hand, these approaches are also limited on classic pair support-confidence [1] which produces a high number of association rules whose interest is not always guaranteed.

3 Basic concepts

In association rules problem, a Database (cf. Table 1) is a triplet $\mathcal{B} = \mathcal{T}, \mathcal{I}, \mathcal{R}$. $\mathcal{T}$ and $\mathcal{I}$ are finite sets of transactions and items respectively. $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$ is a binary relation between $\mathcal{T}$ and $\mathcal{I}$. A relation $i\mathcal{R}t$ denotes that the item i satisfies the transaction t. Let $X \subseteq \mathcal{I}$, $\overline{X} = \{t \in \mathcal{T} | \exists i \in X : i, t \notin \mathcal{R}\}$ is complementary set of X. A subset $X \subseteq \mathcal{I}$ with $k = |X|$ is called k-itemset, where $|X|$ denotes the cardinality of X. The set $\phi X = X' = \{t \in \mathcal{T} | i\mathcal{R}t, \forall i \in X\}$ is called extension of X. Similarly, the set $\psi Y = Y' = \{i \in \mathcal{I} | i\mathcal{R}t, \forall t \in Y\}$ is intension of Y. Both functions ϕ and ψ form a Galois connection between $\mathcal{PI}$ and $\mathcal{PT}$ [8], where $\mathcal{P}O$ is a power set of O. The composite function $\gamma X = \psi o \phi X$ is called Galois closure operator. Let $X, Y \subseteq \mathcal{I}$, the support of X is defined as $suppX = PX' = \frac{|X'|}{|\mathcal{T}|}$, where P is a discrete probability. The support and confidence [1] of $X \rightarrow Y$ are defined by $suppX \cup Y$ and $confX \rightarrow Y = PY'|X'$ respectively. Let $minsup \in 0, 1$ a minimum support threshold, X is frequent if $suppX \geqslant minsup$. We define $\mathcal{F}$ the set of all frequent in database $\mathcal{B}$ as $\mathcal{F} = \{X \subseteq \mathcal{I} | suppX \geqslant minsup\}$. Let $X, Y \subseteq \mathcal{I}$, X and Y are said to be equivalent, denoted by $X \cong Y$, iff $\gamma X = \gamma Y$. The set of itemsets that are equivalent to X is $X = \{Y \subseteq \mathcal{I} \mid X \cong Y\}$. The item C is closed iff $C = \gamma C$. We define the set $\mathcal{FC}$ of all frequent closed itemsets in database $\mathcal{B}$ as: $\mathcal{FC} = \{C \in \mathcal{I} | C = \gamma C,\ suppC \geqslant minsup\}$. An itemset G is said a minimal generator of a closed C iff $\gamma G = C$ and $g \subseteq \mathcal{I}$ with $g \subseteq G$ such that $\gamma g = C$. We define the set $\mathcal{G}_C$ of all frequent generators as: $\mathcal{G}_C = \{G \in C | C \in \mathcal{FC},\ g \subset G,\ suppG \geqslant minsup\}$. We define $\mathcal{MFC}$ the set of all maximal frequent in database $\mathcal{B}$ as: $\mathcal{MFC} = \{C \in \mathcal{FC} \mid \mathcal{C} \supset C,\ \mathcal{C} \in \mathcal{FC}\}$.

Table 1. Context $\mathcal{B}$

TID	Items
1	ACD
2	BCE
3	ABCE
4	BE
5	ABCE
6	BCE

4 Condensed Representations for PNARs

Our approach is divided into two successive steps: (i) it extracts the set $\mathcal{FC}$, $\mathcal{MFC}$, $\mathcal{G}_{\gamma.}$, and the set $\overline{\mathcal{F}}_{\text{MIN}}$ of minimal infrequent itemsets in $\mathcal{B}$; (ii) it derives from these frequent sets the non-redundant informative rules. An association rule is informative if its premise (resp. conclusion) is minimal (resp. maximal). For lack of space, certain proofs of the Properties are omitted.

4.1 Generating of $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$, $\mathcal{MFC}$ and $\overline{\mathcal{F}}_{\text{MIN}}$

Our main motivation lies in absence of an autonomous approach for mining $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$, $\mathcal{MFC}$ and $\overline{\mathcal{F}}_{\text{MIN}}$. We then propose an efficient algorithm, GC2M, that simultaneously collects these four sets $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$, $\mathcal{MFC}$ and $\overline{\mathcal{F}}_{\text{MIN}}$ in database $\mathcal{B}$. Here we briefly describe GC2M algorithm. It's composed of two algorithms (Algo. 1 and Algo. 2). Its main orginality lies in the effective support counting strategy: *Let X be a frequent k-itemset ($k \geqslant 3$) and $\widetilde{X}$ a $k-1$-subsets of X. Then, X is not a generator iff $suppX = min\{supp\widetilde{X} | \widetilde{X} \subset X\}$* [2], i.e. no access to context $\mathcal{B}$ is made if X is non-generator. On search space pruning, it uses the following properties: *(i) All subsets of a frequent are frequent, (ii) All supersets of an infrequent itemset are infrequent, (iii) All subsets of a generator are also generator, (iv) All supersets of a non-generator are also non-generator* [2]. These results will be synthesized in the algorithm 1. The following Figure 1 shows examplary execution of Algorithm 1 with a small context $\mathcal{B}$ from table 1 and fixed $minsup = 26$. From $\mathcal{MFC}$, we can derive the set $\overline{\mathcal{F}}_{\text{MIN}}$ of minimimal infrequent in $\mathcal{B}$.

Definition 1 (Minimal infrequent itemset) *Let $\mathcal{MFC}$ be the set of maximal frequent, and $\mathcal{F}$ the set of frequent in $\mathcal{B}$. The set $\overline{\mathcal{F}}_{\text{MIN}}$ of minimal infrequent itemsets in $\mathcal{B}$ is defined as :*

$$\overline{\mathcal{F}}_{\text{MIN}} = \{X \in 2^{\mathcal{I}} \backslash \mathcal{MFC} \mid Y \subset X, Y \notin \mathcal{F}\}. \quad (1)$$

Require: A database $\mathcal{B}$, A minimum support threshold $minsup \in 0, 1$.
Ensure: List of $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$ and $\mathcal{MFC}$.
1: $\mathcal{FCC}_1$.GENERATORS $\leftarrow$ {1-itemsts}
2: **for all** $(k \leftarrow 1; \mathcal{FCC}_k$.GENERATORS $\neq \emptyset; k$) **do**
3: $\mathcal{FCC}_k.closure \leftarrow \emptyset$; $\mathcal{FCC}_k.support \leftarrow 0$;
4: $\mathcal{FCC}_k \leftarrow$ GENCLOSURES$\mathcal{FCC}_k$
5: **for all** (candidate itemsets $c \in \mathcal{FCC}_k$) **do**
6: Calculate $suppc$;
7: **if** $(suppc \geqslant minsup)$ **then**
8: $\mathcal{FC}_k \leftarrow \mathcal{FC}_k \cup \{c\}$
9: **end if**
10: **end for**
11: $\mathcal{FCC}_{k1} \leftarrow$ GENGENERATORS$\mathcal{FC}_k$
12: $\mathcal{FCC}_{k1} \leftarrow$ GENMAXIMAL$\mathcal{FC}_k$
13: **end for**
14: $\mathcal{FC} \leftarrow \overset{k-1}{\underset{j=1}{}}\{\mathcal{FC}_j$.CLOSURE, $\mathcal{FC}_j$.support$\}$
15: **return** $\mathcal{FC}$

Algorithm 1: GENERATING $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$ AND $\mathcal{MFC}$

Scan $\mathcal{B}$ $\longrightarrow$

$\mathcal{FCC}_1$		$supp$
A	AC	3/6
B	BE	5/6
C	C	5/6
~~D~~	~~ACD~~	$min36, 16, 16 = 16 < 26$
E	BE	5/6

Pruning the infrequent itemsets $\longrightarrow$

$\mathcal{G}_{\gamma.}$	$\mathcal{FC}$	$\mathcal{MFC}$	$supp$
A	AC	-	3/6
B	BE	-	5/6
C	C	-	5/6
E	BE	-	5/6

Scan $\mathcal{B}$ $\longrightarrow$

$\mathcal{FCC}_2$		$supp$
AB	ABCE	$min26, 26, 26, 46 = 26$
AE	ABCE	$min26, 26, 26, 46 = 26$
BC	BCE	$min46, 56, 46 = 46$
CE	BCE	$min46, 56, 46 = 46$

Pruning the infreq. itemsets $\longrightarrow$

$\mathcal{G}_{\gamma.}$	$\mathcal{FC}$	$\mathcal{MFC}$	$supp$
AB	ABCE	ABCE	2/6
AE	ABCE	ABCE	2/6
BC	BCE	-	4/6
CE	BCE	-	4/6

Figure 1. List of $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$ and $\mathcal{MFC}$ using Algorithm 1 with $minsup = 26$

The Algorithm 2 inputs a database $\mathcal{B}$, a $minsup$, and outputs the set $\overline{\mathcal{F}}_{\text{MIN}}$. Let's take our

Require: $\mathcal{B}$, $\mathcal{MFC}$ and $minsup \in 0, 1$.
Ensure: $\overline{\mathcal{F}}_{\text{MIN}}$ the set of minimal infrequent itemsets.
1: $\overline{\mathcal{F}}_{\text{MIN}} \leftarrow \emptyset$
2: **for all** $(X \in 2^{\mathcal{I}} \backslash \mathcal{MFC})$ **do**
3: **if** $(Y \subset X \mid suppY \leq minsup)$ **then**
4: $\overline{\mathcal{F}}_{\text{MIN}} \leftarrow \overline{\mathcal{F}}_{\text{MIN}} \cup \{X\}$
5: **end if**
6: **end for**
7: **return** $\overline{\mathcal{F}}_{\text{MIN}}$

Algorithm 2: GENERATING MINIMAL INFREQUENT ITEMSETS

example in Figure 1 at $minsup = 26$, we have $D \in 2^{\mathcal{I}} \backslash \mathcal{MFC}$. We see that $D \notin \mathcal{F}$ and $\widetilde{D} \subset D$ such taht $supp\widetilde{D} \leq minsup$. This means D is a minimal infrequent itemsets (i.e. $D \in \overline{\mathcal{F}}_{\text{MIN}}$).

4.2 Generating non-redundant PNARs

This Subsection is based essentially on 5 components : Pruning UAR, Modelization of significant rules, Search space pruning, Pruning redundant PNARs, and CONCISE algorithm.

4.2.1 Pruning Uninteresting Association Rules (UAR).

We first formalize the idea of UAR, and then propose a strategy to prune UAR. Note that the classic support-confidence [1] is not able to prune UAR. Table 2 ullustrates these limits. The information given in this Table 2 can be used to evaluate the association $A \rightarrow B$ and tea $\rightarrow$ coffee. For the pair A, B, we have $suppA \cup B = 0.72$ and $confA \rightarrow B = 0.9$. For the pair (tea,coffee), we have $supp$tea $\cup$ coffee $= 0.2$ and $conf$tea $\rightarrow$ coffee $= 0.8$. The support and confidence are considered fairly high for both rules, i.e. $A \rightarrow B$ and tea $\rightarrow$ coffee are interesting rules. How-

Table 2. Contingency table

	A	$\neg A$			coffee	¬coffee	
B	72	18	90	tea	20	5	25
$\neg B$	8	2	10	¬tea	70	5	75
	80	20	100		90	10	100

ever, $PB'|A' = PB' = 0.9$ and $conf$tea $\rightarrow$ coffee = $0.8 < 0.9 = supp$coffee implie A and B are independent (resp. tea disfavors coffee), i.e. $A \rightarrow B$ and tea $\rightarrow$ coffee are UAR.

Definition 2 (Uninteresting Association Rules (UAR)) *Let $X, Y \subseteq \mathcal{I}$ such that $X \cap Y = \emptyset$. An association rule $X \rightarrow Y$ is said to be uninteresting rule if Y is independent on X (i.e. $PY'|X' = PY'$) or Y is negatively dependent on X (i.e. $PY'|X' < PY'$).*

We then propose an UAR pruning strategy by measuring the degree dependency of X and Y, denoted $\Delta_{X,Y} = PY'|X' - PY'$. We then use M_{GK} measure [7], defined as :

$$M_{GK}X \rightarrow Y = \begin{cases} \frac{PY'|X' - PY'}{1 - PY'}, \text{if } \Delta_{X,Y} > 0 \\ \frac{PY'|X' - PY'}{PY'}, \text{if } \Delta_{X,Y} \leqslant 0. \end{cases} \quad (2)$$

The M_{GK} refers to dependencies between the antecedent and consequent of an association rule. Values in $-1, 0$ show that there is a negative dependence between X and Y. Values in $0, 1$ show that there is a positive dependence between X and Y. Value equal 0 show that Y independent on X. We recall, rules with M_{GK} equal to 1 are called Exact Association Rules, and rules with M_{GK} less than 1 are called Approximate Rules. Theorem 1 below states that value of UARs defined by Definition 2 will be statistically **null** or **negative**.

Theorem 1 ([2) *] Let $X, Y \subseteq \mathcal{I}$. (1) If $PY'|X' \leqslant PY'$, we have $-1 \leqslant M_{GK}X \rightarrow Y \leqslant 0$. (2) If $PY'|X' > PY'$, then $0 < M_{GK}X \rightarrow Y \leqslant 1$.*

From the same example of table 2, we have $M_{GK}A \rightarrow B = \frac{0.9-0.9}{1-0.9} = 0$, this verifies that A and B are independent. So, $A \rightarrow B$ is UAR. We also obtain M_{GK}tea $\rightarrow$ coffee = $\frac{0.8-0.9}{1-0.9} = -1 < 0$, this means that `coffee` and `tea` are negatively dependent. In other words `tea` $\rightarrow$ `coffee` is UAR. As result, the UARs are systematically pruned using M_{GK}.

4.2.2 Modelization of significant rules using M_{GK}.

Note that the first component of M_{GK} (Eq. (2)) is implicative but the second is not, only the first will be active in modelization. We introduce the quantities $n = |\mathcal{T}|$, $n_X = |\phi X|$, $n_Y = |\phi Y|$, $n_{X \wedge Y} = |\phi X \cup Y|$ and $n_{X \wedge \overline{Y}} = |\phi X \cup \overline{Y}|$. The quantity $N_{X \wedge \overline{Y}}$ indicates a random variable which generates $n_{X \wedge \overline{Y}}$, and $N_{X \wedge Y}$ generates $n_{X \wedge Y}$. In that case, the Eq. (2) can be rewritten :

$$M_{GK}X \rightarrow Y = 1 - \frac{n n_{X \wedge \overline{Y}}}{n_X n_{\overline{Y}}} \quad (3)$$

The current versions [7] are based on critical value γ_α defined as

$$\gamma_\alpha = \sqrt{\frac{1}{n} \frac{n - n_X}{n_X} \frac{n_Y}{n - n_Y} \chi^2 \alpha}, \quad (4)$$

where α a real in the interval $0, 1$ and $\chi^2\alpha$ is a Chi-square statistic of a single degree of freedom. This means that $X \rightarrow Y$ will be valid if $M_{GK}X \rightarrow Y \geqslant \gamma_\alpha$. However, this critical value can nevertheless present some limits. Indeed, a low α leads to a high critical value which rapidly exceeds the M_{GK} value. This rejects certain robust rules. Conversely, a large value of α leads to a very low critical value. This accepts certain very weak rules (i.e. independent rules).

To overcome these limits, we define a new model based on the test H_0 independence hypothesis of X and Y in the face of a positive dependence hypothesis H_1, of the rule $X \rightarrow Y$. We then model, under H_0 independence hypothesis, the probability between the random variable $N_{X \wedge \overline{Y}}$ and the observed counter-examples $n_{X \wedge \overline{Y}}$ using measure M_{GK}. We notice that

the sensitivity of this measure M_{GK} to variations in the occurrences of the observed counter-examples $n_{X\wedge\overline{Y}}$ reades with the partial derivative given in the following Equation (5):

$$\frac{\partial M_{GK}}{\partial n_{X\wedge\overline{Y}}} = -\frac{1}{\frac{n_X n_{\overline{Y}}}{n}} \tag{5}$$

This shows that M_{GK} decreases when the number $n_{X\wedge\overline{Y}}$ increases and all the more quickly as the quantity $\frac{n_X n_{\overline{Y}}}{n}$ is low. In other words, M_{GK} grows when $n_{X\wedge\overline{Y}}$ decreases, which is semantically acceptable, but the rate of variation is constant, independent of the rate of decrease of this number, variations of n_Y. Consider $\widetilde{M_{GK}}$ as the realization of a variable M_{GK}, defined as:

$$\widetilde{M_{GK}}X \to Y = -\widetilde{M_{GK}}X \to \overline{Y} = -\frac{n_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{\frac{n_X n_{\overline{Y}}}{n}}}\sqrt{\frac{n}{n_X n_{\overline{Y}}}} \tag{6}$$

It is the opposite of the directed contribution of the cell $X \cup \overline{Y}$ to the $\frac{\chi^2}{n}$ except for a constant. In practice, it is quite common to observe a few transactions which contain X and not Y without having the general trend to have Y when X is present contested. Therefore, $n_{X\wedge\overline{Y}}$ must be taken into account to statistically accept to retain or not the rule $X \to Y$. Suppose we draw at random two subsets $U, Z \subseteq \mathcal{I}$ which contain n_X and n_Y respectively, i.e. $N_{X\wedge\overline{Y}} = |\phi U \cup \overline{Z}|$. This variable $N_{X\wedge\overline{Y}}$ follows a Poisson law with parameter $\frac{n_X n_{\overline{Y}}}{n}$ [9]. We then measure the smallness of random variable $N_{X\wedge\overline{Y}}$ expected to the number $n_{X\wedge\overline{Y}}$ under H_0 independence hypothesis between X and Y. Such an association rule $X \to Y$ is then said to be admissible at the threshold $\alpha \in 0, 1$ if the probability that the random variable $N_{X\wedge\overline{Y}}$ is lower than that observed number $n_{X\wedge\overline{Y}}$ under H_0 independence hypothesis on X and Y is relatively low :

$$PN_{X\wedge\overline{Y}} \leqslant n_{X\wedge\overline{Y}}|H_0 \leqslant \alpha. \tag{7}$$

We then have :

$$PN_{X\wedge\overline{Y}} \leqslant n_{X\wedge\overline{Y}}|H_0 = P\left(\frac{N_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{n_X n_{\overline{Y}}}}\frac{n}{\sqrt{n_X n_{\overline{Y}}}} \leqslant \widetilde{M_{GK}}X \to \overline{Y}\right)$$

Noting Φ. the standard normal distribution, we have $\frac{N_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{n_X n_{\overline{Y}}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}0, 1$, and $\frac{n}{\sqrt{n_X n_{\overline{Y}}}} \xrightarrow[n\to\infty]{p.s} 1 \Rightarrow \frac{n}{\sqrt{n_X n_{\overline{Y}}}} \xrightarrow[n\to\infty]{\mathbb{P}} 1 \Rightarrow \mathcal{K}X, \overline{Y} = \frac{N_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{n_X n_{\overline{Y}}}} \times \frac{n}{\sqrt{n_X n_{\overline{Y}}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}0, 1$. Finally, we have:

$$\begin{aligned} PN_{X\wedge\overline{Y}} \leqslant n_{X\wedge\overline{Y}}|H_0 &= P\left(\mathcal{K}X, \overline{Y} \leqslant \widetilde{M_{GK}}X \to \overline{Y}\right) \\ &\overset{TCL}{\simeq} \int_{-\infty}^{\widetilde{M_{GK}}X\to\overline{Y}} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t = \Phi\widetilde{M_{GK}}X \to \overline{Y}. \end{aligned}$$

Our model of significant association rules is given in the following Definition 3.

Definition 3 (Valid association rule) *Given a minimum threshold $\alpha \in 0, 1$. An association rule $X \to Y$ is said to be valid at level confidence $1 - \alpha$, called $1 - \alpha$-valid, if only if :*

$$p_{XY} = 1 - \Phi - \widetilde{M_{GK}}X \to \overline{Y} \geqslant 1 - \alpha \tag{8}$$

For example, from Table 1, consider a rule $A \to \overline{BCE}$ with $\alpha = 1\%$. Here, $n_A = 2$, $n_B = 5$ and $n_{BCE} = 4$. From $n_A = 2$, $n_B = 5$, $n_{BCE} = 4$ and $\frac{n_A n_{BCE}}{n} = 0 < 3$ (Gaussian hypothesis, cf. [9]), we have $p_{A\overline{BCE}} = 0.5 < 0.99 \Rightarrow A \to \overline{BCE}$ is 99%-invalid (i.e. it's not valid at $\alpha = 1\%$).

4.2.3 Search space pruning.

Pasquier's approach [14] is the most popular approach for generating of non-redundant rules. However, no methods for search space pruning of significant valid rules is used by this approach. While it is possible to restrict the search space by partitioning into 2 the 8 ($X \rightarrow Y$, $Y \rightarrow X$, $\overline{X} \rightarrow \overline{Y}$, $\overline{Y} \rightarrow \overline{X}$, $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $\overline{Y} \rightarrow X$ and $Y \rightarrow \overline{X}$) candidates in database $\mathcal{B}$.

We then explain this restriction. In [2], we demonstrated that if X favors Y (i.e. $PY'|X' > PY'$), then these are the four association rules $X \rightarrow Y$, $Y \rightarrow X$, $\overline{X} \rightarrow \overline{Y}$ and $\overline{Y} \rightarrow \overline{X}$, which will be studied. If X disfavors Y (i.e. $PY'|X' \leqslant PY'$), then these are the four contrary association rules $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $\overline{Y} \rightarrow X$ and $Y \rightarrow \overline{X}$ which will be studied. We then obtain two classes: Class of rules ($X \rightarrow Y$, $Y \rightarrow X$, $\overline{X} \rightarrow \overline{Y}$, $\overline{Y} \rightarrow \overline{X}$), denoted $\mathscr{C}_1$, and Class of rules ($X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $\overline{Y} \rightarrow X$, $Y \rightarrow \overline{X}$), denoted $\mathscr{C}_2$. We also demonstrated that all rules of $\mathscr{C}_1$ can be derived from $X \rightarrow Y$, and all rules of $\mathscr{C}_2$ can be derived from $X \rightarrow \overline{Y}$. So, we only study two rules such as $X \rightarrow Y$ and $X \rightarrow \overline{Y}$. This gives the reduction space 100(8-2)/8=75%.

4.2.4 Pruning redundant PNARs.

The most popular method to prune redundant rules is the base of rules that is a set of reduced size rules that do not contain any redundant rule. Definition 4 defines a redundant rule.

Definition 4 (Redundant rule) *The rule $r_1 : X_1 \rightarrow Y_1$ is redundant if $\exists r_2 : X_2 \rightarrow Y_2$, where $X_1 \supset X_2$, $Y_1 \subset Y_2$ such that $suppr_1 = suppr_2$ and $M_{GK}r_1 = M_{GK}r_2$.*

Corresponding to the three popular approaches [7], [14], [16], we propose three more efficient bases called Concise Bases as defined in Definitions 6, 7 and 8. In addition, we define a base for Positive Exact Rules using M_{GK}, called CBE (cf. Definition 5). More precisely, CBE basis is similar of *Base for Exact Rules* defined in [14], because an exact rule of Confidence is also exact of M_{GK} (cf. [2]). We prove that these concise bases are a lossless representation of non-redundant rules since all valid rules can be derived from these (cf. Theorems 2, 3, 4, 5).

Definition 5 (CBE Basis) *Let $\mathcal{FC}$ be the set of frequent closed itemsets. For each $\mathcal{C} \in \mathcal{FC}$, let $\mathcal{G}_\mathcal{C}$ be the set of minimal generators of $\mathcal{C}$, we have:*

$$CBE = \{G \rightarrow \mathcal{C} \backslash G \mid G \in \mathcal{G}_\mathcal{C},\ \mathcal{C} \in \mathcal{FC},\ G \neq \mathcal{C}\} \tag{9}$$

Theorem 2 *(i) All valid positive exact rules and their supports can be derived from to CBE basis. (ii) All rules in CBE are non-redudant exact rules.*

Proof 1 *i Let $r_1 : X_1 \rightarrow Y_1 \backslash X_1$ be the exact positive rule between two frequents X_1 and Y_1 such that $X_1 \subset Y_1$. Let $\mathcal{C}$ be a frequent closed itemset in $\mathcal{B}$ (i.e. $\mathcal{C} \in \mathcal{FC}$). Since $M_{GK}r_1 = 1$, we have $suppX_1 = suppY_1$. From $suppX_1 = suppY_1$, we derived that $supp\gamma X_1 = supp\gamma Y_1 \Rightarrow \gamma X_1 = \gamma Y_1 = \mathcal{C}$. Obviously, there exists a rule $r_2 : G \rightarrow \mathcal{C} \backslash G \in CBE$ such that G is a generator of $\mathcal{C}$ for which $G \subseteq X_1$ and $G \subseteq Y_1$. We show that the rule r_1 and its supports can be derived from the rule r_2 and its supports. From $\gamma X_1 = \gamma Y_1 = \mathcal{C}$ and $\gamma G = \mathcal{C}$, we then have $suppr_1 = supp\gamma X_1 = supp\gamma Y_1 = supp\mathcal{C} = suppr_2$, and deduce that $M_{GK}r_1 = M_{GK}r_2$. This explains that r_1 can be derived from r_2, and is a redundant rule of r_2, so it's pruned in CBE base.*

ii Let $r_2 : G \rightarrow \mathcal{C} \backslash G \in CBE$, we then have $G \in \mathcal{G}_\mathcal{C}$ and $\mathcal{C} \in \mathcal{FC}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow Y_3 \backslash X_3 \in CBE$ such as $suppr_3 = suppr_2$, $M_{GK}r_3 = M_{GK}r_2$, $X_3 \subseteq G$ and $\mathcal{C} \subseteq Y_3$. If $X_3 \subseteq G$, we then have $\gamma X_3 \subseteq \gamma G = \mathcal{C}$. We deduce that $X_3 \notin \mathcal{G}_\mathcal{C} \Rightarrow r_3 \notin CBE$. If $\mathcal{C} \subseteq Y_3$, we then have $\mathcal{C} = \gamma\mathcal{C} = \gamma G \subset Y_3 = \gamma Y_3 \Rightarrow G \notin \mathcal{G}_{Y_3}$. In other words, r_2 is non-redundant. This proves that CBE is a non-redundant base.

Definition 6 (CBA Basis) *Let $\mathcal{FC}$ be the set of frequent closed. For each $\mathcal{C} \in \mathcal{FC}$, let $\mathcal{G}_{\mathcal{C}}$ be the set of generators of $\mathcal{C}$. Consider $0 < \alpha \leqslant 1$, we have:*

$$CBA = \{G \rightarrow \mathcal{C}\backslash G | G, \mathcal{C} \in \mathcal{G}_{\gamma G} \times \mathcal{FC}, \gamma G \subset \mathcal{C}, P\mathcal{C}' | G' > P\mathcal{C}', p_{G\mathcal{C}} \geqslant 1 - \alpha\} \tag{10}$$

Theorem 3 *(i) All valid positive approximate association rules, their supports and M_{GK}, can be derived from the rules of CBA. (ii) All association rules in the CBA basis are non-redundant approximate association rules.*

Proof 2 *i Let $r_1 : X_1 \rightarrow Y_1\backslash X_1 \in CBA$ such that $X_1 \subset Y_1$. For any X_1 and Y_1, there is a generator G_1 such that $G_1 \subset X_1 \subseteq \gamma X_1 = \gamma G_1$ and a generator G_2 such that $G_2 \subset Y_1 \subseteq \gamma Y_1 = \gamma G_2$. Since $X_1 \subset Y_1$, we have $X_1 \subseteq \gamma G_1 \subset Y_1 \subseteq \gamma G_2$ and the rule $r_2 : G_1 \rightarrow \gamma G_2\backslash G_1 \in CBA$. We show that r_1 can be derived from r_2. Since $G_1 \subset X_1 \subseteq \gamma X_1 = \gamma G_1$ and $G_2 \subset Y_1 \subseteq \gamma Y_1 = \gamma G_2$, we have $suppG_1 = suppX_1$ and $suppG_2 = suppY_1 = supp\gamma G_2$. This gives that $suppr_1 = suppr_2$ and $M_{GK}r_1 = M_{GK}r_2$, in other words, r_1 can be derived from r_2 and therefore, r_1 is a redundant rule of r_2.*

ii Let $r_2 : G \rightarrow \mathcal{C}\backslash G \in CBA$, we then have $\mathcal{C} \in \mathcal{FC}$ and $G \in \mathcal{G}_{\mathcal{C}}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow Y_3\backslash X_3 \in CBA$ such as $suppr_3 = suppr_2$, $M_{GK}r_3 = M_{GK}r_2$, $X_3 \subseteq G$ and $\mathcal{C} \subseteq Y_3$. If $X_3 \subseteq G$, we then have $\gamma X_3 \subset \gamma G = \mathcal{C} \Rightarrow X_3 \notin \mathcal{G}_C$. If $\mathcal{C} \subseteq Y_3$, we then have $\mathcal{C} = \gamma\mathcal{C} \subset Y_3 = \gamma Y_3$. As result, $G \notin \mathcal{G}_{Y_3} \Rightarrow r_3 \notin CBA$, in other words, r_2 is a non-redundant rule. This proves that CBA is a non-redundant base.

Definition 7 (CBE^- Basis) *Let $\mathcal{MFC}$ be the set of maximal frequent itemsets, $\overline{\mathcal{F}}_{\text{MIN}}$ the set of minimal infrequent on database $\mathcal{B}$. For each $\mathcal{M} \in \mathcal{MFC}$, let $\mathcal{G}_{\mathcal{M}}$ be the set of minimal generators of $\mathcal{M}$, we have:*

$$CBE^- = \{G \rightarrow \overline{y} \mid G \in \mathcal{G}_{\mathcal{M}},\ \mathcal{M} \in \mathcal{MFC},\ y \in \overline{\mathcal{F}}_{\text{MIN}}\} \tag{11}$$

Theorem 4 *(i) All valid negative exact association rules, their supports and M_{GK}, can be derived from the rules of the CBE^- basis. (ii) All association rules in the CBE^- basis are non-redudant negative exact association rules.*

Proof 3 *i Let $r_1 : X_1 \rightarrow \overline{Y_1}\backslash X_1 \in CBE^-$ such that $X_1 \subset \overline{Y_1} \subseteq \mathcal{M}$ where $\mathcal{M} \in \mathcal{MFC}$. Since $M_{GK}r_1 = 1$, we have $X_1 \cong \overline{Y}_1 \Rightarrow suppX_1 = supp\overline{Y_1}$. Since $suppX_1 = supp\overline{Y_1}$, we have $supp\gamma X_1 \cup \overline{Y_1} = supp\gamma X_1 = supp\gamma\overline{Y_1} \Rightarrow \gamma X_1 \cup \overline{Y}_1 = \gamma X_1 = \gamma\overline{Y_1} = \mathcal{M}$ a. Obviously, $\exists r_2 : G \rightarrow \overline{y}\backslash G \in CBE^-$ such that $G \in \mathcal{G}_{\mathcal{M}}$ for which $G \subseteq X_1$ and $G \subseteq \overline{Y_1}$, and thus $G \subseteq \overline{y}$ (by Definition 7). We show that the rule r_1 can be derived from r_2. Since $r_2 : G \rightarrow \overline{y}\backslash G \in CBE^-$, we have $suppG \cup \overline{y} = suppG$. From $suppG \cup \overline{y} = suppG$, we have $supp\gamma G \cup \overline{y} = supp\gamma G = supp\gamma\overline{y} \Rightarrow \gamma G \cup \overline{y} = \gamma G = \gamma\overline{y} = \mathcal{M}$ a'. From relations a and a', we have $\gamma G \cup \overline{y} = \gamma X_1 \cup \overline{Y_1} \Leftrightarrow suppr_1 = suppr_2$. Since $G \subseteq X_1 \subset \overline{Y_1} \subset \overline{y} \subseteq \gamma G = \mathcal{M}$, we have $suppG = suppX_1 = supp\overline{Y_1} = supp\overline{y} = supp\mathcal{M} \Rightarrow M_{GK}r_1 = M_{GK}r_2$. These results explain that r_1 can be derived from r_2, and is a redundant rule w.r.t r_2.*

ii Let $r_2 : G \rightarrow \overline{y}\backslash G \in CBE^-$, i.e. $G \in \mathcal{G}_{\mathcal{M}}$ and $y \in \overline{\mathcal{F}}_{\text{MIN}}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow \overline{Y}_3\backslash X_3 \in CBE^-$ such as $suppr_3 = suppr_2$, $M_{GK}r_3 = M_{GK}r_2$, $X_3 \subseteq G$ and $\overline{y} \subseteq \overline{Y_3}$. If $X_3 \subseteq G$, we then have $\gamma X_3 \subseteq \gamma G \subset \gamma\overline{y} = \mathcal{M}$. We deduce that $X_3 \notin \mathcal{G}_{\mathcal{M}}$ and conclude that $r_3 \notin CBE^-$. If $\overline{y} \subseteq \overline{Y_3}$, we then have $\gamma G \subset \gamma\overline{y} \subseteq \gamma\overline{Y_3} = \mathcal{M}$. We deduce that $G \notin \mathcal{G}_{\overline{Y}_3}$ and conclude that $r_3 \notin CBE^-$. This implies that r_2 is a non-redundant rule, and proves that CBE^- is a non-redundant base.

Definition 8 (CBA^- Basis) *Let $\mathcal{FC}$ be the set of frequent closeds. For each $\mathcal{C} \in \mathcal{FC}$, let $\mathcal{G}_{\mathcal{C}}$ be the set of generators of $\mathcal{C}$. Consider $0 < \alpha \leqslant 1$, we have:*

$$CBA^- = \{G \rightarrow \overline{g} | G, g \in \mathcal{G}_{\gamma G} \times \mathcal{G}_{\gamma g}, \gamma G \subsetneq \gamma g, PG' | \overline{g'} > P\overline{g'}, p_{G\overline{g}} \geqslant 1 - \alpha\} \tag{12}$$

Theorem 5 *(i) All valid negative approximate association rules, their supports and M_{GK}, can be derived from the rules of CBA^-. (ii) All association rules in the CBA^- are non-redudant negative approximate association rules.*

Proof 4 *i Let $r_1 : X_1 \rightarrow \overline{Y_1}\backslash X_1 \in CBA^-$ with $X_1 \subset \overline{Y_1}$. For any frequent X_1 and Y_1, there is a generator G_1 such that $G_1 \subset X_1 \subseteq \gamma X_1 = \gamma G_1$ and a generator G_2 such that $G_2 \subset Y_1 \subseteq \gamma Y_1 = \gamma G_2$. Since $X_1 \subset \overline{Y_1}$, we have $X_1 \subseteq \gamma G_1 \subset \overline{Y_1} \subset \overline{G_2} \subseteq \gamma\overline{Y_1} = \gamma\overline{G_2}$. Obviously, $\exists r_2 : G_1 \rightarrow \overline{G_2}\backslash G_1 \in CBA^-$ such that $\gamma G \subsetneq \gamma g$ (by Definition 8). We show that r_1 can be derived from r_2. From $G_1 \subset X_1 \subseteq \gamma G_1$ and $G_2 \subset Y_1 \subseteq \gamma G_2$, we then have $G_1 \cong X_1$ and $\overline{Y_1} \cong \overline{G_2} \Rightarrow suppX_1 \cup \overline{Y_1} = suppG_1 \cup \overline{G_2}$ and $M_{GK}X_1 \rightarrow \overline{Y_1} = M_{GK}G_1 \rightarrow \overline{G_2}$. This explains that r_1 can be derived from r_2, and is a redundant rule w.r.t. r_2. ii Let $r_2 : G \rightarrow \overline{g}\backslash G \in CBA^-$, i.e. $G \in \mathcal{G}_C$ and $g \in \mathcal{G}_\mathcal{C}$ such that $\gamma G \subsetneq \gamma g$ (i.e. $C \subsetneq \mathcal{C}$). We demonstrate that there is no other rule $r_3 : X_3 \rightarrow \overline{Y_3}\backslash X_3 \in CBA^-$ such that $suppr_3 = suppr_2$, $M_{GK}r_3 = M_{GK}r_2$, $X_3 \subset G$ and $\overline{Y_3} \supset \overline{g}$. If $X_3 \subset G$, we then have $\gamma X_3 \subset \gamma G = C \Rightarrow X_3 \notin \mathcal{G}_C$. Since $X_3 \subset G$, we have $suppX_3 > suppG \Rightarrow M_{GK}r_3 < M_{GK}r_2$. If $\overline{g} \subset \overline{Y_3}$, we then have $supp\overline{g} > supp\overline{Y_3} \Rightarrow M_{GK}r_2 > M_{GK}r_3$. This means that r_2 is a non-redundant rule, and proves that CBE^- is a non-redundant base.*

4.2.5 CONCISE algorithm.

CONCISE is composed of three algorithms (Algo. 1, Algo. 2,Algo. 3). The principal procedure (Algo. 3) takes as input $\mathcal{G}_\mathcal{C}$, $\mathcal{FC}$, $\mathcal{MFC}$, $\overline{\mathcal{F}}_{\text{MIN}}$, $minsup$ and α. It returns all non-redundant PNARs.

Require: $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$, $\mathcal{MFC}$, $\overline{\mathcal{F}}_{\text{MIN}}$, $minsup$ and α.
Ensure: $\mathcal{CB}$, A Concise Base of Non-Redundant PNARs.

$\mathcal{CB} = \emptyset$;
for all ($\mathcal{C} \in \mathcal{FC}$) **do**
  for all ($G \in \mathcal{G}_\mathcal{C}$) **do**
    if ($P\mathcal{C}'|G' > P\mathcal{C}'$) **then**
      if ($\gamma G = \mathcal{C}$) **then**
        if ($G \neq \gamma G$ & $suppG \cup \mathcal{C} \geq minsup$) **then**
          $\mathcal{CB} \leftarrow \mathcal{CB} \cup \{G \rightarrow \mathcal{C}\backslash G\}$; /* CBE Basis */
        end if
      else if ($\gamma G \subset \mathcal{C}$) **then**
        if ($suppG \cup \mathcal{C} \geq minsup$ & $p_{GC} \geqslant 1 - \alpha$) **then**
          $\mathcal{CB} \leftarrow \mathcal{CB} \cup \{G \rightarrow \mathcal{C}\backslash G\}$; /* CBA Basis */
        end if
      end if
    else if ($P\mathcal{C}'|G' < P\mathcal{C}'$) **then**
      for all ($\mathcal{M} \in \mathcal{MFC}$) **do**
        for all ($G \in \mathcal{G}_\mathcal{M}$) **do**
          for all ($y \in \overline{\mathcal{F}}_{\text{MIN}}$) **do**
            if ($suppG \cup \overline{y} \geq minsup$) **then**
              $\mathcal{CB} \leftarrow \mathcal{CB} \cup \{G \rightarrow \overline{y}\backslash G\}$; /* CBE^- Basis */
            end if
          end for
        end for
      end for
      for all ($g \in \mathcal{G}_{\gamma g}|\gamma G \subsetneq \gamma g$ & $Pg'|G' < Pg'$) **do**
        if ($suppG \cup \overline{g} \geq minsup$ & $p_{G\overline{g}} \geqslant 1 - \alpha$) **then**
          $\mathcal{CB} \leftarrow \mathcal{CB} \cup \{G \rightarrow \overline{g}\backslash G\}$; /* Base CBA^- */
        end if
      end for
    end if
  end for
end for
return $\mathcal{CB}$

Algorithm 3: GENERATING NON-REDUNDANT ASSOCIATION RULES

5 Experimental evaluation

We evaluate CONCISE with two comparable baseline approaches Pasquier's approach [14] and Feno's approach [7]. All algorithms are implemented in R. All the experiments are run on a PC Core i3-2350M with 4CPUs and 4GB memory on the Windows 7. We compare their number of valid rules and computational costs on different databases (cf. Table 3): T10I4D100K[1], T20I6D100K (cf. footnote 1), C20D10K[2] and MUSHROOMS (cf. footnote 2). We make CONCISE and Feno's approach to follow the same constraint $\alpha = 5\%$. For Pasquier's approach, consider a minimal confidence $minconf = 80\%$. The number of extracted rules for the three algorithms, by varying the $minsup$, is shown in Table 4. For this, E (resp. A) indicates the positive exact (resp. approximate) rules. E^- (resp. A^-) indicates the negative exact (resp. approximate) rules. We also denote by "-" a subset which could not generated. We observe that no negative association rules are generated by Pasquier's approach. For each algorithm, no E and A^- are generated on T10I4D100K when $minsup \leq 30\%$. The reason is that all frequent are closed itemsets. On other databases, Feno's approach represents a number smaller than Pasquier's approach and CONCISE. The explanation is that Feno's approach uses the set of pseudo-closed [13] which returns a reduced number of frequent itemsets and thus, it is the same for number of rules generated, but it's not informative. Whereas Pasquier's approach and CONCISE algorithm generate the more informative non-redundant association rules.

Table 3. Data characteristics

Database	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{I}\rvert$	Avg. size
T10I4D100K	100 000	1 000	10
T20I6D100K	100 000	1 000	20
C20D10K	10 000	386	20
MUSHROOMS	8 416	128	23

Table 4. Number of all valide non-redundant association rules

Dataset	$minsup$	Pasquier's approach				Feno's approach				CONCISE			
		$\lvert E\rvert$	$\lvert A\rvert$	$\lvert E^-\rvert$	$\lvert A^-\rvert$	$\lvert E\rvert$	$\lvert E^-\rvert$	$\lvert A\rvert$	$\lvert A^-\rvert$	$\lvert E\rvert$	$\lvert E^-\rvert$	$\lvert A\rvert$	$\lvert A^-\rvert$
T10I4D100K	10%	0	11625	-	-	0	0	10555	1256	0	0	725	52
	20%	0	8545	-	-	0	0	6656	1058	0	0	545	34
	30%	0	3555	-	-	0	0	2785	954	0	0	355	25
T20I6D100K	10%	115	71324	-	-	95	98	51899	3897	115	103	1804	56
	20%	76	57336	-	-	66	91	35560	2705	76	95	1403	38
	30%	58	45684	-	-	43	63	21784	1887	58	63	1175	27
C20D10K	10%	1125	33950	-	-	975	255	28588	11705	1125	285	1856	182
	20%	997	23821	-	-	657	135	19582	8789	997	185	1453	123
	30%	967	18899	-	-	567	98	11581	4800	967	101	1221	97
MUSHROOMS	10%	958	4465	-	-	758	289	3850	3887	958	304	1540	89
	20%	663	3354	-	-	554	178	2144	2845	663	198	1100	78
	30%	543	2961	-	-	444	109	1140	1987	543	115	998	39

On dense databases (C20D10K and MUSHROOMS), CONCISE algorithm is much more selective than Pasquier's and Feno's approaches for all $minsup$. For example, on C20D10K database and less $minsup = 1\%$, Pasquier's (resp. Feno's) approch contains 33950 (resp. 28588) positive approximate rules as showed in Table 4, while the CONCISE contains 1856 positive approximate rules; this gives the reduction ratio 94.5% and 93.51% respectively. In this case, 32094 (resp. 26732) positive approximate rules can be deduced either from the Pasquier's (resp. Feno's) approach or from the CONCISE algorithm. The main reason is associated to the different techniques to prune both UARs and redundant association rules.

We present in the following the execution times of CONCISE compared to those existing. However, this comparison is still very difficult, for several reasons. First, Feno's approach is not comparable to CONCISE, because it ignores the big phase for generating $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$ and $\mathcal{MFC}$. Pasquier's approach could not generate the negative rules. We partialy compare CONCISE and Pasquier's approach on execution times of E and A. The results will be represented in Fig. 2 by varying the $minsup$ at fixed $\alpha = 0.05$ and $minconf = 0.6$. On sparse databases (T10I4D100K and T20I6D100K), CONCISE and Pasquier's approach are almost identical for positive exact rules E for all $minsup$ (cf. Fig. 2a and 2b). On approximate rules A, it is very obvious that CONCISE is better than Pasquier's approach (cf. Fig. 2a, 2b). The explanation is that all frequent are closed itemsets, that complicates the task of Pasquier's approach who performs more operations than CONCISE for counting frequent closed itemsets.

[1] http://www.almaden.ibm.com/cs/quest/syndata.html

[2] http://kdd.ics.uci.edu/

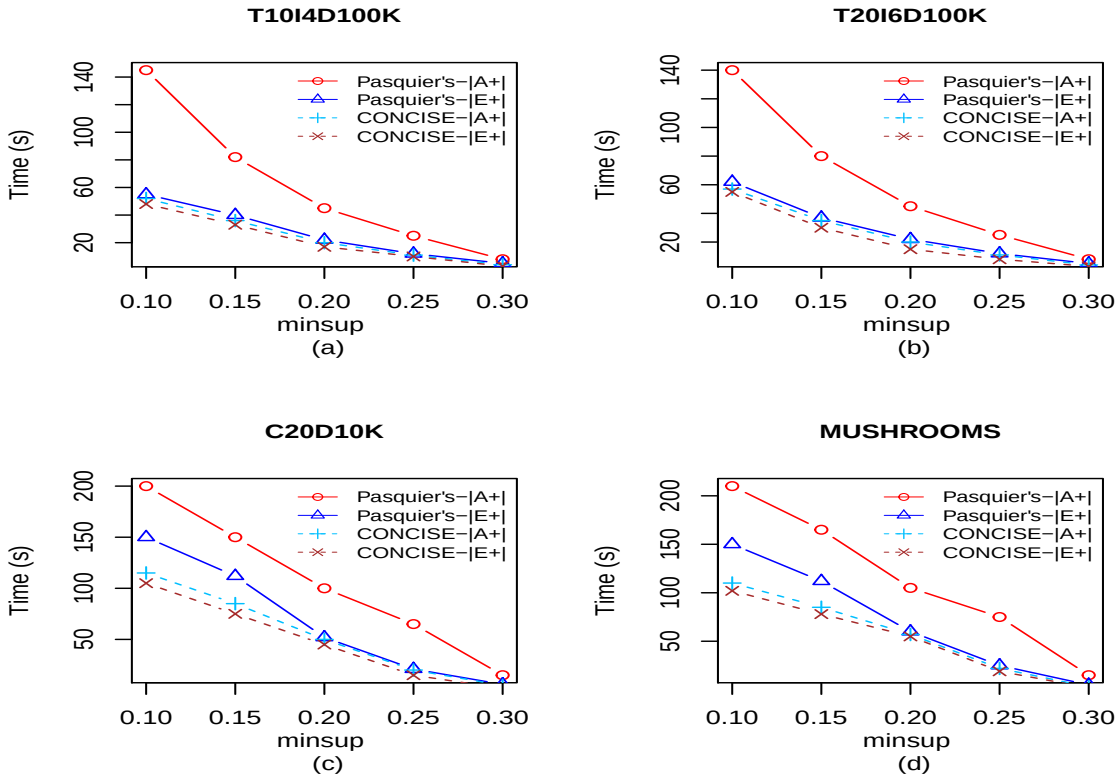


Figure 2. Response times by varying $minsup$ at fixed $\alpha = 0.05$ and $minconf = 0.6$

On dense databases (C20D10K and MUSHROOMS), CONCISE algorithm leads to significant average time compared to Pasquier's approach for all $minsup$ (cf. Figure 2c and Figure 2d). The main reason is associated to the technique for pruning search space of valid positive/negative association rules. Thanks to the different optimizations as defined on Subsection 4.2.3, CONCISE algorithm can reduce considerable amount the execution time for all minimum support threshold $minsup$, it is not the case for Pasquier's approach. The latter obtains the least performance. This is mainly due to the lack of techniques for pruning the search space for valid association rules. This obviously affects its execution time. However, Pasquier's approach joins CONCISE algorithm for the E execution times, when $minsup$ is 20% to 30%.

6 Conclusion

In this paper, we presented and evaluated a condensed representation for association rules. It is an efficient method for representing non-redundant positive and negative rules. We theoretically proved and experimentally confirmed that our approach can eliminate considerable amount of redundancy and uninteresting rules. Compared to the Pasquier's and Feno's approaches, our approach is not only a concise but also a lossless extraction of positive and negative association rules. From this, all informative association rules can be deduced. The perspective would be to extend this proposal in Graphs and Classification problems.

References

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

1. R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules". In Proceedings of 20th VLDB Conference, pp. 487–499. Santiago, Chile (1994).
2. Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, "Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets". In CL'2000 international conference Computational Logic, pp. 972–986 (2000).
3. P. Bemarisika, and A. Totohasina, An Informative Base of Positive and Negative Association Rules on Big Data. In Proc. of BigData, pp. 2428–2437 (2019).
4. L. Cao, X. Dong, and Z. Zheng, "E-NSP: Efficient negative sequential pattern mining". In Artificial Intelligence, pp. 156–182 (2016).
5. X. Dong, H. Hao, L. Zhao, and T. Xu, "An efficient method for pruning redundant negative and positive association rules". In NEUCOM 2018. `https://doi.org/10.1016/j.neucom.2018.09.108` (2018).
6. X. Dong, G. Yongshun, and L. Cao, "F-NSP: A Fast Negative Sequential Patterns Mining Method with Self-adaption Data Storage Strategy". Pattern Rec.(2018).
7. D. Feno, J. Diatta, and A. Totohasina, "Galois Lattices and Based for M_{GK}-valid Association Rules". In Ben Yahia et al. (Eds.), CLA 2006, pp. 186–197, (2006).
8. B. Ganter, and R. Wille, "Formal concept analysis: Mathematical foundations". In Springer Verlag (1999).
9. R. Gras, J-C. Régnier, C. Marinica, and F. Guillet, "L'ASI, Méthode exploratoire et confirmatoire la recherche de causalités". In Cépadus Editions, pp. 11–40 (2013).
10. J.L. Guigues, and V. Duquenne, "Familles minimales d'implications informatives résultant d'un tableau de donnés binaires". Maths et Sci. Humaines, 5–18 (1986).
11. T. Guyet, and R. Quiniou, "NegPSpan: efficient extraction of negative sequential patterns with embedding constraints". `https://hal.inria.fr/hal-01743975v2` (2018).
12. M. Mannila, and H. Toivonen, "Levelwise Search and Borders of Theories in Knowledge Discovery". In Data Mining Knowledge Discovery, pp. 241–258 (1997).
13. N. Pasquier, R. Taouil, and Y. Bastide, G. Stumme, and L. Lakhal, "Generating a condensed representation for association rules". In J. of Intell. Info. Syst., pp. 29–60 (2005).
14. N. Pasquier, "Frequent Closed Itemsets Based Condensed Representations for Association Rules". In Tech. for Eff. Knowl. Extraction, pp. 248–273 (2009).
15. T. Xu, T. Li, and X. Dong, "Efficient High Utility Negative Sequential Patterns Mining in Smart Campusy". In IEEE Access, pp. 23839–23846, (2018).
16. Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules". In Data and Knowledge Engineering, pp. 555–575 (2011).

24th International Conference on Business Information Systems (BIS 2021)

Artificial Intelligence

https://doi.org/10.52825/bis.v1i.57

Published: 02 July 2021

Supporting an Expert-centric Process of New Product Introduction with Statistical Machine Learning

Shima Zahmatkesh[1][https://orcid.org/0000-0002-7832-0288], Alessio Bernardo[1][https://orcid.org/0000-0002-3492-0345], Emanuele Falzone[1][https://orcid.org/0000-0002-2699-2357], Edgardo Di Nicola Carena[2][https://orcid.org/0000-0001-6856-6957], and Emanuele Della Valle[1][https://orcid.org/0000-0002-5176-5885]

[1]DEIB, Politecnico di Milano, Italy

[2]Abstract s.r.l., Milano, Italy

Abstract. Industries that sell products with short-term or seasonal life cycles must regularly introduce new products. Forecasting the demand for New Product Introduction (NPI) can be challenging due to the fluctuations of many factors such as trend, seasonality, or other external and unpredictable phenomena (e.g., COVID-19 pandemic). Traditionally, NPI is an expert-centric process. This paper presents a study on automating the forecast of NPI demands using statistical Machine Learning (namely, Gradient Boosting and XGBoost). We show how to overcome shortcomings of the traditional data preparation that underpins the manual process. Moreover, we illustrate the role of cross-validation techniques for the hyper-parameter tuning and the validation of the models. Finally, we provide empirical evidence that statistical Machine Learning can forecast NPI demand better than experts.

Keywords: Demand Forecasting, New Product Introduction, Statistical Machine Learning, Gradient Boosting, XGBoost

Introduction

In several industries (e.g., Fashion), the period in which the products are saleable is likely to be short and seasonal. Since the products of these industries are replaced every new season by new ones, there are little relevant historical data available. Moreover, the industries may carry on many products with various futures corresponding to stock-keeping units (SKUs). Demand for these products is hardly stable or linear. It may be influenced by the fluctuations of many factors like weather conditions, holidays, marketing strategy, fashion trends, films, or even by celebrities and footballers. These factors make it challenging to forecast the demands for New Product Introductions (NPI). The major part of the companies uses manual efforts to predict the demands for NPI, the sales for existing manufactured goods, and the budget quantity or to benchmark the various products among them. Being manual processes, they can lead to inaccurate predictions, and so, failing in forecasting supply chain demands can cause under-staffing or over-staffing, incorrect operation budgeting, loss of credibility, failure in customer experience, economic loss in expenses, and waste of unsold products/services.

Several studies investigate NPI demand forecast in the literature, but most are not tailored for short life cycle environments. Most of the researches are based on diffusion theory, mainly including Bass [1] and Norton[2] models. Several methods have been developed based on these models, and different approaches are proposed in which analogical approaches are the

most important category. In this group of approaches, the assumption is that the diffusion patterns of the new products are similar to the analog products.

However, these approaches have some limitations. For example, experts define the similarity between products, and they often struggle to find a suitable benchmark. Some studies tried to overcome these limitations by using statistical Machine Learning [3], and Deep Learning [4] approaches focusing on optimizing the parameters of the Bass model.

In this paper, we aim to improve the judgment of human experts in forecasting NPI by utilizing statistical Machine Learning approaches (ML). In particular, we exploit statistical ML methods to automatically find the similarities between products and use them for forecasting NPI demands. In this way, the process no longer depends on expert judgments for the selection of similar products. As a result, we improve the forecast accuracy beyond the traditional methods, helping the short life cycle industries in making fast adaptations to their products to compete successfully in the market.

To this extent, we investigate the following research question:

RQ. *Utilizing the statistical Machine Learning, is it possible to improve the demand forecast for New Product Introduction done by the experts?*

In more details, the main contributions of this paper are:

- A characteristics analysis of a typical dataset collected in a short life-cycle industry;
- A data exploration task focusing on understanding and trying to enrich the data;
- The proposal of seven different approaches for predicting sales quantity using the Gradient Boosting and XGBoost statistical ML models;
- The exploration of the sensitivity to different cross-validation methods, hyper-parameters values, and feature encoding options; and
- A comparison, using the MAPE metric, of the results achieved by the proposed approaches with both a baseline and manual predictions, positively answering to **RQ**.

The remainder of the paper is structured as follows. Section Background presents the details of the statistical ML techniques used. Section NPI predictive analytics introduces the dataset used in this study and discusses the pre-processing analysis and the details of proposed approaches. Section Experimental Settings introduces the hypothesis tested and focuses on the design of the experiments that are carried out within this work, while Section Results and Discussion shows and discusses the results achieved. At the end, Section Related Work reviews the related work, and Section Conclusion concludes the paper.

Background

To solve the NPI problem and predict the future sale quantities, we utilized two different statistical Machine Learning models: Gradient Boosting [5] and XGBoost [6]. We introduce them in the following two sections. Moreover, to evaluate our models, we used two different cross-validation techniques. We present them in the last section.

Gradient Boosting

Gradient Boosting [5] refers to a class of ensemble ML algorithms that can be used for regression predictive modeling problems. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, ”gradient boosting,” as the loss gradient is minimized as the model is fit, much like a Neural Network in Deep Learning. One way to produce a weighted combination of ensembles that optimizes the cost function is by gradient descent in function space. After calculating the loss, we must add a tree to the model

that reduces the loss (i.e., follow the gradient) to perform the gradient descent procedure. We do this by parameterizing the tree, then modifying the tree's parameters and moving in the right direction by reducing the residual loss. The output for the new tree is then added to the output of the existing sequence of trees to correct or improve the final output of the model.

Naïve gradient boosting is a greedy algorithm and can overfit the training dataset quickly. However, it can benefit from regularization methods that penalize various parts of the algorithm and generally improve the algorithm's performance by reducing overfitting. There are three types of enhancements to naïve gradient boosting that can improve performance:

- *Tree constraints*: the weak learners have skill but remain weak. There are several ways in which the trees can be constrained, such as the number of trees used in the ensemble, the depth of each tree, the minimum number of samples required to split an internal node, or the minimum number of samples required to be at a leaf node.
- *Weighted updates*: the predictions of each tree are added together sequentially. The contribution of each tree to this sum can be weighted to slow down the algorithm's learning. This weighting is called a learning rate.
- *Random sampling*: a considerable insight into bagging ensembles and random forests was allowing trees to be greedily created from sub-samples of the training dataset. This approach also reduces the correlation between the trees in the sequence.

There are some advantages in using the Gradient Boosting algorithm:

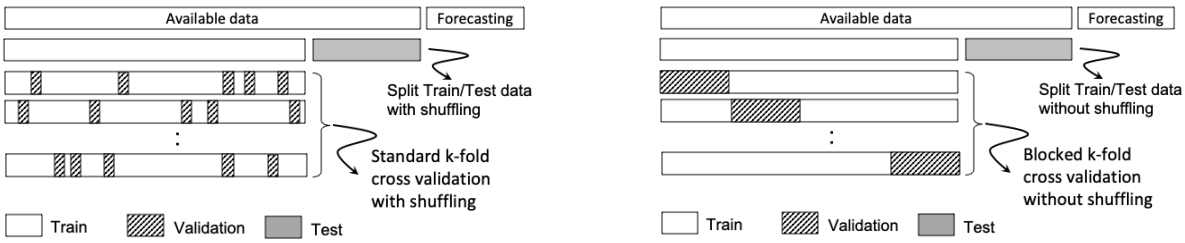
- *Better accuracy*: Gradient Boosting, compared with other regression techniques like Linear Regression, generally provides better accuracy. This is why it is used in most online hackathons and competitions.
- *Less pre-processing*: data pre-processing is one of the vital steps in ML workflow because it affects the model accuracy. However, Gradient Boosting requires minimal data pre-processing, which helps in implementing this model faster with lesser complexity.
- *Higher flexibility*: Gradient Boosting offers a wide range of hyper-parameters and loss functions. This makes the model flexible and usable for solving a wide variety of problems.
- *Missing data*: Gradient Boosting handles missing data[1] on its own. During the tree building phase, splitting decisions for a node are decided by minimizing the loss function and treating missing values as a separate category that can go either left or right.

XGBoost

eXtreme Gradient Boosting [6] is an optimized and distributed version of the Gradient Boosting algorithm. It improves upon the base Gradient Boosting framework through systems optimization and algorithmic enhancements. In particular, the system is optimized as follows.

- *Parallelization*: XGBoost approaches the process of sequential tree building using parallelized implementation.
- *Tree pruning*: the stopping criterion for tree splitting within the Gradient Boosting framework is greedy and depends on the negative loss criterion at the split point. XGBoost uses the *max-depth* parameter as specified instead of criterion first and starts pruning trees backward. This *depth-first* approach improves computational performance significantly.
- *Hardware optimization*: this algorithm has been designed to make efficient use of hardware resources. This is accomplished by cache awareness by allocating internal buffers in each thread to store gradient statistics. Further enhancements such as out-of-core computing optimize available disk space while handling big data-frames that do not fit into memory.

[1] In NPI forecasting, missing data is a severe problem that traditionally experts solve by selecting products similar to the one we want to predict demand for.



(a) Standard k-fold CV

(b) Blocked k-fold CV

Figure 1. Cross-validation approaches

While, algorithms are enhanced as follows.

- *Regularization*: it penalizes more complex models through both LASSO (L1) and Ridge (L2) regularization to prevent overfitting.
- *Sparsity awareness*: XGBoost naturally admits sparse features for inputs by automatically learning the best missing values depending on training loss and handles different types of sparsity patterns in the data more efficiently.
- *Weighted quantile sketch*: XGBoost employs the distributed weighted quantile sketch algorithm to find the optimal split points among weighted datasets effectively.
- Cross-validation: the algorithm comes with a built-in cross-validation method at each iteration, taking away the need to explicitly program this search and to specify the exact number of boosting iterations required in a single run.

Cross-Validation

One way to tune ML models hyperparameters is to split data into train and validation sets (keeping the test set only for the final evaluation). The models can be trained on a smaller train set, and the evaluation set can be used for evaluating the models. The standard k-fold cross-validation [7] technique, shown in Fig. 1(a), firstly randomly splits the data into k distinct subsets, called folds, and then it trains and evaluates the model k times, each time selecting one of the folds as the validation set and the rest of them as the training set. Then, it saves the evaluation score, and it discards the model. The scores are averaged over the rounds to give an estimate of the model's predictive performance. The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias [8] and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset).

However, in the literature, some authors have raised some theoretical problems about using the k-fold cross-validation approach in time series prediction[2]. This led to the introduction of new methods to overcome these problems. They can be divided into three categories: 1) cross-validation based on the last block such as forward validation [9], [10], 2) cross-validation with omission of dependent data [7], and 3) cross-validation with blocked subsets [11].

In this study, besides the standard k-fold cross-validation, we also test the blocked k-fold cross-validation approach [11], shown in Fig. 1(b), in which the k-fold cross-validation is done without shuffling the training set at the beginning, i.e., the training set is temporally sorted and the order of products on time is preserved. Unlike the standard k-fold CV, this new version, avoiding the initial shuffling and using blocks of data contiguous in time as validation sets, does not use obvious temporal dependencies in the short term.

[2]Indeed, the typical datasets used in NPI demand forecast are time-series.

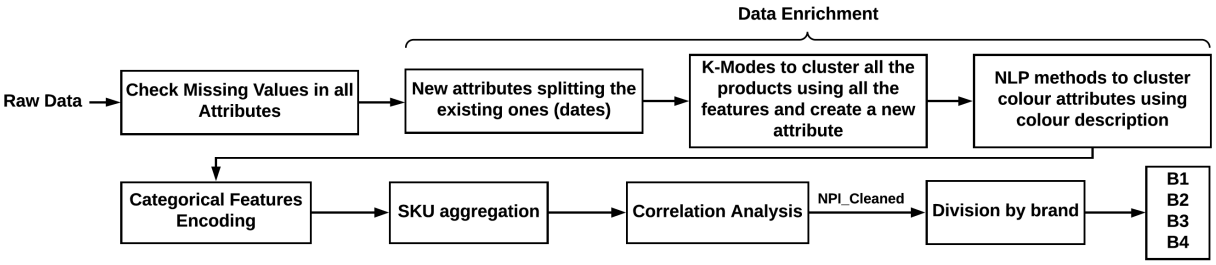


Figure 2. Data preparation pipeline

New Product Introduction predictive analytics

In this section, we present a dataset of past NPI sale quantities, and we describe the proposed solution for the NPI problem in detail. We divide it into two parts: 1) the data ingestion and preparation phase, and 2) the modeling phase.

NPI dataset

NPI dataset contains $30,750$ products introduced (or to be introduced) on the fashion market from 2017 to 2019. It aims at estimating new products' number of sales before introducing them on the market in 2019. From a business perspective, the company shall use the estimation to decide if it is worth or not introducing those products on the market. Each product has *67* features, and the label (ORDER-QTY) represents the number of sales in the first three months after the introduction. The features are divided into four groups:

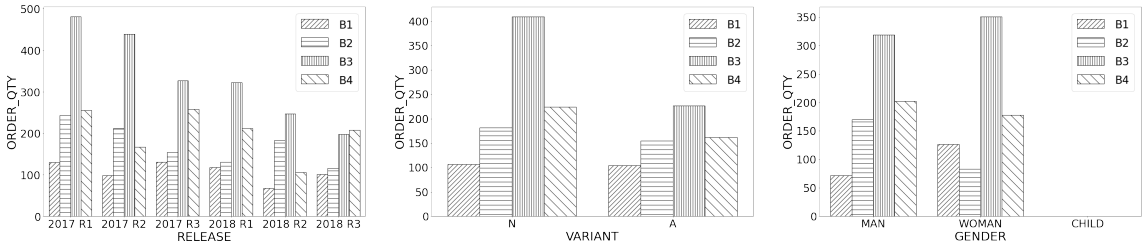
- features related to the product's characteristic, i.e., brand, model, color, material, price,
- time-related features, i.e., main release, and first availability date,
- benchmark features representing similar products from past collections *manually selected by experts*, i.e., benchmark-1, benchmark-2, and
- aggregated features, i.e., number of models by brand release, number of colors by model, and number of sizes by model.

Moreover, there is also an attribute named BGT-QTY representing the manual estimations of the experts.

Data ingestion and preparation

Fig. 2 shows a series of operations done during the data ingestion and preparation phase. Starting from the previous section's dataset, we first performed a missing values analysis all over the attributes finding seven attributes (five of them are aggregated features) having a lot of missing values. So we replaced them with ND, standing for "not defined," in case of nominal features, with the mean of the not null values in case of numerical features, and with a meaningful date format, to avoid formatting errors, in case of date features.

Moreover, we studied the ORDER-QTY demands distribution w.r.t. the different features. Fig. 3 shows the ORDER-QTY demands divided for brand respect to the RELEASE, VARIANT and GENDER attributes. In particular, from Fig. 3(a), we can notice different sales trend for brands. In brand B3, there is a decreasing trend in sales, while in brand B1 the sales are stable across the years. In brand B4, the sales increase in the last quarters of the year and then they decrease, while in brand B2, the sales slightly decrease during the years quarters. Fig. 3(b) shows that, for all brands, there were introduced more new models than variants of already existing ones. In particular, the most introduced models are from brand B3, followed by brand B4. Finally, Fig. 3(c) shows that brands B3 and B4 sold similarly among men and women. Brand B2 sold more among men, while brand B1 sold more among women. The second task of the pipeline refers to a data enrichment operation. For example, we split the *main release*, and the *alternative release* date attributes into, respectively, main release year, main release



(a) ORDER-QTY for Release (b) ORDER-QTY for Variant (c) ORDER-QTY for Gender

Figure 3. The ORDER-QTY divided for BRAND respect to the RELEASE, VARIANT and GENDER attributes

quarter, alternative release year, and alternative release quarter. We did the same with the *first availability date week*, *benchmark launch 1* and *benchmark launch 2* attributes. After that, we also decided to add another attribute to group similar products into the same cluster. Since the dataset has mixed numerical and nominal features, we used the K-Modes [12] algorithm to create some product clusters, and, through the Elbow curve [13], we selected the exact number of clusters to use (17 in our case). The last step consisted of adding 3 other attributes that cluster the products based on some features' descriptions.

In the next step, we performed a nominal feature encoding. We used the label encoder as a classical approach to encoding the nominal features into numerical values. In many ML approaches, numerical encoding would not perform very well as the nominal features do not necessarily have the ordinal relationship introduced when assigning them a number. However, for the ML methods that we use, this approach does not affect the models' performances since they inherently perform very well on nominal features. However, in some of the experiments, we applied the one-hot encoding procedure (detailed in the next section) to avoid the ML algorithm to learn in-existing patterns out of the ordinal relationship. The next step aggregated the products by the *SKU* attribute that combines in one attribute the *model*, *size* and *colour*. Then, we performed a correlation analysis to know the most correlated features to the label. Unfortunately, we discovered that there were no highly correlated features to the label. For this reason, we used all the features to train and test the models. The result of all these tasks was the so-called *NPI-Cleaned* dataset. The last task split it by brand (B1, B2, B3, B4), creating four other datasets that will be all used to train and test the models.

Data Modeling

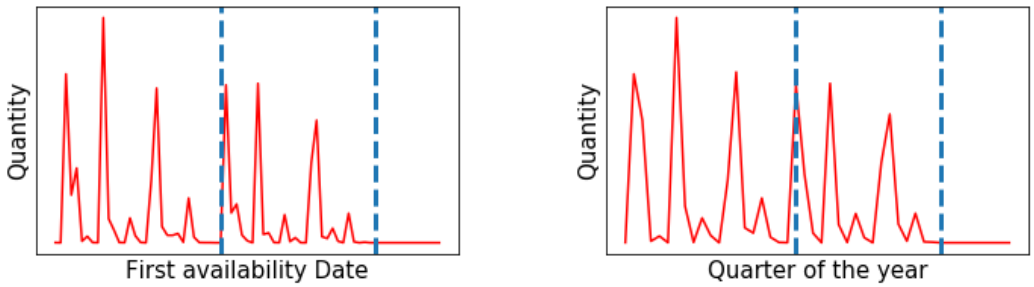
Starting from the five datasets created, we introduce seven approaches for each couple of learners (Gradient Boosting, XGBoost). They combine different cross-validation methods, hyper-parameters tuning values, and encoding. In the following, we introduce them in detail.

Baseline with cross-validation

Here, the standard 5-fold cross-validation approach over the train set was applied. So, we had 5 Gradient Boosting and 5 XGBoost models. Then, they were validated and tested on the respective validation and test sets. In the case of the two models trained on the *NPI-Cleaned* dataset, they were validated and tested, respectively, on the validation and test sets generated from the main dataset. Then the prediction results were divided by brand. Since this approach represents a first try that a typical data scientist would test, we used it as a baseline.

Partial one-hot encoding with K-fold CV (POH) and with blocked CV (POH-BCV)

In this approach, we applied some other feature engineering tasks w.r.t. the baseline. Firstly, we looked at the presence of the seasonality phenomena on the number of sales. Fig. 4 shows the number of products sold grouped by the *first availability date week*. Separating them by year (Fig. 4(a)), we can notice a sort of seasonality, but the peaks are not aligned over the



(a) ORDER-QTY grouped by the first availability date week

(b) ORDER-QTY grouped by a bi-weekly period

Figure 4. Seasonality phenomena in ORDER-QTY attribute

years. Instead, grouping them by every two weeks (Fig. 4(b)), we can notice that the peaks are more aligned over the years than before. So, we added a *bi-weekly* feature, and we removed all the other time-related features. Then, we also evaluated the benefits of using the *benchmarks* features selected by the experts. We inspected how often the NPI SKUs referred to the same benchmark SKUs, finding out that *Benchmark-1* is more present than *Benchmark-2* and that the experts often used the same few benchmarks, while a large number of benchmarks are used less often. We also inspected if there were any NPI SKUs that referred to older NPI SKUs as benchmark SKUs, finding out that only the $15 - 20\%$ of the products, later on, became benchmarks. It happened on average 15 times and at most 100 times. Moreover, the experts selected only a few products frequently or for a long period of time. Half of the products were seldom used, and still, others were never used. Our conclusion was that the *benchmarks* might be useless in the predictions, and they may even disturb the learning. So, we removed all the *benchmarks* features, and all the *aggregated* features, too. The last task was one-hot encoding the *categorical* features before training the two models. Since the major part of the categorical features had less than *10* distinct values, for this approach, we encoded only the features having more than *10* distinct values. At this point, we distinguished two approaches: POH and POH-BCV. In the former, we divided the datasets, and we applied the 5-fold cross-validation as in the baseline. Instead, in the latter, we used the 5-fold blocked cross-validation. Moreover, in both approaches, we explored a maximum tree depth between 6 and 12.

POH-BCV with one test set

This approach is based on the POH-BCV one, but instead of using the test and the validation sets, it uses only one test set. The whole training set was used to apply the 5-fold blocked cross-validation, so having more data points available during the training phase. The validation error is the mean of the errors during the 5-fold blocked cross-validation process.

Other approaches (POH-D4-6, OH, OH-D4-6)

We also tested three other approaches similar to the previously described ones: POH-D4-6 is the same as the POH approach, but it explores a maximum tree depth between 4 and 6; OH is based on the POH one, but it encodes all the categorical features, and OH-D4-6 is the same as the previous one, but it explores a maximum tree depth between 4 and 6.

Experimental Settings

This section firstly introduces the hypotheses to test, then, i) discusses the datasets splitting criteria applied, ii) proposes all the parameters used to train the models, iii) introduces the evaluation metric used for comparing the performances between the proposed approaches and the experts' prediction, and iv) describes the experimental environment.

Research Hypotheses

We formulated our hypotheses as follows:

- *Hp. 1*: Since the introduced Baseline with cross-validation approach is considered just as a baseline, the other approaches that use one-hot encoding, different types of cross-validations and perform more analysis on the data outperform baseline approach and generate more accurate predictions in terms of MAPE metric.
- *Hp. 2*: Applying statistical Machine Learning models, we can improve the forecasting accuracy with respect to the experts' prediction in terms of MAPE metric.

Splitting Criteria

In total, we have five datasets: one is the main dataset containing all the product sales, while the others are related to the specific brands (named B1- B4). We first discarded the 2019 data since they represent the products not already introduced on the market, so the products we need to predict the number of sales (forecasting). Then, we considered as the training set the data sold from 2017 R1 to 2018 R2 quarters and as the test set the data sold in the 2018 R3 quarter. Finally, we split the training set into the train and validation (70%-30%) sets before each method applies the two k-fold cross-validation approaches proposed.

Hyper-parameter Tuning

To improve the performance accuracy of our models, we used the two cross-validation approaches in a grid search to find the optimal values for each model's parameters. Grid search[3] fits different models using the range of defined values for the selected parameters and chooses the model parameter values that minimize the loss. Applying cross-validation to the grid search will help to avoid over-fitting.

We tested the following for the Gradient Boosting and XGBoost parameters:

- number of estimators: 1000, 2000;
- minimum number of samples to split: 5, 15;
- minimum number of samples to be at a leaf node: 5, 15;
- learning rate: 0.05, 0.01; and
- sub-sample rate: 0.5, 0.8;

For the XGBoost parameters in grid search, we also consider the following parameters:

- sub-sample ratio of columns: 0.5, 0.8;
- L1 regularization term on weights: 0.1, 0.9;
- L2 regularization term on weights: 0.1, 0.9; and
- minimum loss reduction gamma: 0.5, 0.8.

Evaluation Metric

All the approaches' predictive performances were evaluated with the Mean Average Percentage Error (MAPE)[4] metric. It measures the accuracy as a percentage and can be calculated as the average absolute percent error for each actual value minus the forecasted one divided by the actual value. Where A_t is the actual value, and F_t is the forecast value, this is given by:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|A_t - F_t|}{A_t} * 100 \tag{1}$$

In our case, A_t were the ORDER-QTY values, while F_t were the predicted values.

Moreover, we also calculated the MAPE between the ORDER-QTY values and the experts' predictions BGT-QTY done by the company, and we compared it with our MAPE.

[3] https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

[4] https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

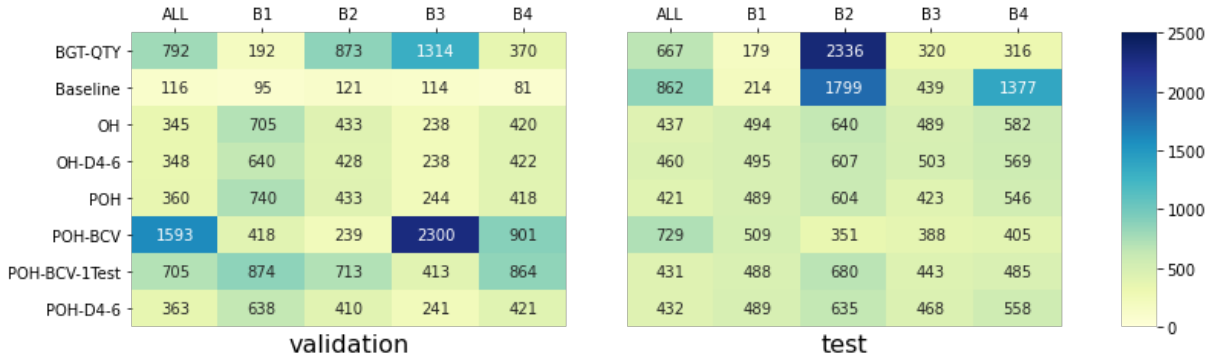


Figure 5. Result of Gradient Boosting Experiments based on MAPE metric

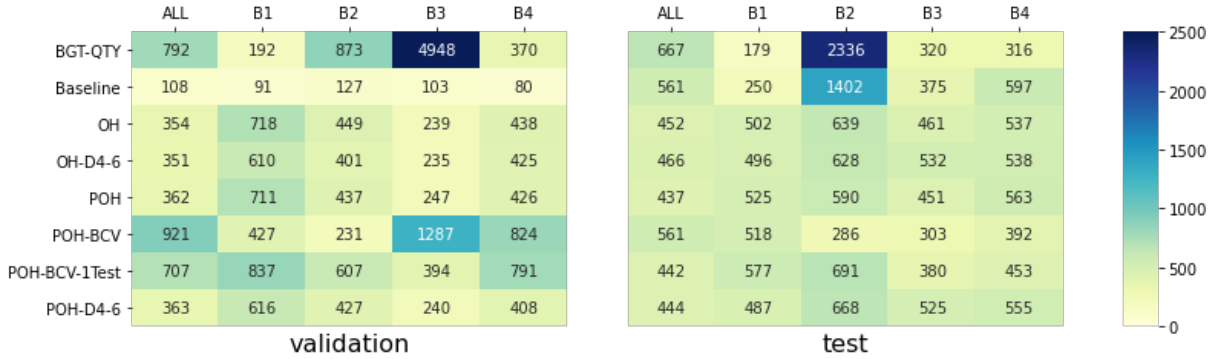


Figure 6. Result of XGBoost Experiments based on MAPE metric

Experimental Environment

All the tests were run on a machine having a CPU with 12 Core/24 Thread, 128GB RAM, and a GPU NVIDIA QUADRO P6000 with 24GB of RAM. In particular, we used Python 3 and the Scikit-learn[5] library for all the methods and approaches presented.

Results and Discussion

This section shows the results achieved by each proposed approach divided by the statistical ML model adopted.

Fig. 5 shows the Gradient Boosting MAPE results for the validation and test sets, using every single brand and for all brands. The validation results show that the Baseline approach achieved the best MAPE among different approaches. However, this is not the case for test results, which show that the baseline approach overfitted the data. In the test results, the POH approach was the best for all the brands. The POH-BCV approach was the best approach among proposed approaches for brands B2, B3, and B4, but the experts' predictions BGT-QTY were better in the case of brands B3, and B4. Also, for brand B1, the best approach was the experts' one. It is worth noting that the validation MAPE in the case of experts' approach corresponding to each proposed approach had different values due to the way we generated the validation set for each approach. In Fig. 5 we put the minimum value between all the achieved results. Fig. 6 shows the XGBoost MAPE results for the validation and test sets, using every single brand and for all brands. Also here, the validation results show that the Baseline approach outperformed the other approaches, but, in the test results, it had the best performance between proposed approaches only for the brand B1, proving that the XGBoost model overfitted the data. The POH approach was the best approach for all brands, while the POH-BCV approach

[5] https://scikit-learn.org/stable/

outperformed other proposed approaches in brands B2, B3, and B4. Comparing the BGT-QTY experts' predictions to the proposed approaches, the former outperformed the latter in the case of B1 and B4 brands.

So, using both statistical ML models, we can say that the *Hp. 1* hypothesis is verified, while, about the *Hp. 2* one, we can say that in 3 cases out of 5 our approaches are better than the experts' one and so verify *Hp. 2*, too.

Related Work

Different studies on demand forecasting for the new product introduction have been investigated in the literature, but most of them are not tailored for short life cycle environments. The most commonly used market demand forecasting methods are based on diffusion theory, mainly including Bass [1] and Norton [2] models. Several methods have been developed based on these models, in which analogical approaches are the most critical category. This approach assumes that a new product will behave as similar products do. Mik et al. [14] introduced a survey on existing approaches for demand forecasting proposing various types of NPI and showing which approach performs better in which situation.

Beheshti-Kashi et al. [15] introduced a survey of sale forecasting and NPI mainly in fashion markets. In particular, for NPI approaches, they reviewed seven studies in fashion markets focused on pre-processing of time series and sales data, color forecasting, E-commerce, and fast fashion sales forecasting, but none of the works in this survey focused on ML approaches. Therefore, in the following, we review some of the studies with a focus on ML approaches.

Lee et al. [3] utilized statistical ML approaches to overcome the challenges existing in analogical approaches and predict the parameters of the Bass model. They created a reliable relationship between the attributes and diffusion characteristics of the existing products, so similar products are automatically selected without any human manipulation and used to forecast new products. Their experimental validation showed that most single prediction models and the ensemble model outperform the conventional analogical method. However, in our study, statistical ML approaches are directly used to finding the similarities between products.

Steenbergen et al. [16] proposed a novel new product forecasting method called DemandForest, which combines K-means, Random Forest, and Quantile Regression Forest. Their approach utilizes the historical demand of existing products and the product characteristics of both new and existing products to make pre-launch forecasts. Furthermore, the Quantile Regression Forest (QRF) algorithm quantifies the uncertainty of the demand and can be used to construct prediction intervals. Their proposed methods are evaluated on a synthetic dataset and five real-life datasets. They showed that DemandForest is a generalizable computational approach that also provides the uncertainty of demand for new products. Their focus on the uncertainty of the demand is out of the scope of this study.

Loureiro et al. [17] focused on the fashion retail industry and explored the use of a deep learning approach to forecasting sales, predicting the sales of new individual products in future seasons. They also compared the sales predictions obtained with the deep learning approach with other ML techniques such as Decision Trees and Random Forests. Their results demonstrated that the use of Deep Neural Network and other data mining techniques for performing sales forecasting in the fashion retail industry is auspicious.

Yin et al. [4] proposed a hybrid model for sale forecasting based on product similarity, which is measured through applying a quantitative similarity measurement method. Also, they employed an ensemble deep learning method to improve the low prediction accuracy caused by insufficient consideration of factors affecting the demand for the new product. Their empirical results proved that the forecasting accuracy could be improved using the deep learning method.

To conclude, several works in market demand forecasting have considered methods based on diffusion theory, such as the Bass model and similarities between products. Instead, in our work, we utilized the statistical ML approaches to automatically finding the similarities between products. Furthermore, the studies done in fashion markets have focused on different methodologies such as time series prediction or fast fashion sales forecasting. In contrast, in this work, we have exploited statistical ML approaches for the NPI problem to improve the forecast accuracy beyond the traditional methods. Lee et al. [3] also used statistical ML approaches, but unlike our approach, they applied it to predict the parameters of the Bass model. However, few works have applied other ML approaches, but none of them have our focus. For instance, Yin et al. [4] applied deep learning methods for sale forecasting.

Conclusion

This study proposes a statistical ML solution for demand forecasting of the New Product Introduction. Using different data pre-processing methods such as data clustering, encoding, grid search, cross-validation approaches, and ML models like Gradient Boosting and XGBoost, we proposed seven approaches for predicting sales quantity in short life-cycle industries (e.g., Fashion). We compared our approaches with both a baseline and the experts' predictions from an industrial partner of ours.

As a result, we found that the baseline approach, which is the first try that a typical data scientist will test, suffers from overfitting. However, for all brands, our proposed approach with partial one-hot encoding and standard 5-fold cross-validation achieved the lower MAPE values in the prediction of validation and test sets (**RQ**). Furthermore, considering single brand evaluation, for brands B2 and B3, the PHO-BCV approach outperforms the experts' prediction. Moreover, although splitting data to different brands can help experts have a more accurate prediction, statistical ML approaches are not the case. Table 1 reports the winner approach comparing the MAPE values of all the experiments and the experts' prediction. In conclusion, we showed that combining statistical ML methods with different data pre-processing tasks can improve the experts' manual predictions and, more in general, can help industries to predict the demands for New Product Introductions better.

As future work, we intend to investigate the ensemble of different ML models and check if we can achieve more accurate predictions. Using ensemble forecasting, we can apply multiple forecast methods independently and finally come to the final forecast. Moreover, in this study, we only focused on the features of the products. However, it is worth paying attention to the customer-generated content. The intentions of potential customers may have some predictive value which can help to improve the prediction. We took into account neither any temporal dependency nor fluctuations nor not-stationarity among the data, too. Future versions of this work should take into account them.

Table 1. Summary of the winner approach

Model	Data	Brand				
		ALL	B1	B2	B3	B4
Gradient Boosting	Validation	Baseline	Baseline	Baseline	Baseline	Baseline
	Test	POH	Experts'	POH-BCV	POH-BCV	Experts'
XGBoost	Validation	Baseline	Baseline	Experts'	Baseline	Baseline
	Test	POH	Experts'	POH-BCV	POH-BCV	Experts'

References

[1] F. M. Bass, “A new product growth for model consumer durables,” *Management science*, vol. 15, no. 5, pp. 215–227, 1969.

[2] J. A. Norton and F. M. Bass, “A diffusion theory model of adoption and substitution for successive generations of high-technology products,” *Management science*, vol. 33, no. 9, pp. 1069–1086, 1987.

[3] H. Lee, S. G. Kim, H.-w. Park, and P. Kang, “Pre-launch new product demand forecasting using the bass model: A statistical and machine learning-based approach,” *Technological Forecasting and Social Change*, vol. 86, pp. 49–64, 2014.

[4] P. Yin, G. Dou, X. Lin, and L. Liu, “A hybrid method for forecasting new product sales based on fuzzy clustering and deep learning,” *Kybernetes*, 2020.

[5] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[6] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[7] S. Arlot, A. Celisse, *et al.*, “A survey of cross-validation procedures for model selection,” *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[8] G. C. Cawley and N. L. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[9] U. Hjorth and U. Hjort, “Model selection and forward validation,” *Scandinavian Journal of Statistics*, pp. 95–105, 1982.

[10] J. U. Hjorth, *Computer intensive statistical methods: Validation, model selection, and bootstrap*. CRC Press, 1993.

[11] J. Racine, “Consistent cross-validatory model-selection for dependent data: Hv-block cross-validation,” *Journal of econometrics*, vol. 99, no. 1, pp. 39–61, 2000.

[12] Z. Huang, “Clustering large data sets with mixed numeric and categorical values,” in *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, Citeseer, 1997, pp. 21–34.

[13] T. M. Kodinariya and P. R. Makwana, “Review on determining number of cluster in k-means clustering,” *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[14] E. C. Mik and G. Koole, “New product demand forecasting,” *Vrije Universiteit Amsterdam, Amsterdam*, 2019.

[15] S. Beheshti-Kashi, H. R. Karimi, K.-D. Thoben, M. Lütjen, and M. Teucke, “A survey on retail sales forecasting and prediction in fashion markets,” *Systems Science & Control Engineering*, vol. 3, no. 1, pp. 154–161, 2015.

[16] R. van Steenbergen and M. Mes, “Forecasting demand profiles of new products,” *Decision support systems*, vol. 139, p. 113 401, 2020.

[17] A. L. Loureiro, V. L. Miguéis, and L. F. da Silva, “Exploring the use of deep neural networks for sales forecasting in fashion retail,” *Decision Support Systems*, vol. 114, pp. 81–93, 2018.

24th International Conference on Business Information Systems (BIS 2021)

Artificial Intelligence

https://doi.org/10.52825/bis.v1i.58

Published: 02 July 2021

Evaluating the New AI and Data Driven Insurance Business Models for Incumbents and Disruptors: Is there Convergence?

Alex Zarifis [1][https://orcid.org/0000-0003-3103-4601], and Xusen Cheng [2][https://orcid.org/0000-0001-7614-6514]

[1] Loughborough University, UK

a.zarifis@lboro.ac.uk

[2] Renmin University of China, China

xusen.cheng@ruc.edu.cn

Abstract. AI and data technologies are a catalyst for fundamental changes to insurance business models. The current upheaval is seeing some incumbent insurers trying to do the same more effectively, while others evolve to fully utilize the new capabilities and users these new technologies bring. At the same time, technologically advanced organizations from outside the sector are entering and disrupting it. Within this upheaval however, there are signs of a convergence towards an ideal and prevailing business model. This research identifies one exemplar incumbent and one disruptor and evaluates whether their models are converging and will become similar eventually. The findings support a high degree of convergence, but some differences are likely to remain even after this transitionary period. The differences identified are firstly in the evaluation of risk and secondly that traditional insurers prioritize revenue generation from what is their primary activity, while new entrants prioritize expanding their user base.

Keywords: Artificial Intelligence, Machine Learning, Business Model, Insurance.

Introduction

Several sectors of the economy are facing a digital disruption driven by technologies like Artificial Intelligence (AI), cloud computing, big data, blockchain, the Internet of Things (IoT), 5G networks and social media. While this digitization has been ongoing for decades, the recent enhanced capabilities of AI make this change faster and more fundamental [1]. AI, particularly machine learning, is used to enhance business operations and offers a step change in speed, reliability, accuracy, new insights and new capabilities. These enhancements AI offers lead to improvements in the organization's performance, profitability, revenue and customer satisfaction [2]. However, because AI and algorithms that learn are making fundamental changes to the employee roles, processes and business models of organizations, there are also challenges [3], 'known unknowns' for which answers are being pursued and 'unknown-unknowns' where the right questions are yet to be formulated. This research follows a widely used agenda for researching digital business models in information systems that suggests insights on their transformative nature as one of the three best ways to make a contribution [4].

There is a convergence of social, mobile, analytics and cloud computing [5] that is accelerated with AI. This raises the question if this wave of technologies will create a radical change in insurance. The insurance sector is widely considered to be traditional and risk averse. Insurers are not usually early adopters in technology. While it is not a low-tech

industry, it seems to set the threshold high for a new innovation to be adopted and have a transformative impact. Therefore, it is not just a question of what the potential is for AI in insurance, but how and when that potential will be reached.

While the insurance sector was always knowledge intensive the increase in data through social media, IoT, mobile and cloud computing means the insight possible cannot be fully reached only with humans manually. Seamless, intelligent automated processes using AI are needed. This increasing role of AI as an enabler of big data utilization removes the bottleneck of humans in the process. Humans need to be in the loop in some processes but not others. Incumbent insurers are enhancing their capabilities in AI and big data while technology companies like Alibaba, Tencent and Tesla with existing AI and big data capabilities are providing insurance. This raises the primary research question:

Are incumbent and disruptor insurance business models converging to one ideal business model?

This research first compared how ten insurers are utilizing AI in their current processes and how they are creating new ones. The case studies covered five incumbent insurers and five new entrants. There was one incumbent and one new entrant from each continent. These ten cases cover the world geographically and allow ecosystems and networks of interest to emerge beyond the boundaries of the organizations used as cases. From this first phase one exemplar incumbent and one disruptor were evaluated in depth with interviews to evaluate whether their models are converging and whether they will become similar after this transitionary period. The interviews were with insurance professionals and insurance consumers. A high degree of convergence was found but some differences are likely to remain even after this transitionary period. These differences are firstly in the evaluation of risk and secondly that traditional insurers prioritize revenue generation from what is their primary activity, while new entrants don't necessarily. The rest of the paper covers the literature review and methodology followed by the analysis of the two phases and finally the conclusion.

Literature Review

2.1 Business Models, Multichannel Retail and Route to Market

General business models that were identified at the start of the internet [6] and were relevant for many years, no longer cover the new online models and ecosystems. Researchers are therefore exploring the distribution channels used today. Distribution channels have been included in some prominent business model ontologies, [7, 8]. Fritscher and Pigneur [7] identify them as one of nine building blocks of a business model. The nine components identified are customer segments, customer relationships, distribution channels, revenue flows, value proposition, key activities, partner network, key resources and cost structure. For insurance, the evaluation of risk is the primary activity which has the largest influence on survival and prosperity. It is necessary to understand what an insurer's strategy is and how they intend their business model to operate, but it is equally important to understand how their users perceive it.

2.2 The Relative Advantage of Each Channel

The consumer of insurance, like other consumers, can access insurance through several channels. These include physical retail stores, online websites and mobile applications. Each channel has different characteristics and can utilize AI differently. Similarly, employees within the organization have several ways of accessing the information they need and fulfilling the processes they need to. Therefore, the relative advantage of each channel spans across consumer facing and back office processes. The application of AI and data technologies is influencing the channel, the form of interaction and subsequently the relationship. For example, an interaction by voice, utilizing natural language processing and sentiment analysis is popular with those wanting to get insurance [9]. Consumers have an

understanding of the relative advantage of each channel across the four steps of their purchase, which are requirements determination, vendor selection, purchase and after sales service [10]. Consumers increasingly expect customized and personalized services and the channel that can achieve this best, has a significant relative advantage. In B2C insurance, this customization can be enhanced extensively by mobile applications.

2.3 Simple and Complex Services

The nature of the service, how complex or simple it is, influences the approach, and expectations of the consumer [10]. Related concepts are that certain services have a low involvement, while others have a high involvement [11]. In retail insurance, there are relatively simple standardized services such as vehicle insurance, and more sophisticated complicated services like health insurance. Beyond retail insurance (B2C), the complexity of the insurance service can increase substantially, but that is outside the scope of this research. This research focuses on the front and back office processes that influence the relationship between the insurer and an individual user.

2.4 Constructs Identified and Theoretic Foundation

The literature review identified the significance of understanding the insurers intended strategy, particularly how they evaluate risk but also how the consumers perceive it. The consumers perception is influenced by the difference between the channels of offline, 2D websites, 3D virtual worlds and mobile across the four stages of the purchase process. Furthermore, for the consumer it is important to distinguish between simple services with limited engagement and complex services with extensive engagement. Therefore, the evaluation of the influence of the new AI and data driven value chain in insurance will be framed by these constructs.

Methodology

This research applied the method of comparative case study analysis with a critical realist epistemological perspective [12]. There were two phases, the first is broad and exploratory to capture all the issues and the second is more focused and in-depth to understand them better. In phase one the case studies covered the five incumbent insurers and the five new entrants listed in table 1.

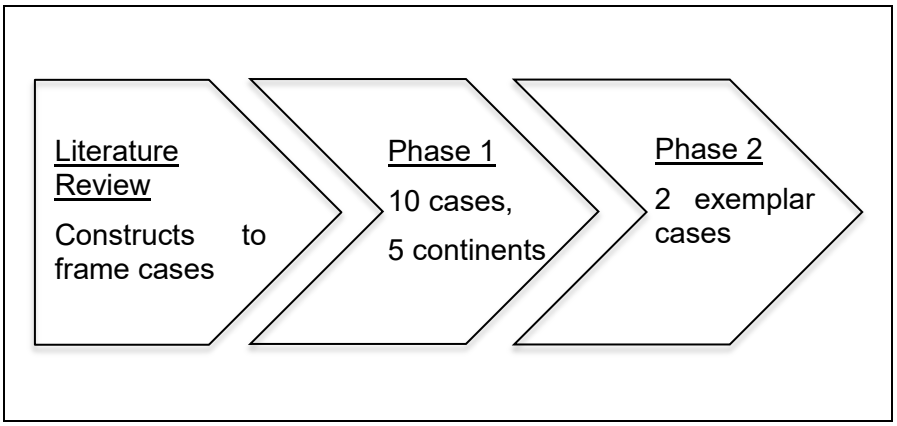


Figure 1. Research Method

There was one incumbent and one new entrant from each continent. The number was chosen for two primary reasons: Firstly, because it was the smallest number that could cover the world geographically. Secondly, the ten case studies would also allow ecosystems and

networks of interest to emerge beyond the boundaries of the organizations used as cases. Thirdly, ten case studies were the largest number that could be presented in a research paper. The companies were selected because they embodied the characteristics of an incumbent or new entrant well. The first phase involved desk-based research covering existing research, reputable industry press like the Financial Times, Intelligent Insurer and Insurance Post, the websites and reports of the insurer, suppliers and partners.

Table 1. 10 case studies by continent

Continent	Incumbent insurers	New entrants (Disruptors)
Europe	Bupa	Wrisk
America	Manulife	TESLA (Insure MyTesla)
Asia	Ping An	Zhong An
Australia	TAL Life Limited	Huddle Insurance
Africa	Old Mutual	Naked Insurance

Table 2. Demographic information of the sample group

Measure	Item	TESLA	Ping An
Gender	Female	5	5
	Male	5	5
Age	Under 18	0	0
	18-24	4	5
	25-39	3	4
	40-59	2	1
	60 or older	0	0
Educational level	Without education level	0	0
	High school graduate	4	3
	University bachelor’s degree	6	7
	University master or doctorate	0	0
Income (in Euro per month)	No regular income	0	1
	400-1200	2	2
	1201-3000	7	6
	3001-5000	1	2
	>5000	0	0

After the initial evaluation of 10 cases, 2 exemplar cases were selected for empirical data collection and in-depth analysis. For the two exemplar cases TESLA and Ping An, ten interviews were carried out for each of them. The interview participants were two managers in the companies and eight consumers. The constructs chosen to frame the business models covered both internal processes and the consumers beliefs, so interviews needed to be

carried out with both employees and consumers. From each company one manager involved in technology implementation and one in marketing and sales were interviewed. The participants that were consumers of insurance had purchased services from one of the two companies. This was a requirement to participate.

Analysis

The desk-based research of the 10 insurers identified that AI and data technologies are adopted at two different speeds. Those adopting the technologies gradually have two approaches: The first is to focus and cover a smaller part of the insurance value chain, reducing cost so that their services are appealing to an ecosystem or platform like Comparethemarket.com. An example of an organization doing this effectively with some of its services is AVIVA. The second approach is to keep the same business model and use AI and data to intelligently automate and improve efficiency. An example of an organization doing this effectively with some of its services is TAL Life Limited. The insurers that are adopting AI and data technologies faster and more extensively are transforming their business model to fully utilize these technologies. An example of incumbent insurers changing their business model for these reasons are BUPA Health Insurance, Manulife and Ping An. An example of new disrupting insurers changing their business model for these reasons are TESLA, Wrisk and Zhong An. For the second phase of the analysis, 20 interviews were implemented with TESLA and Ping An managers and consumers.

4.1 Strategy, Simple and Complex Services and Revenue

The nature of the service, how simple or complex it is, influences the approach and expectations of the consumer [10]. Ping An offers both complex and simple, B2B and B2C, insurance services. They use big data and AI to understand the customers better based on the data collected on them and improve customization and satisfaction. They use this superior understanding to cross sell, create new services and gain new customers. AI driven claims processing is implemented in the Ping An Auto Insurance Smart Claim App (via Ping An), further improving customization.

AI is used for complicated claims such as a claim on the Hong Kong, Zhuhai and Macau bridge where three different insurance covers are required as there are three different jurisdictions with different insurance regulation. They also use the Voiceprint recognition system which is a facial recognition system that identifies their customers. This improves security and efficiency. One interviewee told us ‘we use AI to automate administrative tasks, evaluate claims and interact with voice recognition…’, ‘…we want to understand them (consumers) better than anyone (competitors)…’. While chatbots and natural language processing are ubiquitous in insurance now, the insurers with more and better data can train their algorithms better for both the insight and interaction.

TESLA only offers simple B2C services that can be implemented using current technologies, data and regulation. Providing insurance is not a priority in terms of revenue but complementary to their products. New innovations like self-driving cars face hurdles including insurance so having some capabilities in insurance can overcome these barriers. TESLA takes a proactive approach in shaping the insurance their cars need and collaborates with traditional insurers to deliver it in different countries. One interviewee stated: ‘…we spend a-lot of time on insurance and regulation for our self-driving cars…’.

4.2 Risk Estimation for Underwriting and Claims Payout

Insurers primarily add value by identifying and analyzing risk. In addition to the risks created by the person being insured there are also risks from medical malpractice, cyberthreats and viruses. It was found that the risk estimation that informed the underwriting was one of the major differences between the incumbent insurers and the technology companies entering insurance. Ping An had information of high relevance, quality and reliability on its consumers

but this information was not updated regularly in the past: 'We have the age, education, profession, income, previous claims…we can underwrite the claim with the risk profile we want…', '…our data is more accurate than others (data) because we have taken it for decades…we can avoid premium leakage better'.

The ability of AI to process structured, semi-structured or unstructured datasets also improves the detection and reduction of fraud such as subrogation and multiple unlawful claims: '…it (AI) finds discrepancies by bringing together diverse data including travel information…one fraud claim stopped because the person was abroad (at the time they claimed they had an accident)'.

TESLA uses the data from cameras, and other sensors in their vehicles. Currently this data is aggregated and used in an anonymous way but in the future, it could be used to reward or punish drivers based on how they drive. When an accident happens, TESLA has extensive information on how the driver and the technology performed. Unlike traditional insurers, TESLA can proactively reduce risk by improving their hardware and software. Software updates can be delivered to all TESLA vehicles automatically. Therefore, real time data and close to real time, proactive, risk reduction influence the way risk shapes underwriting, repair cost payouts and the vehicles. One manager at Tesla stated: 'When we first updated all our cars at the push of the button this surprised a-lot of people in the industry…', 'If you look at the Cybertruck it has security features based on our experience, we did not put them in for the insurance but it will keep people safe and reduce claims…'. The security features include autopilot, stability control, anti-theft systems and bullet-resistant steel. However, as TESLA's better understanding of risk only applies to their vehicles, they may not expand beyond insuring their own vehicles.

4.3 Engagement and Relationship with the User Across Different Channels

While it is important to understand the intentions of insurers, it is also important to understand how their services are perceived by the consumer. The interactions and relationships between the insurer and the consumer are more frequent and nuanced than in the past. The logic behind the decision AI makes is not always transparent so this increases ambiguity.

The relative advantage of each channel for the consumer: Ping An is an incumbent insurer that offers the full range of insurance services across several distribution channels. They attempt to fully utilize new technologies by having several business models in parallel. They offer several services directly and they also have alliances with technology companies. An example of this is Zhong An, a partnership between Ping An, Alibaba and Tencent. With Ping An's original model, the engagement with the user was not frequent and did not allow for real time data collection. The partnerships with technology companies increase this engagement. In their traditional model the human interaction with the user is during sales, underwriting, post-sales support and claims management. In the new services there is increased use of chatbots and analysis of behavioral data. For example, one interviewee stated: '…there are some relationships surprise you. How long people take to fill the form in the app is linked to how risky they are…'.

AI and data technologies offer several benefits to the interaction with the user. These include (1) better automated interaction including facial recognition and sentiment analysis through the user's voice, (2) fast offers, often under 60 seconds, (3) quicker claim processing with automated features such as AI evaluating damage from pictures, (4) more organized interactions with better scheduling of meetings, (5) new services such as the consumer paying according to the distance travelled, (6) adapting to local conditions and regulations such as applying adaptive pricing only where it is allowed. One interviewee emphasized the importance of the interaction and how decisive this can be '…there are apps that list all the insurers…my friends like them…they are quick… but I use this one (Ping An)'.

Table 3. Incumbent insurers and tech companies AI driven business models

Difference	Incumbent Insurer	Tech company offering insurance
Service complexity and revenue	Complex and simple service, B2B and B2C service Understand the consumers better, improve customization and satisfaction of cover and interaction	Simple standardised B2C services Revenue not a priority but complementary to their products
Risk estimation and claims payout	Information of a high relevance, quality and reliability, but this information is not always updated regularly	Real time information on behaviour Proactively influence behaviour (e.g. premium changes based on driving behaviour) Proactively reduce risk (e.g. warning or software update)
Engagement with the user	Several business models in parallel to utilize different technologies Alliance with tech companies to access their user base Benefits to the interaction with the user: Better automated interaction, fast offers, quicker claim processing, better scheduling of meetings, new services, adapting to local conditions and regulations	Bundle insurance with existing services Remove the hurdle of insurance both for the organization and the consumer Use existing access to user data so no additional privacy concerns

There is an effort by this insurer to be proactive and promote a healthy lifestyle. This can be basic, offering advice and fitness trackers. It is also more advanced offering remote healthcare technology with an application that enables virtual consultations with AI powered virtual doctors. Prescriptions are provided within the application. The proactive effort reduces the risk of health problems and brings in new information to further reduce risk.

TESLA engages with consumers and makes their vehicles available for purchase both online and offline. When ordering a vehicle, a quote is given either directly from TESLA or its partners. While a TESLA is a luxury vehicle, consumers prioritized the price and simplicity of the insurance when making a decision: '…we saved 20-30% taking the insurance out directly form TESLA…'. Most interviewees did not have serious concerns over their data being used as this was already happening because they were using the vehicle: '…I love TESLA for the technology so why not use it for insurance too…'.

Some comments from users of both insurers showed that AI and data technologies can also strain the relationship and raise privacy, trust and ethical concerns. As the role of AI increases shifting from a tool augmenting human intelligence, to an automated insurance process without direct human involvement, some consumers are concerned about the lack of transparency and empathy: 'I don't want to talk to a machine about problems, where is the information going…', '…when I am waiting I think my internet my phone is stuck or hacking…'. Some interviewees felt the information asymmetry was increasing as their intelligence and knowledge could not compete with AI: '…is my driving good I do not know,

maybe they have a different idea about how someone should drive...'. One interviewee referred to requests for additional information compared to what was needed in the past as '…mission creep…' suggesting the insurer is expanding their role over the years.

Findings

This research answers the call to explore the transformative nature of digital business models in information systems [4] which is particularly pertinent in insurance now. Furthermore, this research follows that call and takes into account the increasing and multifaceted influence of the user and how new ICT create value and shape the business model structure and purpose [1, 4].

5.1 How AI and Data Technologies are Changing Insurance Models

The first stage focuses on efficiency and replacing an existing process with AI. Once an existing process has been replaced by AI, expanding the use of AI is easier because machine learning can learn quickly. For example, once the AI supported virtual assistant is operating, it is easier for the virtual assistant to sell new, additional products and services.

While most large incumbent insurers are active in several countries and need to adapt their AI implementation to different regulations, many of the disruptors focus on one market at a time, keeping the simplicity of implementation low. The approach of the disruptors avoids many of the challenges to AI implementation identified such as the need to have a deep understanding of all the business processes [2]. This research supports previous findings that insurers are adapting their business models to counter competition from other markets [13].

The startups offered simple services in a fully automated way with the help of AI. The incumbents also offered some of their simpler services in this way. More complex services were supported with AI and data technologies but in most cases an expert made the final decision. An example of this are the audits for fraud, where the AI identifies unusual patters and cases for an expert to evaluate. While it is easy to find examples of simple and complex services in insurance, it is not easy to know where exactly to draw the line to separate all insurance services into simple and fully automatable, or complex. This is however an important question facing insurers. An example to illustrate this, is Manulife allowing its automated AI supported underwriter to underwrite life insurance from 18 to 45, which is considered easier and less risky to evaluate compared to older ages.

The consumer's expectations are expanding in the insurance sector. The convenience, flexibility and customization the consumer receives from other sectors such as banking is now expected in insurance also. This 'pull' from the consumer is met by the incumbents and the new entrants using mobile applications, AI and APIs to the necessary data outside the organization. Several of the startups do not offer their services offline. Furthermore, some steps that speed up the process need mobile applications and their functionalities such as cameras, GPS and verification by sending a text. Therefore, it can be concluded that the mobile channel currently amplifies the capabilities offered by AI best.

5.2 AI and Data Technology Insurance Models of Incumbents and Disruptors

AI is changing the insurance value chain causing upheaval, as illustrated in figure 2. There are several signs of convergence between the models of incumbents and disruptors. Firstly, there is convergence in technologies such as, for example, the use of chatbots and IoT. Secondly there is a convergence in processes, for example the interaction with the consumer. Thirdly there is convergence in the strategy on costs and pricing, for example insurers like AVIVA focus on cost cutting to compete with incumbents like WRISK, Huddle Insurance and Naked Insurance. There are however two areas where there seems to be a

limit on the convergence that seems to suggest the business models of the incumbent and the disruptor will remain distinct:

(1) Evaluating risk: AI and data technologies are changing the way risk is evaluated by an insurer. These new ways of using data and technology to assess risk in turn generate new insurance services. This is true for both forward looking incumbents and disruptors. Incumbents like Ping An are creating new services from the new data but they are also creating new services to gain new data.

For disruptors like TESLA risk is calculated in three ways: Firstly, data is collected from the cameras and sensors in the vehicle providing insight on real time behavior both at the individual and aggregate level. Secondly, broader analysis of individuals with hundreds of variables is implemented and new algorithms that evaluate risk accurately are pursued. Thirdly, the impact on risk of the new technologies is constantly monitored and, in some cases, influenced.

(2) Cost of attracting the user and profitability: Access to user data through new technologies. The technology company offering insurance like TESLA have some advantages in terms of the cost of attracting new consumers and the profits they generate. While the incumbent insurer must spend on marketing to attract consumers to their insurance services the technology company uses existing users. Furthermore, while the insurer is dependent on their revenues from insurance services the technology company can draw profits from other services and provide insurance without any profit.

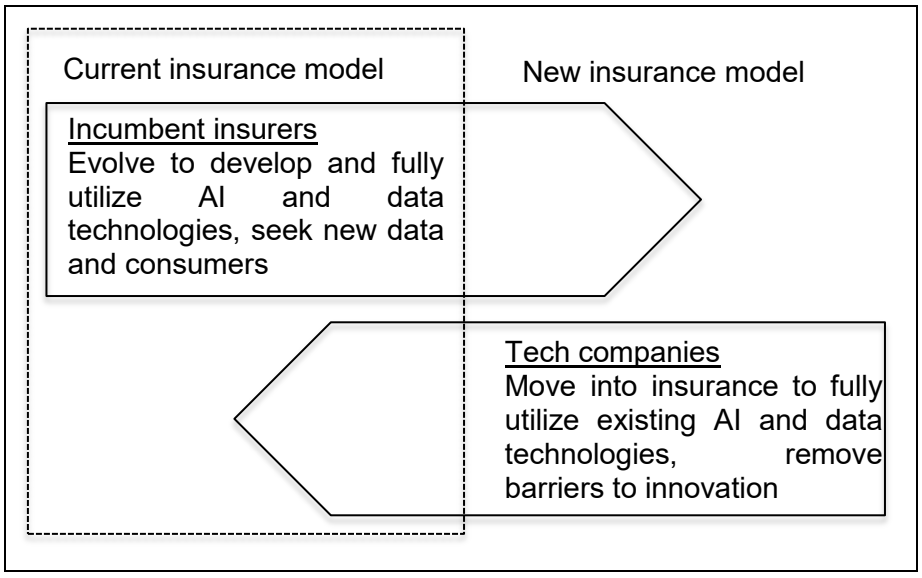


Figure 2. Partial convergence of incumbent and disrupting insurers

Conclusion

AI and data technologies are acting as a catalyst for broad changes in insurance. This research identified and evaluated two exemplar cases of an incumbent and disrupter creating business models to fully utilize these technologies. A comparative case study analysis explored these changes by interviewing insurance professionals and consumers. While there was some convergence between the two models it is not expected that they will merge into one after the transitionary period. This is because, despite the convergence, the two models have some distinct competitive advantages. While the incumbents no longer monopolize the capability of providing insurance, they still have the existing user base and use it to evaluate risk. Technology companies that offer insurance now, have their own forms of engagement with their users, use different methods to evaluate risk due to their access to real time data,

and do not seem to prioritize generating revenue but instead utilize insurance to increase their user base, overcome barriers and reduce the overall cost of their products and services.

The limitation of this research is that despite the care taken to choose two representative cases to study, other cases may have differences. Future research can test the findings with additional in-depth case studies of incumbents and disruptors from across the world. It may also be beneficial to have a longitudinal study to further verify the convergence of the insurance business models that fully utilize AI and data technologies.

Statement on competing interests: The authors declare that there is no conflict of interest.

References

1. Alt, R., Leimeister, J.M., Priemuth, T., Sachse, S., Urbach, N., Wunderlich, N.: Software-Defined Business: Implications for IT Management. Bus. Inf. Syst. Eng. 62, 609–621 (2020).
2. Tarafdar, M., Beath, C.M., Ross, J.W.: Using AI to Enhance Business Operations. MIT Sloan Manag. Rev. 60, 10 (2019).
3. Faraj, S., Pachidi, S., Sayegh, K.: Working and organizing in the age of the learning algorithm. Inf. Organ. 28, 62–70 (2018).
4. Veit, D., Clemons, E., Benlian, A., Buxmann, P., Hess, T., Kundisch, D., Leimeister, J.M., Loos, P., Spann, M.: Business models: An information systems research agenda. Bus. Inf. Syst. Eng. 6, 45–53 (2014).
5. Legner, C., Eymann, T., Hess, T., Matt, C., Böhmann, T., Drews, P., Mädche, A., Urbach, N., Ahlemann, F.: Digitalization: Opportunity and Challenge for the Business and Information Systems Engineering Community. Bus. Inf. Syst. Eng. 59, 301–308 (2017).
6. Timmers, P.: Business Models for Electronic Markets. Electron. Mark. 8, 3–8 (1998).
7. Fritscher, B., Pigneur, Y.: Supporting business model modelling: A compromise between creativity and constraints. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 5963 LNCS, 28–43 (2010).
8. Oliveira, M.A.-Y., Ferreira, J.J.P.: Book Review: Business Model Generation: A handbook for visionaries, game changers and challengers. John Wiley and Sons, Hoboken, New Jersey (2010).
9. Maull, R., Collomosse, J., Brewer, S., Bordon, A., Jones, K., Breeze, J.: Taking control: Artificial intelligence and insurance. (2019).
10. Choudhury, V., Karahanna, E.: The Relative Advantage of Electronic Channels: A Multidimensional View. Manag. Inf. Syst. Q. 32, 179–200 (2008).
11. Zaichkowsky, J.L.: Measuring the Involvement Construct. J. Consum. Res. 12, 341–352 (1985).
12. Eisenhardt, K.M.: Building Theories from Case Study Research. Acad. Manag. Rev. 14, 532–550 (1989).
13. Eling, M., Lehmann, M.: The Impact of Digitalization on the Insurance Value Chain and the Insurability of Risks. Geneva Pap. Risk Insur. Issues Pract. 43, 359–396 (2018).

24th International Conference on Business Information Systems (BIS 2021)

Artificial Intelligence

https://doi.org/10.52825/bis.v1i.59

Published: 02 July 2021

Evaluation of Deep Learning Instance Segmentation models for Pig Precision Livestock Farming

Jan-Hendrik Witte[1], Johann Gerberding[1], Christian Melching[1] and Jorge Marx Gómez[1]

[1] Universität Oldenburg, GER

Abstract. In this paper, the deep learning instance segmentation architectures DetectoRS, SOLOv2, DETR and Mask R-CNN were applied to data from the field of Pig Precision Livestock Farming to investigate whether these models can address the specific challenges of this domain. For this purpose, we created a custom dataset consisting of 731 images with high heterogeneity and high-quality segmentation masks. For evaluation, the standard metric for benchmarking instance segmentation models in computer vision, the mean average precision, was used. The results show that all tested models can be applied to the considered domain in terms of prediction accuracy. With a mAP of 0.848, DetectoRS achieves the best results on the test set, but is also the largest model with the greatest hardware requirements. It turns out that increasing model complexity and size does not have a large impact on prediction accuracy for instance segmentation of pigs. DETR, SOLOv2, and Mask R-CNN achieve similar results to DetectoRS with a parameter count almost three times smaller. Visual evaluation of predictions shows quality differences in terms of accuracy of segmentation masks. DetectoRS generates the best masks overall, while DETR has advantages in correctly segmenting the tail region. However, it can be observed that each of the tested models has problems in assigning segmentation masks correctly once a pig is overlapped. The results demonstrate the potential of deep learning instance segmentation models in Pig Precision Livestock Farming and lay the foundation for future research in this area.

Keywords: Precision Livestock Farming, Instance Segmentation, Computer Vision, Deep Learning, Pig.

Introduction

Structures of modern pig livestock farming, and pork production have been undergoing major changes in recent years. Data from the Federal Statistical Office show the opposite trend of a steadily decreasing number of farms [1] with simultaneously increasing numbers of animals per farm [2] and a continuously decreasing slaughter price[1], which poses and will continue to pose great challenges for the farmer. At the same time, politics and society alike are calling for more sustainable and more animal-friendly husbandry [3], which puts additional pressure on the farmer and makes economically profitable pig livestock farming increasingly difficult. These challenges cannot be met with conventional methods, which is why new and innovative solutions are needed. As a result, research in the domain of Precision Livestock Farming (PLF) has increased in recent years. PLF describes systems that utilizes modern camera and sensor technologies to enable automatic real-time monitoring in livestock production to supervise animal health, welfare and behaviour [3], [4]. This involves the automated acquisition, processing, analysis and evaluation of sensor-based data like temperature, humidity or CO2-concentration [5] and image and video data [6], [7]. Based on

[1] https://www.bmel-statistik.de/preise/preise-fleisch/

this information, systems can be created that support the farmer in his daily work and help him adapt to the changing conditions in pig livestock farming. To enable such systems, methods are first needed that allow the automated processing of these different data streams in the form of image, video, and sensor data. In the case of image data, methods are required that can be used for automated recognition and localization of individual pigs within

the pen. Due to the specific conditions in pigsties, these tasks pose a particular challenge.

Figure 1. Example image from pigsty.

Fig. 1 illustrates some of these problems. Piling, crowded areas, the overlapping of pigs as well as their various orientations and alignments make it difficult to automatically detect and locate individual pigs within the pen. In addition, there are constantly changing factors such as varying light conditions, soiling of the animals and the pen and occlusions caused by objects in the pen. In literature, similar use cases such as automated pedestrian detection in crowded areas have been successfully addressed using deep learning (DL) methods [8]. For pig detection and localisation, there are two approaches that are used in DL domain: (1) Object detection and (2) instance segmentation. Object detection (1) describes the classification and localisation of objects with the help of bounding boxes [9]. The algorithm surrounds each classified object within the image with a bounding box, which can then be used to determine the object's position in the image. Instance segmentation is a combination of object recognition and semantic segmentation. After object detection, semantic segmentation is used to classify and map each pixel of the detected object to a corresponding class or category. This results in detailed masks for each detected object where each mask can be considered as an independent instance [9].

A variety of different object detection methods from the field of DL have already been used in PLF. Cang et al. applied Faster-RCNN for detection and weight estimation of pigs based on image data [10], Nasirahmadi et al. utilized Single Shot Multibox detector, Region-Based Fully Convolutional Neural Network and Faster R-CNN for posture detection of individual pigs [11] and YOLO was applied by Sa et al. for pig detection under various illumination conditions [12]. However, bounding boxes are not able to capture the contours of objects, which is why valuable information could be lost when only using bounding boxes [13]. For some PLF related use cases like the prediction of tail biting in grouped house pigs, this information could be insufficient. In a report by the BMEL[2], in which various indicators for the early detection of tail biting were summarised, it emerges that activities such as tail-in-mouth behaviour or generally manipulative chewing behaviour on pen objects can increasingly be observed before tail biting events [14]. For this type of use case, the much more precise instance segmentation masks could be beneficial. In the DigiSchwein[3] project, the automated early detection of tail biting is a central objective, which we intend to explore

[2] Bundesministerium für Ernährung und Landwirtschaft

[3] https://www.lwk-niedersachsen.de/index.cfm/portal/1/nav/1093/article/35309.html

further with the help of modern deep learning methods. For this reason, this paper investigates whether common instance segmentation methods from the field of DL can be applied to data from Pig PLF. Using defined selection criteria, four different instance segmentation methods are identified in DL literature and tested and evaluated based on a custom dataset. The goal is to evaluate whether the applied methods can deal with the specific challenges of the data from the Pig PLF domain and how their quality is in terms of prediction accuracy and speed.

This paper is structured as follows. First, we examine how instance segmentation has been applied in the context of Pig PLF and which methods have been used. Based on these results, the selection criteria for the instance segmentation methods are presented, followed by a brief description of the selected models. The presentation of the results is done using a quantitative and qualitative analysis. The model evaluation is performed using the mean average precision (mAP) based on our test set that we extracted from our annotated dataset. The qualitative analysis is based on a visual evaluation of the predicted masks by the different models and is used to discuss potential problems and remaining challenges. The interpretation of the results discusses the insights gained from the quantitative and qualitative evaluation. The conclusion and outlook summarize the results and describe how they can be used in future research.

Related Work

Instance segmentation use cases identified in literature can be divided into two different categories: (1) segmentation without DL and (2) segmentation with DL [15]. Segmentation techniques without DL (1) are characterised by using thresholding for image binarization, separating background from foreground [16]. Otsu's method is a popular example for this, which has been used in a variety of use cases [17], [18], [19], [20]. This type of segmentation is usually only done as a pre-processing step, based on which the actual detection of the objects in the foreground takes place. Nasirahmadi et al. use the results of Otsu's binarization to locate pigs on image data using an ellipse fitting algorithm [21]. The approach for the identification and localisation of grouped-house pigs by Huang et al. has a similar structure, but Gabor filters are used for segmentation and feature extraction. The subsequent detection is done with Support Vector Machines [22]. However, methods like Otsu's do not perform instance segmentation, but semantic segmentation since the entire content of an image is segmented according to the defined threshold rather than individual regions or instances. Due to the definition of a threshold for image segmentation, such solutions are also vulnerable to structural changes within the image like changing light conditions, occlusion or dirt [20]. To address these problems and enable actual instance segmentation of objects, DL methods can be applied.

During literature research, only three papers were identified that applied instance segmentation methods from the field of DL to Pig PLF. Seo et al. conclude that the predictions of the Mask R-CNN are insufficient for the use case of separating touching pigs in image data. They describe that the segmentation accuracy of the predicted masks is not satisfactory, as some pigs in overcrowded areas are not recognised correctly or are completely missed out [23]. On the other hand, Li et al. successfully use Mask R-CNN for instance segmentation of pigs. They use the information provided by the segmentation masks to automatically recognise mounting behaviour. The presented model was fine-tuned on pre-trained COCO model and a ResNet50 backbone [24]. Tu et al. tested and evaluated Mask Scoring R-CNN, an adaptation of Mask R-CNN, to improve instance segmentation performance for grouped-housed pigs [15]. Mask Scoring R-CNN improves instance segmentation performance by adding a network block that learns the quality of the predicted instance masks and feeds this information back into the network during training [25].

During analysis of these papers, we noticed some problems in the way the evaluation of instance segmentation models was conducted. In research, the COCO data format has become the standard format to train and evaluate instance segmentation models [26], [27].

Most of the instance segmentation models and methods found in literature, including Mask R-CNN, use the COCO evaluation format to benchmark model performance against other architectures [9]. The commonly used metric for evaluation and benchmarking is the mAP, which is the mean of the Average Precision (AP) based on a set of different Intersection over Union (IoU) thresholds [9]. IoU is the most used evaluation metric for object detection and instance segmentation tasks. It is defined as the similarity between the ground truth segmentation and the predicted segmentation present in the image and is determined by dividing the intersection with the area of union [28]. We noticed that none of the identified paper uses this metric to evaluate their model. Tu et al. use Precision, Recall and F1-Score for evaluation [15], Seo at al. did not state a general model performance for instance segmentation at all [23] and Li et al. used mean Pixel Accuracy (mPA) metric [24] which is normally used to evaluate performance in semantic segmentation tasks and not instance segmentation [29]. Pixel accuracy describes the amount or percentage of pixels which are classified correctly by the model. This can be problematic if, for example, there is a class imbalance in the data used, in which the number of pixels in the image that do not belong to any class greatly exceeds the number of pixels that belong to a class and vice versa, thus enabling a classification performance based on the class imbalance with a priori knowledge [29]. This results in two problems: (1) A comparison of the performance of the respective results is not possible due to the different and partly inappropriate evaluation metrics and (2) No evaluation of instance segmentation models on data from the Pig PLF domain has been conducted yet based on the standard evaluation metric mAP. These research gaps are also addressed in this paper.

Materials and methods

Instance segmentation model selection

Instance segmentation methods were chosen based on defined selection criteria. The definition of the selection criteria was made based on different aspects. On the one hand, the requirements for PLF systems mentioned in the literature should be considered. On the other hand, the selection criteria should serve to answer the research question of this paper regarding prediction accuracy and speed of the model on data from the Pig PLF domain. The following criteria were defined:

1. **Prediction Accuracy**: The prediction of the respective model should be as accurate as possible [30].
2. **Prediction Speed**: Model inference should be in real time [18].

Prediction speed in real time refers to the requirement that the respective algorithm should deliver a result within milliseconds. Cost-effectiveness is a criteria that is mentioned in literature as well regarding PLF systems [18] but has been ignored in the context of this paper as it does not contribute to answering the initial research question. To set a baseline for which models to compare and to allow consideration of innovative approaches for instance segmentation, two additional selection criteria were added:

3. **Innovation**: Architectures with new or innovative approaches are to be examined for suitability.
4. **Usage in PLF research**: The recently used instance segmentation architectures in PLF literature should be considered for comparison.

The website paperswithcode[4] provides an overview of all published instance segmentation architectures and their benchmark results on the COCO test-dev, a dataset on which model performance is evaluated and benchmarked. Since we also use the COCO format for training and evaluation in this paper, this overview serves as a basis for selecting the instance segmentation models based on our criteria. For each defined criterion, an instance

[4] https://paperswithcode.com/task/instance-segmentation

segmentation model was chosen. The following models were selected and evaluated in this paper:

DetectoRS: DetectoRS achieves state of the art (SOTA) performance on COCO test-dev for instance segmentation [31], which is why it was selected for this paper based on criterion 1. Inspired by the human mechanism of looking and thinking twice, the authors tried to implement similar mechanisms into their architecture at both the macro and the micro level. At the macro level, a Recursive Feature Pyramid (RFP) is proposed which builds on top of a Feature Pyramid Network (FPN). The RFP incorporates additional feedback connections from the FPN layers into the bottom-up backbone layers which creates a recursive operation. At the micro level, Switchable Atrous Convolutions (SAC) are incorporated, which convolves the same input feature with different atrous rates and gathers the results using switch functions. The incorporation of this mechanism into both levels improved the mAP on COCO test-dev by up to 4.3%.

SOLOv2: SOLO is an anchorless instance segmentation approach introduced by Wang et al. [32]. It divides an image into a uniform grid with each grid cell being responsible for the detection of an object if its centre is placed in it. Class probabilities and a global binary mask are computed for each cell individually. This restricts each cell to only predict one object or class. The model combines a FPN with a category and mask branch and reduces the problem of instance segmentation to the question of which cell and category a pixel belongs to. SOLOv2 extends this idea by improving the mask branch of the model [33]. Two new branches are introduced, the kernel branch and the feature branch. The kernel branch is responsible for generating a kernel for each cell of the grid while the feature branch generates multiple different prototype masks. Finally, the kernels are used in a convolution operation on the prototype masks to generate the final predictions. SOLOv2 archives SOTA performance in real time inference and is selected based on criteria 2.

DETR: DETR describes a novel object detection method introduced by Carion et al. [34]. The method offers a new approach to object detection as it uses a feature extractor in combination with Transformers [35]. The model uses the feature vectors extracted by a Convolutional Neural Network (CNN) backbone as an input for the encoder and its attention heads. The decoder then generates a defined number of predictions in parallel, each of which is assigned to a class either from the given dataset or an additional class like the background class. While its initial implementation focusses on the prediction of bounding boxes, the authors also demonstrate the adaptation to instance segmentation. This is done by utilizing the output of the decoders in combination with the multi-head-attention values und multiple convolutional layers to generate upscaled binary masks. Due to the innovative approach of instance segmentation with transformers, DETR was selected based on criteria 3.

Mask R-CNN: Mask R-CNN extends on Faster R-CNN by adding an additional branch for masks prediction that works parallel to the branch for bounding box prediction. Thus, the pixel wise prediction of the individual segmentation masks is decoupled from the actual classification of the object. A RoIAlign layer is introduced to improve bounding box alignment after RoIPooling and preserves exact spatial locations. For segmentation mask prediction, a Fully Convolutional Network (FCN) is used on each of the extracted bounding boxes [9]. Since Mask R-CNN is the most current instance segmentation architecture in PLF literature and no evaluation in COCO format has been performed yet, Mask R-CNN was selected based on criteria 4.

Dataset description

To evaluate the performance of selected instance segmentation architectures, a dataset consisting of a total of 731 images with high quality segmentation masks was created. The open source tool Labelme was used to annotate the images and convert them into the COCO format [36]. To ensure heterogeneity in the data, the dataset was compiled from a combination of samples from several datasets. Fig. 2 shows exemplary images that

illustrates the data heterogeneity. When creating the dataset, we tried to cover as many different backgrounds, camera perspectives, shooting lenses and lighting conditions as possible. Care was also taken to include PLF specific challenges such as piling, occlusion or overlapping of pigs in the dataset. Psota et al. published a dataset of a total of 2000 keypoint

Figure 2. Images from the dataset

Figure 3. Mask example

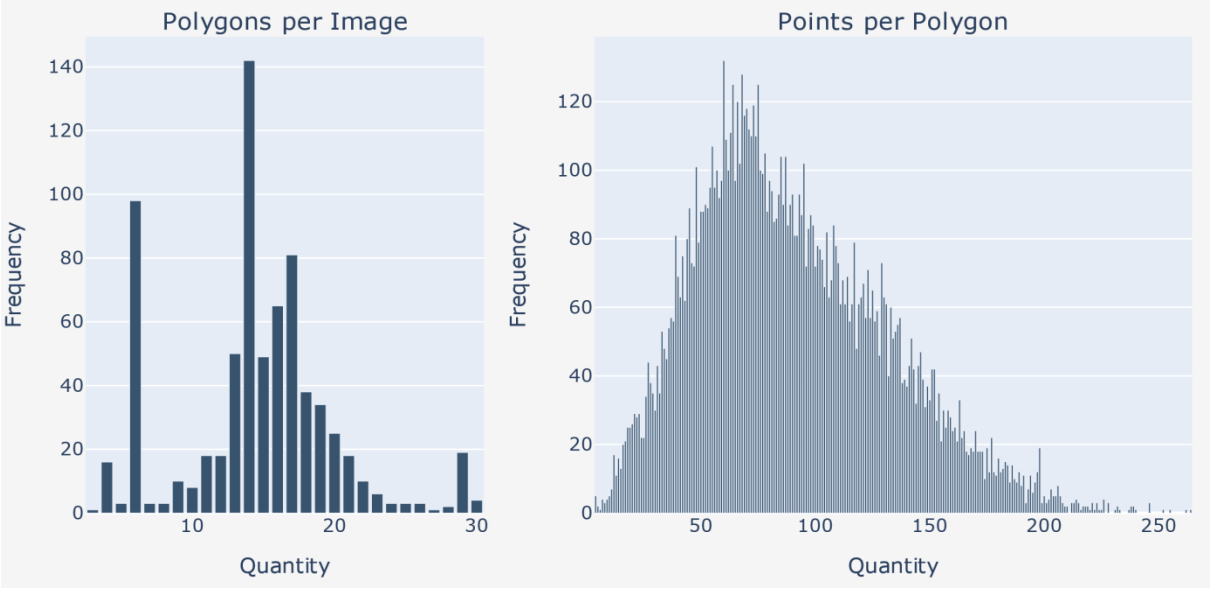


Figure 4. Descriptive statistics of the dataset.

annotated images from 17 different locations [37], of which a total of 631 images were extracted and annotated. The remaining 94 images were provided by the KoVeSch[5] project of the Lower Saxony Chamber of Agriculture, which includes 60 pictures from piglet rearing and 34 pictures with fattening pigs. All images were annotated by hand. Three people were involved in the labelling process, with each annotated image being checked by another person to assure quality and correctness of the annotations. Fig. 4 shows some descriptive statistics about the dataset. Each image contains between 3 and 30 pigs, while the average number of pigs per image is 14.5. The average number of coordinate points per mask is about 90, while a few masks can consist of up to 264 points. Fig. 3 shows an example of

[5] https://www.lwk-niedersachsen.de/index.cfm/portal/1/nav/1093/article/34849.html

annotated masks. Compared to the other datasets, our dataset has a higher number of pigs per image and a higher variation of data within the set, as both different locations and different camera angles were considered. The data set was divided into a training and test set, with 75% being used for training and 25% being used for testing and benchmarking.

Test environment and setup

Model training was performed on a desktop workstation with two Nvidia RTX 3090 with 24 GB VRAM each, a Threadripper 3960X and 64 GB RAM. The MMDetection framework was used to train and evaluate Mask-RCNN and DetectoRS [38], while the AdelaiDet toolbox was used for SOLOv2 [39]. As the DETR implementation in MMDetection did not provide instance segmentation, the original implementation of the authors[6] was used instead. To check the suitability of the models for instance segmentation in the field of Pig PLF, it was decided to use the default configuration for each model and to not make any adjustments to the parameters. For each training job, we fine-tuned the respective architecture using models pre-trained on the COCO dataset. A Resnet50 backbone was used for each of the tested models, so that, apart from a few deviations in the respective configuration file, all models were trained and evaluated on the same baseline. Each model was fine-tuned over 30 epochs.

Results

Quantitative evaluation

The models were evaluated based on their mAP including $AP^{IoU=0.50}$ and $AP^{IoU=0.75}$, inference speed on GPU and CPU, and number of parameters. The number of parameters describes the model size and can affect the required hardware to train and operationalize the respective model. The IoU threshold specifies the threshold at which a prediction is classified as true positive, while the mAP represents the average of all determined APs. The mAP was calculated from the results for IoU thresholds in the range 0.5 to 0.95 with a step size of 0.05 represented as AP@[.5:.05:.95] [27].

Table 1. Results of the evaluation on the test set.

Model	Average precision			Inference time in s		# Parameters
	mAP	$AP^{IoU=0.50}$	$AP^{IoU=0.75}$	GPU	CPU	
DetectoRS	**0.848**	0.978	**0.947**	0.147	-	131,648,615
SOLOv2	0.831	**0.980**	0.946	0.097	**0.905**	46,175,681
DETR	0.830	0.976	0.933	0.122	2.262	**42,613,152**
Mask R-CNN	0.822	0.978	0.946	**0.066**	1.574	43.971,158

As seen in Tab. 1, the best performance in prediction accuracy was achieved using DetectoRS. The model achieves a mAP of 0.848 and is thus slightly better than the competition, but also has the highest resource requirements. With 131 million parameters, DetectoRS is by far the biggest model compared to the others, although inference on GPU is only slightly slower than the more lightweight SOLOv2. For DetectoRS, inference on CPU was not possible because it was not supported by the used framework. On GPU, Mask R-CNN was the fastest among the tested models with an inference time of 0.06s per image, but provides the lowest mAP compared to the other models. However, measured by $AP^{IoU=0.50}$, it can be observed that Mask R-CNN gives a better result than DETR and achieves identical performance to DetectoRS. This is also true for $AP^{IoU=0.75}$, although the difference between Mask R-CNN and DETR is even greater here. SOLOv2 is slightly slower than Mask R-CNN

[6] https://github.com/facebookresearch/detr

but has the best overall result for $AP^{IoU=0.50}$. It is also the fastest model when tested on CPU. DETR represents the smallest model with a total of 42 million parameters but is also the slowest on CPU. Overall, SOLOv2, DETR, and Mask R-CNN do not differ significantly in their parameter size. In general, it is noticeable that the results for $AP^{IoU=0.50}$ of all tested models are similar. Differences in performance are only noticeable at higher thresholds.

Qualitative evaluation

For the qualitative evaluation of the predicted masks, we selected an example image from the test set that included as many of the mentioned visual challenges in Pig PLF as possible. For visualization purposes, the visualization method provided by the respective package was used, which explains the different coloured masks of the respective visualizations. Fig. 5 shows the predictions of the different models on the selected example image. As with the measured mAPs, the quality of the predicted masks is at a similar level for all models. Different coloured points were placed on the image to mark specific areas where the quality of the predicted masks sometimes deviated significantly. The red points indicate that in some cases, DETR and SOLOv2 have problems assigning pixels to the correct instance. While DETR does not assign the pixels to any instance at all, SOLOv2 tends to assign them to the wrong one. However, these errors can be observed in any of the tested architectures. This inconsistency can be illustrated with the help of the white points that focus on the pig's tail.

Figure 5. Predictions on test image.

Although DetectoRS produces the highest overall quality masks in comparison, correct segmentation of the tail is possible in only 2 out of 5 cases. In this task, DETR generates the best segmentations for the tail region. A problem that every tested model struggles with is highlighted by the orange dots. Due to the overlap of the pigs, the legs cannot be assigned to the correct mask by any model. The weaknesses of the models highlighted by the different points can be found consistently in other test images as well.

After comparing the individual models and identifying the best model based on mAP and visual evaluation, the generalization ability of the model should also be demonstrated. For this purpose, a series of images was selected from the test set containing images with as many different camera positions, shooting angles, camera lenses and light conditions as possible. Fig. 6 shows an overview of these images. As can be seen in the images, the segmentation of the individual masks also succeeds in completely different scenarios.

Although there are similar detail errors as in Fig. 5, the general quality of the masks is already promising. Of all the images, there is only one in which a pig was not provided with a mask, found in red rectangle. This problem occurred in about 10% of the tested images and if present, always occurs in crowded areas. This is similar to the problem mentioned by Seo et al. and is caused by the Non-Maximum Suppression (NMS) algorithm, which He et al. also pointed out in their introduction of the Mask-R CNN architecture [9]. Adjusting the threshold of the NMS could therefore lead to better results. It is also be seen that in theory the model can be applied in different pig compartments as well. Both piglets and fattening pigs can be correctly segmented by the model.

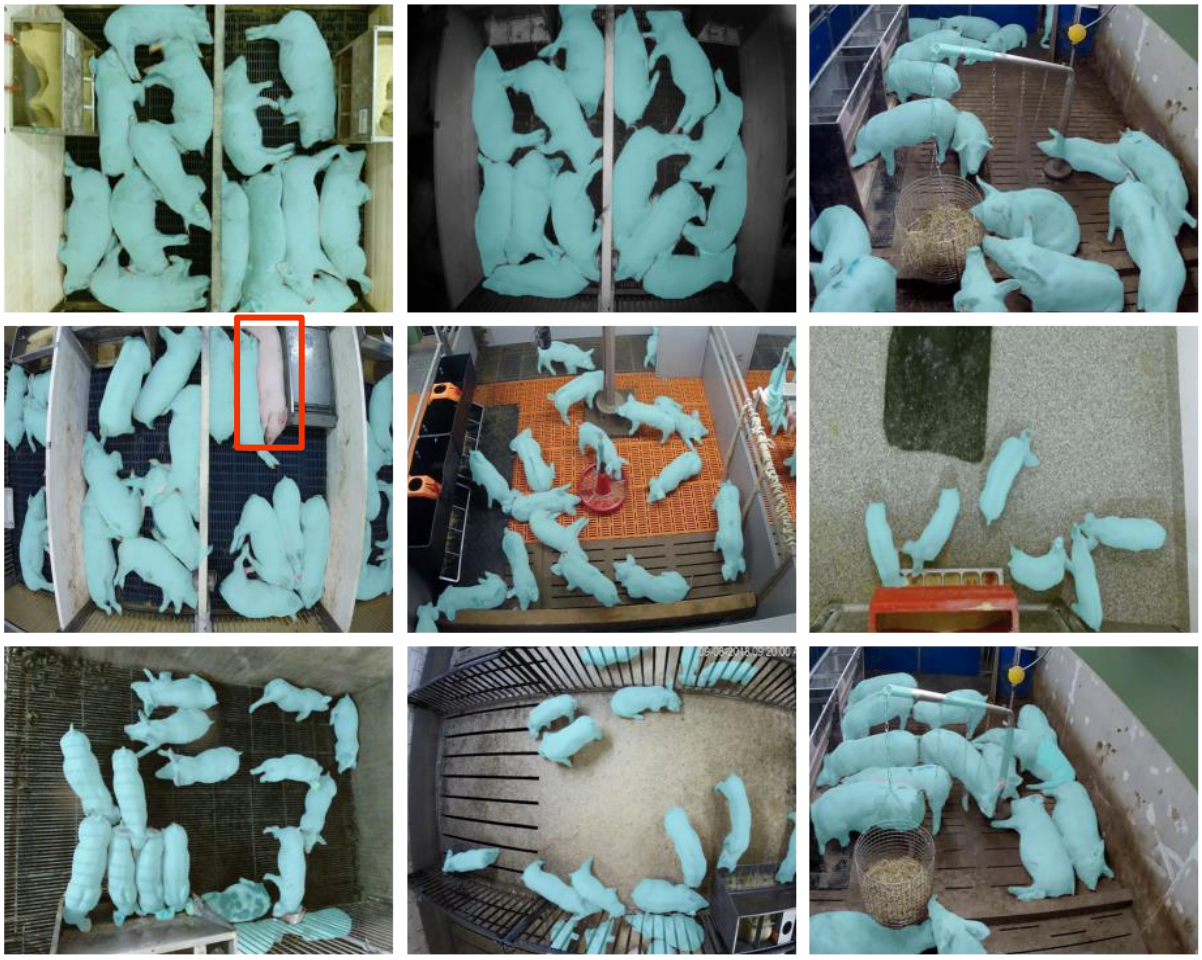


Figure 6. Predictions of DetectoRS.

Interpretation of the results

Both the quantitative and the qualitative evaluation show that the results of the models do not differ significantly in terms of prediction accuracy. In this context, it can also be stated that each of the four instance segmentation models we tested is applicable to Pig PLF data when it comes to prediction accuracy. The fact that the prediction accuracy of the models is so close may be related to the fact that in this case of the Pig PLF, only one class needs to be predicted, whereas the COCO dataset on which the models evaluated here were trained on, contains 80 different classes. Although the number of parameters of DetectoRS is almost three times larger than that of the other models, it does not result in a significantly better mAP. This means that increasing the size and complexity of the model does not necessarily have a significant effect on improving the mAP in this type of use case. Accordingly, the problems identified in the qualitative evaluation for instance segmentation of pigs cannot be solved by scaling the models vertically in depth but require other approaches to improve segmentation accuracy. Taking the prediction speed and the number of parameters into account, it can be stated that SoloV2, DETR and MASK R-CNN are better suited than

DetectoRS for potential use cases in this domain under current circumstances. This is due to the fact that fewer resources are required to operationalize these models, allowing them to be deployed on less expensive hardware. With respect to the prediction accuracy, it is also necessary to differentiate for which use case the respective models should be used and what the requirements for accuracy are in this case. Based on the results, it can be stated that Mask R-CNN can be used as a baseline due to its fast execution time, small size, and its accuracy. For use cases such as tail bite detection, where more precise masks might be needed, DETR could be used as a baseline.

Conclusion and outlook

In this paper, we demonstrated the usability of the deep learning instance segmentation models DetectoRS, SOLOv2, DETR and Mask R-CNN when being applied to data from the field of PLF for instance segmentation of pigs. The standard evaluation metric mAP was also applied for the first time to uniformly evaluate deep learning instance segmentation models in the Pig PLF domain to make performance more comparable. The results show that, in terms of prediction accuracy, each of the tested models can in principle be applied for instance segmentation of pigs. We observed that for instance segmentation in this context, the complexity and size of the model does not have a significant impact on the mAP, as the less complex models Mask R-CNN, SOLOv2 and DETR achieve similar prediction accuracy compared to DetectoRS. Accordingly, the identified problems in pig instance segmentation such as incorrect mask assignment when pigs overlap cannot be solved by vertically scaling models in depth but require other approaches or improvements. Based on this work, future research in this domain will focus on the aspect of cost efficiency when evaluating instance segmentation models for PLF systems. For example, it could be investigated which of the tested models can be deployed and operationalized on low-cost hardware or edge devices. Since in the context of this paper only the default configuration of each framework was applied to create the training jobs, the optimization of the configuration parameters could also be a direction for future research. Here, methods such as hyperparameter tuning could be applied to find an optimal configuration of the respective models to investigate the influence of these on the mAP. Alternative instance segmentation models and architectures such as YOLACT could also be explored in this domain for suitability in future research.

References

[1] Statisitsches Bundesamt (Destatis), *Betriebe: Deutschland, Jahre, Tierarten.* [Online]. Available: https://www-genesis.destatis.de/genesis/online, Code: 41311-0003 (accessed: Feb. 17 2021).

[2] Statisitsches Bundesamt (Destatis), *Gehaltene Tiere: Deutschland, Jahre, Tierarten.* [Online]. Available: https://www-genesis.destatis.de/genesis/online, Code: 41311-0001 (accessed: Feb. 17 2021).

[3] D. Berckmans, “Precision livestock farming technologies for welfare management in intensive livestock systems,” *Revue scientifique et technique (International Office of Epizootics)*, vol. 33, no. 1, pp. 189–196, 2014, doi: 10.20506/rst.33.1.2273.

[4] R. B. D'Eath *et al.,* “Automatic early warning of tail biting in pigs: 3D cameras can detect lowered tail posture before an outbreak,” *PloS one*, vol. 13, no. 4, e0194524, 2018, doi: 10.1371/journal.pone.0194524.

[5] J. Cowton, I. Kyriazakis, T. Plötz, and J. Bacardit, “A Combined Deep Learning GRU-Autoencoder for the Early Detection of Respiratory Disease in Pigs Using Multiple Environmental Sensors,” *Sensors (Basel, Switzerland)*, vol. 18, no. 8, p. 2521, 2018, doi: 10.3390/s18082521.

[6] C. Chen *et al.,* “Recognition of aggressive episodes of pigs based on convolutional neural network and long short-term memory,” *Computers and Electronics in Agriculture*, vol. 169, p. 105166, 2020, doi: 10.1016/j.compag.2019.105166.

[7] C. Chijioke Ojukwu, Y. Feng, G. Jia, H. Zhao, and H. Ta, “Development of a computer vision system to detect inactivity in group-housed pigs,” *International Journal of Agricultural and Biological Engineering*, vol. 13, no. 1, pp. 42–46, 2020, doi: 10.25165/j.ijabe.20201301.5030.

[8] S. Zhang, J. Yang, and B. Schiele, “Occluded Pedestrian Detection Through Guided Attention in CNNs,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition: CVPR 2018 : proceedings : 18-22 June 2018, Salt Lake City, Utah*, Salt Lake City, UT, 2018, pp. 6995–7003, doi: 10.1109/CVPR.2018.00731

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” Mar. 2017. Accessed: May 28 2020. [Online]. Available: http://arxiv.org/pdf/1703.06870v3

[10] Y. Cang, H. He, and Y. Qiao, “An Intelligent Pig Weights Estimate Method Based on Deep Learning in Sow Stall Environments,” *IEEE Access*, vol. 7, no. 99, pp. 164867–164875, 2019, doi: 10.1109/ACCESS.2019.2953099.

[11] A. Nasirahmadi *et al.,* “Deep Learning and Machine Vision Approaches for Posture Detection of Individual Pigs,” *Sensors (Basel, Switzerland)*, vol. 19, no. 17, 2019, doi: 10.3390/s19173738.

[12] J. Sa, Y. Choi, H. Lee, Y. Chung, D. Park, and J. Cho, “Fast Pig Detection with a Top-View Camera under Various Illumination Conditions,” *Symmetry*, vol. 11, no. 2, p. 266, 2019, doi: 10.3390/sym11020266.

[13] S. Küster, M. Kardel, S. Ammer, J. Brünger, R. Koch, and I. Traulsen, “Usage of computer vision analysis for automatic detection of activity changes in sows during final gestation,” *Computers and Electronics in Agriculture*, vol. 169, p. 105177, 2020, doi: 10.1016/j.compag.2019.105177.

[14] S. Schukat and H. Heise, “Indikatoren für die Früherkennung von Schwanzbeißen bei Schweinen – eine Metaanalyse,” (in de) *Berichte über Landwirtschaft - Zeitschrift für Agrarpolitik und Landwirtschaft*, vol. 11, no, 22, 2019, doi: 10.12767/BUEL.V97I3.249.

[15] S. Tu *et al.,* “Instance Segmentation Based on Mask Scoring R-CNN for Group-housed Pigs,” in *2020 International Conference on Computer Engineering and Application: ICCEA 2020 : 27-29 March 2020, Guangzhou, China : proceedings*, Guangzhou, China, 2020, pp. 458–462, doi: 10.1109/ICCEA50009.2020.00105

[16] B. Li, L. Liu, M. Shen, Y. Sun, and M. Lu, “Group-housed pig detection in video surveillance of overhead views using multi-feature template matching,” *Biosystems Engineering*, vol. 181, pp. 28–39, 2019, doi: 10.1016/j.biosystemseng.2019.02.018.

[17] A. Nasirahmadi, S. A. Edwards, S. M. Matheson, and B. Sturm, “Using automated image analysis in pig behavioural research: Assessment of the influence of enrichment substrate provision on lying behaviour,” *Applied Animal Behaviour Science*, vol. 196, pp. 30–35, 2017, doi: 10.1016/j.applanim.2017.06.015.

[18] S. Lee, H. Ahn, J. Seo, Y. Chung, D. Park, and S. Pan, “Practical Monitoring of Undergrown Pigs for IoT-Based Large-Scale Smart Farm,” *IEEE Access*, vol. 7, pp. 173796–173810, 2019, doi: 10.1109/ACCESS.2019.2955761.

[19] J. Kim *et al.,* “Depth-Based Detection of Standing-Pigs in Moving Noise Environments,” *Sensors (Basel, Switzerland)*, vol. 17, no. 12, 2017, doi: 10.3390/s17122757.

[20] K. Jun, S. J. Kim, and H. W. Ji, “Estimating pig weights from images without constraint on posture and illumination,” *Computers and Electronics in Agriculture*, vol. 153, pp. 169–176, 2018, doi: 10.1016/j.compag.2018.08.006.

[21] A. Nasirahmadi, O. Hensel, S. A. Edwards, and B. Sturm, “Automatic detection of mounting behaviours among pigs using image analysis,” *Computers and Electronics in Agriculture*, vol. 124, pp. 295–302, 2016, doi: 10.1016/j.compag.2016.04.022.

[22] W. Huang, W. Zhu, C. Ma, Y. Guo, and C. Chen, “Identification of group-housed pigs based on Gabor and Local Binary Pattern features,” *Biosystems Engineering*, vol. 166, pp. 90–100, 2018, doi: 10.1016/j.biosystemseng.2017.11.007.

[23] J. Seo, J. Sa, Y. Choi, Y. Chung, D. Park, and H. Kim, “A YOLO-based Separation of Touching-Pigs for Smart Pig Farm Applications,” in *2019 21st International Conference on Advanced Communication Technology (ICACT)*, 2019, pp. 395–401, doi: 10.23919/ICACT.2019.8701968.

[24] D. Li, Y. Chen, K. Zhang, and Z. Li, “Mounting Behaviour Recognition for Pigs Based on Deep Learning,” *Sensors (Basel, Switzerland)*, vol. 19, no. 22, 2019, doi: 10.3390/s19224924.

[25] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, “Mask Scoring R-CNN,” Mar. 2019. [Online]. Available: https://arxiv.org/pdf/1903.00241

[26] T.-Y. Lin *et al.,* “Microsoft COCO: Common Objects in Context,” May. 2014. [Online]. Available: https://arxiv.org/pdf/1405.0312

[27] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, “A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit,” *Electronics*, vol. 10, no. 3, p. 279, 2021, doi: 10.3390/electronics10030279.

[28] M. A. Rahman and Y. Wang, “Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation,” in *ISVC*, 2016, doi: 10.1007/978-3-319-50835-1_22

[29] M. Thoma, “A Survey of Semantic Segmentation,” Feb. 2016. [Online]. Available: https://arxiv.org/pdf/1602.06541

[30] T. Norton, C. Chen, M. L. V. Larsen, and D. Berckmans, “Review: Precision livestock farming: building 'digital representations' to bring the animals closer to the farmer,” *animal*, vol. 13, no. 12, pp. 3009–3017, 2019, doi: 10.1017/S175173111900199X.

[31] S. Qiao, L.-C. Chen, and A. Yuille, “DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution,” Jun. 2020. [Online]. Available: https://arxiv.org/pdf/2006.02334

[32] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, “SOLO: Segmenting Objects by Locations,” Dec. 2019. [Online]. Available: https://arxiv.org/pdf/1912.04488

[33] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, “SOLOv2: Dynamic and Fast Instance Segmentation,” Mar. 2020. [Online]. Available: https://arxiv.org/pdf/2003.10152

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” May. 2020. Accessed: May 28 2020. [Online]. Available: http://arxiv.org/pdf/2005.12872v2

[35] A. Vaswani *et al.,* “Attention Is All You Need,” Jun. 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762

[36] Kentaro Wada, *labelme: Image Polygonal Annotation with Python*.

[37] E. T. Psota, M. Mittek, L. C. Pérez, T. Schmidt, and B. Mote, “Multi-Pig Part Detection and Association with a Fully-Convolutional Network,” *Sensors (Basel, Switzerland)*, vol. 19, no. 4, p. 852, 2019, doi: 10.3390/s19040852.

[38] K. Chen *et al.,* “MMDetection: Open MMLab Detection Toolbox and Benchmark,” *arXiv preprint arXiv:1906.07155*, 2019.

[39] Z. Tian, H. Chen, X. Wang, Y. Liu, and C. Shen, *AdelaiDet: A Toolbox for Instance-level Recognition Tasks*.

24th International Conference on Business Information Systems (BIS 2021)

Artificial Intelligence

https://doi.org/10.52825/bis.v1i.62

Published: 02 July 2021

Text-Aware Predictive Monitoring of Business Processes

Marco Pegoraro[1][https://orcid.org/0000-0002-8997-7517], Merih Seran Uysal[1][https://orcid.org/0000-0003-1115-6601], David Benedikt Georgi[1], and Wil M.P. van der Aalst[1][https://orcid.org/0000-0002-0955-6940]

[1]Process and Data Science chair, RWTH Aachen University, Aachen, Germany
{pegoraro,uysal,wvdaalst}@pads.rwth-aachen.de david. georgi@rwth-aachen.de

Abstract. The real-time prediction of business processes using historical event data is an important capability of modern business process monitoring systems. Existing process prediction methods are able to also exploit the data perspective of recorded events, in addition to the control-flow perspective. However, while well-structured numerical or categorical attributes are considered in many prediction techniques, almost no technique is able to utilize text documents written in natural language, which can hold information critical to the prediction task. In this paper, we illustrate the design, implementation, and evaluation of a novel *text-aware process prediction model* based on *Long Short-Term Memory* (LSTM) neural networks and natural language models. The proposed model can take categorical, numerical and textual attributes in event data into account to predict the activity and timestamp of the next event, the outcome, and the cycle time of a running process instance. Experiments show that the text-aware model is able to outperform state-of-the-art process prediction methods on simulated and real-world event logs containing textual data.

Keywords: Predictive Monitoring, Process Mining, Natural Language Processing, LSTM Neural Networks

1 Introduction

In recent years, a progressive and rapid tendency to digital transformation has become apparent in most aspects of industrial production, provision of services, science, education, and leisure. This has, in turn, caused the widespread adoption of new technologies to support human activities. A significant number of these technologies specialize in the management of enterprise business processes.

The need of analysis and compliance in business processes, united to a larger and larger availability of historical event data have stimulated the birth and growth of the scientific discipline of *process mining*. Process mining enables the discovery of process models from historical execution data, the measurement of compliance between data and a process model, and the enhancement of process models with additional information extracted from complete process cases.

Advancements in process mining and other branches of data science have also enabled the possibility of adopting *prediction* techniques, algorithms that train a mathematical model from known data instances and are able to perform accurate estimates of various features of future instances. In the specific context of process mining, *predictive monitoring* is the task of

predicting features of *partial process instances*, i.e., cases of the process still in execution, on the basis of recorded information regarding complete process instances. Examples of valuable information on partial process instances are the next activity in the process to be executed for the case, the time until the next activity, the completion time of the entire process instance, and the last activity in the case (outcome). If accurately estimated, these case features can guide process owners in making vital decisions, and improve operations within the organization that hosts the process; as a result, accurate predictive monitoring techniques are widely desirable and a precious asset for companies and organizations.

Existing predictive monitoring techniques typically operate at the merging point between process mining and machine learning, and are able to consider not only the control-flow perspective of event data (i.e., the activity, the case identifier, and the timestamp), but also additional data associated with them. However, few prediction techniques are able to exploit attributes in the form of text associated with events and cases. These textual attributes can hold crucial information regarding a case and its status within the workflow of a process. A general framework describing the problem is shown in Figure 1.

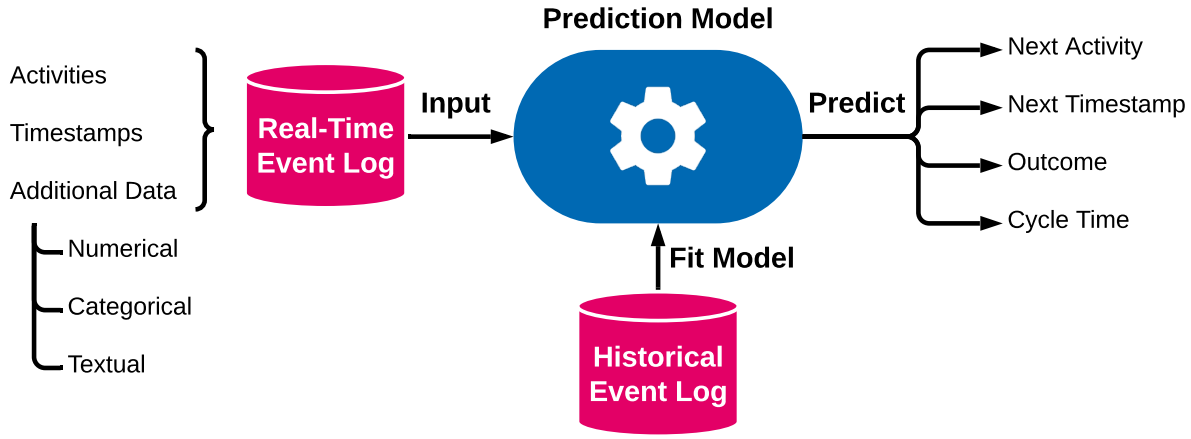


Figure 1. Problem overview: a general predictive monitoring model. The aim is predicting features of running process instances based on historical data, by exploiting numerical, categorical, and textual data.

The aim of this paper is to assess the extent to which textual information can influence predictive monitoring. To this end, we present a novel predictive monitoring approach able to exploit numerical, categorical, and textual attributes associated with events, as well as control-flow information. Our prediction model estimates features of cases in execution by combining a set of techniques for sequential and textual data encoding with predictions from an LSTM neural network, a machine learning technique particularly effective on sequential data such as process traces. Validation through experiments on real-life event logs shows that our approach is effective in extracting additional information from textual data, and outperforms state-of-the-art approaches for predictive monitoring.

The remainder of the paper is structured as follows. Section 2 discusses some recent work related to predictive monitoring. Section 3 presents some preliminary definitions. Section 4 illustrates the details and architecture of our text-aware predictive monitoring technique. Section 5 presents the evaluation of the predictor and the results of the experiments. Section 6 concludes the paper.

2 Related Work

The intersection of process mining and machine learning is a rich and influential field of research. Among the numerous applications of machine learning in process mining, feature prediction on partial process traces based on historical complete traces (i.e., predictive monitoring)

is particularly prominent. Accurately predicting features such as cycle time and bottlenecks leads to valuable insights in many domains, e.g. production processes [17].

Earlier techniques for prediction in process mining focused on white-box and human-interpretable models, largely drawn from statistics. Many proposals have been put forward to compute an estimate of the cycle time of a process instance, including decision trees [6] and simulation through stochastic Petri nets [13]. Additionally, Teinemaa et al. [16] proposed a process outcome prediction method based on random forests and logistic regression. Van der Aalst et al. [1] exploit process discovery as a step of the prediction process, obtaining estimations through replay on an annotated transition system; this technique is then extended by Polato et al. [12] by annotating a discovered transition system with an ensemble of naïve Bayes and support vector regressors, allowing for the data-aware prediction of cycle time and next activity.

The second half of the 2010s saw a sharp turn from ensemble learning to single prediction models, and from white-box to black-box models – specifically, recurrent neural networks. This is due to the fact that recurrent neural networks have been shown to be very accurate in learning from sequential data. However, they are not interpretable, and the training efficiency is often lower.

This family of prediction methods employs LSTM neural networks to estimate process instance features. Evermann et al. [7] proposed the use of LSTMs for next activity prediction; Tax et al. [15] trained LSTMs to predict cycle time of process instances. Navarin et al. [9] extended this approach by feeding additional attributes in the LSTM, attaining data-aware prediction. More recently, Park and Song [10] merged system-level information from a process model with a compact trace representation based on deep neural networks to attain performance prediction.

No existing predictive monitoring technique, to the best of our knowledge, incorporates information from free text, recorded as event or trace attribute, with the control-flow perspective of the process into a state-of-the-art LSTM neural network model for predictive monitoring: this motivates the approach we present in this paper.

3 Preliminaries

Let us first introduce some preliminary definitions and notations.

Definition 1 (Sequence) *A* sequence *of length $n \in \mathbb{N}_0$ over a set X is an ordered collection of elements defined by a function $\sigma\colon \{1,\dots,n\} \to X$, which assigns each index an element of X. A sequence of length n is represented explicitly as $\sigma = \langle x_1, x_2, \dots, x_n \rangle$ with $x_i \in X$ for $1 \leq i \leq n$. In addition, $\langle\ \rangle$ is the empty sequence of length 0. Over the sequence σ we define $|\sigma| = n$, $\sigma(i) = x_i$, and $x \in \sigma \Leftrightarrow \exists_{1 \leq i \leq n}\colon x = x_i$. X^* denotes the set of all sequences over X.*

The function $hd^k\colon X^* \to X^*$ gives the head or prefix of length k of σ for $0 \leq k \leq n$: $hd^k(\sigma) = \langle x_1, x_2, \dots, x_k \rangle$. For instance, $hd^2(\sigma) = \langle x_1, x_2 \rangle$.

Definition 2 (Event, Trace, Event Log, Prefix Log) *Let $\mathcal{A}$ be the universe of* activity labels*. Let $\mathcal{T}$ be the closed under subtraction and totally ordered universe of* timestamps*. Let $\mathcal{D}_1$, $\mathcal{D}_2, \dots, \mathcal{D}_m$ be the domains of* additional attributes*. An* event *is a tuple $e = (a, t, d_1, \dots, d_m) \in \mathcal{A} \times \mathcal{T} \times \mathcal{D}_1 \times \dots \times \mathcal{D}_m = \mathcal{E}$. Over an event e we define the projection functions $\pi_{\mathcal{A}}(e) = a$, $\pi_{\mathcal{T}}(e) = t$, and $\pi_{\mathcal{D}_i}(e) = d_i$. A* trace *$\sigma \in \mathcal{E}^*$ is a sequence of events such that timestamps are non-decreasing: $\pi_{\mathcal{T}}(e_i) \leq \pi_{\mathcal{T}}(e_j)$ for $1 \leq i < j \leq |\sigma|$. An* event log *$L \in \mathcal{B}(\mathcal{E}^*)$ is a multiset of traces. Given an event log L, we define the* prefix log *$\mathbb{L} = \{hd^k(\sigma) \mid \sigma \in L \wedge 1 \leq k \leq |\sigma|\}$.*

Additional attributes $d_i \in \mathcal{D}_i$ may be in the form of text, i.e., its domain is the set of sequences $\mathcal{D}_i = \Sigma^*$ from a fixed and known alphabet Σ.

Next, let us define the target functions for our predictions:

Definition 3 (Target Functions) *Let $\sigma \in \mathcal{E}^*$ be a non-empty trace, and let $1 \leq k \leq |\sigma|$. The* next activity *function $f_{\mathrm{a}} \colon \mathcal{E}^* \times \mathbb{N} \to \mathcal{A} \cup \{\blacksquare\}$ returns the activity of the next event, or an artificial activity $\blacksquare$ if the given trace is complete:*

$$f_{\mathrm{a}}(\sigma, k) = \begin{cases} \blacksquare & \text{if } k = |\sigma| \\ \pi_{\mathcal{A}}(\sigma(k+1)) & \text{else} \end{cases}$$

The next timestamp *function $f_{\mathrm{t}} \colon \mathcal{E}^* \times \mathbb{N} \to \mathcal{T}$ returns the time difference between the next event and last event in the prefix:*

$$f_{\mathrm{t}}(\sigma, k) = \begin{cases} 0 & \text{if } k = |\sigma| \\ \pi_{\mathcal{T}}(\sigma(k+1)) - \pi_{\mathcal{T}}(\sigma(k)) & \text{else} \end{cases}$$

The case outcome *function $f_{\mathrm{o}} \colon \mathcal{E}^* \to \mathcal{A}$ returns the last activity of the trace: $f_{\mathrm{o}}(\sigma) = \pi_{\mathcal{A}}(\sigma(|\sigma|))$. The* cycle time *function $f_{\mathrm{c}} \colon \mathcal{E}^* \to \mathcal{T}$ returns the total duration of the case, i.e., the time difference between the first and the last event of the trace: $f_{\mathrm{c}}(\sigma) = \pi_{\mathcal{T}}(\sigma(|\sigma|)) - \pi_{\mathcal{T}}(\sigma(1))$.*

The prediction techniques we show include the information contained in textual attributes of events. In order to be readable by a prediction model, the text needs to be processed by a *text model*. Text models rely on a *text corpus*, a collection of text fragments called *documents*. Before computing the text model, the documents in the corpus are preprocessed with a number of normalization steps: conversion to lowercase, tokenization (separation in distinct terms), lemmatization (mapping words with similar meaning, such as “diagnose” and “diagnosed” into a single lemma), and stop word removal (deletion of uninformative parts of speech, such as articles and adverbs). These transformation steps are shown in Table 1.

Step	Transformation	Example Document
0	Original	“The patient has been diagnosed with high blood pressure.”
1	Lowercase	“the patient has been diagnosed with high blood pressure.”
2	Tokenization	⟨“the”, “patient”, “has”, “been”, “diagnosed”, “with”, “high”, “blood”, “pressure”, “.”⟩
3	Lemmatization	⟨“the”, “patient”, “have”, “be”, “diagnose”, “with”, “high”, “blood”, “pressure”, “.”⟩
4	Stop word filtering	⟨“patient”, “diagnose”, “high”, “blood”, “pressure”⟩

Table 1. Text preprocessing transformation of an example document containing a single sentence.

In order to represent text in a structured way, we consider four different text models:

Bag of Words (BoW) [5]: a model where, given a vocabulary V, we encode a document with a vector of length $|V|$ where the i-th component is the *term frequency* (tf), the number of occurrences of the i-th term in the vocabulary, normalized with its *inverse document frequency* (idf), the inverse of the number of documents that contain the term. This tf-idf score accounts for term specificity and rare terms in the corpus. This model disregards the order between words.

Bag of N-Grams (BoNG) [5]: this model is a generalization of the BoW model. Instead of one term, the vocabulary consists of n-tuples of consecutive terms in the corpus. The unigram model ($n = 1$) is equivalent to the BoW model. For the bigram model ($n = 2$), the vocabulary consists of pairs of words that appear next to each other in the documents. The documents are encoded with the td-idf scores of their n-grams. This model is able to account for word order.

Paragraph Vector (Doc2Vec) [8]: in this model, a feedforward neural network is trained to predict one-hot encodings of words from their context, i.e., words that appear before or after the target word in the training documents. An additional vector, of a chosen size and unique for each document, is trained together with the word vectors. When the network converges, the additional vector carries information regarding the words in the corresponding document and their relationship, and is thus a fixed-length representation of the document.

Latent Dirichlet Allocation (LDA) [4]: a generative statistical text model, representing documents as a set of topics, which size is fixed and specified a priori. Topics are multinomial (i.e., categorical) probability distributions over all words in the vocabulary and are learned by the model in an unsupervised manner. The underlying assumption of the LDA model is that the text documents were created by a statistical process that first samples topic from a multinomial distribution associated with a document, then sample words from the sampled topics. Using the LDA model, a document is encoded as a vector by its topic distribution: each component indicates the probability that the corresponding topic was chosen to sample a word in the document. LDA does not account for word order.

In the next section, we will describe the use of text models in an architecture allowing to process a log to obtain a data- and text-aware prediction model.

4 Prediction Model Architecture

The goal of predictive monitoring is to estimate a target feature of a running process instance based on historical execution data. In order to do so, predictive monitoring algorithms examine *partial traces*, which are the events related to a process case at a certain point throughout its execution. Obtaining partial traces for an event log is equivalent to computing the set of all prefixes for the traces in the log. Prefix logs will be the basis for training our predictive model.

In this paper, we specifically address the challenge of managing additional attributes that are textual in nature. In order to account for textual information, we need to define a construction method for fixed-length vectors that encode activity labels, timestamps, and numerical, categorical, and textual attributes.

Given an event $e = (a, t, d_1, \ldots, d_m)$, its activity label a is represented by a vector $\vec{a}$ using *one-hot encoding*. Given the set of possible activity labels $\mathcal{A}$, an arbitrary but fixed ordering over $\mathcal{A}$ is introduced with a bijective index function $index_{\mathcal{A}} \colon \mathcal{A} \to \{1, \ldots, |\mathcal{A}|\}$. Using this function, the activity is encoded as a vector of size $|\mathcal{A}|$, where the component $index_{\mathcal{A}}(\pi_{\mathcal{A}}(e))$ has value 1 and all the other components have value 0. The function $\mathbb{1}_{\mathcal{A}} \colon \mathcal{A} \to \{0, 1\}^{\mathcal{A}}$ is used to describe the realization of such one-hot encoding $\vec{a} = \mathbb{1}_{\mathcal{A}}(\pi_{\mathcal{A}}(e))$ for the activity label of the event e.

In order to capture time-related correlations, a set of time-based features is utilized to encode the timestamp t of the event. We compute a time vector $\vec{t} = (\hat{t}_1, \hat{t}_2, \hat{t}_3, \hat{t}_4, \hat{t}_5, \hat{t}_6)$ of min-max normalized time features, where t_1 is the time since the previous event, t_2 is the time since the first event of the case, t_3 is the time since the first event of the log, t_4 is the time since midnight, t_5 is the time since previous Monday, and t_6 is the time since the first of January. The min-max normalization is obtained through the formula

$$\hat{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where $\min(x)$ is the lowest and $\max(x)$ is the highest value for the attribute x.

Every additional attribute d_i of e is encoded in a vector $\vec{d_i}$ as follows:

$$\vec{d_i} = \begin{cases} \mathbb{1}_{\mathcal{D}_i}(d_i) & \text{if } \mathcal{D}_i \text{ is categorical} \\ \hat{d}_i & \text{if } \mathcal{D}_i \text{ is numerical} \\ \textsc{TextModel}(d_1) & \text{if } \mathcal{D}_i \text{ is textual} \end{cases}$$

The encoding technique depends on the type of the attribute. Categorical attributes are one-hot encoded similarly to the activity label. Numerical attributes are min-max normalized: if the minimum and maximum are not bounded conceptually, the lowest or highest value of the attribute in the historical event log is used for scaling. Finally, if $\mathcal{D}_i$ is a textual model, it is encoded in a fixed-length vector with one of the four text models presented in Section 3; the documents in the text corpus for the text model consist of all instances of the textual attribute $\mathcal{D}_i$ contained in the historical log. This technique allows to build a complete fixed-length encoding for the event $e = (a, t, d_1, \ldots, d_m)$, which we indicate with the tuple of vectors $\textit{enc}(e) = (\vec{a}, \vec{t}, \vec{d_1}, \ldots, \vec{d_m})$.

This encoding procedure allows us to build a training set for the prediction of the target functions presented in Section 3 utilizing an LSTM neural network.

Figure 2 illustrates the entire encoding architecture, and the fit/predict pipeline for our final LSTM model. The schematic distinguishes between the *offline* (fitting) phase, where we train the LSTM with encoded historical event data, and the *online* (real-time prediction) phase, where we utilize the trained model to estimate the four target features on running process instances. Given an event log L, the structure of the training set is based on the partial traces in its prefix log $\mathbb{L} = \{hd^k(\sigma) \mid \sigma \in L \wedge 1 \leq k \leq |\sigma|\}$. For each $\sigma = \langle e_1, e_2, \ldots e_n \rangle \in L$ and $1 \leq k \leq n$, we build an instance of the LSTM training set. The network input $\langle vecx_1, \vec{x_2}, \ldots, \vec{x_k} \rangle$ is given by the event encodings $vecx_1 = \textit{enc}(e_1)$, $vecx_2 = \textit{enc}(e_2)$, through $vecx_k = \textit{enc}(e_k)$. The targets $(\vec{y_a}, y_t, \vec{y_o}, y_c)$ are given by $\vec{y_a} = f_{\mathrm{a}}(\sigma, k)$, $\vec{y_t} = f_{\mathrm{t}}(\sigma, k)$, $\vec{y_o} = f_{\mathrm{o}}(\sigma)$, and $y_c = f_{\mathrm{c}}(\sigma, k)$.

Figure 3 shows the topology of the network. The training utilizes gradient descent and backpropagation through time (BPTT). The loss for numerical prediction values $\hat{y}$ and the true value y is the absolute error $\mathrm{AE}(\hat{y}, y) = |\hat{y} - y|$, while the loss for categorical prediction values is computed using the categorical cross-entropy error $\mathrm{CE}(\vec{\hat{y}}, \vec{y}) = -\sum_{i=1}^{k} y_i \cdot \log \hat{y}_i$.

5 Evaluation

The predictive monitoring approach presented in this paper has been implemented for validation, utilizing a Python-based, fully open-source technological stack. PM4Py [3] is a process mining Python tool developed by Fraunhofer FIT. It is used for event log parsing and its internal event log representation. The neural network framework Tensorflow [2], originally developed by Google, and its API Keras[1] were utilized to implement the final LSTM model. Furthermore, the libraries Scikit-learn [11], NLTK[2], and Gensim[3] provided the natural language processing capabilities required to preprocess and normalize text, as well as build and train the text models.

The text-aware model is compared to two other process prediction methods. First, the pure LSTM approach based on the ideas of Navarin et al. [9] is considered, which only uses the activity, timestamp, and additional non-textual attributes of each event. This approach can be considered the state of the art in predictive monitoring with respect to prediction accuracy. The second baseline is the process model-based prediction method originally presented by van der Aalst et al. [1]. This approach builds an annotated transition system for a log using a sequence, bag, or set abstraction. Each state of the transition system is annotated with measurements of historical traces that can be used to predict target values for unseen traces. During the prediction phase, running traces are mapped to the corresponding state of the transition system, and the measurements of the state are used to compute a prediction. We adopt the improvement of this method described in [14] to apply it to classification tasks and obtain the next activity and outcome predictions. The first 8 events of a trace are considered for the construction of the state space. Experiments with different horizon lengths (1, 2, 4, 16) mostly led to inferior results, and are thus not reported.

[1] https://keras.io/
[2] https://nltk.org/
[3] https://radimrehurek.com/gensim/

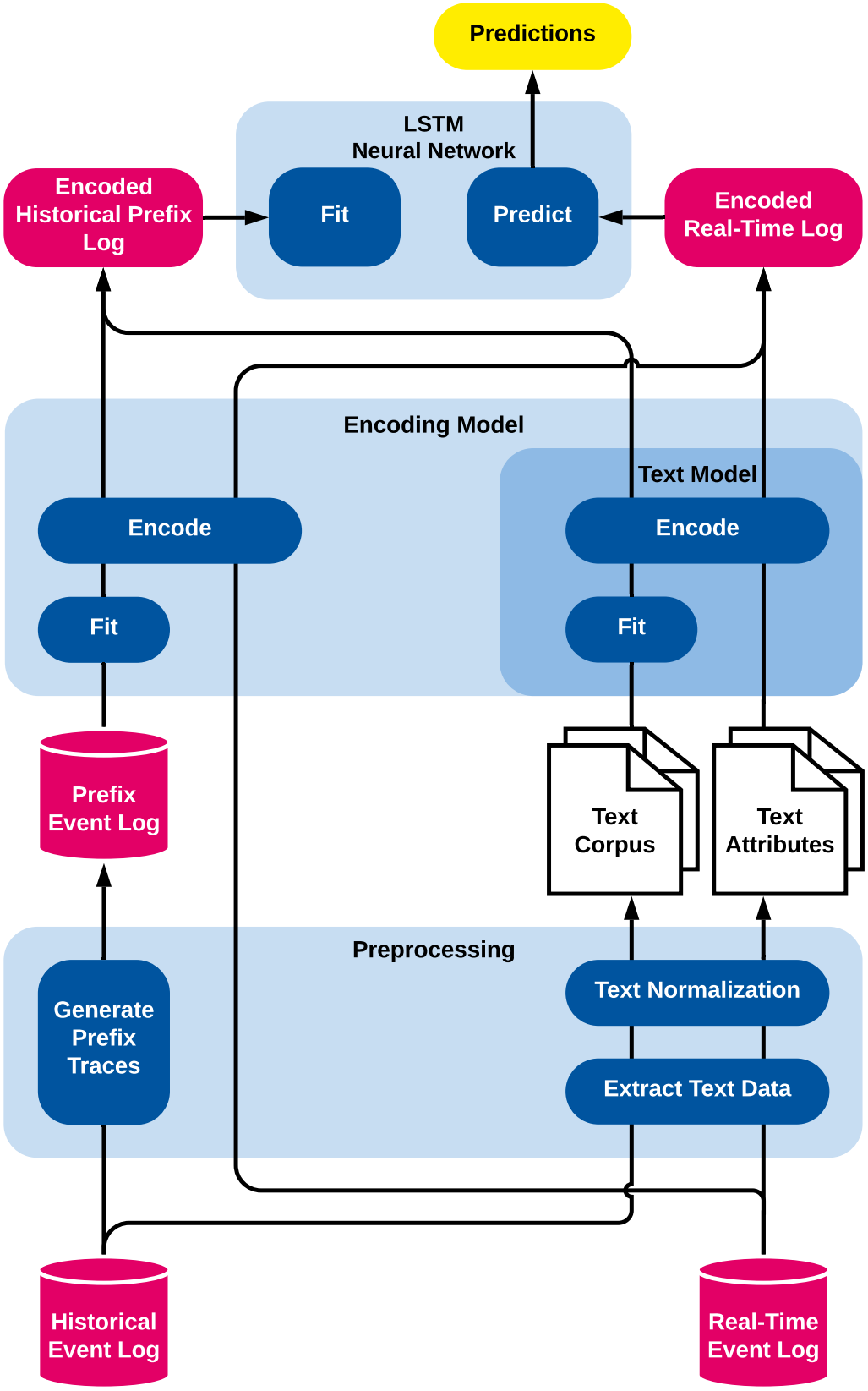


Figure 2. Overview of the text-aware process prediction model. Predictions for real-time processes are realized by an LSTM model that is fitted using an encoded representation of all prefixes of the historical event log. The encoding of textual attributes is realized by a text preprocessing pipeline and an exchangeable text encoding model.

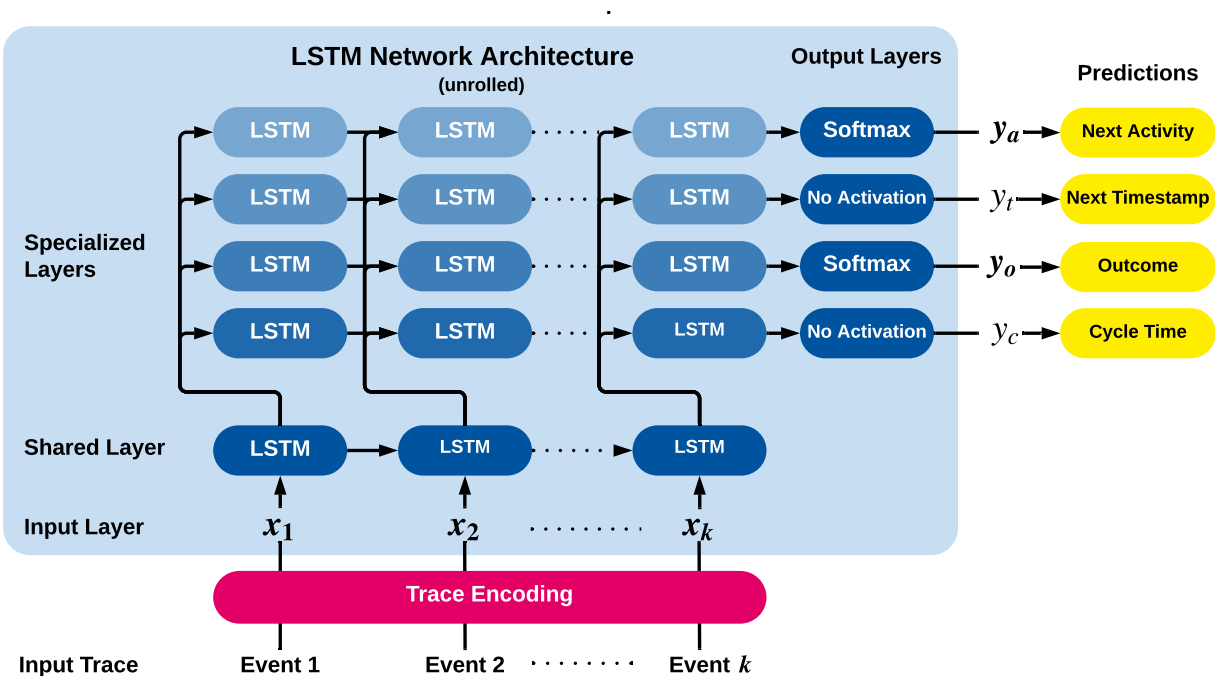


Figure 3. LSTM model architecture to simultaneously predict the next activity ($\vec{y_a}$), next event time (y_t), outcome ($\vec{y_o}$) and cycle time (y_c) for an encoded prefix trace $\vec{x_1}, \vec{x_2}, \ldots, \vec{x_k}$.

We evaluate the two baseline methods against our approach considering all four text models presented here, with a varying dimension of vector size (50, 100 and 500 for BoW and BoNG, 10, 20 and 100 for PV and LDA). The BoNG model is built with bigrams ($n = 2$). Of the four target functions presented in Section 3, classification tasks (next activity and outcome) are evaluated with a weighted-average class-wise F_1 score; regression tasks (next timestamp and cycle time) are evaluated on Mean Absolute Error (MAE). The first 2/3 of the chronologically ordered traces is used to fit the prediction model to the historical event data. The remaining 1/3 of traces are used to measure the prediction performance.

Event Log	Customer Journey	Hospital Admission
Cases	15 001	46 520
Trace variants	1001	2784
Events	55 220	117 952
Events per case (mean)	3.681	2.536
Median case duration (days)	0.224	7.579
Mean case duration (days)	0.713	121.154
Activities	18	26
Words before preprocessing	247 010	171 938
Words after preprocessing	98 915	165 285
Vocabulary before preprocessing	1203	4973
Vocabulary after preprocessing	817	4633
Text attribute	Customer question	Diagnosis
Additional non-textual attributes	Gender Age	Admission type Insurance

Table 2. Overview of the evaluated event logs with their key properties.

The process prediction models are evaluated on two real-world event logs, of which the general characteristics are given in Table 2. Additionally, snippets of the datasets are shown in Tables 3 and 4. The first describes the customer journeys of the Employee Insurance Agency

Case	Activity	Timestamp	Age	Gender	Message
40154127	question	2015/12/15 12:24:42.000	50-65	M	Can you send me a copy of the decision?
40154127	taken	2015/12/30 15:39:36.000	50-65	M	
40154127	mijn_sollicitaties	2015/12/30 15:39:42.000	50-65	M	
40154127	taken	2015/12/30 15:39:46.000	50-65	M	
40154127	home	2015/12/30 15:39:51.000	50-65	M	
23245109	question	2015/07/21 09:49:32.000	50-65	M	Law: How is the GAA (Average Number of Labor)?
23245109	question	2015/07/21 09:54:28.000	50-65	M	Dismissal Procedure: Stops my contract automatically after two years of illness?
23245109	question	2015/07/21 10:05:43.000	50-65	M	Dismissal: Am I entitled to a transitional allowance?
23245109	question	2015/07/21 10:05:56.000	50-65	M	Chain Determination: How often may be extended a fixed-term contract?
23245109	mijn_werkmap	2015/07/27 09:54:03.000	50-65	M	
23245109	mijn_berichten	2015/07/27 09:54:13.000	50-65	M	
23245109	mijn_cv	2015/07/27 10:04:20.000	50-65	M	
21537056	taken	2015/10/30 13:16:48.000	50-65	M	
21537056	question	2015/10/30 13:22:00.000	50-65	M	How can I add a document/share with my consultant work through the workbook?
21537056	taken	2015/10/30 13:23:24.000	50-65	M	
21537056	mijn_werkmap	2015/10/30 13:24:39.000	50-65	M	
19290768	question	2015/09/21 12:41:21.000	30-39	V	Filling: What should I do if I made a mistake when filling out the Income Problem?
19290768	home	2015/09/22 10:09:53.000	30-39	V	
19290768	taken	2015/09/22 10:10:14.000	30-39	V	
19290768	home	2015/09/22 10:11:12.000	30-39	V	
53244594	mijn_berichten	2016/02/25 09:10:40.000	40-49	M	
53244594	question	2016/02/25 13:27:38.000	40-49	M	When is/are transferred my unemployment benefits?
53244594	question	2016/02/29 10:04:23.000	40-49	M	Problem: I have to pay sv € 0 and further fill only the amount of holiday pay. What should I do if I get an error?
53244594	question	2016/02/29 10:10:52.000	40-49	M	Why did you change the amount of my payment?

Table 3. Snippet from the customer journey log.

commissioned by the Dutch Ministry of Social Affairs and Employment. The log is aggregated from two anonymized data sets provided in the BPI Challenge 2016, containing click data of

Case	Activity	Timestamp	Admission Type	Insurance	Diagnosis
8	PHYS REFERRAL/NORMAL DELI	2117-11-20 10:22:00	NEWBORN	Private	NEWBORN
8	HOME	2117-11-24 14:20:00	NEWBORN	Private	
9	EMERGENCY ROOM ADMIT	2149-11-09 13:06:00	EMERGENCY	Medicaid	HEMORRHAGIC CVA
9	DEAD/EXPIRED	2149-11-14 10:15:00	EMERGENCY	Medicaid	
10	PHYS REFERRAL/NORMAL DELI	2103-06-28 11:36:00	NEWBORN	Medicaid	NEWBORN
10	SHORT TERM HOSPITAL	2103-07-06 12:10:00	NEWBORN	Medicaid	
11	EMERGENCY ROOM ADMIT	2178-04-16 06:18:00	EMERGENCY	Private	BRAIN MASS
11	HOME HEALTH CARE	2178-05-11 19:00:00	EMERGENCY	Private	
12	PHYS REFERRAL/NORMAL DELI	2104-08-07 10:15:00	ELECTIVE	Medicare	PANCREATIC CANCER SDA
12	DEAD/EXPIRED	2104-08-20 02:57:00	ELECTIVE	Medicare	
13	TRANSFER FROM HOSP/EXTRAM	2167-01-08 18:43:00	EMERGENCY	Medicaid	CORONARY ARTERY DISEASE
13	HOME HEALTH CARE	2167-01-15 15:15:00	EMERGENCY	Medicaid	
16	PHYS REFERRAL/NORMAL DELI	2178-02-03 06:35:00	NEWBORN	Private	NEWBORN
16	HOME	2178-02-05 10:51:00	NEWBORN	Private	
17	PHYS REFERRAL/NORMAL DELI	2134-12-27 07:15:00	ELECTIVE	Private	PATIENT FORAMEN OVALE PATENT FORAMEN OVALE MINIMALLY INVASIVE SDA
17	HOME HEALTH CARE	2134-12-31 16:05:00	ELECTIVE	Private	
17	EMERGENCY ROOM ADMIT	2135-05-09 14:11:00	EMERGENCY	Private	PERICARDIAL EFFUSION
17	HOME HEALTH CARE	2135-05-13 14:40:00	EMERGENCY	Private	
18	PHYS REFERRAL/NORMAL DELI	2167-10-02 11:18:00	EMERGENCY	Private	HYPOGLYCEMIA SEIZURES
18	HOME	2167-10-04 16:15:00	EMERGENCY	Private	
19	EMERGENCY ROOM ADMIT	2108-08-05 16:25:00	EMERGENCY	Medicare	C 2 FRACTURE
19	REHAB/DISTINCT PART HOSP	2108-08-11 11:29:00	EMERGENCY	Medicare	
20	PHYS REFERRAL/NORMAL DELI	2183-04-28 09:45:00	ELECTIVE	Medicare	CORONARY ARTERY DISEASE CORONARY ARTERY BYPASS GRAFT SDA
20	HOME	183-05-03 14:45:00	ELECTIVE	Medicare	

Table 4. Snippet from the hospital admission log.

customers logged in the official website werk.nl and phone call data from their call center.

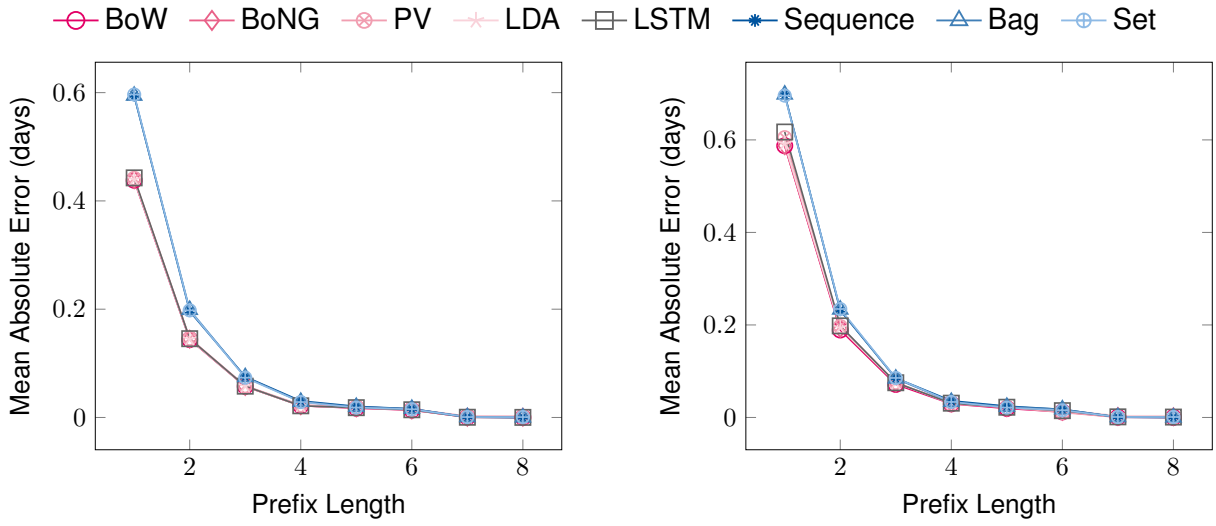
The second log is generated from the MIMIC-III (Medical Information Mart for Intensive Care) database and contains hospital admission and discharge events of patients in the Beth Israel Deaconess Medical Center between 2001 and 2012.

The results of the experiments are shown in Table 5. The next activity prediction shows an improvement of 2.83% and 4.09% on the two logs, respectively, showing that text can carry information on the next task in the process. While the impact of our method on next timestamp prediction is negligible in the customer journey log, it lowers the absolute error by approximately 11 hours in the hospital admission log. The improvement shown in the outcome prediction is small but present: 1.52% in the customer journey log and 2.11% in the hospital admission log. Finally, the improvement in cycle time prediction is particularly notable in the hospital admission log, where the error decreases by 27.63 hours. In general, compared to the baseline approaches, the text-aware model can improve the predictions on both event logs with at least one parametrization.

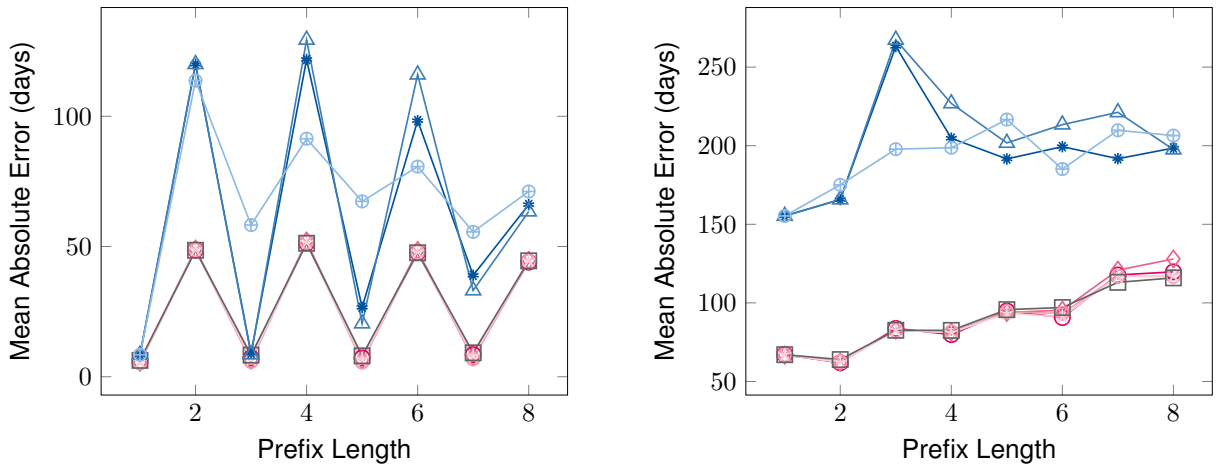
In addition, the prediction performance is evaluated per prefix length for each event log.

Table 5. Experimental results for the next activity, next timestamp, outcome, and cycle time prediction. All MAE scores are in days.

		BPIC2016 Customer Journey				MIMIC-III Hospital Admission			
Text Model	Text Vect. Size	Activity F_1	Time MAE	Outcome F_1	Cycle MAE	Activity F_1	Time MAE	Outcome F_1	Cycle MAE
Text-Aware Process Prediction (LSTM + Text Model)									
BoW	50	0.4251	0.1764	0.4732	0.2357	0.5389	29.0819	0.6120	69.2953
BoW	100	0.4304	**0.1763**	0.4690	**0.2337**	0.5487	31.4378	**0.6187**	70.9488
BoW	500	**0.4312**	0.1798	0.4690	0.2354	**0.5596**	27.5495	0.6050	70.1084
BoNG	50	0.4270	0.1767	0.4789	0.2365	0.5309	27.5397	0.6099	69.4456
BoNG	100	0.4237	0.1770	0.4819	0.2373	0.5450	28.3293	0.6094	69.3619
BoNG	500	0.4272	0.1773	0.4692	0.2358	0.5503	27.9720	0.6052	70.6906
PV	10	0.4112	0.1812	0.4670	0.2424	0.5265	29.4610	0.6007	73.5219
PV	20	0.4134	0.1785	0.4732	0.2417	0.5239	**27.2902**	0.5962	69.6191
PV	100	0.4162	0.1789	0.4707	0.2416	0.5292	28.2369	0.6058	69.4793
LDA	10	0.4239	0.1786	0.4755	0.2394	0.5252	28.8553	0.6017	**69.1465**
LDA	20	0.4168	0.1767	0.4747	0.2375	0.5348	27.8830	0.6071	69.6269
LDA	100	0.4264	0.1777	**0.4825**	0.2374	0.5418	27.5084	0.6106	69.3189
LSTM Model Prediction Baseline									
LSTM [9]		0.4029	0.1781	0.4673	0.2455	0.5187	27.7571	0.5976	70.2978
Process Model Prediction Baseline (Annotated Transition System)									
Sequence [1, 13]		0.4005	0.2387	0.4669	0.2799	0.4657	64.0161	0.5479	171.5684
Bag [1, 13]		0.3634	0.2389	0.4394	0.2797	0.4681	64.6567	0.5451	173.7963
Set [1, 13]		0.3565	0.2389	0.4381	0.2796	0.4397	63.2042	0.5588	171.4487



(a) BPIC2016 Customer Journey event log: next timestamp (left), cycle time (right).



(b) MIMIC-III Hospital Admission event log: next timestamp (left), cycle time (right).

Figure 4. Prediction performance on selected metrics and logs, shown by length of trace prefix.

Figure 4 shows the F_1 score and next timestamp MAE for every prefix trace of length $1 \leq k \leq 8$

on a selection of prediction tasks. Note that the results on shorter traces are supported by a much larger set of traces due to prefix generation. For text-aware models, only the best encoding size is shown.

On the customer journey log, the performance of all models correlates positively with the available prefix length of the trace. All text-aware prediction models surpass the baseline approaches on very short prefix traces of length 3 or shorter, for next activity and outcome prediction: we hypothesize that the cause for this is a combination of higher availability of textual attributes in earlier events in the traces, and the high number of training samples of short lengths, which allow text models to generalize. The next timestamp and cycle time predictions show no difference between text-aware models and the LSTM baseline, although they systematically outperform transition system-based methods.

The hospital admission log is characterized by the alternation of admission and discharge events. Therefore, the prediction accuracy varies between odd and even prefix lengths. The text-aware prediction models generate slightly better predictions on admission events since only these contain the diagnosis as text attribute. Regarding the next timestamp prediction, higher errors after discharge events and low errors after admission events are observed. This can be explained by the short hospital stays compared to longer time between two hospitalizations.

6 Conclusion

The prediction of the future course of business processes is a major challenge in business process mining and process monitoring. When textual artifacts in a natural language like emails or documents hold critical information, purely control-flow-oriented approaches are limited in delivering accurate predictions.

To overcome these limitations, we propose a text-aware process predictive monitoring approach. Our model encodes process traces of historical process executions to sequences of meaningful event vectors using the control flow, timestamp, textual, and non-textual data attributes of the events. Given an encoded prefix log of historical process executions, an LSTM neural network is trained to predict the activity and timestamp of the next event, and the outcome and cycle time of a running process instance. The proposed concept of text-aware predictive monitoring has been implemented and evaluated on real-world event data. We show that our approach is able to outperform state-of-the-art methods using insights from textual data.

The intersection between the fields of natural language processing and process mining is a promising avenue of research. Besides validating our approach on more datasets, future research also includes the design of a model able to learn text-aware trace and event embeddings, and the adoption of privacy-preserving analysis techniques able to avoid the disclosure of sensitive information contained in textual attributes.

Acknowledgements

We thank the Alexander von Humboldt (AvH) Stiftung for supporting our research interactions.

References

1. van der Aalst, Wil M.P., M. Helen Schonenberg, and Minseok Song. "Time prediction based on process mining." *Information Systems* 36.2 (2011): 450-475.
2. Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin et al. "Tensorflow: A system for large-scale machine learning." *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 2016.
3. Berti, Alessandro, Sebastiaan J. van Zelst, and Wil M.P. van der Aalst. "Process Mining for

Python (PM4Py): Bridging the Gap Between Process-and Data Science." In *International Conference on Process Mining (ICPM) Demo Track (CEUR 2374)*. pp. 13–16 (2019)
4. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *the Journal of Machine Learning Research* 3 (2003): 993-1022.
5. Brown, Peter F., Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. "Class-based n-gram models of natural language." *Computational linguistics* 18.4 (1992): 467-480.
6. Ceci, Michelangelo, Pasqua Fabiana Lanotte, Fabio Fumarola, Dario Pietro Cavallo, and Donato Malerba. "Completion time and next activity prediction of processes using sequential pattern mining." In *International Conference on Discovery Science*, pp. 49-61. Springer, Cham, 2014.
7. Evermann, Joerg, Jana-Rebecca Rehse, and Peter Fettke. "A deep learning approach for predicting process behaviour at runtime." In *International Conference on Business Process Management*, pp. 327-338. Springer, Cham, 2016.
8. Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." In *International Conference on Machine Learning*, pp. 1188-1196. PMLR, 2014.
9. Navarin, Nicolò, Beatrice Vincenzi, Mirko Polato, and Alessandro Sperduti. "LSTM networks for data-aware remaining time prediction of business process instances." In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1-7. IEEE, 2017.
10. Park, Gyunam, and Minseok Song. "Predicting performances in business processes using deep neural networks." *Decision Support Systems* 129 (2020): 113191.
11. Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." *the Journal of Machine Learning Research* 12 (2011): 2825-2830.
12. Polato, Mirko, Alessandro Sperduti, Andrea Burattin, and Massimiliano de Leoni. "Time and activity sequence prediction of business process instances." *Computing* 100.9 (2018): 1005-1031.
13. Rogge-Solti, Andreas, and Mathias Weske. "Prediction of remaining service execution time using stochastic Petri nets with arbitrary firing delays." In *International Conference on Service-Oriented Computing*, pp. 389-403. Springer, Berlin, Heidelberg, 2013.
14. Tax, Niek, Irene Teinemaa, and Sebastiaan J. van Zelst. "An interdisciplinary comparison of sequence modeling methods for next-element prediction." *Software and Systems Modeling* 19.6 (2020): 1345-1365.
15. Tax, Niek, Ilya Verenich, Marcello La Rosa, and Marlon Dumas. "Predictive business process monitoring with LSTM neural networks." In *International Conference on Advanced Information Systems Engineering*, pp. 477-492. Springer, Cham, 2017.
16. Teinemaa, Irene, Marlon Dumas, Fabrizio Maria Maggi, and Chiara Di Francescomarino. "Predictive business process monitoring with structured and unstructured data." In *International Conference on Business Process Management*, pp. 401-417. Springer, Cham, 2016.
17. Uysal, Merih Seran, Sebastiaan J. van Zelst and Tobias Brockhoff and Anahita Farhang Ghahfarokhi and Mahsa Pourbafrani and Ruben Schumacher and Sebastian Junglas and Günther Schuh and Wil M.P. van der Aalst. "Process Mining for Production Processes in the Automotive Industry" In *Industry Forum at BPM 2020* co-located with *International Conference on Business Process Management*, Springer, 2020.

24th International Conference on Business Information Systems (BIS 2021)
Social Media
https://doi.org/10.52825/bis.v1i.37

Published: 02 July 2021

Predicting E-commerce Item Sales with Web Environment Temporal Background

Yihong Zhang[1] and Takahiro Hara[1]

[1] Osaka University, Japan

Abstract. In this paper, we study the effect of Web environment temporal background in predicting e-commerce item sales, especially those in temporary sales. Temporary sales nowadays are a popular strategy for quickly clearing inventories. For traditional recommender systems, predicting the sales of an item is done based on its past purchase records. For temporary sales items, however, such records are not available. In order to make recommendation for such items, contextual information, such as product descriptions, is usually used. We investigate whether temporal background in the Web environment can be additional useful contextual information in recommender systems. It is assumed that items consistent with the temporal background would have higher demands. We propose a method for representing the temporal background using word embeddings of e-commerce activities and social media data, and evaluate their effect on sales prediction. Through empirical analysis with real-world data, we found that temporal background does have positive effects for sales prediction. The findings in this paper can be conveniently incorporated into future recommender system designs.

Keywords: e-commerce, recommender, social media

1 Introduction

Predicting item sales is an important and challenging problem in e-commerce and marketing. Potentially, knowing the outcome of sales before putting the item on the shelf help sellers better manage inventories. And e-commerce websites can also use this prediction to make more accurate recommendations. This is particularly true for temporary sales, in some situations can also be called flash sales, for which the main purpose of the campaign is to clear a certain amount of inventories [DL18]. Running a successful temporary sales campaign will involve several considerations, including choosing the right product to offer, promoting ahead of time, and using the right word for campaign descriptions. Among them, the timing to start the campaign is of utter importance. It would be much easier to sell the item when the timing is right. For example, it is known that it is much easier to sell air conditioners in early June in Japan as talks about the summer holiday start to appear. Continuous information that reflects on such moments can be considered as the *temporal background*. In this paper, our aim are two folds. First, we would quantify temporal background in Web environment, representing it in a way that can be processed computationally. Then, we would investigate to what extent temporal background can influence item sales through empirical analysis.

We have a similar goal with recommender systems, which in recent years have attracted significant research efforts [SKKR01]. Typically, a recommender system suggests a ranking of available products that the user may purchase in the future. For items in temporary sales that have no previous record of being purchased, however, such recommendation systems based

on past transactions are not useful. This is known as the *cold start problem* [LVLD08]. The cold start problem provides a challenge to recommend new items to users, and the typical solution is to use the contextual information associated with the item or the user that are available before there is any transaction [LLK14]. In this paper, we are not proposing a solution to the cold start problem. Instead, we focus on temporal background as a factor that can potentially help solving the cold start problem. The outcome of our study can reveal to what extent temporal background can help predicting sales of items that have no previous sales records. If there is a clear influence, then future cold start recommender systems can incorporate temporal background as an additional contextual information.

To be discussed in detail in Section 3, we build representations of temporal backgrounds from two data sources, including purchases records of an e-commerce website, and text messages of a social media platform. This specific e-commerce website hosts exclusively temporary sales that are usually available for a period between 7 to 14 days, and thus can be easily influenced by the temporal background. We are provided by this website with all all purchase records that occurred during a period of one year. These purchase records will be used as both the target to be predicted and the data for building the temporal background. The prediction is thus done for the number of item purchases in this e-commerce website, based on the temporal background constructed from the same purchase records and the social media data. The temporal background built from purchases records has a local and closer association with the temporal aspects of products, while that from social media represents a broader environment that reflects the social interest of the moment.

To summarize, our main contribution with this paper are three folds. First, we propose a method to represent temporal background from e-commerce and social media data. Second, we propose a method to predict product sales based on temporal background. Third, using real-world datasets, we verify our approach and reveal the answer about to what extent temporal background can be used to predict item sales. In Section 3 we will introduce our method and in Section 4 we will present experimental results.

2 Related Work

In this paper, we propose the concept of temporal background that can be generated using e-commerce purchase records and social media data. While we consider this to be a new concept, there are a number of previous works already studied the predictive relationships between social media and product sales. Gruhl et al. for example, proposed to use online blogs to predict book sales performance, which was quantified as sale ranks, published by Amazon [GGK+05]. They first studied the correlation between sales rank and blog mentions. From selected top-ranked books, they extracted book titles and author names, as the queries to generate mention frequencies, then the correlations were calculated. Asur and Huberman proposed a similar approach to predict movie revenues from discussions on Twitter [AH10]. They selected a number of movies and extracted related tweets using keywords present in the movie title. Based on the correlation with the Hollywood Stock Exchange index, they studied two statistics in the tweets, URLs and retweets representing promotional material, and rate of tweet mentions. Zhang and Pennacchiotti conducted a study of predicting product sales on eBay from Facebook data [ZP13]. They used a database containing eBay users who connect their accounts to Facebook. Based on the fact that there are similar categories on eBay and Facebook, they found that there is a strong correlation between liking Facebook pages and product purchase of the same category. Their prediction fell in the setting of recommendation systems, instead of the separation of past and future, and they claimed that the use of social media data can solve the cold-start problem for recommendation. Pai and Liu proposed a method to predict vehicle sales from tweets and stock market values [PL18]. They collected tweets mentioning brand names and conducted sentiment analysis to find correlation between tweets and sales.

Lassen et al. conducted a study of predicting iPhone sales from tweets [LMV14]. They collected tweets containing the word iPhone and then primarily conducted sentiment analysis on these tweets. They used a linear model to find correlation between tweet sentiment and iPhone sales, which were divided into quarters. In these previous works, a common drawback is that they relied on keywords that can be associated with the product. While this is feasible for products such as books and movies, in many real-world e-commerce scenarios, such associations may not be present. Our work on the other hand generalizes product and social media data into embeddings, so that the prediction does not require keyword association.

There is also a number of works that predicts continuous sales such as stocks from a textual background. For example, Bollen et al. conducted a sentiment analysis of tweets with regard to changes in Dow Jones Industrial Average index [BMZ11]. They extracted sentiment expressions in tweets using a dictionary that has six mood categories. These tweets are not necessarily related to stock market. However, as they concluded, the collective sentiments could be indicators of stock market changes. Particularly, the mood "calm" seems to be a strong indicator of stock market changes in three or four days. Moat et al. attempted to use Wikipedia activities to predict stock market changes [MCA+13]. They found evidences of increases in the number of page views of articles relating to companies or other financial topics before stock market falls. These works, however, rely on the continuous association between the item in question and the background. In our case of temporary deals, such association may not be available. In this aspect, our method is more general in that it can be applied to the case with or without continuous association between specific item and temporal background.

3 Methodology

We aim to develop a method that predicts future item sales from temporal background. The overview of our method is shown in Fig. 1. Our method consists of two main parts. In the first part, we represent temporal background using embeddings, which are vectors of real numbers that can be processed computationally. In the second part, we make prediction of item sales by comparing the temporal background and the new item description, which are projected to a same embedding space. Evaluating prediction results can thus reveal to us to what extent temporal background can influence item sales. In this section, we will present our method in detail.

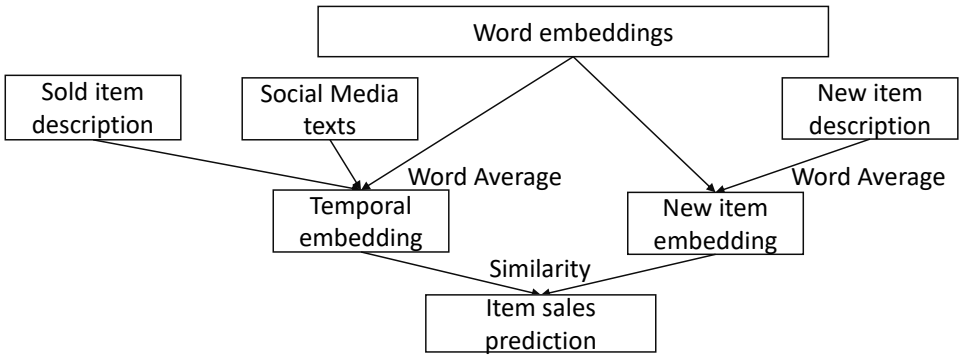


Figure 1. Overview of the method for predicting future item sales from temporal background

3.1 Representing Temporal Background

We found two sources that may contain information about the temporal environment. The first source is past records of e-commerce sales. As we have discussed in the introduction, products in our e-commerce website are temporary deals that are available for a short period of time. As such, the products sold in the past may not connect directly to the products currently on the

market. But the descriptions of products sold in the past may nevertheless contain information about selling aspects given a certain time period. The second source is social media. A social media platform such as Twitter posts millions of messages each day, which contain all kinds of topics that are worth public discussion. Although rarely connect directly to products in an e-commerce site, these messages certainly contain temporal information that reflects interesting aspects occur in the time period. Our first task at hand is to generate a temporal representation of these background data sources.

Given a day, when something interesting or stimulating happens, some topics in social media may become trending, and some products with certain aspects may see sudden rise in sales. When this happens, we say certain temporal aspects are *emerging*. We capture this emergence by observing the change of word frequency. Both the product description and social media messages are text data and can be represented as words, whose frequency we can count. The frequency of a product description word is counted as $\sum sales_i$ where $sales_i$ is the number of sales of product i whose description contains the word. The frequency of the social media words is simply the count of messages that contains the word. Thus we obtain the frequency table of product description and social media words.

We then devise a method for emergence detection based on word frequencies. Similar to some previous works on social media event detection [CALC13], our method involves a foreground and a background. Suppose the period for foreground is fp, and for background is bp, so that word frequencies in these periods are $F_{fp} = \{f_{t-fp}, ..., f_{t-1}\}$ and $F_{bp} = \{f_{t-fp-bp}, ..., f_{t-fp-1}\}$. We set inc_{fp} to $True$, if $f_{t-1} > \mu(F_{fp})$, where $\mu(\cdot)$ is the mean function, i.e., the frequency in the last day in the foreground period increases compared to the mean of foreground period, and $False$ otherwise. Similarly we set inc_{bp} for the background period. Finally the emergence e_t of the word at time t is set as:

$$e_t = \{\, 1\ , \quad \text{if } inc_{fp} \textit{ OR } (inc_{bp} \textit{ AND } \mu(F_{fp}) > \mu(F_{bp}))0, \quad \text{otherwise}$$

With this formula, we aim to capture two phases of surges of words in social media. First, inc_{fp} captures a new surge. Second, inc_{bp} *AND* $\mu(F_{fp}) > \mu(F_{bp})$ captures the sustenance of a previous surge. Both phases can be considered as a part of an emergence. With this calculation, we obtain for each time unit the emerging words in product sales and social media.

From here, we can follow a naive approach and represent each time unit as bag-of-words, which are the emerging words. However, this approach does not consider the meaning of words, which may cause error, for example, when two words of similar meaning are counted separately. Considering this, we would like to generalize the words into meanings. More specifically, we use distributed representation of words, also called word embeddings, which nowadays are commonly used in text-based analysis. Based on an implementation made available online [1], we learn a set of Japanese word embeddings using Wikipedia. The result word embeddings have 50 dimensions[2], each represent a certain semantic aspect of a word. To represent a group of words, we take the average vector of embeddings for these words. To represent a time unit which consists of emerging words from two sources, we concatenate the group embedding of the product words and social media words, and as the result, we have a vector of 100 dimensions to represent the time unit. This is our representation of the temporal background.

3.2 Predicting Future Temporal Background from Past Data

We have shown how embeddings representing the temporal background can be generated from e-commerce and social media activities in the time unit. However, as a prediction problem, data for the current time unit is not known before hand, and embeddings of current time unit need to be somehow generated from past data. We can consider embeddings across time as a

[1] https://github.com/philipperemy/japanese-words-to-vectors

[2] Although one can train word embeddings with different dimensions, we chose 50 to balance the complexity and generality

multivariate time series, and the task can be considered as a time series forecasting problem, for which many solutions have been proposed [WB04]. Since our focus is on the association between temporal background and item sales, we will only discuss two simple solutions here. The first solution is simply taking the embedding from the previous time unit as the prediction. This method can potentially work because major trends in e-commerce and social media only change gradually. There may be a problem, though, when some abnormal event happens and disrupts the continuity of embeddings across time.

The more popular method for forecasting time series nowadays, however, is through neural networks [LCYL18]. For our task, we build a simple recurrent neural network (RNN), which takes input of h embeddings from previous time units and outputs embedding of current day time unit. Between input and output there are two hidden layers, one contains 48 long short term memory (LSTM) nodes, and the other one contains 30 fully connected nodes. When training this neural network, predictions are iteratively compared with the target embedding, and the mean absolute error (MAE) is used to update neurons through back propagation. The LSTM layer essentially learns how a number of past values lead up to the current value in the time series. The fully connected layer is expected to capture interaction between two data sources, which cannot be captured by simply using the previous day values. We set h to 3 in our experiments, but different values such as 4 or 5 result in similar forecasting performance.

3.3 Evaluating the Effect of Temporal Background in Item Sales Prediction

After obtaining the temporal background representations, we need to find a connection between them and item sales. The popularity of an item will depend on many factors, including inherent quality, brand awareness, discount rate, and so on, and temporal background may be just one among them. Nevertheless, we consider this hypothesis:

Hypothesis (Temporal background consistency) *An item that is more consistent with the temporal background tends to have higher demand.*

As an example scenario, in Japan autumn is strongly associated with appetite. When people actively talk about food in autumn in social media, food products in e-commerce sites are so expected to have higher sales. Although temporal background cannot always be associated with products in this way, for example, when people in social media are talking about a recent political event, we argue that the consistency between product and temporal background can always to some degree influence the product sales.

To measure the consistency between product and temporal background, we apply the cosine similarity. Given a product embedding v_p and the temporal background embedding v_t, which are real value vectors, the consistency between them is calculated as:

$$c(v_p, v_t) = cos\_sim(v_p, v_t) = \frac{\sum_i v_{pi} v_{ti}}{\sqrt{\sum_i v_{pi}^2}\sqrt{\sum_i v_{ti}^2}}$$

After quantifying the consistency between the item and temporal background we can compare the ranking based on it and the actual sales number ranking. There are several measurements we can take. One example is Recall@k, which is calculated as

$$Recall@k = \frac{TP@k}{TP@k + FN@k}$$

where $TP@k$ is the number of actual top items in the selected k suggestions (True Positives), and $FN@k$ is the number of actual top items not in the selected k suggestions (False Negatives). Recall@k tells the ability the prediction method has to find top items given a certain number of choices.

Another possible metric is *Average Precision* (AP). First we get Precision@k as

$$Precision@k = \frac{TP@k}{TP@k + FP@k}$$

where $FP@k$ is the number of items that are not actual top items, among k suggestions (False Positives). Then AP is calculated as

$$AP = \sum_{k=1}^{K} (Recall@k - Recall@(k-1)) * Precision@k$$

Essentially, a higher AP tells that the top items are more concentrated in the top suggestions by the method.

4 Experimental Analysis

4.1 Dataset Preparation

We obtain a product sales dataset from a Japanese e-commerce website. The products, called $deals$ by the site, are discount coupons that are made available for a limited period of time, usually between 7 and 14 days. Customers who bought these deals can exchange them with real products. The products include several categories of items, including food, cosmetics, home appliances, hobby classes, travel packages, and so on. The dataset provided to us are of a period between October 2016 and August 2017. In total, there are 68,271 products made available and sold at least once during this period, and attracted about 1.6 million purchases. The number of available deals each day is about 1,000 on average. Each deal in the dataset is associated with a textual description written mostly in Japanese.

We obtain a social media dataset by collecting Japanese tweets through Twitter API[3]. To align with the period of e-commerce dataset, we develop a procedure to search past tweets without monitoring Stream API. In addition to time requirement, it is also desirable that the tweets are talking about Japanese domestic affairs, which reflects the background in which the e-commence business was operated. Our procedure is thus as the following. First, we collect a list of Japanese politician Twitter accounts[4]. From them we remove a few top politician accounts such as Abe Shinzo as they would attract foreign followers. Next we collect the follower of these politicians, who are expected to be Japanese citizens. Then we select from these citizen accounts whose earliest tweets are dated earlier than October 1st, 2016. This is to ensure that the accounts are active during the entire period of e-commerce dataset. Finally, we collect tweets in the said period from these selected accounts. These tweets become our social media source of temporal background in this experimental analysis. In total this dataset contains about 1.7 million tweets from 11,673 accounts.

We use the natural language processing package kuromoji[5] to process the Japanese text in the e-commerce and social media datasets. The package can effectively perform segmentation and part-of-speech (POS) tagging for Japanese text. After POS tagging, we select only nouns to represent the information in the text. These nouns are converted to temporal background embeddings following the method described in Section 3.

We use day as the time unit. Some of the components in our method such as the RNN model for embedding prediction require training data. We thus split our dataset into a training set and a testing set. The training set consists of 300 days of data, and the testing set consists of 20 days of data.

[3]https://developer.twitter.com/en/docs

[4]Such a list can be found online as political social media accounts are usually public. An example list is provided by the website Meyou with the url https://meyou.jp/group/category/politician/

[5]https://github.com/atilika/kuromoji

4.2 Direct Prediction of Sales from Product Description

The method proposed in this paper uses two steps to predict the product sales, first embeddings are generated for the product and temporal background, and then sales are predicted by comparing these embeddings. It is also possible, however, to learn a model that directly projects product description to sales, i.e., without the intermediate step of comparing it with temporal background. In this experimental analysis, we implement and test such a method. Using the training set described above, we train the model by setting the response variable as the daily sales number of the product, and the explanatory variable as the 50-dimension word embedding of the product description.

From here there are many possible machine learning techniques that can be applied, for example, linear regression or support vector machines. Since we expect non-linear relationship between dimensions in the word embedding and the sales number, we choose random forest (RF) as our learning model. In previous works, it has been shown that random forest can effectively predict product sales with token-based social media timing signals [ZHS20]. We train a random forest model using the training dataset, and then apply it to the testing dataset, by producing one sales number prediction for each product. The predicted sales numbers are ranked for each day in the test period and evaluated in the same way as we evaluate our method.

4.3 Embedding Prediction Accuracy

As discussed in the methodology section, we use two methods to predict the temporal background embedding of the current day using the embeddings in the past. First we use simply the embeddings of the previous day (P1), then we train a RNN model to forecast current embedding using embeddings of past h days. It would be interesting to see their prediction accuracy for the current embedding, which is used to predict product sales. After investigation, we found that the mean absolute error (MAE) for P1 method is 0.139, and for RNN method is 0.113. The root mean square error (RMSE) are 0.176 and 0.154 for P1 and RNN methods, respectively. Therefore, it is evident that RNN produces a forecast closer to the actual embeddings to be found for the current day.

4.4 Results and Discussions

We test different methods for predicting item sales ranks, and the accuracy measured as Recall@K and mean AP (mAP@K) is shown in Table 1.We test two K values of 50 and 100. The accuracy for the random method is based on theoretical values. For methods based on temporal background, we made two separate predictions, comparing the new item embedding first with the item part of the temporal background embedding, then with the tweet part. We tested three temporal background embeddings, namely, now, P1, and RNN. “Now” is taken as the current day embedding. It is not predicted and cannot be known before hand, but a comparison between it and predicted embeddings can be interesting. These results are averaged over the 20-day testing period. For each day, we pick top 20 items from all the items available for the day according to actual sales amounts, and then make 100 predictions. The theoretical $Recall@k$ for the random method is thus $100/a$ regardless of k, where a is the number of available items of the day.

Table 1. Average accuracy of prediction methods

	random	RF item	item now	tweet now	item P1	tweet P1	item RNN	tweet RNN
Recall@50	0.055	0.053	0.108	0.090	0.065	0.083	0.063	0.070
Recall@100	0.111	0.140	0.185	0.158	0.155	0.158	0.158	0.165
mAP@50	0.001	0.005	0.013	0.007	0.010	0.006	0.009	0.005
mAP@100	0.002	0.008	0.017	0.010	0.014	0.009	0.014	0.009

There are several insights we can draw from the results. First we look at the comparison with the random method. We can see that all prediction methods are better than the random method, which indicating that both item description and temporal background contain positive clues for predicting item sales. We can also see that using temporal background achieves better prediction accuracy than using the item description, indicating stronger predictiveness.

Second, comparing "now" embedding the predicted embeddings, we can see that using current day embedding achieves a higher accuracy. Even though it is not a prediction, we can see from it how correctly predicted temporal background can improve item sales accuracy. This also explains why RNN-predicted embedding is better than using previous day embedding. Since RNN predicted an embedding based on the embeddings in the last few days, it tends to predict a value that lays between the values in the previous day and the current day. As the result, its sales rank prediction accuracy also tends to lay between those using the previous day and the current day embeddings.

Last we compare between predictions using item and tweet embeddings. According to the result, when measuring Recall@k, tweet-based prediction is better than the item-based prediction. But when measuring AP, item-based prediction is better. It means that tweet-based prediction can generally find more top items, but item-based prediction can give higher rank to found items even though they are fewer. Similar tendencies are observed for both K values of 50 and 100.

4.5 Item Analysis

In order to get a closer view of what exactly happens within the prediction process, we analyze some concrete cases. We first pick the first day of test data and collect the emerging social media words most consistent with the temporal embedding of the day. Top 20 words collected and their cosine similarity scores with the temporal background are shown in Table 2[6].

Table 2. Consistent emerging social media words on test day 1

entirety 0.708	event 0.695	this 0.688	everyone 0.687	answer 0.685
every time 0.671	treatment 0.667	national 0.658	everyday 0.655	target 0.644
hope 0.639	choice 0.638	opposition 0.634	procedure 0.632	expectation 0.632
report 0.629	purpose 0.627	only one 0.626	today 0.625	investigation 0.623

We can roughly guess that the trending social media topic of the day is about some national events and something that involves report and investigation. Next we pick some items in top positions of the rank predicted by RNN tweet method, which is shown to be the best prediction method. More specifically, we pick one true positive and one false positive items by comparing the predicted ranked and actual sales ranks. The true positive item is ranked 17th by prediction and 6th by actual sales. The false positive item is ranked 1st by prediction and 273rd by actual sales. The descriptions and words most consistent with the temporal embedding of the day are shown in Table 3.

Comparing item description words with social media words, we see that both items ranked high because their descriptions contain words related to trouble, reporting and investigation, which are trending semantics in social media. However, item 2 is a false positive mostly because other factors cause low sales for this item. From these examples, we can see how temporal background influences items sales predictions and its limitations.

[6]Words and sentences shown in this section are translated from Japanese by authors.

Table 3. Description and consistent words for selected items

item 1				
Description: [experience interview / July 21st new open commemoration / existing account OK] all hand 100% placenta concentrate undiluted introduction worries raised with age are approached. CURACION introduction 60 minutes. Highly anticipated pure placenta specialty store				
interview 0.667	experience 0.641	existing 0.537	approach 0.479	quality 0.414
introduction 0.359	specialty 0.325	street 0.320	age 0.316	anticipated 0.300
item 2				
Description: [second anniversary / specialty shop for food trouble/ better know more about prevention and care] 97.2% satisfaction with German foot care, corns and keratin on the soles of the feet. Painless care as a solution to your concern (original prevention care set menu)				
solution 0.645	trouble 0.618	concern 0.593	satisfaction 0.585	menu 0.511
pain 0.490	prevention 0.482	set 0.426	original 0.380	German 0.347

5 Conclusion

Our aim with this paper is to discover the effect of Web environment temporal background in predicting e-commerce item sales. In particular, we would like to verify the hypothesis that items more consistent with the temporal background would have higher demands. For this purpose, we propose a method to generate embeddings for temporal backgrounds from e-commerce and social media activities, and make prediction of item sales based on them. By testing the accuracy of the predictions made using our method, and comparing it to the random baseline, we would be able to tell whether temporal background has positive effects on item sales prediction. Experimental analysis done using real-world data does show this positive effect. However, with item-level analysis, we can see some limitations of temporal background-based prediction. Initially this work is developed to support cold-start recommendation systems. Future works can be done on cleaning and filtering social media data so that its content can be more relevant to e-commerce items and potentially produce stronger positive effects.

Acknowledgement

Statement on competing interests: The authors declare that there is no conflict of interest. This research is partially supported by JST CREST Grant Number JPMJCR21F2.

[AH10] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, IEEE, vol. 1, 2010, pp. 492–499.

[BMZ11] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[CALC13] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua, "Emerging topic detection for organizations from microblogs," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2013, pp. 43–52.

[DL18] F. Dilmé and F. Li, "Revenue management without commitment: Dynamic pricing and periodic flash sales," *The Review of Economic Studies*, 2018.

[GGK+05] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, 2005, pp. 78–87.

[LCYL18] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, “Modeling long-and short-term temporal patterns with deep neural networks,” in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 95–104.

[LLK14] G. Ling, M. R. Lyu, and I. King, “Ratings meet reviews, a combined approach to recommend,” in *Proceedings of the 8th ACM Conference on Recommender systems*, ACM, 2014, pp. 105–112.

[LMV14] N. B. Lassen, R. Madsen, and R. Vatrapu, “Predicting iphone sales from iphone tweets,” in *Proceeding of the 18th International Enterprise Distributed Object Computing Conference*, IEEE, 2014, pp. 81–90.

[LVLD08] X. N. Lam, T. Vu, T. D. Le, and A. D. Duong, “Addressing cold-start problem in recommendation systems,” in *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, 2008, pp. 208–211.

[MCA+13] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis, “Quantifying wikipedia usage patterns before stock market moves,” *Scientific reports*, vol. 3, p. 1801, 2013.

[PL18] P.-F. Pai and C.-H. Liu, “Predicting vehicle sales by sentiment analysis of twitter data and stock market values,” *IEEE Access*, vol. 6, pp. 57 655–57 662, 2018.

[SKKR01] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in *Proceedings of the 10th International Conference on World Wide Web*, ACM, 2001, pp. 285–295.

[WB04] Z. Wang and D. A. Bessler, “Forecasting performance of multivariate time series models with full and reduced rank: An empirical examination,” *International Journal of Forecasting*, vol. 20, no. 4, pp. 683–695, 2004.

[ZHS20] Y. Zhang, T. Hara, and M. Shirakawa, “Discovering social media timing signals for predicting temporary deal success.,” in *Proceedings of the 28th European Conference on Information Systems*, 2020.

[ZP13] Y. Zhang and M. Pennacchiotti, “Predicting purchase behaviors from social media,” in *Proceedings of the 22nd international conference on World Wide Web*, ACM, 2013, pp. 1521–1532.

24th International Conference on Business Information Systems (BIS 2021)
Social Media
https://doi.org/10.52825/bis.v1i.55

Published: 02 July 2021

Social Media Crisis Communication Model for Building Public Resilience: A Preliminary Study

Umar Ali Bukar[1][https://orcid.org/0000-0002-3983-6919], Marzanah A. Jabar[1][https://orcid.org/0000-0002-3619-5028], and Fatimah Sidi[2][https://orcid.org/0000-0001-9556-9045], and RNH Binti Nor[1][https://orcid.org/0000-0001-5099-6692], and Salfarina Abdullah[1][https://orcid.org/0000-0002-8522-9175]

[1]Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia

[2]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Serdang 43400, Malaysia

Abstract. There is an ongoing discussion about the effectiveness of social media usage on the ability of people to recover from the crisis. However, the existing social media crisis communication models could not address the dynamic feature of social media users and the crisis, respectively. Therefore, the objective of this study is to conduct a preliminary investigation of the social media crisis communication model for building public resilience. Thus, 34 items were generated from the literature concerning the crisis, crisis response, social interaction, and resilience. The items were validated by three experts via content validity index and modified kappa statistics. After passing the validation test, the instruments were pre-tested by 32 participants. The reliability of the items was analyzed using Cronbach's alpha. Also, the model fits and mediation were examined by the regression model, and the hypotheses were independently assessed in process macro models. Based on the result obtained, each of the constructs satisfied the internal consistency requirement; crisis (0.743), crisis response (0.724), social media interaction (0.716), and resilience (0.827). Furthermore, the result also indicates that the regression model is a good fit for the data. The independent variables statistically significantly predict the dependent variable, $p < 0.05$. Also, the result of the process macro models indicates that all the hypotheses are independently supported.

Keywords: Crisis communication, Social media, Resilience, Social interaction, Crisis response

Introduction

The widespread adoption of technology has enabled crisis response and humanitarian development to be considered the future of human progress and well-being (1). Crisis management is driven by advances in computing, communications, storage, processing, and analysis. Technology-driven emergency management is continuously evolving as a new research field. Each step to improve methods or tools can make a significant contribution to saving human lives and resources. Emergency management, disaster management, and crisis management are often used interchangeably (2). Their role is to coordinate efficient actions related to information dissemination, security, supply, lodging in a highly dynamic and uncertain environment (3). The occurrence of a crisis, in particular, a disaster, is hard to predict, but its effects can be minimized through enabling technologies (4).

Crises come with a rapid increase in communication and a decrease in physical interactions. Coronavirus (Covid-19) is an example of a crisis that has become a threat to interaction

as everyone tries to have little interaction as possible (5)(World Health Organisation [WHO]. Information about the virus is spread on social networking sites which assists situational awareness, stakeholders interactions, and crisis responses (6). Social media enabled groups and individuals to collaborate and engage in crisis communication which also enabled digital convergence of people, information, and resources during crises (7). As a result, the people are far from being passive receivers; they actively seek out crisis information and exchange views with others (8). Hence, the dependence of formal and informal stakeholders (management and public) response is an established requisite for effective crisis communication and management (9; 10). Social media hypothetically intensifies the influence of the public's responses (10).

Therefore, this study aims to investigates the impact of the crisis, crisis response, and social media interaction on community resilience as part of an ongoing study on modeling social media crisis communication for building resilience. Therefore, the section "related work" discusses a few works that study the interaction among various stakeholders involved in crisis response on social media, the existing theories, and the conceptual model. The section "methodology" discusses the data and measurement method. The section "result and findings" presents the result of the content validity index, modified kappa, Cronbach's alpha, and regression model. The section "discussion" presents our argument, and finally, the section "conclusion" presents our concluding remarks and future research.

Related Work

The most dominant and commonly used theory to examine crisis management and communication strategies is the situational crisis communication theory (SCCT) (11; 12). The social-mediated crisis communication (SMCC) model classified social media public as influencers (content creators), followers, and inactive users (13; 14). The integrated crisis mapping (ICM) conceptualizes stakeholders' emotions (15). Reference (16) integrated the SCCT, SMCC, and ICM to link stakeholders' emotions and response strategies. Furthermore, (17) proposed the STREMII model, particularly for social media crisis management, is dynamic and cyclic, but the model is limited to response and recovery (18). Also, a new integrated crisis mapping approach based on ICM provides a general approach and directions for building a crisis communication model and a direct way of handling crisis response for effective reactions of public emotions (19). Moreover, the social media disaster resilience (SMDR) model has demonstrated how social media usage improves community resilience during and after crisis (20). Additionally, reference (21) introduced an interactive crisis communication model (ICCM) based on SCCT, SMCC, and traditional crisis communication strategies (CCS). ICCM is the first model that provides an integrated strategy toolkit that synthesizes SCCT and CCS crisis response strategies. The major weakness of ICCM is its failure to investigate the impact of interaction among stakeholders. The theoretical models and hypotheses are discuss as follows;

Situational Crisis Communication Theory (SCCT)

Situational crisis communication theory (SCCT) was developed in 1995, refined, and renamed in 2002 (11). The SCCT was motivated by the absence of a model to connects crisis to crisis response strategies (what crisis communicators say and do during a crisis) and crisis situations. The SCCT connects crisis and crisis response strategies and crisis types through the lens of attribution theory. Through crisis response, individuals seek to comprehend why a crisis event occurred. The work by (11) highlighted the importance of attributions, stating that they influence how people feel and react to an event. According to SCCT, crises are adverse events that cause people to judge the crisis management authorities, and a timely response safeguards the crisis management reputations (22). The digital environment, mainly social media platforms, enables individuals to challenge crisis management (11) due to public engagement in crisis response via social media, which also aids in the recovery process. On social media, the crisis has an impact on how crisis response is formed and led. Thus, crisis response and social media

interaction are used to address the situation when the crisis occurs. This demonstrated the effectiveness of crisis response and social media interaction in assisting people recovering from a crisis. Thus, the research framework (Figure 2) indicates that crisis significantly affects resilience, social media interaction, and crisis response, as the following hypothesis suggests.

- Hypothesis 1 (H1): Crisis has a significant impact on resilience.
- Hypothesis 2 (H2): Crisis has a significant impact on the crisis response.
- Hypothesis 3 (H3): Crisis has a significant impact on social media interaction.

Additionally, crisis response refers to the response of stakeholders (both public and management) to a crisis. This is discussed in the social-mediated crisis communication model (SMCC), which focuses on the types of public, the sources of information, and the information format. The crisis response enables stakeholders to create content that expresses their views on/about the crisis or the entity managing it. Understanding the full range of public emotions enhances the effectiveness of crisis response strategies (14; 16; 17), all of which have an effect on the public's ability to recover. The crisis management literature frequently refers to three stages of crisis management as para-crisis, crisis, and post-crisis (14; 18; 21; 23). As is the case with crisis responses, the nature of the para-crisis is considered to determine the para-crisis response that will most effectively mitigate the crisis risk (11). Thus, crisis response has an effect on resilience and social media interaction. Hence, the crisis-resilience relationship is mediated by crisis response and social media interaction. Therefore, this study implies that crisis response acts as a mediator between crisis and social interaction, as well as between crisis and resilience, as described in the hypothesis below.

- Hypothesis 4 (H4): Crisis response has a significant impact on resilience.
- Hypothesis 5 (H5): Crisis response has a significant impact on social media interaction.
- Hypothesis 7 (H7): The relationship between crisis and social media interaction is mediated by crisis response.
- Hypothesis 8 (H8): The relationship between crisis and resilience is mediated by crisis response.

Interactive Crisis Communication Model (ICCM)

The interactive crisis communication model (ICCM), introduced by (21), is relatively new in the crisis communication literature, which is base on SCCT, SMCC, and traditional crisis communication strategies (CCS). The ICCM is built for social media, demonstrating and representing the total interaction of stakeholders in a digital environment. Since everyone participates in crisis response, the model reaffirms the SMCC classification of public engagement and interactions. The ICCM demonstrates the importance of social interaction by demonstrating its capacity to provide one of the four gratifications identified in the uses and gratification theory (UGT) (24; 25). The entire ICCM is referred to as an interactive model, as it illustrates the fundamental elements of crisis management's interaction with the public. Similarly, the interaction is a fundamental component of the STREMII model (17). The ICCM demonstrates why social interaction is critical in social media crisis response. According to the ICCM model, since social media is an object or environment that enables groups and individuals to collaborate, the content can take the form of text, visual, audio, or a combination of these, which referred to as the interaction's content (14; 17; 20; 21). The contents are from two sources: crisis management and public response. The purpose of this study is to examine the effect of social interaction on resilience. Thus, this study implies that there is a relationship between social media interaction and resilience. As demonstrated in the hypothesis below, social media interaction serves as a mediator between crisis and resilience and crisis response and resilience.

- Hypothesis 6 (H6): Social media interaction has a significant impact on resilience.
- Hypothesis 9 (H9): The relationship between crisis and resilience is mediated by social media interaction.

- Hypothesis 10 (H10): The relationship between crisis response and resilience is mediated by social media interaction.

Social Media Disaster Resilience (SMDR) Model

The social-mediated disaster resilience (SMDR) model was introduced by (20), who demonstrated how social media usage is integrated into resilience building and discusses its potential for increasing hotel resilience. The study links resilience and disaster management literature using the revised 3Rs (robustness, rapidity, and redundancy) resilience model. Then, discussed social media as a robust technology to be used in crisis(13), to increase the speed of communication and information distribution (rapidity) (26; 27), and to redistribute the targeted information to a larger crowd via crowdsourcing (redundancy). In general, there is enough literature on how social media detects and document disasters (28), send and receive assistance (29), spread warnings (30), and solicit donations and volunteerism (29; 30). Thus, the robustness of social media has enabled the public to participate in crisis communication discussions, establishing them as vital resources (31). Additionally, the information can be quickly and widely distributed via social media crisis response and social interaction. While resistance is the first ideal outcome following a crisis, robustness, rapidity, and redundancy (3Rs) are critical for increasing resilience to the adverse effects of a crisis or disaster (20). Hence, it is essential to investigate the impact of crisis management efforts since their role is to strengthen relationships and improves the community's resilience (14; 20). Accordingly, crisis response and social media interaction mediates the crisis-resilience relationship as the following hypothesis suggest.

- Hypothesis 11 (H11): The relationship between crisis and resilience is mediated by crisis response and social media interaction.

Conceptual Model

According to reference (12), in the effort to represent the nature of the interactions between various stakeholders and how this study concept is different from existing studies of crisis communication. The study reiterates this concept and introduces the ICCM model representing the interaction between crisis management and the public on the social media environment (21) as reported in the STREMII model (17). Then adopted the resilience concept from the SMDR model that shows how social media usage improves community resilience (20).

First, since the crisis is the trigger that allows crisis response (32) to take place on social media, the nature of the crisis and crisis response are factors influencing stakeholder's formation on social media. These reflect the unmet challenge of integrating qualitative and quantitative modeling to understand how interaction, leadership, and social structure are represented in electronic trace data generated due to crisis and stakeholder's crisis responses (33). Second, the online contexts in which stakeholders interact as sociotechnical interaction places as represented in ICCM (21). This indicates where peoples interact as groups for a specific purpose and mediate consistent and meaningful aspects of their activity through technology that generates electronic trace data. The entire ICCM represents social interaction between stakeholders involves in crisis responses. Third, the SMDR proved the use of social media for building resilience but does not investigate the impact of social media usage or interaction through quantitative means. Therefore, the conceptual model (presented in Figure 2 along with the result) intends to show how social media interactions can improve community resilience since crisis management actions aim to improve the relationship and increase community resilience. The interactions between stakeholders (management and public) due to crisis can show the intensity of crisis response on social media(14; 20).

Methods

Data and Measures

The questionnaire's items are derived from existing literature (20; 25). The expert evaluation was developed based on these items, including discussing each construct, item, and how the existing theories were linked to form the relationship between the variables. Then, the evaluation document was sent to the expert for content and face validity. There are several recommendations for performing content validity (25; 34; 35). In this study, the validation uses three experts in crisis informatics and crisis communication, which have over 50years of combined working experience.

The expert evaluation form consist of 34 items distributed among the four construct; crisis = 6, crisis response = 6, social media interaction = 7, and resilience = 15. For content validation, several studies are encouraging using a 4-point Likert scale to avoid aggregations problems (34; 35).Therefore, each item was measured using 4-point Likert scales ranging from 1 (not relevant) to 4 (highly relevant) for the expert evaluation phase to examine if the items are appropriate to measure what it intends to measure. Notably, the expert encouraged the use of a 5-point Likert for further study. After passing the early validation phase, the questionnaire was designed via an online platform (google form) on 5-point Likert scales with a range from 1 (strongly agree) to 5 (strongly disagree). The answers were received within one week, with 32 completed responses. Study participants must have experienced Covid-19 lockdown and have observed social distancing rules to avoid the crowded area, self-isolate, or quarantine. The respondents were mainly males (approximately 85%) and are highly educated; 94% held a postgraduate or master's degree. In terms of age, most respondents were in the age range of 35–44 (51.5%) or 25–34 (33.3%). Approximately 12.1% were between 45 and above years old, and about 3.1% were between 18 - 24 years old. The sample was considered appropriate for the preliminary test (pilot study) of the study.

Method of Analysis

The analysis methods are considered threefold; the content validity method, the reliability, and the assessment model fit, and the mediation fit model. Therefore, the content validity index (CVI) is considered the most widely used method to validate research items (34; 36). The CVI expressed and classified the proportion of agreement into either 0 or 1. Although the significant weakness of CVI is its failure to adjust for chance agreement and the researchers suggest the modified kappa statistics to address chance agreement (34; 35; 37). Studies encouraging the use of a 4-point Likert scale for CVI specified that the rating of 1 and 2 are considered content invalid while those from 3 and 4 to be content valid (34; 35; 36; 37). Therefore, the 4-point rating is collapsed into dichotomous categories of responses. Any response that falls between 1 and 2 is assigned 0, and those that fall in 3 and 4 are assigned 1. Secondly, the instrument's reliability and the relationship between the independent variables (IVs) and the dependent variable (DV) were measured using Cronbach's alpha and regression model. Thirdly, the hypotheses were independently assessed in PROCESS macro models (38), and the indirect effect was tested using a percentile bootstrap estimation approach with 5000 samples (39) on SPSS version 24.

Result and Findings

Content Validation

Several studies have provided information on how to sufficiently measure the value that indicates a high level of agreement among the raters (34; 35; 37; 36). (34) emphasized that an average value of 70% is necessary for agreement, 80% for adequate agreement, and 90% for a good agreement. In comparison, other researchers recommend values between 70 and 79 percent to be considered for revision while less than 70% to be eliminated (35; 37). Also, few

suggestions are also employed to measure the strength of agreement for kappa which classified as less than 0.40 as poor, 0.40 – 0.59 as fair, 0.60 - 0.74 as good, and 0.75 – 1.00 as excellent (34; 37). The result of CVI and modified kappa statistics are excellent for this study. All the items indicate an excellent level of agreement among the expert. The value for I-CVI and Modified kappa for each item is 1.

Reliability and Consistency of the Instrument

The Cronbach's alpha assessed scale reliability, and the alpha for each construct is presented in Table 1.

Table 1. Crombah Alpha Reliability Parameters of the Constructs

Constructs	No of items	Crombach Alpha
Crisis	6	0.743
Crisis response	6-1 (5)	0.724
Social media interaction	7	0.716
Resilience	15	0.827

Model Analysis

Multiple regression was applied to predict resilience from the crisis, crisis response, and social media interaction. These variables statistically significantly predicted resilience, F(3, 95) = 15.141, p $<$ 0.05, R^2 = 0.619. The standard error of the estimate determines how the model fits the data. A rule of thumb R-value is considered a measure of predicting the DV; in this case, resilience. A value of 0.787 indicates a good level of model fits. Also, the R Square value, which is 0.619, shows that the IVs explain 61.9% of the variability of the DV.

In Table 2, the overall model is tested to determine if the regression model is a good fit for the data. The result of the analysis of variance (ANOVA) indicates that the IVs statistically significantly predicts the DV, p $<$ 0.05. The results indicated that the overall model is a good fit for the data.

Table 2. Analysis of Variance (ANOVA)

Model 1	Sum of Squares	Df	Mean Square	F	Sig.
Regression	3.056	3	1.019	15.141	.000[b]
Residual	1.884	28	0.067		
s Total	4.94	31			

a. Dependent Variable: Resilience

b. Predictors: (Constant), Interaction, CrisisResponse, Crisis

Figure 1 presents the results for multivariate normality and linear relationship between the IVs and the DV. The results in Figure 1 (left) indicate that the multivariate normality is present in the data, normally distributed. Secondly, in Figure 1 (right), the results suggested an excellent linear relationship between the IVs and the DV.

Path and Mediation Analysis

Results are obtained concerning the relationship between the IVs and the DV and whether social media interaction acted as a mediator between crisis and resilience and between crisis response and resilience. Additionally, results are obtained for whether crisis response acted as a mediator between crisis and resilience. Figure 2 presents the impact of the crisis, crisis response, and social media interaction on resilience.

The results indicated that crisis is a significant predictor of resilience, β = 0.3155, SE = 0.1318, p $<$ 0.05, crisis response, β = 0.5480, SE = 0.1795, p $<$ 0.05, and social media interaction, β = 0.6636, SE = 0.1413, p $<$ 0.05, indicating support for H1, H2, and H3. Similarly,

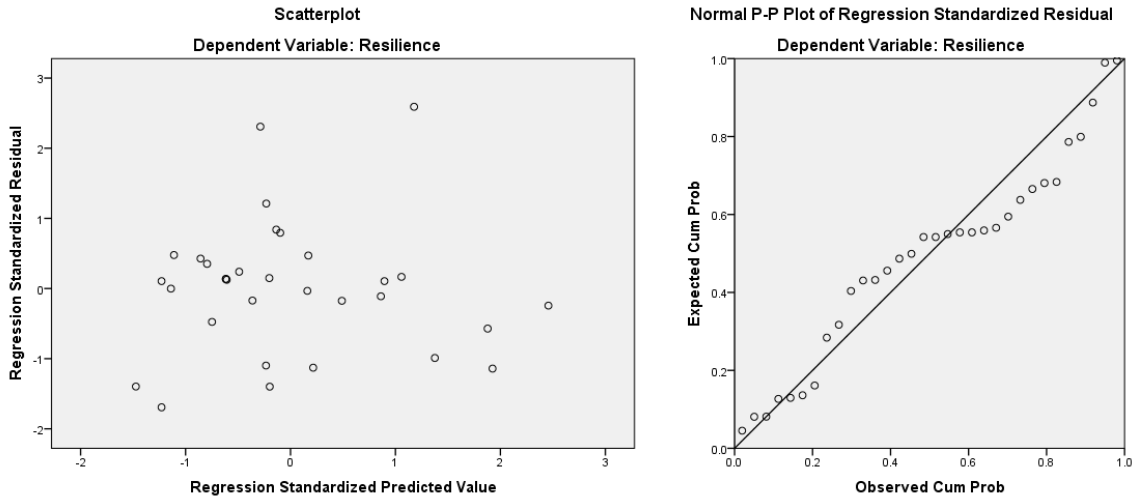


Figure 1. Result for Multivariate Normality of IV and DV (left) and Relationship Between IV and DV (right)

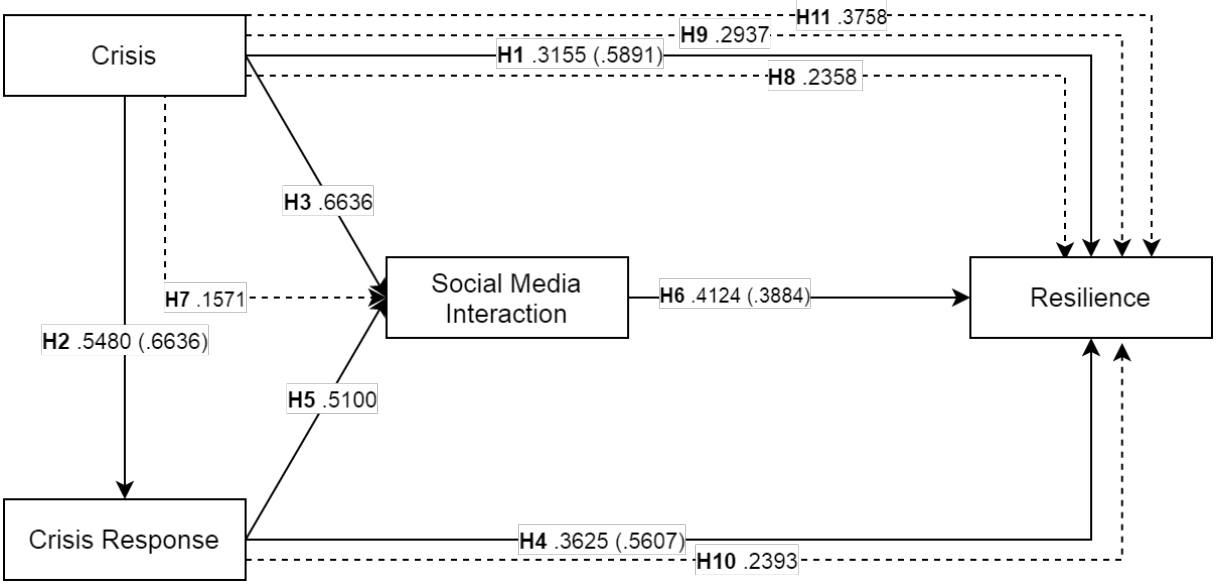


Figure 2. A Model for Social Media Crisis Communication for Resilience

the results indicated that crisis response is a significant predictor of resilience, β = 0.3625, SE = 0.1112, p $<$ 0.05 and social media interaction, β = 0.5100, SE = 0.1367, p $<$ 0.05, indicating support for H4 and H5. Also, social media interaction is a significant predictor of resilience, β = 0.4124, SE = 0.1553, p $<$ 0.05, indicating support for H6.

The results of mediation was observed. The results for the indirect effect of social media interaction on crisis-crisis response relationship indicated that the indirect coefficient was significant, β = 0.1571, SE = 0.0969, 95% CI = 0.0145, 0.3898, indicating support for H7. The proportion of the total effect of crisis on crisis response is β = 0.6636, SE = 0.1413, p $<$ 0.05, R^2 = 0.4237, is indirectly significant by 42.37%. Moreover, the results for the indirect effect of crisis response on crisis-resilience relationship indicated that the indirect coefficient was significant, β = 0.2358, SE = 0.0958, 95% CI = 0.0625, 0.4465, indicating support for H8. The proportion of the total effect of crisis on resilience is β = 0.5891, SE = 0.1318, p $<$ 0.05, R^2 = 0.3998, is indirectly significant by 39.98%. Moreover, the results for the indirect effect of social media interaction on crisis-resilience relationship indicated that the indirect coefficient was significant, β = 0.2937, SE = 0.1598, 95% CI = 0.0834, 0.7250, indicating support for H9. The proportion

of the total effect of crisis on resilience is β = 0.5891, SE = 0.1318, p $<$ 0.05, R^2 = 0.3998, is indirectly significant by 39.98%. Additionally, the results for the indirect effect of social media interaction on crisis response-resilience relationship indicated that the indirect coefficient was significant, β = 0.2393, SE = 0.1152, 95% CI = 0.0576, 0.5003, indicating support H10. The proportion of the total effect of crisis response on resilience is β = 0.5607, SE = 0.1112, p $<$ 0.05, R^2 = 0.4587, is indirectly significant by 45.87%. Finally, the results for the indirect effect of social media interaction and crisis response on crisis-resilience relationship indicated that the indirect coefficient was significant, β = 0.3758, SE = 0.1385, 95% CI = 0.1499, 0.7269, indicating support for H11. The proportion of the total effect of crisis on resilience is β = 0.5891, SE = 0.1318, p $<$ 0.05, R^2 = 0.3993, is indirectly significant by 39.93%.

Result Summary

The expert validation results show that all items indicate excellent agreement both from CVI and modified kappa statistics. The scale reliability and internal consistency were measures using Cronbach's alpha, and the results observed that one of the items has less scored by the participants. This item was not considered in the linear regression model and process macro model, which is left for further study with a large sample size. Furthermore, the mediation analysis based on the regression coefficients supported all the hypotheses, see Table 3.

Table 3. Hypothesis Assessment of the Direct and Indirect Effects

S/N	Hypotheses	Decision
H1	Crisis $->$ Resilience	Supported
H2	Crisis $->$ Crisis response	Supported
H3	Crisis $->$ Social media interaction	Supported
H4	Crisis response $->$ Resilience	Supported
H5	Crisis response $->$ Social media interaction	Supported
H6	Social media interaction $->$ Resilience	Supported
H7	Crisis $->$ Crisis response $->$ Social media interaction	Supported
H8	Crisis $->$ Crisis response $->$ Resilience	Supported
H9	Crisis $->$ Social media interaction $->$ Resilience	Supported
H10	Crisis response $->$ Social media interaction $->$ Resilience	Supported
H11	Crisis $->$ Crisis response $->$ Social media interaction $->$ Resilience	Supported

Discussion

The discussion about the potentials of social media to improve stakeholders' relationships and increase community resilience called for empirical investigation of this research. As the central aspect of this work, the study conducted a preliminary investigation for social media crisis communication and resilience that offer the basics to determine effective social media-based crisis communication and management. Thus, a particular discussion is about the results and findings achieved from the evaluation of the pilot study. According to the findings, all the research hypotheses were supported. The findings indicated that the public ability to recover from a crisis could be improved through social media-based crisis communication by meaningful engagement in crisis response and social interaction on social media.

Firstly, the findings of this study support existing works (21; 20; 17), whose conceptualizations of a social media-based crisis communication model emphasized the critical nature of crisis response and social interaction via social media use. These studies laid the groundwork for the model proposed in this paper. However, a few studies have examined the various stakeholders involved in crisis communication on social media (40; 41; 42) with focused on actors in the issue arena, knowledge about stakeholders relationships, and content of crisis communication. Reasonably, this study may serve as evidence that the goal of crisis informatics is to comprehend the interaction of stakeholders involved in crisis communication (43). Thus, this

study contributes to the advancement of social media use for crisis communication.

Secondly, reference (44) stated that the long-term impact of the Covid-19 pandemic is unknown, but it is undoubtedly going to last longer than anticipated. One of the pandemic's most significant and immediate effects is how it has shattered the public relationship. Restriction of movement creates a stressful awareness that one's well-being is dependent on others. The pandemic severely harmed individuals' capacity to have actual close relationships with other people, severely harmed the extreme human need for contact, and discouraging any other physical affection and connection. Whereas the explosion of electronic-mediated interaction in recent decades, most notably via social media, has aided people in recovering from crises, demonstrating how critical in-person interaction via social media is building public resilience during the pandemic.

Thirdly, individuals were unable to participate in social activities during the pandemic period. Thus, this study responds to a call for additional research into the short and long-term consequences of critical life-course changes. The crisis and physical separation have the potential to have both immediate and long-term effects on relationships. Some people are susceptible to changes in their environment, especially children. When isolation began, there was considerable anxiety about the infection entering and spreading through the families, and household members struggled to express affection securely physically. Interestingly, social media activities (crisis response and social interaction) have aided significantly in effectively managing physical distancing in these circumstances. Thus, the proposed model demonstrates the future direction of crisis communication to assist affected citizens during such times.

Additionally, when schools are closed, children and adolescents cannot interact face-to-face with their friends. Peer gatherings and fellowships are critical for teenagers' character and identity exploration. Social media enabled friendships to endure and engagement to occur. Similarly, despite the availability of various platforms to mitigate the impact of restricted human movement, there were still concerns about the public's emotions, needs for fulfillment, and relationship survival. As a result, the model validated in this study established the requirements and framework for providing effective crisis communication to address these concerns. Finally, the pandemic displaced many people from their jobs, forcing them to collaborate and maintain associations remotely and distantly. Surprisingly, the relationships between the variables examined in this study may aid crisis management organizations in re-positioning their crisis communication activities more effectively.

Conclusion

This study reported the preliminary test of the impact of the crisis, crisis response, and social media interaction on resilience building. CVI was used for content validation, while Cronbach alpha and regression analysis was incorporated to check the reliability, model fits, and mediation fits. The findings show that the content validity shows excellent agreement among the raters. The internal consistency of each variable meets the required minimum for acceptance, and the regression model is significant.

This study is not without limitations, and the major one is that the test sample is very homogeneous. The respondents were mainly male and mainly from a particular age group. This makes it hard to generalize the finding to women or other age groups, who are also affected by crises. In the future, we intend to collect data from large sample size to investigates the impact of the crisis, crisis response, and social media interaction. Also since the model suggested some attributes of social media interaction such as time distance, type (mode), content, frequency, duration, and intensity. Further work could collect the content of crisis responses from social networking sites based on these attributes. The content can be analyzed using sentiment analysis to understand the intensity of the interaction among stakeholders.

Acknowledgement

The authors would like to express gratitude for the financial support provided under the Fundamental Research Grant Scheme (FRGS) through the Grant Cost Centre under FRGS/1/2019/ICT04/UPM/02/5, Vote Number: 5540287

References

[1] J. Qadir, A. Ali, R. ur Rasool, A. Zwitter, A. Sathiaseelan, and J. Crowcroft, "Crisis analytics: big data-driven crisis response," *Journal of International Humanitarian Action*, vol. 1, no. 1, p. 12, 2016.

[2] V. Mijović, N. Tomašević, V. Janev, and S. Vraneš, "Event-driven decision support system for intelligent emergency management in critical infrastructures."

[3] M. De Brito, L. Thévin, C. Garbay, O. Boissier, and J. F. Hübner, "Supporting flexible regulation of crisis management by means of situated artificial institution," *Frontiers of Information Technology & Electronic Engineering*, vol. 17, no. 4, pp. 309–324, 2016.

[4] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on the red: what hazards threat reveals about the social life of microblogged information," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 2010, pp. 241–250.

[5] W. H. Organization *et al.*, "Novel coronavirus (2019-ncov): situation report, 3," 2020.

[6] T. Onorati, P. Díaz, and B. Carrion, "From social networks to emergency operation centers: A semantic visualization approach," *Future Generation Computer Systems*, vol. 95, pp. 829–840, 2019.

[7] L. Palen and S. B. Liu, "Citizen communications in crisis: anticipating a future of ict-supported public participation," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 727–736.

[8] Y. Ji and S. Kim, "Communication-mediated psychological mechanisms of chinese publics' post-crisis corporate associations and government associations," *Journal of Contingencies and Crisis Management*, vol. 27, no. 2, pp. 182–194, 2019.

[9] L. Palen, K. M. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald, "A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters," *ACM-BCS Visions of Computer Science 2010*, pp. 1–12, 2010.

[10] H. Purohit, A. Hampton, S. Bhatt, V. L. Shalin, A. P. Sheth, and J. M. Flach, "Identifying seekers and suppliers in social media communities to support crisis coordination," *Computer Supported Cooperative Work (CSCW)*, vol. 23, no. 4-6, pp. 513–545, 2014.

[11] W. T. Coombs, "Revising situational crisis communication theory," *Social media and crisis communication*, vol. 1, pp. 21–37, 2017.

[12] U. A. Bukar, M. A. Jabar, F. Sidi, R. N. H. Nor, S. Abdullah, and M. Othman, "Crisis informatics in the context of social media crisis communication: Theoretical models, taxonomy, and open issues," *IEEE Access*, vol. 8, pp. 1–1, 2020.

[13] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, "Using social media to enhance emergency situation awareness," *IEEE Annals of the History of Computing*, vol. 27, no. 06, pp. 52–59, 2012.

[14] B. F. Liu, L. Austin, and Y. Jin, “How publics respond to crisis communication strategies: The interplay of information form and source,” *Public Relations Review*, vol. 37, no. 4, pp. 345–353, 2011.

[15] Y. Jin, A. Pang, and G. T. Cameron, “Toward a publics-driven, emotion-based conceptualization in crisis communication: Unearthing dominant emotions in multi-staged testing of the integrated crisis mapping (icm) model,” *Journal of Public Relations Research*, vol. 24, no. 3, pp. 266–298, 2012.

[16] C. Vignal Lambret and E. Barki, “Social media crisis management: Aligning corporate response strategies with stakeholders’ emotions online,” *Journal of Contingencies and Crisis Management*, vol. 26, no. 2, pp. 295–305, 2018.

[17] M. C. Stewart and B. G. Wilson, “The dynamic role of social media during hurricane# sandy: An introduction of the stremii model to weather the storm of the crisis lifecycle,” *Computers in Human Behavior*, vol. 54, pp. 639–646, 2016.

[18] R. Syed, “Enterprise reputation threats on social media: A case of data breach framing,” *The Journal of Strategic Information Systems*, vol. 28, no. 3, pp. 257–274, 2019.

[19] Y. Jin, J.-S. Lin, B. Gilbreath, and Y.-I. Lee, “Motivations, consumption emotions, and temporal orientations in social media use: A strategic approach to engaging stakeholders across platforms,” *International Journal of Strategic Communication*, vol. 11, no. 2, pp. 115–132, 2017.

[20] C. Möller, J. Wang, and H. T. Nguyen, “# strongerthanwinston: Tourism and crisis communication through facebook following tropical cyclones in fiji,” *Tourism Management*, vol. 69, pp. 272–284, 2018.

[21] Y. Cheng, “How social media is changing crisis communication strategies: Evidence from the updated literature,” *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 58–68, 2018.

[22] W. T. Coombs, “Attribution theory as a guide for post-crisis communication research,” *Public Relations Review*, vol. 33, no. 2, pp. 135–139, 2007.

[23] W. T. Coombs and J. S. Holladay, “The paracrisis: The challenges created by publicly managing crisis prevention,” *Public Relations Review*, vol. 38, no. 3, pp. 408–415, 2012.

[24] A. Whiting and D. Williams, “Why people use social media: a uses and gratifications approach,” *Qualitative Market Research: An International Journal*, 2013.

[25] Y. Li, S. Yang, S. Zhang, and W. Zhang, “Mobile social media use intention in emergencies among gen y in china: An integrative framework of gratifications, task-technology fit, and media dependency,” *Telematics and Informatics*, vol. 42, p. 101244, 2019.

[26] C. Ehnis and D. Bunker, “Social media in disaster response: Queensland police service-public engagement during the 2011 floods,” 2012.

[27] M. Irons, ““we can help”: an australian case study of post-disaster online convergence and community resilience,” Ph.D. dissertation, University of Tasmania, 2015.

[28] C. Reuter and M.-A. Kaufhold, “Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics,” *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 41–57, 2018.

[29] M. Taylor, G. Wells, G. Howell, and B. Raphael, “The role of social media as psychological first aid as a support to community resilience building,” *Australian Journal of Emergency Management, The*, vol. 27, no. 1, pp. 20–26, 2012.

[30] C. Reuter, T. Ludwig, M.-A. Kaufhold, and T. Spielhofer, "Emergency services attitudes towards social media: A quantitative and qualitative survey across europe," *International Journal of Human-Computer Studies*, vol. 95, pp. 96–111, 2016.

[31] M. Irons and D. Paton, "Social media and emergent groups: The impact of high functionality on community resilience," *Disaster Resilience: An Integrated Approach, 2nd ed.; Paton, D., Johnston, DM, Eds*, pp. 194–211, 2017.

[32] W. T. Coombs, "Protecting organization reputations during a crisis: The development and application of situational crisis communication theory," *Corporate reputation review*, vol. 10, no. 3, pp. 163–176, 2007.

[33] S. P. Goggins, C. Mascaro, and G. Valetto, "Group informatics: A methodological approach and ontology for sociotechnical group research," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 3, pp. 516–539, 2013.

[34] C. A. Wynd, B. Schmidt, and M. A. Schaefer, "Two quantitative approaches for estimating content validity," *Western journal of nursing research*, vol. 25, no. 5, pp. 508–518, 2003.

[35] D. F. Polit and C. T. Beck, "The content validity index: are you sure you know what's being reported? critique and recommendations," *Research in nursing & health*, vol. 29, no. 5, pp. 489–497, 2006.

[36] M. S. B. Yusoff, "Abc of content validation and content validity index calculation." *Education in Medicine Journal*, vol. 11, no. 2, 2019.

[37] D. F. Polit, C. T. Beck, and S. V. Owen, "Is the cvi an acceptable indicator of content validity? appraisal and recommendations," *Research in nursing & health*, vol. 30, no. 4, pp. 459–467, 2007.

[38] A. F. Hayes, *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications, 2017.

[39] P. E. Shrout and N. Bolger, "Mediation in experimental and nonexperimental studies: new procedures and recommendations." *Psychological methods*, vol. 7, no. 4, p. 422, 2002.

[40] I. Hellsten, S. Jacobs, and A. Wonneberger, "Active and passive stakeholders in issue arenas: A communication network approach to the bird flu debate on twitter," *Public Relations Review*, vol. 45, no. 1, pp. 35–48, 2019.

[41] T. G. van der Meer, P. Verhoeven, H. W. Beentjes, and R. Vliegenthart, "Communication in times of crisis: The stakeholder relationship under pressure," *Public Relations Review*, vol. 43, no. 2, pp. 426–440, 2017.

[42] C. Du Plessis, "Social media crisis communication: Enhancing a discourse of renewal through dialogic content," *Public relations review*, vol. 44, no. 5, pp. 829–838, 2018.

[43] M. L. Tan, R. Prasanna, K. Stock, E. Hudson-Doyle, G. Leonard, and D. Johnston, "Mobile applications in crisis informatics literature: A systematic review," *International journal of disaster risk reduction*, vol. 24, pp. 297–311, 2017.

[44] A. Schleicher, "The impact of covid-19 on education insights from education at a glance 2020," *Retrieved from oecd. org website: https://www. oecd. org/education/the-impact-of-covid-19-on-education-insights-education-at-a-glance-2020. pdf*, 2020.

24th International Conference on Business Information Systems (BIS 2021)

Applications

https://doi.org/10.52825/bis.v1i.41

Published: 02 July 2021

Stream Processing Tools for Analyzing Objects in Motion Sending High-Volume Location Data

Krzysztof Węcel[1][https://orcid.org/0000-0001-5641-3160], Marcin Szmydt[1][https://orcid.org/0000-0003-4392-0205], and Milena Stróżyna[1][https://orcid.org/0000-0001-7603-7369]

[1] Poznań University of Economics, Poland

Abstract. Recently we observe a significant increase in the amount of easily accessible data on transport and mobility. This data is mostly massive streams of high velocity, magnitude, and heterogeneity, which represent a flow of goods, shipments and the movements of fleet. It is therefore necessary to develop a scalable framework and apply tools capable of handling these streams. In the paper we propose an approach for the selection of software for stream processing solutions that may be used in the transportation domain. We provide an overview of potential stream processing technologies, followed by the method for choosing the selected software for real-time analysis of data streams coming from objects in motion. We have selected two solutions: Apache Spark Streaming and Apache Flink, and benchmarked them on a real-world task. We identified the caveats and challenges when it comes to implementation of the solution in practice.

Keywords: stream processing, location data, transport, mobility, AIS

Introduction

The recent rapid development of communication and detection technologies, the emergence of low-cost and widespread smart sensors, and a significant drop in data storage costs have all contributed to a significant increase in the amount of easily accessible data on transport and mobility. The volume and speed at which sensor data is generated, processed, and stored is unprecedented [1].

The advent of Big Data, including massive streams of real-time data of high velocity, magnitude, and heterogeneity, have triggered changes in many fields. One of them is the transport sector that manages a massive flow of goods and at the same time creates large data sets. These data streams concern inter alia mil-lions of shipments and the movements of fleet that are tracked every day [2]. This data is then considered as a source for context-aware applications and intelligent services, aiming to improve traffic efficiency, safety, and security [3] as well as last-mile delivery optimization, route optimization, fleet management, and detection of anomalous behavior [2] These services are applied in various transport modes, such as road, maritime or public transport as well as in various forms of shared transport (e.g., carsharing, bike-sharing). For example, data extracted from computers embedded in vehicles, that concern an object in motion, i.e., its origin, destination, content, and location, can be used to better understand and predict the flow of goods and people in real time. These data streams can be further combined with other data gathered from navigational systems, mobile phones (e.g., location, activities), environmental sensors (e.g., pollution levels), or social net-works (e.g., people's preferences and relationships).

Nevertheless, the potential of this massive amount of data can be exploited only if there are tools and models able to efficiently extract, detect and analyze relevant information from the data streams [4]. However, most of the approaches or methods applied in the existing

transportation systems and applications do not fit the paradigm of Big Data analytics [1]. The challenge now is not only to collect the data but to draw conclusions from them. Therefore, it is required to develop modern system architectures that would allow to efficiently process large data streams. As indicated by [5] in their systematic over-view of big data stream analysis, a significant amount of previous research has been directed to real-time analysis of big data streams and not much attention has been given to the preprocessing stage. Moreover, only a few big data streaming technologies offer a possibility to do both batch, streaming and iterative jobs. Therefore [5] suggest that research effort should be directed towards developing scalable frameworks and architectures able to accommodate data stream computing mode, effective resource allocation strategy, and parallelization issues.

In order to design such a scalable and efficient architecture, a number of choices have to be made by designers, including the selection of proper software and hardware. In response to this challenge, we propose an approach for the se-lection of software for stream processing solutions that may be used in the transportation domain (analysis of data streams generated by objects in motion). The aim is to show steps that need to be followed while designing system architecture for data stream analysis. To achieve this goal, we provide an overview and characterization of some stream processing technologies, followed by proposing a method for comparing the selected software for real-time analysis of data streams coming from objects in motion along with identification of the caveats and challenges when it comes to implementation of the software in practice.

We designed and implemented in the selected technologies a Minimum Working Example (MWE) in which an exemplary data stream with location of moving objects is used. The stream is roughly 10 MB per minute, which gives around 15 GB for a 24h time window. The idea is to perform the real-time anomaly detection on a sliding window using the smallest time interval possible. This naturally depends on the size of the data, the complexity of the used algorithms, and avail-able processing power.

The paper is structured as follows: in section 2 a comparison of the batch and streaming processing is provided, followed by software selection options (section 3) and description of the method (section 4). Section 5 and 6 present the results of benchmarking of the selected tools and discussion of the result. The paper ends with conclusions and indication of future work.

Background

The focus of this paper is stream processing. It is then necessary to explain the characteristic features of this approach to processing compared to more traditional batch processing.

Batch processing is the processing of data in a group, referred to as batch. This means that data has to be collected first and then processed on-demand or based on a schedule. In this case we know in advance the volume of data to work on and the system we use for calculations can plan processing steps accordingly. Algorithms can also be designed in a way to allow optimization of the resources used. The batch processing is dedicated for large quantities of data that is usually not time-sensitive.

In the **stream processing** there is no collection of data for “future” use. Data is sent to analytical modules immediately. Such an approach allows addressing the real-time or near-real-time scenarios. This also entails additional requirements for the processing system. It is also more difficult to design algorithms that need to consider a constant update of the results when new records arrive in a stream.

There are then several features specific for the stream processing [6].

- Buffering past input as streaming state.

Data is not stored explicitly but this does not mean that it is forgotten. Arriving data changes the way we would interpret the overall situation or data to come. Thus, it changes the so-

called streaming state. The typical case in the studied domain would be to remember the last known position of a moving object, along with a timestamp. The state can also be remembered as values accumulated over time. Many algorithms can be implemented this way, where future input can be matched with past input.

- Joining streams with streams

It is very common in the SQL world to join tables to perform a query. The same can apply to the stream processing but here instead of static tables we have dynamic streams. Joining streams thus requires an additional temporal dimension to be considered. For example, we can detect if tracked moving objects were close to each other in a specific time frame.

- Streaming aggregations

Another very common query type in SQL is aggregation. A big number of rows is reduced to a small number of unique entities with assigned totals. The same can also apply to streams. One of the options is an accumulation of values over the whole history. The most demanding, however, is the accumulation of values in a specific time frame. For example, one can request the total number of moving objects observed in a monitored area in the past 24 hours. Results of the query will differ depending on the moment of its execution. In the stream processing, we need to have the answer at any moment, which represents the current state of the world characterized by the collected data. The totals in aggregations change when new data arrives.

- Handling late and out-of-order data

This issue is a consequence of how data is collected. We do not have any control over the objects, whether they send data immediately or need to cache because of connectivity issues. Data cannot be sorted either because it is not physically stored. Storing it locally would mean constant reprocessing because of out-of-order events and that the overall would be counterproductive. Therefore, additional measures were undertaken to handle it. First of all, the most distinctive for the stream processing is watermarking, which allows tracking of events (data accompanied by a timestamp) that still need to be processed although being late. The out-of-order events are not problematic as the stream processing considers them in specific time frames.

Software selection

The analysis of moving objects is a task that is best performed using stream processing. For this purpose, we need to choose appropriate software and this is what the paper is about. Currently, on the market, there are a couple of computation engines that support stream processing. The literature review reveals few studies focused on benchmarking streaming computation engines [7, 8, 9, 10]. However, those studies show different use cases and there is no clear winner. Therefore it is necessary to design our criteria for a specific case. Thus, let us discuss the selection of criteria for such software.

Criteria for Selection

There are many software solutions supporting this use case. We, therefore, de-fined criteria to support designers in making the choice. They are as follows:

- Horizontal scalability, that is the scalability of a cluster of many processing nodes. The horizontal scalability allows going beyond what a single machine, used in vertical scalability, may provide, as, depending on the requirements, one may distribute the processing to hundreds of thousands of nodes, which delivers resources many times greater than any vertical scaling may provide.
- Maturity of the solution and commercial as well as community support. While there are multiple big data stream processing solutions, we recommend picking the ones which are more widely adopted and available for not less than several years. This is

due to the fact that streaming applications running on clusters are notoriously difficult to debug. While some brand-new cutting-edge technologies may provide some benefits, without proper community support (on some web forums, etc.) it may be very difficult to find root causes for different issues that we may encounter along the way.

- Suitable licensing for a commercial project.
- Support by cloud providers. Some software solutions are well adopted on cloud platforms, which greatly increases ease of deployment and subsequent management of the system in production. Theoretically, we may use a cloud platform only in the Infrastructure as a Service (IaaS) model and install all the software manually on the hardware operated by the cloud platform. Then, we may use any software we like. Still, this does not allow us to utilize some of the most important advantages of the cloud. The cloud platforms provide more managed solutions, which deliver both the hardware and pre-configured soft-ware distributions, along with functions of monitoring, encryption, etc. Such managed solutions are nevertheless available only for limited software solutions and, for example, not all stream processing engines may be used this way.

Our next step was the selection of software candidates based on a keyword search for “stream processing” and “stream analytics”, along with “big data”.

First Round Candidates

As one of our primary requirements was the stream processing, we have focused on software providing this capability. For each solution, we filled in details concerning an implementation language, a license, and stream features. We also con-fronted the features with our criteria for selection and presented the major pros and cons. Below we shortly characterize main solutions in this area.

Apache Spark[1] is implemented in Scala and can be used from Scala, Java, and Python. It is based on the Apache license. It supports the stream storage and the stream processing. As an advantage, we can mention tight integration with the batch processing framework and ML library and the capability for the structured streaming. It is also production-ready. On the negative side, it is a near-real time solution (up to a few seconds of delay). Continuous processing with low (~1 ms) end-to-end latency is still in the experimental phase, hence, not recommended for production environments.

Apache Kafka[2] is written in Scala and Java. It can be integrated with many languages, including C++, Haskell, Node.js, OCaml, PHP, Python, Ruby, C#/.NET. It is based on the Apache license. It supports the stream storage and the stream processing. On the positive side, it is very fast, mostly due to the simplicity – it is more a low-level publish-subscribe message broker than a data processing pipe-line.

Apache Flink[3] is also implemented in Java and Scala and additionally pro-vides interfaces to Python. It is based on the Apache license. It supports the stream storage and the stream processing. Its main advantage is speed – it is faster than Spark, and seems to be reliable and powerful for complex stream processing.

Apache Pulsar[4] is dubbed as a cloud-native, distributed messaging and streaming platform. It can be bound with Java, C++, Python, and Go. It supports the stream storage but not the stream processing. It is a publish-subscribe messaging system, and in this respect, it is very similar to Kafka. Still, Kafka is more efficient, supports more features and is easier to use.

[1] http://spark.apache.org
[2] https://kafka.apache.org
[3] https://flink.apache.org
[4] https://pulsar.apache.org

Apache Storm[5] can be used with JVM-based languages (predominantly Clojure), JavaScript, Python, and Ruby. It is based on the Apache license. It supports the stream processing but does not support the stream storage. Its advantage is speed. The drawback is that it is not so convenient for more complex tasks.

Apache Samza[6] is implemented in Java and Scala. It is based on the Apache license. It supports the stream processing but does not support the stream storage. There are not any advantages related to our task. It is not suitable for more complex tasks and depends on Kafka. It is not actively developed (the recent stable version was launched almost 2 years ago).

In **Table 1** we provide the comparison of the main features of the studied solutions.

Table 1. The comparison of the main features

Solution	Horizontal scalability	Maturity/ community	Suitable licensing	Cloud support
Apache Kafka	++	++	++	++
Apache Spark	++	++	++	++[7]
Apache Flink	++	+	++	++[8]
Apache Storm	++	-	++	-
Apache Samza	++	-	++	-

It is important to note that both Apache Flink and Apache Spark[9] are supported by cloud providers as a managed solution. An example of a good integration with the cloud is EMR[10]. It is the industry-leading cloud big data platform for processing vast amounts of data using open source tools such as Apache Spark and Apache Flink.

Apache Kafka is mentioned as an underlying solution both for Flink and Spark. It is not a fully-fledged stream processor but merely a streaming engine. It can be considered as a kind of database that needs higher-level tools to work with. Therefore, in the next sections, we will study the short-listed candidates, i.e., Apache Spark and Apache Flink.

Apache Spark

According to the documentation, Apache Spark is a unified analytics engine for large-scale data processing[11]. It was originally designed for the static datasets batch processing (Spark generic) and the streaming data processing (Spark Structured Streaming) was added later.[12]

Spark can be run using its standalone cluster mode or on the already existing clusters. Its engine can access diverse data sources. The important feature of Spark is being scalable using Hadoop YARN, Mesos, or K8s. Spark (generic) has also support for powerful data engineering and machine learning libraries (e.g. MLlib). Originally, Spark was written in Scala. However, there are also APIs available in Java and Python. It is an open-source solution and has a huge community. Spark Structured Streaming supports the stream processing using a micro-batch processing engine (100 milliseconds latencies) and continuous processing (still in an experimental phase). It ensures an end-to-end exactly-once fault-tolerance guarantee.

[5] http://storm.apache.org
[6] https://samza.apache.org
[7] https://aws.amazon.com/emr/features/spark/
[8] https://aws.amazon.com/about-aws/whats-new/2019/11/you-can-now-run-fully-managed-apache-flink-applications-with-apache-kafka/
[9] https://aws.amazon.com/emr/features/spark/
[10] https://aws.amazon.com/emr/
[11] https://spark.apache.org/docs/latest/
[12] https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html

From a practical point of view, a stream processed by Apache Spark Structured Streaming is treated as an unbounded structured table and it is possible to perform the same operations as in static Spark SQL. Then, the streaming output can be saved on various sinks such as Kafka or file system files.

Apache Flink

According to the authors of the solution, Apache Flink is "a framework and distributed processing engine for stateful computations over unbounded and bounded data streams."[13] What is referred to as bounded stream processing is in fact equivalent to the batch processing. Unbounded streams have a start but no defined end and are equivalent to the stream processing.

It is important to emphasize that Apache Flink was designed with data streams in mind – it is referred to as a stream processor. The developers also paid attention to flexibility in programming, hence Flink allows for integration with SQL and Table API. Finally, in the context of the moving object analysis scenarios, one of the interesting features could be CEP (Complex Event Processing) for pattern discovery.

The main feature of the Flink approach is the abstraction of streams. They are perceived as temporary tables, also referred to as dynamic tables. The dynamic tables are changing over time. They can be queried like static batch tables but using a continuous query. Such a query never terminates and produces a dynamic table as a result. It is important to note that the result of the continuous query is always semantically equivalent to the result of the same query being executed in the batch mode on a snapshot of the input tables.

Summarizing, the following steps are performed in Flink to return the analysis results. 1. A stream is converted into a dynamic table. 2. A continuous query is evaluated on the dynamic table yielding a new dynamic table. 3. The resulting dynamic table is converted back into a stream.

Method

In order to choose the most appropriate tools for the analysis of objects in motion that send high-volume location data we have decided to get hands-on and test solutions on real data streams. This gave an extra opportunity for the identification of additional obstacles that are not clearly stated in the documentation. Also, we could assess the community involvement which is also very important as companies usually struggle with the implementation of more complex scenarios than the ones presented in tutorials.

Minimum Working Example

Based on data streams available for this study, we have defined a simple scenario. Two teams were implementing the same functionality: one in Apache Spark and the other in Apache Flink. They were supposed to come up with a so-called minimum working example (MWE). We then measured performance in terms of consumption of system resources (CPU, RAM). Please note that we did not consider the running time as we were operating on a data stream, so there is no such notion as the end of processing.

Implementation of MWE served several purposes. First, we checked the capabilities of the analyzed solutions. Second, it provided hints as to how difficult it is to implement a specific scenario, what are the challenges and obstacles. Third, we could compare the efficiency of solutions measured by CPU and consumption memory requirements.

The scenario for analysis consisted of counting the number of messages sent by moving objects, grouped by type of the object. There were two additional variants: counting all

[13] https://flink.apache.org/flink-architecture.html

messages for the whole history of observation and counting messages within the 5-minutes tumbling windows.

There were also additional non-functional requirements for the implementations. The streams to be used were served by Apache Kafka in CSV format. There were several streams and they had to be combined and deduplicated before any further processing. After the processing, the resulting streams should have been sent back in Avro format and stored in a database. Whereas the second variant could use “insert” queries (aggregation within window), the first required “upsert” queries (both insert and update), and message keys were necessary (global aggregation).

Getting System-wide Measurements

In order to compare both implementations, system-wide performance metrics were designed and then monitored. For the purpose of the monitoring, no other significant processes were running on the host machine at the time of those measurements. The main focus in the study was put on memory, processor load, and message processing lag. Therefore, the following metrics were used in the experiment:

- Average CPU (15 min, 12h) and Max CPU (1 min, 12h)

The server average CPU load during the 15 minutes sliding window was measured every minute for the period of 12 hours. The final value for this metric is a simple mean of those averages. Similarly, it was possible to calculate the maximum CPU usage. However, for the maximum CPU load, the one-minute sliding window was considered. The bash script used for recording the measurement is following:

```
while true; do uptime >> uptime.log; sleep 1; done
```

- Average RAM (1s, 12h) and Max RAM (1s, 12h)

The average RAM usage across all server cores was measured every second during the period of 12 hours. The final metric was an average of the recorded measurements. Similarly, the max RAM metrics refer to the single maximal average RAM usage value across all cores during the period of 12 hours. The bash script used for recording the measurement is following:

```
while true; do
  echo "$(date)" `cat /proc/meminfo | grep Active: | sed
's/Active: //g'`/`cat /proc/meminfo | grep MemTotal: | sed
's/MemTotal: //g'` >> ram_monitoring.txt
  sleep 1
done
```

- Average 5 min lag (12h), Max 5 min lag (12h), and Percentiles 5 min lag (12h)

The experiment of counting and grouping received messages for the last 5 minutes (moving windows) with 5 minutes watermark involved inserting results to the relational database. The lag metrics refer to the time difference between the count result database insert timestamp and the timestamp of the sliding window period (end window timestamp). Those metrics can be considered as an average, maximum, and percentiles of event processing delay during the period of 12 hours. The metrics were calculated by creating appropriate SQL query for the relational database containing the results.

Results

We have conducted our benchmarking on the following hardware.

Processor	2 x Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz 16 cores
Memory	DRAM 256 GB
Disk Type	HDD, 4x 1.2TB 10K RPM SAS, transfer 6Gb/s, configured as RAID 5, interface SATA 6Gb/s / SAS 12Gb/s, PCIe 3.0
Network Adapter	4 x Broadcom Limited NetXtreme BCM5720 Gigabit Ethernet PCIe, 2x10Gb BT + 2x1Gb BT
Operating System	CentOS Linux (Version 7)

Below we present the results achieved with the two tools: Apache Spark Structured Streaming and Apache Flink. We compare them based on the system-wide measurements presented in the previous section.

Apache Spark Structured Streaming

This section presents the results obtained using single-machine (standalone mode) with Spark Structured Streaming (Scala) to deploy MWE (grouping and counting received messages in the 5 min window frames and global counts). From **Figure 1** and **Figure 2** we can infer that the results are stable throughout the whole measurement period. **Table 2** show average results.

Table 2. Spark Structured Streaming - MWE Performance Metrics

KPI	Value
Average CPU (15 min, 6h)	32.46%
Max CPU (1 min, 6h)	43.99%
Average RAM (6h)	66.96 [GB]
Max RAM (6h)	72.96 [GB]
Average 5 min lag (6h)	8.36 min
Max 5 min lag (6h)	12 min
95 percentile 5 min lag (6h)	10 min
Median 5 min lag (6h)	8 min
5 percentile 5 min lag (6h)	7 min

Source: Own work.

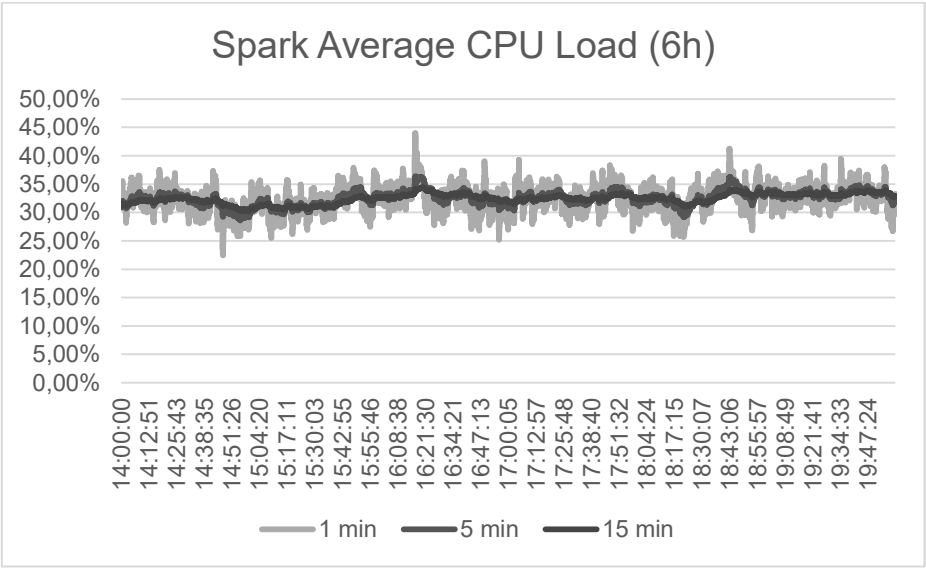


Figure 1. Average CPU load in Spark in various intervals

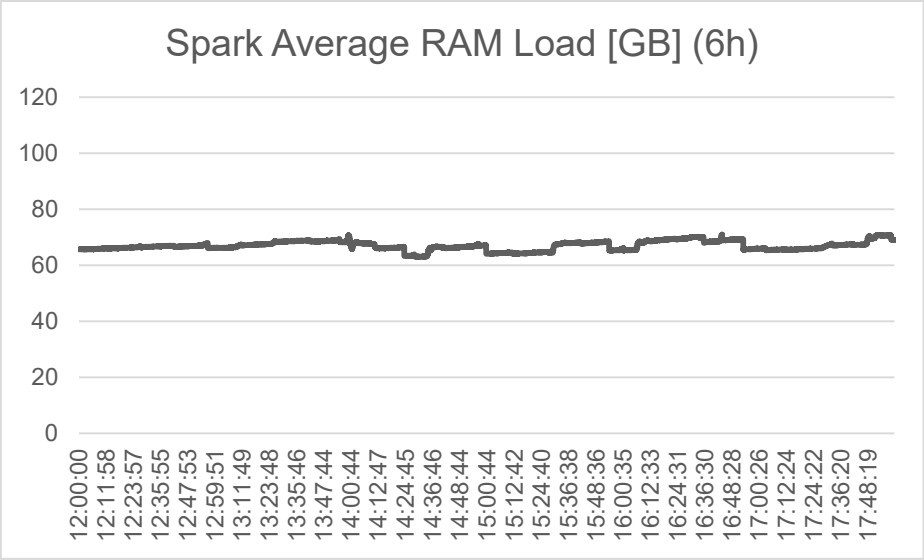


Figure 2. RAM usage in Spark

Apache Flink

The Flink job of making aggregations according to the described scenario was responsible for counting messages sent by objects in the 5 min window frames. The global counts were running during the measurement of the first metric. Below we present the summary statistics (**Table 3**), analogously to Apache Spark. **Figure 3** presents CPU load and **Figure 4** memory usage.

Table 3. Flink MWE Performance Metrics

KPI	Value
Average CPU (15 min, 6h)	4.20%
Max CPU (1 min, 6h)	7.50%
Average RAM (6h)	102.00 GB
Max RAM (6h)	108.69 GB
Average 5 min lag (6h)	11.10 min
Max 5 min lag (6h)	22 min
95 percentile 5 min lag (6h)	21 min
Median 5 min lag (6h)	9 min
5 percentile 5 min lag (6h)	6 min

Source: Own work.

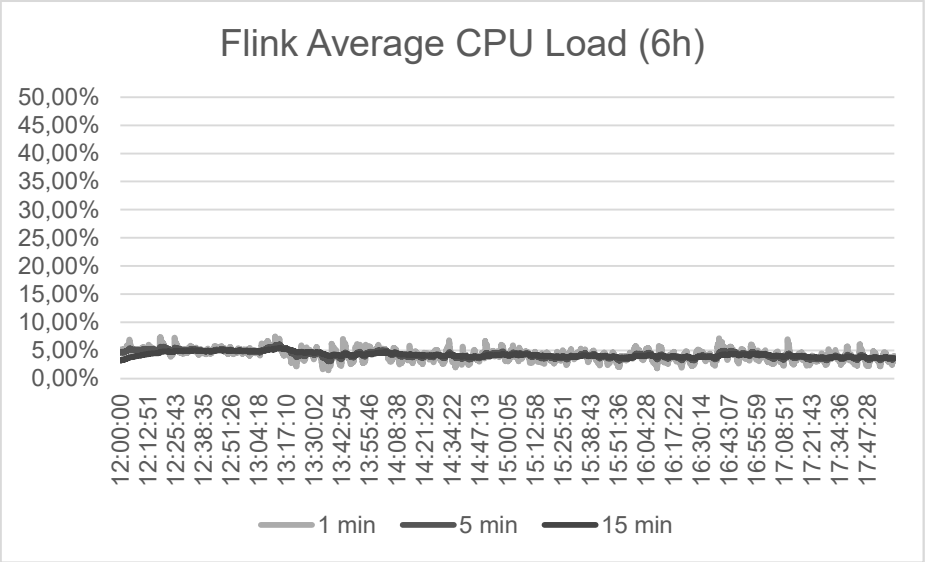


Figure 3. Average CPU load in Flink in various intervals

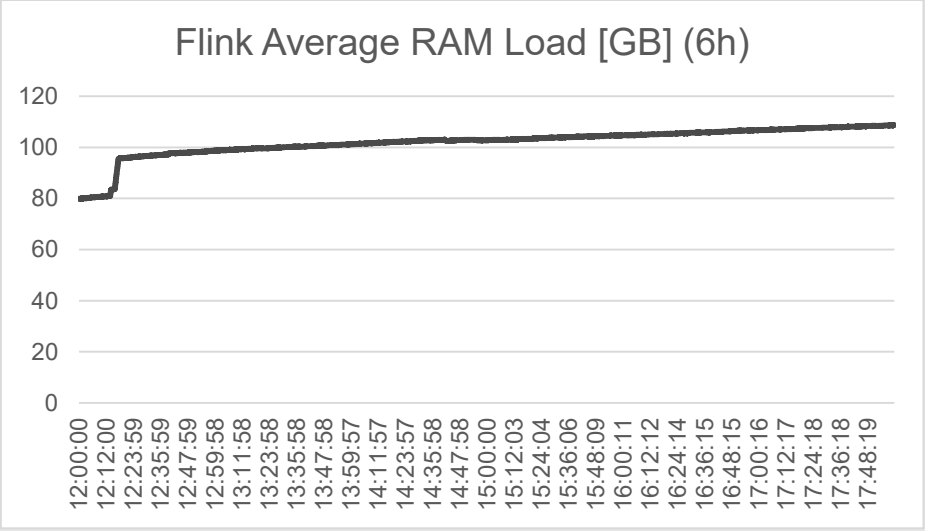


Figure 4. RAM usage in Flink

Discussion

In this section, we share our experience when it comes to the implementation of MWE in Spark and in Flink, as both provided a lot of technical challenges. The most cumbersome was understanding how the implementations work under the hood. Code examples available on the Internet were not always useful as they were fragmentary and did not mention any API version.

Initially, we assumed the transfer of data from streams to the stream processor via CSV files. Files were placed on RAID and mounted with NFS. Reading from files (FileSource) had a huge cold-start problem - around 60 seconds per query on data collected for only one day. Kafka was much faster to deal with this issue.

As Kafka was not planned in the first iteration of MWE implementation, initially we assumed that data will be stored directly in a database. Whereas Flink had no problems with such an approach, the JDBC driver for Spark did not provide update (UPSERT) mode. As a solution, we used Kafka topics to transfer data from Spark to a database. For comparability, Flink followed the same steps. Sending output to Kafka topics and creating a Confluent Kafka connector for saving data to the database worked with UPSERT mode as expected.

Concerning the output, saving Spark output to the Kafka topic in Avro format (instead of JSON) is the most popular way with the best community support. Message exchange between Flink and Kafka can be done in JSON and in Avro but the latter is more efficient.

There were also different approaches to submit a job. Scala does not go well with Jupyter Notebooks. In case of Spark, which was implemented in Scala, we executed jobs using either spark-shell or spark-submit with a prepared JAR file. For submission of job in Flink, which was implemented in Java, we used a compiled JAR file either.

The most important feature in the stream processing is time. Implementations in Python (pyspark and pyflink) were missing some important API calls referring to time windows. Moreover, documentation for Python was very often missing. Although we usually code in Python, we had to switch to Java for Flink and to Scala for Spark. Extending on time, it is very important to verify time zones in each environment (i.e., Docker containers, spark-shell, host machine, database) to prevent inconsistencies.

One of the specific challenges was the discovery of how to perform a specific step to achieve the required functionality. Although many tutorials were present for some solutions, there was always a “magic sauce” that was not revealed to the others. The discovery of an appropriate approach was mainly a trial-and-error process.

For example, in the case of Flink, there were drastic changes of API between 1.9 (when we started), 1.11 (our implementation target), and 1.12. (current stable); version 1.13 is in production. API is in constant reengineering and many methods were deprecated along with changes in classes. Although many examples were available on different fora, they were usually correct for older versions of Flink and were deprecated when we tried to run examples in version 1.11. The Flink project is developed by Alibaba, hence many discussions on problems are in Chinese.

Missing detailed documentation is also characteristic for Spark. In the documentation, there are only code snippets and not fully blown examples. Therefore, expertise in the selected tools is necessary to understand a missing context.

Conclusions and Future Work

Looking for a technology capable of processing streams for analyzing objects in motion that send high-volume location data we came up with a list of potential solutions, which was later restricted to Kafka, Flink, and Spark. The objective of the paper was to narrow the selection and identify the caveats and challenges to be verified in the second phase of our benchmarking effort. So far we have implemented a simple Minimum Working Example to make sure we know the effort necessary to implement a fully-fledged solution. We found that a learning curve is rather steep, especially when we consider more-than-standard requirements, i.e., integration with Confluent’s version of Kafka, storage of intermediate results in a database with key-based updates, exchange of messages in Avro serialization format. To meet these requirements Spark required an external library and for Flink we needed to provided dedicated implementations of some interfaces.

Concerning the efficiency of the solutions, Spark seems to be more memory efficient but at the cost of a higher CPU. Spark job made the machine quite busy with 30% CPU usage; allocated RAM was at the range of 70 GB. Flink was much more efficient – the same task caused a load of only 5%, but RAM consumption reached 110 GB. As we cannot tell which solution is preferred in the mentioned task, for future work we plan to extend MWE with additional scenarios. These scenarios will cover more tasks typical for the analysis of moving objects. In Scenario “Time Difference” we will measure the time passed since the last message was received; we can then alert if the time difference exceeds a certain threshold. In Scenario “Area Count” we will count the number of objects in a given area in a given moment or entering the area. In Scenario “Value Difference” we will measure the difference

between values of a parameter in two consecutive messages, e.g. speed of an object to detect if it is accelerating or slowing down.

Another direction of extension is the improvement of the measurement methodology. So far we have used the system-level measures which are just aggregations and do not help much in the identification of weak points of each solution. In the future, we plan to use the built-in solutions available via additional monitoring API. Both Flink and Spark offer their own interface for monitoring various additional parameters, like for example heap size or separation of CPU usage between manager and workers. We will look for a common denominator to compare Spark and Flink in greater detail. The ultimate goal of the next paper will be confronting the capabilities of the tools with specific algorithms according to requirements to determine architecture choice.

References

[1] Amini S, Gerostathopoulos I, Prehofer C. Big data analytics architecture for real-time traffic control. *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). 2017 06. https://doi.org/10.1109/mtits.2017.8005605

[2] Lekić M, Rogić K, Boldizsár A, Zöldy M, Török Á. Big Data in Logistics. *Periodica Polytechnica Transportation Engineering*. 2019 Dec 17;49(1):60-65. https://doi.org/10.3311/pptr.14589

[3] Xu H, Lin J, Yu W. Smart Transportation Systems: Architecture, Enabling Technologies, and Open Issues. In: Sun Y, Song H, eds. *Secure and Trustworthy Transportation Cyber-Physical Systems*. Vol SpringerBriefs in Computer Science. SpringerBriefs in Computer Science. Singapore: Springer; 2017. https://doi.org/https://doi.org/10.1007/978-981-10-3892-1_2

[4] Nguyen D, Vadaine R, Hajduch G, Garello R, Fablet R. A Multi-Task Deep Learning Architecture for Maritime Surveillance Using AIS Data Streams. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018 Oct. https://doi.org/10.1109/dsaa.2018.00044

[5] Kolajo T, Daramola O, Adebiyi A. Big data stream analysis: a systematic literature review. *Journal of Big Data*. 2019 06 06;6(1). https://doi.org/10.1186/s40537-019-0210-7

[6] Hueske F, Kalavri V. *Stream processing with Apache Flink*. O'Reilly; 2019.

[7] Chintapalli S, Dagit D, Evans B, Farivar R, Graves T, Holderbaugh M, Liu Z, Nusbaum K, Patil K, Peng BJ, Poulosky P. Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming. *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. 2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). 2016 05. https://doi.org/10.1109/ipdpsw.2016.138

[8] Marcu O, Costan A, Antoniu G, Perez-Hernandez MS. Spark Versus Flink: Understanding Performance in Big Data Analytics Frameworks. *2016 IEEE International Conference on Cluster Computing (CLUSTER)*. 2016 IEEE International Conference on Cluster Computing (CLUSTER). 2016 09. https://doi.org/10.1109/cluster.2016.22

[9] Quoc D, Chen R, Bhatotia P, Fetze C, Hilt V, Strufe T. Approximate stream analytics in apache flink and apache spark streaming. *arXiv preprint arXiv:1709.02946*. 2017;

[10] van Dongen G, Van den Poel D. Evaluation of Stream Processing Frameworks. *IEEE Transactions on Parallel and Distributed Systems*. 2020 08 01;31(8):1845-1858. https://doi.org/10.1109/tpds.2020.2978480

24th International Conference on Business Information Systems (BIS 2021)

Applications

https://doi.org/10.52825/bis.v1i.47

Published: 02 July 2021

Enterprise-Wide Metadata Management

An Industry Case on the Current State and Challenges

Rebecca Eichler [1], Corinna Giebler[1], Christoph Gröger[2], Eva Hoos[2], Holger Schwarz[1], and Bernhard Mitschang[1]

[1] University of Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany
{Firstname.Lastname}@ipvs.uni-stuttgart.de

[2] Robert Bosch GmbH, Borsigstraße 4, 70469 Stuttgart, Germany
{Firstname.Lastname}@de.bosch.com

Abstract. Metadata management is a crucial success factor for companies today, as for example, it enables exploiting data value fully or enables legal compliance. With the emergence of new concepts, such as the data lake, and new objectives, such as the enterprise-wide sharing of data, metadata management has evolved and now poses a renewed challenge for companies. In this context, we interviewed a globally active manufacturer to reveal how metadata management is implemented in practice today and what challenges companies are faced with and whether these constitute research gaps. As an outcome, we present the company's metadata management goals and their corresponding solution approaches and challenges. An evaluation of the challenges through a literature and tool review yields three research gaps, which are concerned with the topics: (1) metadata management for data lakes, (2) categorizations and compositions of metadata management tools for comprehensive metadata management, and (3) the use of data marketplaces as metadata-driven exchange platforms within an enterprise. The gaps lay the groundwork for further research activities in the field of metadata management and the industry case represents a starting point for research to realign with real-world industry needs.

Keywords: Metadata Management, Data Sharing, Data Transparency, Data Catalog, Data Marketplace

Introduction

In recent years, metadata management has regained focus in the scientific field and has once more become a topic of discussion in enterprises. Metadata management is important as it constitutes the activities to administrate an organization's data assets through metadata [1]. Without metadata, an organization does not know, for instance, what data it has collected, what it represents or whether it is confidential. Consequently, topics like legal compliance cannot be guaranteed without, e.g., information on confidentiality or the data's value cannot be fully leveraged when its meaning is unclear.

Data value in the form of new insights can be extracted though methods such as data analytics and is of great significance in enterprises as it can provide a competitive advantage [2]. In order to maximise the utilization of data and the extraction of its value, it needs to be made available to a wide range of users. In order to make data available throughout the enterprise, that is, beyond individual systems like data lakes or enterprise resource planning (ERP) systems, enterprise-wide metadata management is needed. Enterprise-wide metadata management encompasses and integrates metadata management

initiatives of both analytical systems such as data lakes and operational systems such as ERP systems and enables the access to and usage of metadata across the enterprise [3], for instance, in the form of a data-asset-inventory across various source systems. Yet, recent research, such as [3],[5],[6],[7],[8], mainly deals with metadata management explicit to data lakes, in which it is a central aspect [9]. Apart from research, there are a number of metadata management tools like data catalogs on the market, designed to solve specific metadata management tasks [3]. However, how to conduct enterprise-wide metadata management, which tasks are involved and which tools are suited, remains unclear.

We have conducted interviews with a globally active manufacturer, to gain insights into the current metadata management strategies and tools used in industry. Based on the case of the global manufacturer we examine metadata management challenges in practice and investigate whether these are solved by scientific research or existing tools. As a result this paper delivers four main contributions: (1) We present metadata management goals and their solution-oriented approaches in practice, based on the case of a large industrial enterprise; (2) We provide insight into current metadata management challenges; (3) We evaluate whether these challenges are sufficiently addressed in scientific literature or by tools from industry or research, and (4) based on this evaluation, we investigate research gaps in metadata management.

The remainder of this paper is structured as follows: The subsequent section illustrates the manufacturer's metadata management goals and the section thereafter presents their approaches for addressing these. Within the next section the manufacturer's challenges in metadata management are highlighted together with associated literature and tool coverage. The second to last section presents research gaps in metadata management and the last section concludes this paper.

The Industry Case and Metadata Management Goals

Metadata management is generally conducted to support data management. Hence, the data management goal needs to be clear in order to set up metadata management. The manufacturer is engaged in various sectors, such as the mobility or industrial sector and operates a global manufacturing network for mass and individual production. In this context, a lot of data is collected and stored, e.g., through the internet of things devices, enterprise resource planning systems and manufacturing execution systems. The manufacturer is pursuing the business strategy to become a data-driven industry 4.0 company. In the context of becoming more data-driven, the manufacturer has implemented novel technologies and concepts such as data lakes, storage repositories for data at scale and analytical purposes [2], and aims to establish an environment in which data can be shared freely and efficiently within the enterprise. With this goal the manufacturer aims to drive innovative data utilization and leverage more data value. For example, the ability to perform more data analysis supports realizing industry 4.0 use cases like predictive maintenance or real-time manufacturing quality analysis [10].

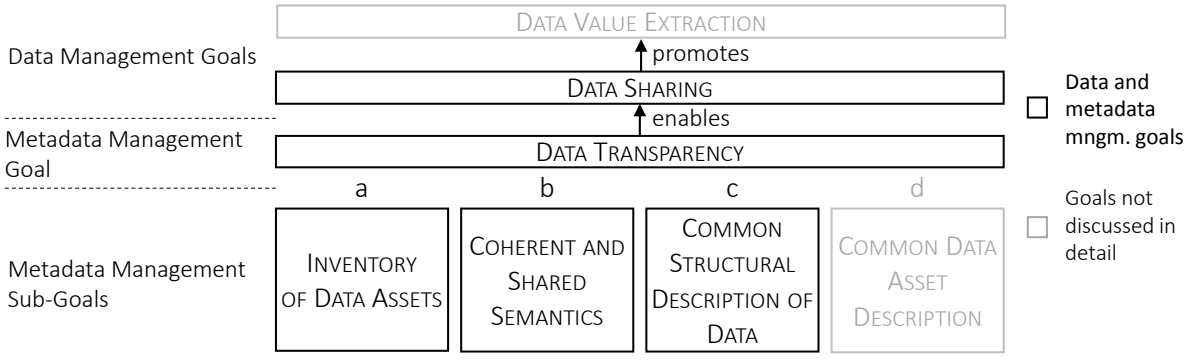


Figure 1. The manufacturer's data management and metadata management goals with sub-goals and corresponding solutions.

Data sharing enables a broader use of data and thus also promotes extracting the data's value. The sharing of data entails one party provisioning data and others accessing and using it. Sharing data 'freely' means the data will be made available to many of its employees and all types of users. To support all user types, the provisioning, access and usage of data should be enabled though self-service functionality without requiring the involvement of IT specialists. Doing so 'efficiently' signifies the process of sharing data should involve little effort. Nonetheless, the data's compliance and trust must be retained.

One prerequisite for sharing data freely and efficiently is *data transparency*. Zhu defines information transparency as the degree to which information is visible and accessible [11]. By our understanding, data transparency involves the ability to find, understand and access the data, which substantially aids data sharing. Data transparency can be ensured through metadata management by, e.g., acquiring sufficient documentation. Hence, the main metadata management goal is to establish data transparency, for which the manufacturer compiled four metadata management sub-goals, as illustrated in Figure 1. In the following we only discuss sub-goals (a) to (c), as the manufacturer has not yet dealt with sub-goal (d) in detail.

- Sub-goal (a) involves taking *inventory of data assets* across the enterprise. To exploit data value, it must first be known what data are available. As there is a multitude of storage systems, like data lakes and ERP systems, it is infeasible for single employees to retain an overview of existent data. For instance, a data scientist searching for customer data would have to know all systems which contain such data. Therefore, an enterprise-wide inventory of data assets is required.
- The second sub-goal, (b), entails the introduction of *coherent and shared semantics* throughout the entire enterprise. By ensuring that people use and understand the same concepts, data can be described and understood without misunderstandings. For example, the concepts 'customer' and 'consumer' are often used interchangeably, although a customer purchases goods and a consumer is the end user of goods, so they do not share the same semantics. Some also have several meanings, like the concept 'right', which can mean 'correct' or refer to a direction. Hence, coherent and shared semantics are needed to clarify data's meaning and avoid misunderstandings.
- Sub-goal (c) aims at establishing a *common structural documentation of data assets*, i.e., a modeling standard. There is a multitude of data models within the enterprise. These are modeled with varying tools, abstraction levels and documentation standards, for instance, as an entity relationship diagram created with tools similar to Visual Paradigm[1]. This makes it difficult to access, understand, integrate and reuse these data models. Therefore, a modeling standard based on, e.g., a common meta-model, must be established to foster model access, understanding, and sharing.

Data findability and consequently accessibility are improved through an inventory. Coherent and shared semantics together with a modeling standard and a standardized data asset description facilitate the understanding required for data transparency. Together these four sub-goals provide the data transparency required for data sharing.

Practical Approaches for Addressing the Metadata Management Sub-Goals

Having discussed the metadata management goals, this section illustrates the manufacturer's approach for reaching the sub-goals (a) to (c) with the solutions listed in Figure 1.

A Data Catalog for Taking Inventory of Data Assets. Sub-goal (a) is attained through the introduction of a commercial *data catalog*. A data catalog is a metadata management

[1] https://www.visual-paradigm.com/

tool, which is essentially a data inventory with documentation on registered data sources and data assets [12]. Alation[2] or the Collibra Data Catalog[3] are examples of such tools. The documentation ranges from business metadata such as the content description 'customer purchases', over technical metadata on, e.g., the data type 'String', to operational metadata describing for instance, the data's access history. The catalog provides a single interface for an enterprise-wide search of data. Beyond that, it provides functionality like enrichment, collaboration and governance features through, e.g., tags, commenting and user roles [12]. With search, documentation and other features, the catalog enables findability and understandability. The data scientist, for instance, can find customer data through the search and ascertain whether it fits their needs through the documentation.

Coherent and Shared Semantics by Means of a Business Glossary. Coherent and shared semantics, which constitute sub-goal (b), are established by compiling a *business glossary*. A business glossary specifies business terms and their definitions together with term relations for all business relevant concepts, such as a customer to product relation [1]. A business glossary tool such as erwin Data Literacy[4] can be used and embedded into the overall application landscape, so the terms do not merely serve as documentation but can be reused in other applications such as an enterprise-knowledge graph [13] or data models. For instance, a data model with a customer entity would refer to the corresponding glossary entry, thereby clarifying what exactly the entity represents. To establish enterprise wide acceptance and trust in the terms, business term management is executed. This entails governing the terms with, e.g., responsibilities and defining term maturity for signifying standardization levels. Hence, introducing a business glossary and conducting business term management is one step to gaining enterprise-wide shared semantics and thus, a basis for shared understanding of data assets.

A Meta-Model and Semantic Modeling for Shared Structural Descriptions. Sub-goal (c) involves creating a modeling standard and is done by introducing a *meta-model* for standardizing the modeling approach, and a tool set for *semantic modeling* [14]. The meta-model differentiates between model abstraction layers like a business-object layer and domain layer. The more abstract business-object layer may, for instance, describe a 'machine' object and the domain layer a 'bench drill'. The higher abstraction layers facilitate the integration of models, which results in an enterprise-wide knowledge graph and support insights into cross-functional usage of business objects. The integration is done through semantic modeling, by connecting more specific model's instances to the more abstract instances. For this, a semantic modelling tool set has to be set up, for example with Protégé[5], and integrated with the glossary, enabling users to search, explore and access semantic models based on business terms. Through the meta-model, semantic modeling tool set and integration with the business glossary, both enterprise-wide model understandability and findability are improved. The data scientist can now, for example, understand and find all data models which contain a customer entity.

With the introduction and integration of the listed tools a metadata management tool-landscape is created, which facilitates data transparency.

Challenges in Practice: A Literature and Tool Review

The manufacturer encountered several metadata-related challenges in the process of achieving their metadata management goal. We focus on three key challenges which we consider to be most relevant for research in metadata management. These pertain to: (1) *metadata management for data lakes*, (2) the *selection and composition of metadata management tool types* and (3) the implementation of *easy data provisioning, access and*

[2] https://www.alation.com/
[3] https://www.collibra.com/data-catalog
[4] https://erwin.com/products/erwin-data-literacy/
[5] https://protege.stanford.edu/

use through data marketplaces. As depicted in Figure 2, challenge one occurred irrespective of the sub-goals and challenge two throughout all the sub-goals. While implementing the sub-goals it was determined that the aspects of data access, usage and provisioning necessary for data sharing and value extraction are not yet sufficiently supported and therefore, this was identified as challenge three and as a missing sub-goal. For each challenge, related literature and tools are examined to find either solutions or research gaps. Although the challenges originate from the manufacturer, we evaluate these independent of the company, so the results are representative and not company-specific.

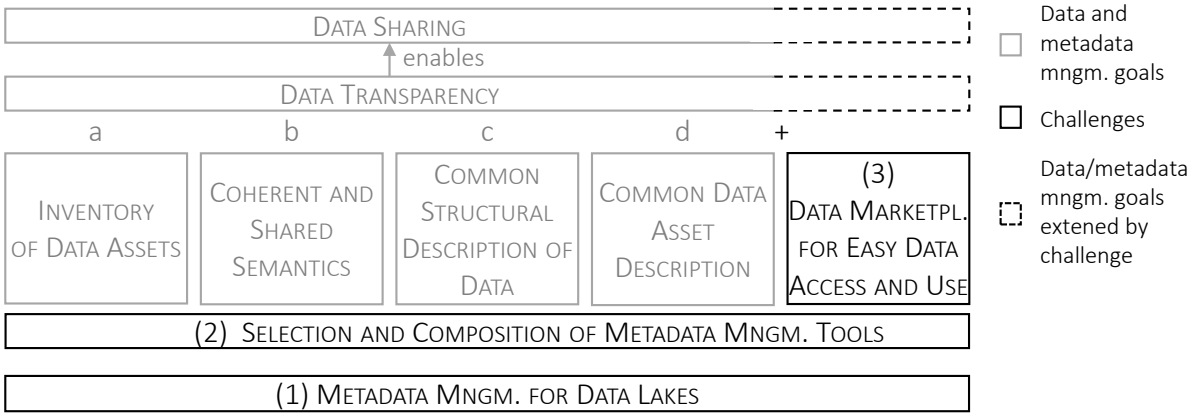


Figure 2. Extract of Challenges Arisen While Implementing the Data Sharing Goal

Challenge 1: Metadata Management for Data Lakes

A data lake is a storage repository for data at scale, which can incorporate data from heterogeneous sources, in its raw format, in various structures and without an overall schema [2], [6]. Data lakes can support the goal of making data available, e.g., by removing data silos [6] and exploiting its value. Metadata management is a critical success factor in data lakes, as it prevents these from turning into a data swamp, an inoperable data lake holding data that is not fit for use [5].

Challenges in Conducting Metadata Management for Data Lakes

Although it is known that metadata management is required in data lakes, it is not clear which combination of metadata tasks are sufficient to prevent the transition to a data swamp. There is a wide range of tasks which are supported by metadata management, to name a few: implementing governance specifications [4], data quality management and query processing [6]. For some of these tasks there is no sufficient insight as to how the corresponding metadata management needs to be implemented for a data lake. This concerns questions such as what metadata needs to be collected, what tools, protocols, and standards are needed for this task and do these exist, what do the processes look like and who is responsible, or also, how is the collected metadata integrated into the enterprise-wide landscape, but especially, how are these aspects different in data lakes. For example, these questions arise for metadata management to support data quality management within data lakes. Lack of clarity about essential tasks and how some of these have to be conducted in a data lake also leads to the challenge of designing and implementing a metadata management system or system landscape for managing the data lake.

Current State in Metadata Management for Data Lakes

A variety of scientific articles address the issue of data swamps and the required metadata management. The data lake system Constance and metadata management system GEMMS counteract a data swamp by collecting semantic and structural metadata, e.g., as annotations and schemata [6], [7]. The authors of Constance state that metadata management is essential for data reasoning, query processing and data quality management [6]. Yet it is not explained what metadata is collected or used for within data

quality management. This is also not clarified in the system CLAMS, designed to bring quality to data lakes [15]. This is an example of a specific metadata management task which' implementation is unclear and in addition is neglected by other systems such as GEMMS, which in contrast to Constance does not address data quality management. Sawadogo and Darmont define six mandatory tasks for metadata management systems ranging from semantic enrichment and data indexing over link generations and data polymorphism to data versioning and usage tracking [5]. However, a system containing all of these has not been implemented and hence not tested yet, and they do not seem to include the same scope of structural metadata, i.e. schemata, as previously mentioned systems nor topics such as data quality. According to Gröger and Hoos, metadata management must support self-service and governance in the lake [4]. There are multiple systems for managing data lakes like GOODS [16], Ground [17] and CoreKG [18] which implement a variety of data and metadata management tasks. Due to the divergences it remains unclear which tasks are strictly necessary to prevent a data swamp, how some of the metadata tasks differ when conducted in data lakes, which tasks are best suited to be integrated in an overall metadata management system for data lakes and which should be outsourced to a specialized system, like a system for data quality management.

Challenge 2: The Multitude of Metadata Management Tool Types

As described in the section on practical approaches the manufacturer is building a metadata management tool-landscape for achieving data transparency and data sharing. However, the selection and combination of tools has become increasingly difficult. In the following, we focus on tool types so the examinations are not dependent on individual tools and their specific characteristics and, therefore, more general observations and statements can be made.

Challenges in Differentiating and Combining Metadata Management Tool Types

There are many different tool types, such as business glossaries, data catalogs, and data marketplaces. The scope of their functionality is unclear and overlaps, for example, data marketplaces, which are platforms for trading data [19], also contain cataloging functionality [20]. Furthermore, the commercial tools are evolving by integrating new functionality, which is sometimes common for other tool types. For example, some data catalog products like the Informatica Catalog[6] have added data preparation features which is also a central aspect in data marketplaces [19]. Moreover, some vendors have rebranded their metadata management tools to, e.g., data catalogs, to capitalize customer interest [12]. There are also new tool types, which might simply be a synonym of another tool type, for instance, the emerging tool type called data hub might be a data marketplace. To identify the suitable tool types for comprehensive metadata management, a categorization of these tools is in order. In this context, information on their functional scope, characteristics, synonyms and subtypes are of interest.

As the tool types' functionality overlap, it is not clear which types to combine so they complement each other. Hence, the manufacturer needs an overview of the tool types' functional building blocks. For instance, data catalogs often contain a business glossary and data marketplaces contain data catalogs. Insights into compositions of these building blocks and how these work together are needed as a guideline for the erection of a compatible and comprehensive metadata management tool-landscape.

In closing, the manufacturer is struggling to select a set of tools, which enable comprehensive metadata management as these are not clearly differentiated, it is not clear which building blocks they contain, which are required and how these work together.

[6] https://www.informatica.com/de/products/data-catalog.html

Current State of Metadata Management Tool Types

There are variously detailed definitions and lists of functionality per tool type in literature. For instance, Zaidi et al. provide a definition on data catalogs [12] and Meisel and Spiekermann supply one for data marketplaces with a list of their functionality [19]. Some sources also specify sub-types of tools, such as Zaidi et al. on data catalogs [12], or Lange et al. [21] and Meisel and Spiekermann [19] for data marketplaces. However, the information on the tools has to be assembled from a multitude scientific articles, white papers and tool webpages and then be compared and evaluated. For example, Bhardwaj et al. define a tool called data hub, which is suspiciously similar to a data marketplace [22].

There are very few comparisons of metadata management tool types in scientific literature. Gröger and Hoos, differentiate the data dictionary, data catalog and data lake management platform through a high-level description [4]. There are various consulting blog articles on tools such as [23] which differentiates business glossaries from data dictionaries. Gartner published a list of metadata management tools by various vendors and presents the vendors' strengths and cautions, not, however, those of the tools [3]. Therefore, there is no comprehensive categorization and differentiation of metadata management tools to help in the tool selection.

In terms of functional building blocks, Zaidi et al. mention that a glossary can be contained in a catalog [12], Wells' framework of a data marketplace contains a data catalog [20] and based on Gröger and Hoos the data lake management platform also has a data catalog [4]. Hence, there is information spread throughout scientific articles from which one can laboriously deduce which tools possibly contain or complement each other. Some vendors' tool suites such as IBM's InfoSphere platform[7] can be used as a reference for combining tools, but these often focus on other topics like data management and not specifically metadata management. We have not found a comprehensive overview of tool types or their building blocks. It follows that there is no proposed building block assembly for comprehensive metadata management.

Challenge 3: Implementing Easy Data Provisioning, Access and Use

Currently, the manufacturer's metadata management tools support finding and understanding, not however, provisioning and accessing data. For this, additional metadata is required, such as metadata on the data owner or technical metadata to build an automated pipeline for provisioning. Having found the data, a user must at present contact the data owner and organize data access. If not available, they must set up the required environment to use it, e.g., for analysis. This process is time-consuming and challenging, especially for non-technical users. Therefore, the manufacturer needs additional metadata-driven tooling as depicted in Figure 2. To enable efficient data sharing, they need a platform that builds on the established metadata management tools and through which data provisioning and access is offered compliantly, via self-service with, e.g., data preparation functionality. Data marketplaces, metadata-driven platforms for sharing data, partially offer such functionality [19]. Optimally, an analysis environment can be obtained with courses, e.g., to learn analytics. For instance, a virtual machine with tools like Tableau[8] could be offered with the data. This would save technical users time and strongly support non-technical users, such as marketing specialists.

[7] https://www.ibm.com/de-de/analytics/information-server
[8] https://www.tableau.com/

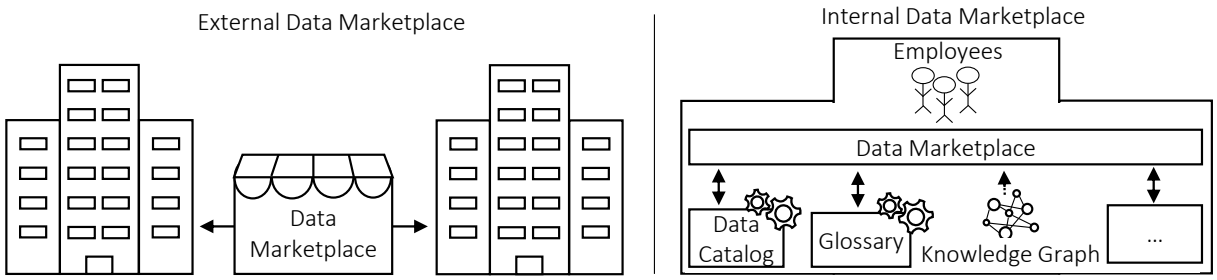


Figure 3. Data Marketplaces for Exchanging Data Between or Within an Enterprise

Challenges in Realizing Efficient Data Sharing Through Data Marketplaces

As shown in Figure 3, the data marketplace concept is initially designed for exchanging data between and not within an enterprise [19]. For the internal use, the marketplace should integrate with the existing tool landscape, tools such as data catalogs so it reuses collected metadata and does not duplicate functionality. But the marketplace tools are designed as standalone solutions and contain tools such as a data catalog. Besides, a marketplace for internal use should offer most of the enterprises data, as opposed to only specifically uploaded or selected data like the external marketplaces. In this context it should not only be able to store the data but also to refer to it, e.g., in the data lake. Furthermore, in the externally used marketplaces, monetary aspects serve as the main motivation to supply data [24]. Internally, the employees need other functionality driving them to supply their data, as monetarization promotes market competition, which is not desired within the enterprise. Recapitulatory, the manufacturer is challenged by extending its tool-landscape to support easy data provisioning, access and usability via self-service functionality through internal data marketplaces.

Current State in Applying Data Marketplaces for Enterprise-Internal use

In terms of acquiring data via self-service, Wells suggests that data marketplace receive a data storefront for shopping experiences like on Amazon [20]. To support data analytics, features for data preparation, curation etc. are envisioned [19], [20]. Moreover, Meisel and Spiekermann define external service providers as users of data marketplaces and suggest that these can provide analytical and infrastructure services as well as courses and services to support non-technical users [19]. However, the conceptual and implementation details of these analytics related features are not described in detail.

Using data marketplaces for an enterprise-internal exchange of data is suggested by Wells [20] and by Tata Consultancy Services [25]. Both provide a framework with functionality, but do not elaborate in detail. Using Commercial marketplace products like the Data Intelligence Hub[9] is difficult, as most are explicitly built as exchange platforms between companies, are often hosted as external platforms, and are therefore unsuited for storing or connecting all company data. Some marketplaces like Chordant[10] can be set up for private use, but Chordant is specialized for specific data on, e.g., autonomous mobility and is therefore not suited for making all company data available. Furthermore, commercial marketplaces have their own tool ecosystem. They double functionality and are not designed to fit into an existing metadata management landscape. Also, marketplaces like Chordant store the data and can not necessarily refer to data in other systems as required. In literature architectural aspects of data marketplaces are addressed in contexts like the trading of cloud of things resources [26], marketplaces in the IoT ecosystem [27] or trusted data marketplaces [28], but not for internal marketplaces. Finally, a multitude of monetization models have been discussed in various sources such as [29], but have not been verified in the context of internal marketplaces. Concluding, commercial marketplaces are so far unsuitable for the internal use and there are no detailed concepts, solutions with architectural proposals and detailed functional scopes for an internal data marketplace in literature.

[9] https://dih.telekom.net/
[10] https://www.chordant.io/solutions

Recapitulatory, the manufacturer had difficulties in implementing metadata management for data lakes, differentiating and combining metadata management tools and finally, enabling easy data provisioning, access and use of data through data marketplaces. For all the challenges, literature and tools from industry and research were reviewed and scanned for solutions, but none were found that fully solve the issues.

Identified Research Gaps

Based on the literature and tool review conducted in the previous section, we have identified three research gaps. As described in the section on challenge 1, it remains unclear which metadata management tasks are strictly necessary to prevent a data swamp, how to execute some of these tasks and how these differ when conducted in data lakes. It is also unclear which of these tasks need to be integrated into a data lake specific metadata management system. Hence, the topic *metadata management for data lakes* contains a research gap.

Second, as shown in the section on challenge 2, there is no work on the metadata management tool's building blocks and how these can be assembled to a comprehensive metadata management landscape. As more and more companies are faced with the challenge of building such a tool landscape, this is a significant topic. Hence, *the categorization and composition of metadata management tools* also constitutes a relevant research gap.

Lastly, the use of *data marketplaces within an enterprise* to foster the internal exchange of data and for these to serve enterprise-internal needs, has only been suggested, but not sufficiently explored, and therefore constitutes research gap three.

Conclusion

Metadata management has evolved in recent years and now presents a new challenge for companies. In this context, interviews were conducted with a globally active manufacturer to find out how metadata management is implemented in practice today, what challenges companies are faced with and whether these present research gaps. It was established, that the manufacturer's overall goal is to share its data freely and efficiently throughout the enterprise. To achieve this, the manufacturer defines data transparency as a prerequisite and derives four metadata management sub-goals. These are namely, taking inventory of data assets, the creation of coherent and shared semantics, the introduction of a common structural and common data asset description. While implementing these sub-goals, the manufacturer encountered several challenges, three of which, we identified as research gaps. The research gaps include: metadata management for data lakes, lacking categorizations and information on compositions of metadata management tools for comprehensive metadata management, and finally, absent research on the use of data marketplaces within an enterprise. Having defined research gaps, the groundwork is laid for further scientific research on metadata management, which will later enable exploiting data value fully and foster innovative data utilization while remaining compliant and within a legal framework.

References

[1] DAMA International, *DAMA-DMBOK: Data Management Body of Knowledge*. Technics Publications, 2017.

[2] C. Giebler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang, "Leveraging the Data Lake: Current State and Challenges," in *Proc. of the 21st International Conference on Big Data Analytics and Knowledge Discovery (DaWaK)*, Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-27520-4_13

[3] G. De Simoni, M. Beyer, and A. Jain, “Magic Quadrant for Metadata Management Solutions,” *Gartner*, 2019.

[4] C. Gröger and E. Hoos, “Ganzheitliches Metadatenmanagement im Data Lake: Anforderungen, IT-Werkzeuge und Herausforderungen in der Praxis,” in *Proc. of the 18. Fachtagung für Datenbanksysteme für Business, Technologie und Web (BTW)*, 2019. https://doi.org/10.18420/btw2019-26

[5] P. Sawadogo and J. Darmont, “On data lake architectures and metadata management,” *J. Intell. Inf. Syst.*, 2020. https://doi.org/10.1007/s10844-020-00608-7

[6] R. Hai, S. Geisler, and C. Quix, “Constance: An intelligent data lake system,” in *Proc. of the 2016 International Conference on Management of Data (SIGMOD)*, 2016. https://doi.org/10.1145/2882903.2899389

[7] C. Quix, R. Hai, and I. Vatov, “Metadata Extraction and Management in Data Lakes With GEMMS,” *Complex Syst. Informatics Model. Q.*, no. 9, 2016. https://doi.org/10.7250/csimq.2016-9.04

[8] R. Eichler, C. Giebler, C. Gröger, H. Schwarz, and B. Mitschang, “Handle - a generic metadata model for data lakes,” in *Proc. of the 22nd International Conference on Big Data Analytics and Knowledge Discovery (DaWaK)*, 2020. https://doi.org/10.1007/978-3-030-59065-9_7

[9] C. Giebler, C. Gröger, E. Hoos, R. Eichler, H. Schwarz, and B. Mitschang, “The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture,” in *Proc. der 19. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW)*, 2021.

[10] C. Gröger, “Building an Industry 4.0 Analytics Platform,” *Datenbank-Spektrum*, vol. 18, no. 1, 2018. https://doi.org/10.1007/s13222-018-0273-1

[11] K. Zhu, “Information Transparency in Electronic Marketplaces: Why Data Transparency May Hinder the Adoption of B2B Exchanges,” *Electron. Mark.*, vol. 12, no. 2, 2002. https://doi.org/10.1080/10196780252844535

[12] E. Zaidi, G. De Simoni, R. Edjlali, and A. D. Duncan, “Data Catalogs Are the New Black in Data Management and Analytics,” *Gartner*, 2017.

[13] L. Ehrlinger and W. Wöß, “Towards a definition of knowledge graphs,” *Semant. (Posters, Demos, SuCCESS)*, vol. 48, 2016.

[14] Y. Svetashova, S. Schmid, and A. Harth, “Towards semantic model extensibility in interoperable IoT data exchange platforms,” in *Proc. of the 2018 Global Internet of Things Summit (GIoTS)*, 2018. https://doi.org/10.1109/GIOTS.2018.8534561

[15] M. Farid, A. Roatiş, I. F. Ilyas, H. F. Hoffmann, and X. Chu, “CLAMS: Bringing quality to data lakes,” in *Proc. of the 2016 International Conference on Management of Data (SIGMOD)*, 2016. https://doi.org/10.1145/2882903.2899391

[16] A. Halevy *et al.*, “Goods : Organizing Google ’ s Datasets,” in *Proc. of the 2016 International Conference on Management of Data (SIGMOD)*, 2016.

[17] J. M. Hellerstein *et al.*, “Ground : A Data Context Service,” in *Proc. of the 8th Biennial Conference on Innovative Data Systems Research (CIDR)*, 2017.

[18] A. Beheshti, B. Benatallah, R. Nouri, and A. Tabebordbar, “CoreKG: A Knowledge Lake Service,” *Proc. VLDB Endow.*, vol. 11, no. 12, Aug. 2018. https://doi.org/10.14778/3229863.3236230

[19] L. Meisel and M. Spiekermann, “Datenmarktplätze - Plattformen für Datenaustausch und Datenmonetarisierung in der Data Economy,” *Fraunhofer ISST*, 2019.

[20] D. Wells, “The Rise of the Data Marketplace: Data as a Service,” *Eckerson Gr.*, 2017.

[21] J. Lange, F. Stahl, and G. Vossen, “Datenmarktplätze in verschiedenen Forschungsdisziplinen: Eine Übersicht,” *Informatik-Spektrum*, vol. 41, no. 3, 2018. https://doi.org/10.1007/s00287-017-1044-3

[22] A. Bhardwaj *et al.*, “Collaborative data analytics with DataHub,” *Proc. VLDB Endowment*, vol. 8, no. 12, 2015. https://doi.org/10.14778/2824032.2824100

[23] R. Lutton, “Data Management 20/20: Business Glossary Best Practices – TDAN.com,” 2019. [Online]. Available: https://tdan.com/data-management-2020-business-glossary-best-practices/25216. [Accessed: 02-Nov-2020].

[24] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, “Data market platforms: Trading data assets to solve data problems,” *Proc. VLDB Endow.*, vol. 13, no. 12, 2020.

[25] S. Saxena, “Enterprise Data Marketplace: Democratizing Data within Organizations,” *Tata Consult. Serv.*, 2018.

[26] A. S. Alrawahi, K. Lee, and A. Lotfi, “AMACoT: A Marketplace Architecture for Trading Cloud of Things Resources,” *IEEE Internet Things J.*, vol. 7, no. 3, 2019. https://doi.org/10.1109/JIOT.2019.2957441

[27] S. Schmid *et al.*, “An architecture for interoperable IoT Ecosystems,” in *Proc. of the 2nd International Workshop on Interoperability and Open-Source Solutions for the Internet of Things (InterOSS-IoT)*, 2016. https://doi.org/10.1007/978-3-319-56877-5_3

[28] D. Roman and G. Stefano, “Towards a reference architecture for trusted data marketplaces: The credit scoring perspective,” in *Proc. of the 2nd International Conference on Open and Big Data (OBD)*, 2016. https://doi.org/10.1109/OBD.2016.21

[29] M. Spiekermann, “Data Marketplaces: Trends and Monetisation of Data Goods,” *Intereconomics*, vol. 54, no. 4, 2019. https://doi.org/10.1007/s10272-019-0826-z

24th International Conference on Business Information Systems (BIS 2021)

Applications

https://doi.org/10.52825/bis.v1i.49

Published: 02 July 2021

A Scoping Review of the Digital Transformation Literature using Scientometric Analysis

Ziboud Van Veldhoven[1] [https://orcid.org/0000-0001-6013-7437], Vedavyas Etikala[1][https://orcid.org/0000-0002-5184-3812], Alexandre Goossens[1][https://orcid.org/0000-0001-8907-330X], and Jan Vanthienen[1][https://orcid.org/0000-0002-3867-7055]

[1] KU Leuven, Belgium

Abstract. Digital transformation is the rapidly expanding research field dealing with the increased impact of digital technologies on both business and society. Due to the large number of papers and the semantic ambiguity surrounding the terminology, covering such a broad topic is difficult. To help researchers gain a better understanding of the knowledge structure of the research field, we conduct a scoping review using scientometrics. We searched for publications dealing with digital transformation on both Scopus and Web of Science. We downloaded their bibliometric data and thoroughly merged and cleaned it using lemmatization and stemmatization. This dataset was analyzed using VOSviewer to create co-author networks and co-word occurrence graphs of the titles, abstracts, and keywords. We also visualized the growth of the research field and retrieved the top conferences and journals based on the number of papers and the number of citations. K-means clustering was performed on the abstracts and keywords to find similar research focuses. These findings highlight the broad scope of the research field, the ambiguity of the terminology, the lack of collaboration, and the absence of research into the impact of digital transformation on society. Moving forward, more research needs to be done to establish the boundaries of digital transformation and to investigate the importance of society in this phenomenon.

Keywords: Digital transformation, scientometrics, literature review, bibliometrics

Introduction

The world is going through a rapid digital evolution. The increased impact digital technologies have on both business and society has been frequently referred to as digital transformation (DT) in both information systems research and the professional world. DT is mostly defined in a business scope, such as ‘the changes in ways of working, roles, and business offering caused by the adoption of digital technologies in an organization or the operating environment of the organization’ [1]. In a broader sense, DT can be understood as the changes in all aspects of human life due to digital technology [2], or as ‘the continuously increasing interaction between digital technologies, business, and society’ [3, p.11]. The term dates back from the year 2000 [4] but it is only since 2015 the term truly gained traction.

The DT research is quickly gaining in popularity over the past few years. Due to the broad impact DT has on all aspects of society and industry, there is a large number of research topics related to DT. In addition, numerous researchers are linking their work with DT even though the connection is not always indistinct. Several authors have pointed out that it is not clear what is included in DT and what not [5]–[7]. Furthermore, there is a lack of consistent theoretical frameworks that can reconcile the literature [8]. This creates a situation in which it is hard to keep track of the research and its boundaries.

Given this outlook, an important scientific activity is to look back and analyze what has been researched so far. A series of literature reviews have already been conducted [6], [9]–[12]. Another method to examine the literature is scientometric analysis [13] that deals with analyzing the bibliometric meta-data surrounding scientific publications, e.g. the keywords, publication year, funding, and authors. Emphasis is placed on investigating the advances and structure of the research field by using data science and visualization techniques. Scientometrics is considered a complement to traditional literature reviews [13] and has several advantages. The results can be considered more objective because they are based on the analysis of bibliometric data and therefore not on the qualitative interpretation of the content of the papers [13]. This method also scales with a large number of papers without slowing down the process. Finally, the visualizations of the literature are easy to understand and can give new insights that are hard to grasp from literature reviews.

Scientometrics has been used in similar information systems topics such as industry 4.0 [14], digital innovation [15], and digital business models [16]. In DT, which can be considered as an overarching concept, only a handful more specific studies have been done. Reis et al. [17] performed a keyword analysis and some quantitative analysis, Schneider and Kokshagina [18] structured the literature based on technology and their impact, and Hausberg et al. investigated the co-citation graphs and research streams [19].

The DT research field is especially interesting to do a scientometric study due to its rapid expansion and size, its broad scope and impact, and the semantic ambiguity surrounding the terminology. Network graphs that can display the entire research field at glance can help to understand the extent, range, and nature of the phenomenon. For these reasons, we conduct a scoping review using scientometrics [13], [20]. Scoping reviews are ‘concerned with contextualizing knowledge in terms of identifying the current state of understanding’ [30, p.10].

- We contribute to the scientometric research by describing a detailed methodology with particular detail to data merging, cleaning, and manipulation using state-of-the-art natural language processing (NLP) algorithms such as lemmatization and stemmatization. This methodology can guide future studies on scientometrics.
- We contribute to the DT research by highlighting the breadth of the DT literature. We described the research growth, the most influential outlets, the research hubs, and the structure of the research field. The latter was done using co-occurrence graphs of the titles, abstracts, and keywords. These graphs can aid researchers to gain a better understanding of the research field without requiring much effort or expertise.
- We analyze and debate how these results add to the discussion of DT and provide several discussion points and a research agenda. The discussion points can help the scientific community to move forward in the DT research field.

The paper proceeds as follows: the next section discusses the applied methodology in detail. In chapter 3, we present our scientometric results followed by the discussion and the research agenda in section 4. Then, the limitations of this study are discussed in section 5. We end the study with a conclusion in section 6 and a link to access the figures used in this paper online in higher quality in section 7.

Methodology

The research aim of this paper is to conduct a scoping review using scientometrics of the DT literature. To do so, we based our methodology on the recommended workflow for mapping research using bibliometric tools proposed by Zupic and Čater [13] while also being guided by the methodology on how to conduct a scoping review of Arksey and O’Malley [20].

We searched both Scopus and Web of Science (WoS), due to their wide and multidisciplinary coverage, for English conference and journal papers published between the years 2000 and 2020. Several queries were evaluated to find a query that retrieved as many relevant articles as possible without including irrelevant articles. We found that searching for

papers with DT as keywords is a good strategy. This way, the obtained dataset has an extremely low false-positive rate (i.e., an article that is not related to DT). However, there are several issues with this search strategy. First, not all journals include keywords. Secondly, there is a higher chance of data quality errors in the keywords indexed by the databases compared to the titles. Thirdly, some papers use different synonyms as keywords to describe DT. Including these synonyms in the search query quickly results in massive datasets. Hence, we decided to compromise by also including papers with DT in the title. This results in a slightly higher false-positive rate, requiring more manual checking, but resolves some of the issues with only searching for keywords while keeping the dataset relatively precise and manageable. The full queries are listed below:

- Scopus: AUTHKEY ("digital transformation") OR TITLE ("digital transformation") AND PUBYEAR > 1999 AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English"))
- WoS: AK=”digital transformation” OR TI=”digital transformation” AND LANGUAGES: (English) Refined by: DOCUMENT TYPES: (ARTICLE OR PROCEEDINGS PAPER) AND TIMESPAN=2000-2020

The query was executed in December 2020 and retrieved 1985 articles in Scopus (1113 conference papers, 872 journal papers) and 1158 articles in WoS (527 conference papers, 638 journal papers). We downloaded all bibliometric data of these articles in CSV files. These include the authors, document type, citation count, references (only for Scopus), the abstract, title, keywords, and the source title. We then removed all duplicates by scanning for the same titles. In total, 716 duplicates were removed. This means that the overlap of WoS on Scopus is about 58%, or in other words, the Scopus dataset got extended by 22%. In total, this gives us a dataset of 2427 papers with 10 features or variables.

Next, we performed data preprocessing in five steps. In the first step, we merged the two datasets from Scopus and WoS using a custom-build python script. While this step is essential to get wide coverage of the literature, it is often overlooked in scientometric studies, e.g. [16], [17], [22]. Merging these datasets is not particularly easy because the formatting in each one is different. Hence, the script standardized all formatting differences such as fixing the punctuation marks between authors’ names and initials and merging the column names.

Secondly, the papers were inspected by two authors to identify irrelevant papers or data errors. The inclusion criteria to assess relevance were based on whether or not the abstract is coherent with changes in one or more aspects of a business, society, or industry due to digital technologies. For example, several papers discussed computer algorithms to transform datatypes. Also, several papers were removed that were wrongly classified as conference or journal papers. In total, we removed 32 irrelevant articles and 28 data errors such as wrongly imported papers or non-English papers bringing the total to 2367 papers.

In the third step, we performed several manual data cleaning manipulations for the title, abstract, and keywords variables. Missing values and spelling errors that occurred in several papers such as wrongly exported characters were fixed. We merged synonyms, such as the fourth industrial revolution and industry 4.0. Additionally, several words were transformed into their acronyms. For example, ‘chief digital officer(s)’ was changed into ‘CDO’. Other maintained acronyms include SMEs (small and medium enterprises), ICT (information and communication technology), ML (machine learning), and (I)IoT ((Industrial)Internet of Things). Finally, acronym variants were merged, such as SMAC-IT and SMACIT.

In the fourth step, we continued cleaning the words in the title, abstract, and keywords by building a Python-based text processor that utilizes the natural language toolkit package. The processor scans the entire text corpus and creates a dictionary made of all acronyms such as IT, IS, and AI. Next, all words except for the acronyms are changed to lower case. Then, all English stop words and non-alphabetical words that do not occur in the constructed dictionary are removed. We then singularized all words and changed British English into American English using the US2GB dictionary. To do so, a dictionary that includes 1,730

British and American words was used. We extended the dictionary with several DT specifics terminologies that were not included yet including digitalization, digitizing, and servitization.

In the fifth step, the Python processor changed each word into the most popular lemma of its stem. A lemma is a canonical form of the processed word. For example, ‘transforms’, ‘transforming’, ‘transformed’ are all forms of the lemma ‘transform’. To do this, the script runs through the corpus several times. In the first iteration, a dictionary was created of all the words their lemmas and their total occurrence count. For the lemmas, we used WordNet Lemmatizer. Next, we generated a second dictionary with the stem of each and their lemma, using the Porterstemmer package. The stem is the root of the word, e.g. ‘connect’ is the stem of ‘connects, ‘connecting’, ‘connected’ but also of ‘connection’. The lemma in this dictionary is chosen from the first dictionary’s most popular lemma of that stem and will represent the entire stem group. For example, if the first dictionary has two lemmas of the stem ‘strateg’, namely ‘strategy’, and ‘strategic’, the lemma with the highest occurrence will be chosen as the lemma of the stem ‘strateg’ for the entire corpus. In the final iteration, the script uses the created dictionary of stems and their lemmas to substitute the words with their respective lemma based on their stem value. Several exceptions were made to prevent cases where the lemma would reduce the meaning of the original word. For example, digitalization was added as an exception so that it cannot be changed into its lemma digital. Transforming the corpus into lemmas based on stem value makes sure that every word is an actual word (not always true for stems). Moreover, this method fixes different spelling styles for the same word. Lastly, this technique reduces the number of lemmas in the text while staying close to the original text, e.g. the number of unique words in the title got reduced from 2545 to 2293.

For the visualization of the data, we used Python for data exploration, cleaning, visualization, and clustering. VOSviewer [23] was used to create co-author and co-text network graphs. The clustering was done in three steps. First, we vectorized the data using tf-idf. We then applied principal component analysis to reduce the feature size to the top features to filter out noise. Lastly, we clustered the papers based on abstract and keywords using k-means. For the description of the topics, latent dirichlet allocation was used. The visualization was done with the Python package Bokeh.

Results

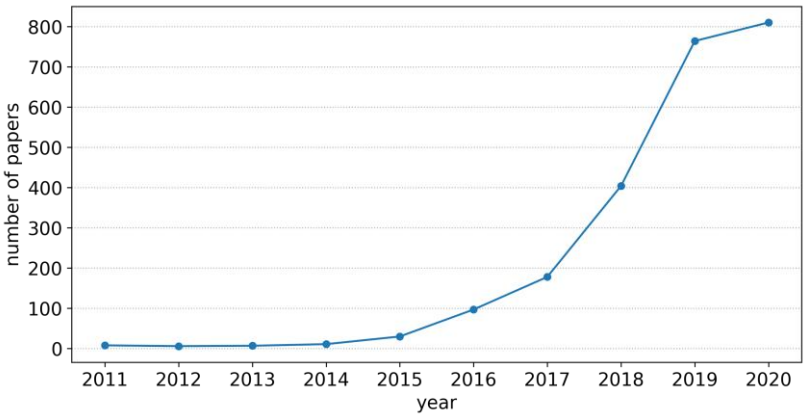


Figure 1. The number of papers published per year.

The research field is quickly expanding. In the past five years, the number of publications with DT as a keyword or in the title doubled annually, as shown in Figure 1. Two major factors could explain this growth. First, there is an increased interest in DT both by business and academics. In business, the importance of digitally transforming is more crucial than ever. Since the year 2000, digital is ‘the main reason just over half of the companies on the Fortune 500 have disappeared’ [24]. This translates itself into more research and interest from the academic world. Secondly, similar terminologies used to describe this phenomenon in the past, such as IT-enabled transformation, digitization, digitalization, or business

transformation, have started to consolidate into DT. Therefore, it is likely that this evolution will continue in the coming years. While the growth seems to be slowing down in 2020, one must consider the publication lag and the covid-19 pandemic as potential factors.

When looking at the publication details, we see that there are many outlets for research in DT. We give an overview of the most influential journals and conferences in the literature by the number of publications, citations, and average citations in Table 1. There are several points worth noticing. First, conferences seem highly important both by the number of papers and citations. They have a higher number of papers but a lower average citation count than journals. Second, the most influential conferences are focused on IS. Research in IS deals with the impact of IT in use by individuals and organizations [25], which fits closely with DT. These conferences usually have a track dedicated to digital transformation and business models (cf. ECIS 2020). When looking at the journals, the publication count is generally lower compared to conferences and compared to other fields. The journals themselves have a wide scope; there is no specific focus on DT itself. This can explain why there are many outlets with a small number of published DT papers. On the other side, journals are likely to suffer from publication lag, i.e. the time between submitting and publishing [26]. The average citation count is generally high due to several well-received papers in the journals. As the research field matures, we expect the journal papers to rise further in importance and new journals that focus on DT to emerge.

Table 1. Overview of important outlets and conferences.

Top outlets by volume	#	**Top outlets by citations**	#	**Top outlets by avg. citations (min. 2 papers)**	#
Sustainability Switzerland	35	MIS quarterly executive	697	European Journal of Information Systems (3)	114
Technological Forecasting and Social Change	16	European Journal of Information Systems	341	International Journal of Production Economics (2)	112
Journal of Business Research	15	Procedia Manufacturing	272	Strategy and Leadership (2)	100
MIS quarterly executive	12	Information Systems Research	262	Journal of Management Information Systems (2)	60
Business Horizons	11	Sustainability Switzerland	241	MIS Quarterly Executive (12)	58
Top conferences by volume	#	**Top conferences by citations**	#	**Top conferences by avg. citations (min. 2 papers)**	#
ACM international conference proceedings	74	ICIS	192	PACIS (2)	18
Advances in Intelligent Systems and Computing	66	Procedia CIRP	115	MKWI (10)	7
IOP conference series materials science and engineering	45	AMCIS	88	ICIS (30)	6
Lecture Notes in Computer Science	43	Lecture notes in business information processing	80	Advances in Information and Communication Technology – IFIP (9)	6
Lecture Notes in Business Information Processing	41	ECIS	70	Procedia CIRP (19)	6

In Figure 2, the evolution of the keywords over time is displayed to see research shifts and focusses. The years 2000 to 2015 are omitted due to the low number of papers. The keyword DT is not included in this graph because it is more or less shown in Figure 1. Emerging keywords were colorized for visual clarity. Several things are worth mentioning: the most popular keyword is digitalization. The second most common keyword is industry 4.0. Albeit its meaning is grounded in manufacturing, industry 4.0 can be considered as the DT of the manufacturing industry [14]. Upcoming keywords highlight the interest and role of big data and artificial intelligence (AI). Another frequent keyword used in combination with DT is the digital economy, which is the level of development of a social production system when digital technologies are implemented systematically [27]. In total, the ten most popular keywords based on frequency are DT (*f*:1575), digitalization (*f*:219), industry 4.0 (*f*:198), digital economy (*f*:88), digital technology (*f*:83), digitization (*f*:74), innovation (*f*:69), business model (*f*:65), IoT (*f*:64), and AI (*f*:23).

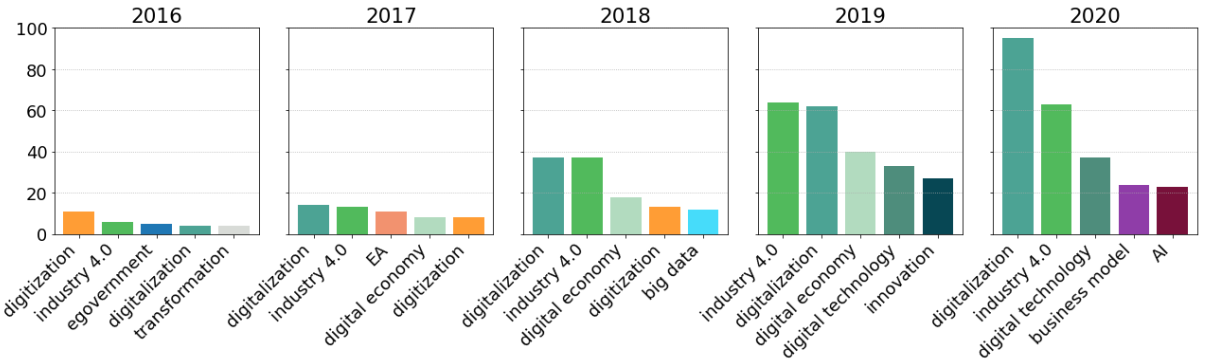


Figure 2. The keyword evolution over time.

In the co-word analysis, words are represented as nodes that are linked when they co-occurred together in one paper's title, keywords, or abstract. The number of times that happens in one paper is inconsequential because binary counting is used. The size of the nodes and links corresponds to the number of times they appeared in different papers. Clusters, determined by relatedness, are tinted in the same color. We performed a co-word analysis on the title, the abstract, and the keywords using VOSviewer. For each variable, the threshold of the number of occurrences needed for a term to be included was adapted so that each figure contained as much information as possible without overloading it. In the title and keywords, fewer frequent terms were found than in the abstract. Hence, a minimum of 10 occurrences per term was chosen for the title and keywords and 60 for the abstract. This means that all terms that appear x or more times are included in the graph and the bigger nodes appearing more times than x.

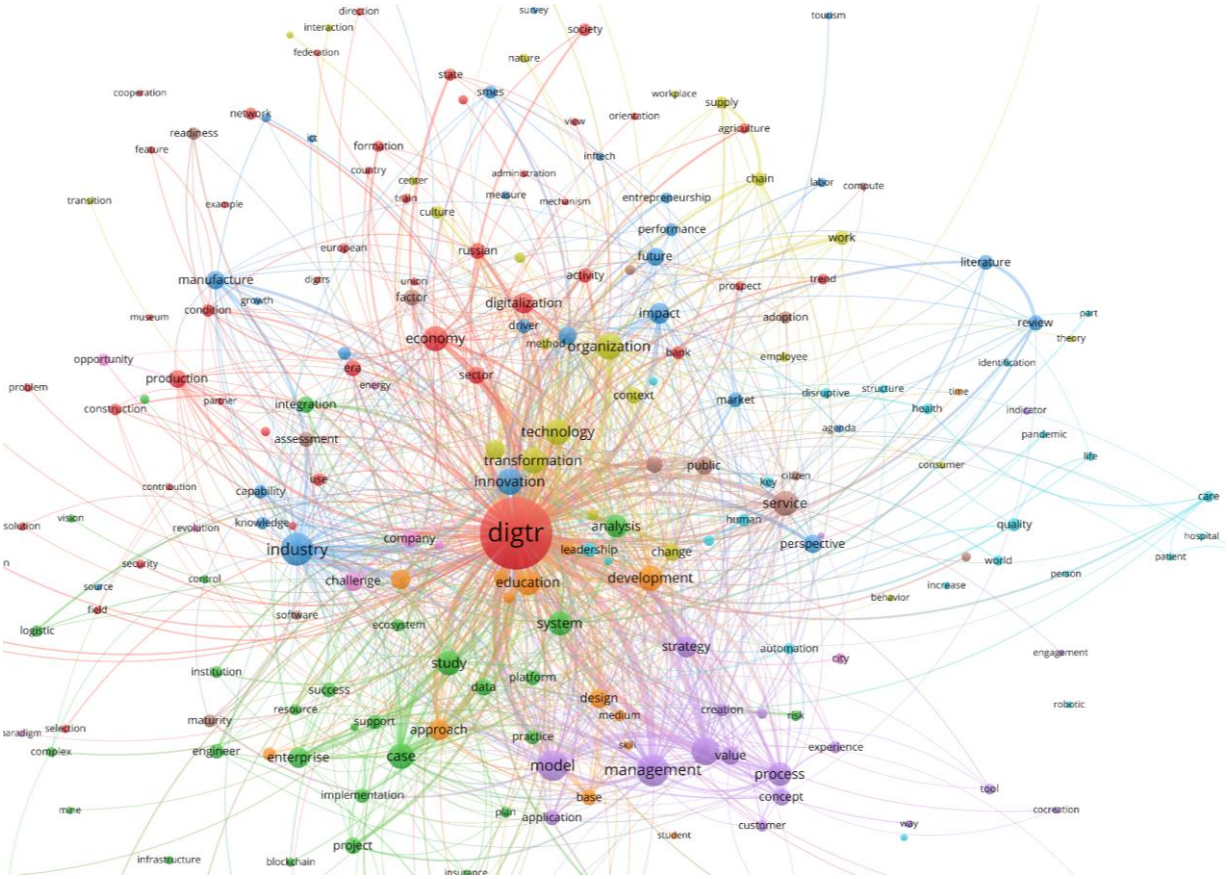


Figure 3. Title co-word network.

The title co-word analysis is shown in Figure 3. Analyzing the titles can be useful to detect the typical paper focuses or research areas. The network clearly shows DT research

is mostly focused on organizations or industries. Particularly, how organizations use and manage technology, innovate, and how industries are transformed. Several typical title structures can be identified in clockwise order: organization and technology research, supply chain research, literature reviews, health care research, public services, management and strategy, education, enterprise case studies, industry case studies, production and manufacturing research, and finally impact on the economy or sector level.

Similarly, several clusters can be identified in the abstract co-word analysis as shown in Figure 4. Irrelevant words like ‘study’ or ‘paper’ were excluded due to their irrelevance. The abstracts tell us more about the content of the papers than the title. Due to the higher word count in the abstract, and thus word overlap, the research hubs are less distinct as shown by the large clusters. If we zoom in on the clusters, we can identify several research areas. In the green cluster, we identify research about manufacturing, production, and supply chains; about the use and change of technology; about applications and solutions; about systems and networks. The blue cluster contains research about the development and implementation of technology; about society and economy; about education; about digitalization; and about analysis. Finally, several research hubs are present in the red cluster about employees and culture; about the impact of DT; about innovation in organizations; and about strategy and management.

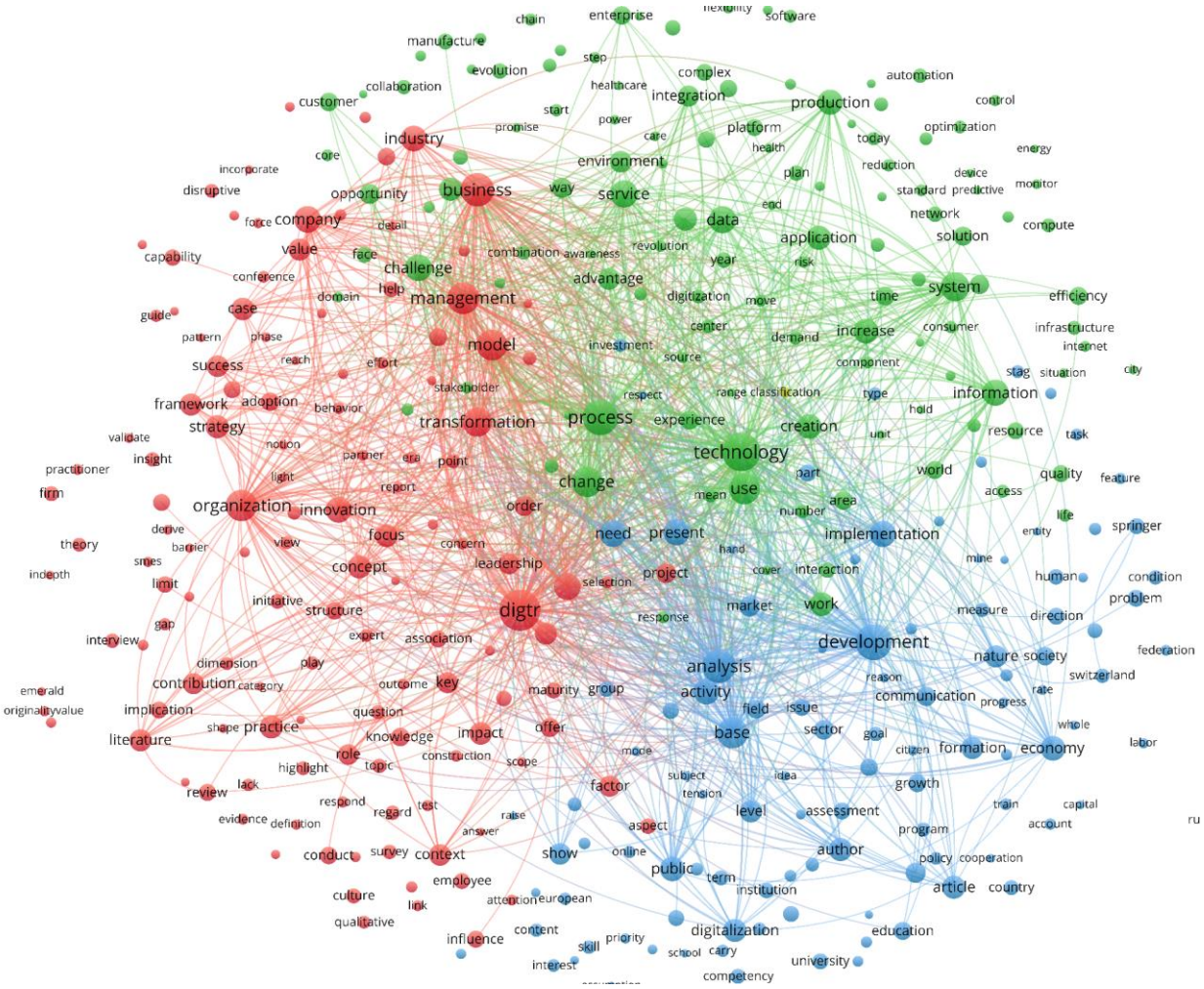


Figure 4. Abstract co-word network.

Lastly, the keyword co-occurrence network is shown in Figure 5. The keywords have the benefit of summarizing the entire paper in several words. This graph can also be used to understand domain closeness and relatedness. From the figure, it can be seen that DT has a broad scope with many distinct aspects. The graph shows that digitalization and industry 4.0 are often used in combination with DT, highlighting the overlapping terminologies. Other

highly connected keywords include the digital economy, innovation, business model, and technology. Several digital technologies are frequently mentioned such as big data, AI, cloud computing, blockchain, and (I)IoT. Once again, several research hubs can be identified. In clockwise order, starting from the top, the following hubs can be identified: leadership and CIOs; business model innovation (such as in transport and the sharing economy); egovernment or digital government and public services; digital ecosystems and platforms; strategy and change management; organizational culture; business agility; cyber-physical systems; transformation in education, production, manufacturing, and SMEs; healthcare; elearning and higher education; IoT and AI; enterprise architecture (EA), enterprise models, and requirements engineering.

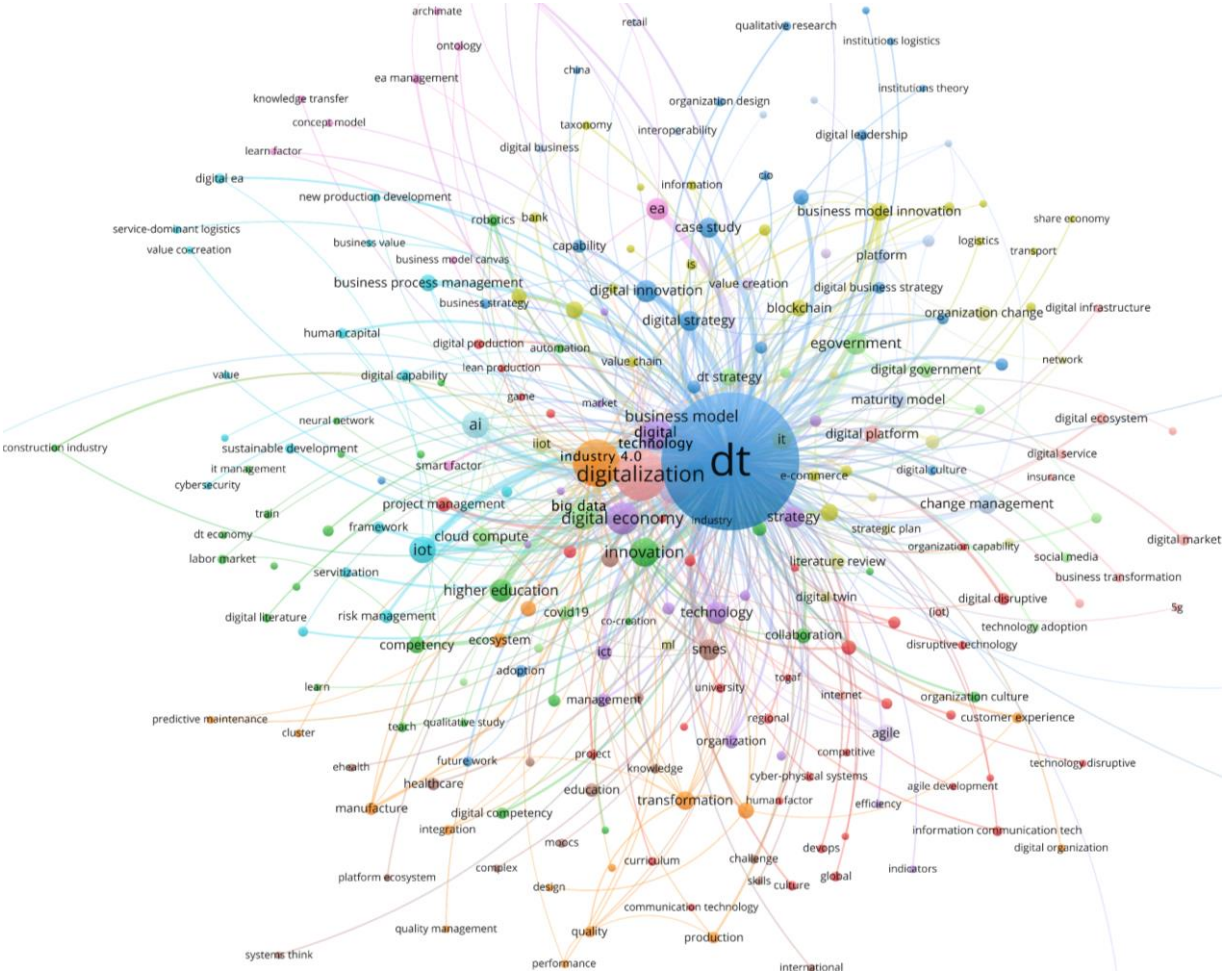


Figure 5. Keyword co-word network.

Using VOSviewer, a co-authorship network graph was constructed. This visualization displays the collaborations in the field by showing researchers as nodes and their collaborations (co-authored papers) as links between the nodes. The largest network of co-authors is shown in Figure 6. The different clusters correspond to several universities. The yellow, green, and brown clusters contain researchers from the Danube University Krems (Austria), the purple and cyan clusters contain researchers from the RWTH Aachen University (The Netherlands), the orange cluster contains the university national du sud (Argentina), and the blue cluster contains researchers from the Warwick business school (United Kingdom). These graphs show the research collaboration at glance.

Surprisingly, the largest co-authorship network is limited to 92 researchers. Furthermore, when we specify a minimum of two papers for an author to be included in the analysis, the network shrinks further to 32 authors. This is a rather unexpected outcome, as collaborations are generally higher in similar fields. This means that most DT research is located in smaller internal research groups. These research groups are plenty. Our analysis identified more

than 100 research teams, of which many important ones can be accredited. To give a few, the research group with the most publications consists of A. Zimmermann, M. Möhring, D. Jugel, and R. Schmidt of Reutlingen University. Another important group with a high number of citations is from the Ludwig-Maximilians University of Munich, with the researchers T. Hess, A. Horlacher, and S. Chanias.

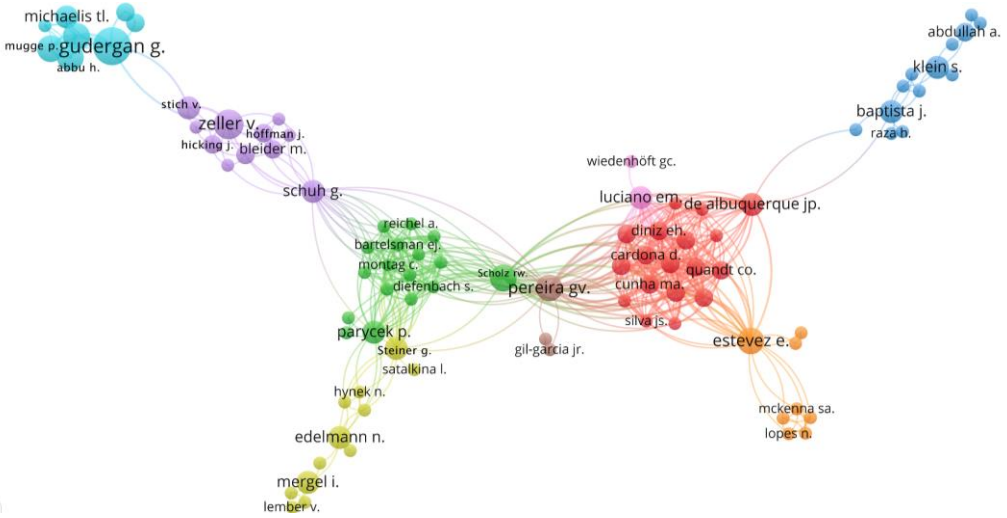


Figure 6. The largest co-authorship network.

One possible question to ask is whether collaboration is low due to different research areas. While the co-word networks above indicate that this is not the case due to the several research hubs in which researchers can collaborate, we further investigate this by applying a k-means clustering algorithm to the papers based on the abstract and keywords. This method groups similar papers together based on their semantic similarity of the keywords and abstracts. The clustering outcome indicates that the papers retrieved in this study have several common themes in which authors can collaborate. There are 10 clusters recognized by the algorithm. We used topic modeling to find the most appropriate keywords for each cluster. In short, the clusters are about: c1: organization, digital, people, technology, research, human, resource, new, online, communication, analysis, process; c2: development, digital, technology, russian, social, consumer, energy, common, regional, information, requirement; c3: technology, industry, company, model, energy, production, manufacture, digitalization, service, development; c4: public, service, digital, technology, egovernment, change; c5: architecture, digital, service, framework, business, design, sustainable, compute, technology, institution, achieve; c6: digital, construction, management, process, implementation, analysis, study, paper, pilot, infrastructure, organization; c7: business, digital, model, customer, organization, service, management, bank, process; c8: learn, digital, university, technology, high, student, engineer, future, virtual, derive, image, elearning, traditional; c9: smart, technology, digital, service, energy, urban, development, management, build, economy, strategy, digitalization, regional; c10: innovation, firm, SMEs, strategy, digital, capability, research, dynamic, transformation. Together, these findings suggest that collaboration in the DT research field can be increased.

Discussion and research agenda

The results of this paper further advance our understanding of DT and provide more insight into the current research. We provide several discussion points, based on further reading and the results, for IS researchers who need to play an active role in the research field moving forward. First, multiple authors mention that DT not only deals with changes in the business but also with changes in society [1], [28], [29], people [2], [17], and societal values [8]. However, these aspects are not visible in our results. Society was barely mentioned in the titles in Figure 3, and apart from culture, customer, and collaboration, the

abstracts in Figure 4 contain limited links to societal changes. Looking at the keywords in Figure 5, we find collaboration, customer experience, and adoption but no other keywords related to society. If it is true that DT impacts society, people, and values, and we believe it does, then more research needs to be conducted on the societal aspect of DT.

Second, the combination of findings and readings highlights the potential issue of ill-defined terminology. In figure 5, it is shown that many papers also use digitalization, digitization, or industry 4.0 in combination with DT. In detail, our dataset reveals that 26% of papers include at least one of these 'substitutes'. This is peculiar because their meaning is considerably different. Digitization is about changing analog information into digital data or information [9] whereas digitalization is about the increased use of digital technologies but in essence still doing the same things [1], [30]. Industry 4.0 is a term to talk about the fourth industrial revolution in manufacturing including smart factories, IoT, robotics, and predictive maintenance [14]. The meaning of DT is still disputed, although it can be agreed upon that DT sketches the bigger picture in which businesses and people are completely changing their ways of working by smartly deploying digital technologies [3].

In addition, this ambiguity is problematic because it creates a situation where the overuse and misuse of DT have weakened its potency [31]. While it is true that DT is a broad concept that contains the changes that technology brings forward in business and society, it does not necessarily follow that DT should be used for all changes that can be classified under this definition. This is a prominent issue for future research. The breadth of the research field shown in Figures 4 and 5 could be too large because of the wrongful inclusion of the term DT. On the other hand, it seems like many definitions of DT are too narrow compared to our results. Given this debate and our first discussion point, we believe that DT is in a unique position and has the academic attention to bring together research in business, technology, and society to study the impact of the increased use of digital technologies. Moving forward, we suggest that researchers and practitioners need continued efforts to keep the DT literature relevant and framed correctly.

Third, the co-authorship network in Figure 6 showed a lack of collaboration in the DT research. This is a rather unexpected outcome given the size of the research field and the number of authors. In comparison with the distinct clusters found in Figure 7 and compared to other fields, the level of collaboration is low. For example, when we perform the same query on Scopus but with business process management notation (BPMN) as a keyword instead of DT, we obtain 1,278 papers from which a co-authorship network of 188 authors can be created that clearly shows distinct research hubs. The challenge is now to promote DT research collaboration. This is important for sharing and merging specialized knowledge and expertise which is the engine behind scientific progression.

Future work could investigate the contribution of the different disciplines and the used methodologies. In addition, the boundaries of DT research with other research fields require more investigations. Acquiring a deeper understanding of what knowledge is missing from this broad research field is another fruitful area for future studies. Furthermore, more research is needed to reconcile the various aspects of DT into a coherent theoretical frame and promoting this construct for framing future DT research. A commonly accepted framework for DT can help to enhance collaboration between researchers by connecting similar research through well-agreed upon terms. The precise aspects of DT, their scope, and their meaning need to be investigated and demarcated more clearly to serve as a connecting means for both practitioners and researchers. Doing this can be beneficial for increasing the specialization of outlets and researchers, which would be a welcome addition.

Limitations

The scientometric analysis performed in this paper is based on the acquired dataset from Scopus and WoS. Several biases exist related to the data extraction such as the included outlets of these databases, the data quality, and the fact that different queries could result in

different findings. The insights from this paper were mainly based on the proxy that the title, abstract, and keywords provide a correct impression of the paper. Investigating the corpus of each paper would provide additional insights.

Conclusion

This research aimed to provide a scoping review of the DT literature using scientometric analysis. We described a detailed methodology on how to thoroughly prepare a dataset for scientometric studies. The results identified the general overview of the research field, including the evolution of papers being published each year, the most influential outlets, and the evolution of keywords. Additionally, we created co-word occurrence graphs of the titles, abstracts, and keywords. The data suggest that DT is not well defined and that there exist many different research hubs. It also suggests that the DT research field will continue to expand which makes fundamental, theoretical work increasingly important. Further work needs to be done on reconciling the literature and providing strict terminology.

Online material

All the figures used in this paper can be accessed online in higher resolution using the following link: https://feb.kuleuven.be/public/u0105262/bis2021/.

References

[1] P. Parviainen, M. Tihinen, J. Kääriäinen, and S. Teppola, “Tackling the digitalization challenge: how to benefit from digitalization in practice,” *Int. J. Inf. Syst. Proj. Manag.*, vol. 5, no. 1, pp. 63–77, 2017, doi: 10.12821/ijispm050104.

[2] E. Stolterman and A. C. Fors, *Information Technology and the Good Life*. Boston, MA: Springer, 2004.

[3] Z. Van Veldhoven and J. Vanthienen, “Digital transformation as an interaction-driven perspective between business , society , and technology,” *Electron. Mark.*, p. 16, 2021, doi: https://doi.org/10.1007/s12525-021-00464-5.

[4] K. Patel and M. P. McCarthy, *Digital Transformation: The Essentials of E-business Leadership*. McGraw-Hill Professional, 2000.

[5] A. Hanelt, E. Piccinini, R. W. Gregory, B. Hildebrandt, and L. M. Kolbe, “Digital Transformation of Primarily Physical Industries - Exploring the Impact of Digital Trends on Business Models of Automobile Manufacturers,” in *International conference on wirtschaftsinformatik*, 2015, pp. 1313–1327.

[6] G. Vial, “Understanding digital transformation: A review and a research agenda,” *J. Strateg. Inf. Syst.*, vol. 28, no. 2, pp. 1–27, 2019, doi: 10.1016/j.jsis.2019.01.003.

[7] M. Wade, “Digital business transformation: a conceptual framework,” *Global Center for Digital Business Transformation*, no. June. pp. 1–16, 2015.

[8] B. Hinings, T. Gegenhuber, and R. Greenwood, “Digital innovation and transformation: An institutional perspective,” *Inf. Organ.*, vol. 28, no. 1, pp. 52–61, 2018, doi: 10.1016/j.infoandorg.2018.02.004.

[9] E. Henriette, M. Feki, and I. Boughzala, “The Shape of Digital Transformation : A Systematic Literature Review,” in *MCIS 2015*, 2015, pp. 431–443.

[10] L. Caluwe and S. De Haes, “Board Level IT Governance: A Scoping Review to set the Research Agenda,” *Inf. Syst. Manag.*, vol. 36, no. 3, pp. 262–283, 2019, doi: 10.1080/10580530.2019.1620505.

[11] B. Horlach, P. Drews, and I. Schirmer, “Bimodal IT : Business-IT alignment in the age of digital transformation,” in *Multikonferenz Wirtschaftsinformatik (MKWI)*, 2016, pp. 1417–1428.

[12] J. Hagberg, M. Sundstrom, and N. Egels-Zandén, “The digitalization of retailing: an exploratory framework,” *Int. J. Retail Distrib. Manag.*, vol. 44, no. 7, pp. 694–712,

2016, doi: 10.1108/IJRDM-09-2015-0140.
[13] I. Zupic and T. Čater, “Bibliometric Methods in Management and Organization,” *Organ. Res. Methods*, vol. 18, no. 3, pp. 429–472, 2015, doi: 10.1177/1094428114562629.
[14] A. Janik and A. Ryszko, “Mapping the field of Industry 4 . 0 based on bibliometric analysis,” in *IBIMA conference*, 2018, pp. 1–15.
[15] R. Kohli and S. Devaraj, “Measuring information technology payoff: A meta-analysis of structural variables in firm-level empirical research,” *Inf. Syst. Res.*, vol. 14, no. 2, 2003, doi: 10.1287/isre.14.2.127.16019.
[16] A. Caputo, S. Pizzi, M. M. Pellegrini, and M. Dabić, “Digitalization and business models: Where are we going? A science map of the field,” *J. Bus. Res.*, vol. 123, no. February 2020, pp. 489–501, 2021, doi: 10.1016/j.jbusres.2020.09.053.
[17] J. Reis, M. Amorim, N. Melão, and P. Matos, “Digital Transformation: A Literature Review and Guidelines for Future Digital Transformation,” in *World Conference on Information Systems and Technologies*, 2018, no. March, pp. 411–421, doi: 10.1007/978-3-319-77703-0.
[18] S. Schneider and O. Kokshagina, “Digital transformation: What we have learned (thus far) and what is next,” *Creat. Innov. Manag.*, no. May 2020, pp. 1–28, 2021, doi: 10.1111/caim.12414.
[19] J. P. Hausberg, K. Liere-netheler, S. Packmohr, S. Pakura, and K. Vogelsang, “Research streams on digital transformation from a holistic business perspective: a systematic literature review and citation network analysis,” *J. Bus. Econ.*, p. 33, 2019, doi: 10.1007/s11573-019-00956-z.
[20] H. Arksey and L. O’Malley, “Scoping studies: Towards a methodological framework,” *Int. J. Soc. Res. Methodol. Theory Pract.*, vol. 8, no. 1, pp. 19–32, 2005, doi: 10.1080/1364557032000119616.
[21] S. Anderson, P. Allen, S. Peckham, and N. Goodwin, “Asking the right questions: Scoping studies in the commissioning of research on the organisation and delivery of health services,” *Heal. Res. Policy Syst.*, vol. 6, pp. 1–12, 2008, doi: 10.1186/1478-4505-6-7.
[22] S. Lozano, L. Calzada-Infante, B. Adenso-Díaz, and S. García, “Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature,” pp. 1–21, 2019, doi: 10.1007/s11192-019-03132-w.
[23] N. J. van Eck and L. Waltman, “Software survey: VOSviewer, a computer program for bibliometric mapping,” *Scientometrics*, vol. 84, no. 2, pp. 523–538, 2010, doi: 10.1007/s11192-009-0146-3.
[24] T. von Leipzig *et al.*, “Initialising Customer-orientated Digital Transformation in Enterprises,” *Procedia Manuf.*, vol. 8, pp. 517–524, 2017, doi: 10.1016/j.promfg.2017.02.066.
[25] J. Recker, *Scientific research in information systems*, vol. 169. 2013.
[26] U. Al and I. Soydal, “Publication lag and early view effects in information science journals,” *Aslib J. Inf. Manag.*, vol. 69, no. 2, pp. 118–130, 2017, doi: 10.1108/AJIM-12-2016-0200.
[27] A. Ulez’ko, P. Demidov, and A. Tolstykh, “The effects of the digital transformation,” in *International Scientific and practical conference “digitization of agriculture - development strategy” (ISPC 2019)*, 2019, vol. 167, pp. 125–129.
[28] H. Demirkan, J. C. Spohrer, and J. J. Welser, “Digital Innovation and Strategic Transformation,” *IT Prof.*, vol. 18, no. 6, pp. 14–18, 2016, doi: 10.1109/MITP.2016.115.
[29] C. Heavin and D. J. Power, “Challenges for digital transformation–towards a conceptual decision support guide for managers,” *J. Decis. Syst.*, vol. 27, no. May, pp. 38–45, 2018, doi: 10.1080/12460125.2018.1468697.
[30] D. A. Skog, H. Wimelius, and J. Sandberg, “Digital Disruption,” *Bus. Inf. Syst. Eng.*, vol. 60, no. 5, pp. 431–437, 2018, doi: 10.1007/s12599-018-0550-4.
[31] Kane, “Digital Maturity , Not Digital Transformation,” *MIT Sloan Management Review*, 2017. http://sloanreview.mit.edu/article/digital-maturity-not-digital-transformation/.

24th International Conference on Business Information Systems (BIS 2021)

Applications

https://doi.org/10.52825/bis.v1i.50

Published: 02 July 2021

Innovating in Circles

A Qualitative Analysis on Cycles of IT Feature Recombinations for Performative and Creative Outcomes

Katharina Ebner[1], Geneviève Bassellier[2], and Stefan Smolnik[1]

[1] University of Hagen, Germany

[2] McGill University, Montreal, Canada

Abstract. Innovations do not emerge in isolation but are at least to some extent recombinations of previously existing building blocks. In this paper, we will build on the recombination processes feature set broadening and deepening to show how individuals innovate with IT. In our understanding, the out-comes of innovative use can be performative (improving existing task performance) or creative (leading to new deliverables). We build on a longitudinal case of stresstracking initially designed to improve meditation, but ultimately increasing work productivity by using the meditation tool in an innovative way. Using a theoretically grounded analysis framework, we were able to derive eight propositions on the attainment of performative and creative outcome of innovative IT use. We postulate that innovation only occurs through repeating cycles of recombination processes. Particularly, we propose that it is instrumental to run through a phase that does not benefit any task-related outcomes to trigger true creative outcomes.

Keywords: IT post-adoption, IT innovation, IT features, recombination, learning cycles

Introduction

The attainment of innovations is a complex, non-linear process [1]. There is a common understanding among scholars that innovations do not emerge in isolation but are at least to some extent resulting from the recombinations of previously existing building blocks [2], [3], [4]. By recombination, we refer to a process that brings existing knowledge, known system features, etc. together in a novel way thus allowing new application scenarios and new knowledge to emerge [4]. Innovation by recombination manifests in two fundamental processes – the acquisition of new knowledge and the transformation of existing knowledge – which have in an information technology (IT) context been conceptualized as scanning and evaluating [5], knowledge acquisition and conversion [6], sensing and experimentation [7], or, when referring to specific features, building blocks or knowledge assets, as deepening and broadening [2], [7], [8]. In this paper, we will build on these two recombination processes, feature set broadening and feature set deepening, to show how individuals innovate with IT. We employ a broad conceptualization of innovative IT use in that we define it as covering all types of IT use aimed at finding new uses on an existing IT. These new uses involve users that optimize task-performance of existing tasks [9], [10] as well as uses that result in novel deliverables, e.g. new processes, new products, [11], [12], [13]. As a result, the outcomes of innovative use can be performative (i.e. improving existing task performance) or creative (i.e. leading to new deliverables, may it be tasks, processes, or products) with performative outcomes being visible rather in the short-term, and creative outcomes in the longer term. Looking at the long-term outcomes of IT use, such as individual innovativeness, has also been the recent subject of call for future research [14]. Following this call for research, we seek to an-

swer the research question: Which patterns of recombination processes occur during individuals' innovating with IT, and how these processes impact different outcomes of innovative IT use?

To address our research objectives, besides recombination research, we build on and extend research on adopting and innovating with IT. We see our work rooted in research on IT use as feature adaptation, [8], [15] [16], [17], [18]. We thus build on the concept of deep structure IT usage introduced by [19] and define IT use as "an individual user's employment of one or more features of a system to perform a task" (p. 231), where a task "is a goal-directed activity performed by a user" (p. 231). In the same vein, we will investigate innovative IT use with a lens on the features of IT and link them to IT-enabled outcomes, i.e. IT use is comprised of a system (represented by its features), a user (represented by the learning associated with feature set broadening and deepening) and a task (addressed by feature set). Given this theoretical embedding, our research is rooted in the post-adoptive stage of IT use where users already adopted a new technology and have been using it for some time. We employ a longitudinal qualitative research setting using a rich case of a self-tracker, who constantly changed his use of a stress tracking device from simple meditation to, eventually, a creative use configuration allowing him to sense stress at work, address prejudicial work-related behavioral patterns, and increase his work-related performance.

Theoretical background and development

Performative IT use outcomes

Papers investigating the performative aspect of individual innovativeness focus on optimizing work performance, i.e., increasing efficiency and effectiveness in completing existing tasks with the objective to improve organizational performance [6], [9], [10], [20]. Following [21] and [22], efficiency is the level of goal attainment for a given level of input and effectiveness is the degree to which the task goals are met. Accordingly, we use the following definitions for efficiency and effectiveness in this paper: *Efficiency is the level of goal-attainment of a task a user performs with an IT for a given level of input (e.g. time, effort). Effectiveness is the degree to which the goals of a task a user performs with an IT are met.*

Hence, an improvement in efficiency relates to performing a task by using less features or using the features with less effort or time or thinking while keeping the task output at the same level. By contrast, effectiveness improves the task output by making it more complete or more correct (e.g. through using different features). IT use associated with an improvement of effectiveness requires a deeper understanding of an IT's features and a higher level of experience. Effectiveness is, therefore, often conceptualized to be preceded by efficiency and related use types [23], [24].

Creative IT use outcomes

Papers investigating the creative aspect of individual innovativeness focus on the emergence of new deliverables by using an existing IT. These new deliverables may involve new tasks, goals, or practices [11], [15], [6]. On a higher level, such new deliverables may not just impact how an IT is used as tool, they may also impact associated roles, processes, and procedures involving an IT's use [23]. Hence, individual innovativeness may have an impact on task performance and eventually lead to "true" innovation in that an organization has gained insights on e.g. how to offer new services to customers [2], [25]. From an individual perspective, it is important to clarify the conditions under which the use of IT can be considered as "novel" [13]. Management scholars frequently measure innovativeness by considering how novel an innovation is to a specific team or individual [25], [26], and purposefully abstract from a larger network perspective. Indeed, even though the use of an IT in a specific way is innovative for a given individual, from a broader perspective, it may be that the individual is only closing up to the knowledge level of colleagues. At the individual level, it remains an innovative use. Following this argumentation, we define the creative outcomes of individual

innovativeness as follows: *Creative outcomes of innovativeness refer to new tasks that can be solved using an IT and that were previously unknown to an individual.*

Feature set broadening and deepening

System features are the functional building blocks of an information system [18], [27]. They reflect the "specific types of rules and resources, or capabilities, offered by the system" [28] and result from both the design process and individual decisions about use [27]. System features may be grouped, or (re-)combined in feature set [8], [18], [27]. Building on the literature above, we define a feature set as a group of features that an individual has associated together and assigned to one task or a group of tasks (see Figures 1 and 2). Accordingly, feature sets are the result of individual cognition processes and may involve all features available in an information system or only parts of them. The configuration of these feature sets depends on available knowledge and cognitive absorption [19], [28] and is, thus, constantly changing and adjusted by the user. Feature sets have been used by various authors who build on the system-user-task structure introduced by [19]. For instance, [8] investigates how a change in the configuration of a user's mental feature subsets related to specific task groups may lead to distal and mid-term task-performance outcomes. Furthermore, [15] describe how reconfigurations of task-related feature set may help gaining higher levels of innovativeness. The idea to employ feature set reconfigurations to explain creative outcomes of IT use can also be found at [6]. Researchers have also investigated how the feature sets of a user change as she either extend her use behavior, or reduces and resists using features of an IT [10], [15], [27]. For example, [18] describes "a user's revisions of which and how system features are used" (p. 453). He distinguishes between four behaviors (trying new features, feature substituting, feature combining, and feature repurposing). Similarly, [15] reflect novel ways of employing IT features that involve either using a formerly unused set of available features, using IT features for additional tasks, or using feature extensions. [10] and [17] have also discussed the concept of feature extensions. In sum, prior research shows that the changes to a user's feature sets operate through the recombination processes feature set broadening and deepening.

As users broaden their feature set, they acquire knowledge on new features and, hence, extend the scope and variety of IT features they can apply for task completion [27]. However, simply using more features alone is not sufficient, since they may be used in an unproductive way [8]. Feature set broadening is the process of "actively extending the basket of IT features that may be used by a particular user to accomplish tasks" [8]. We have summarized conceptualizations of feature set broadening in Table 1. All of the conceptualizations build on the role of features and their relationships with existing and new tasks. The broadening process, hence, tightly links to the learning of *new features*.

Table 1. Conceptualizations of feature set broadening.

[27]	IT sensemaking, during which users consciously include or exclude features into their task solving.
[8]	Obtaining a broad grasp of a system's functionality while actively extending the basket of IT features that may be used to accomplish tasks.
[10], [16]	Learning and using more of the functions available in the IT.
[18]	Extension/adjustment of the content of the features in use to cope with changing environment.
[29]	Learn how to use entirely new IT features.
[30]	Extent to which an individual uses the various features of the IS system in question.

Two different manifestations of feature set broadening exist [15], [17], [18], with each having a different task-related impact (Figure 1). The knowledge acquisition may lead to an extension of the task-related feature set, or to an extension of the features in use without any task association. Accordingly, in the first case the user has more task-solving options and may thus come up with better, situation-tailored solutions due to higher flexibility. In the case of an

extension of the features in use without any task association, the user better grasps the IT's features and capabilities, and may thus come up with new application scenarios of the IT.

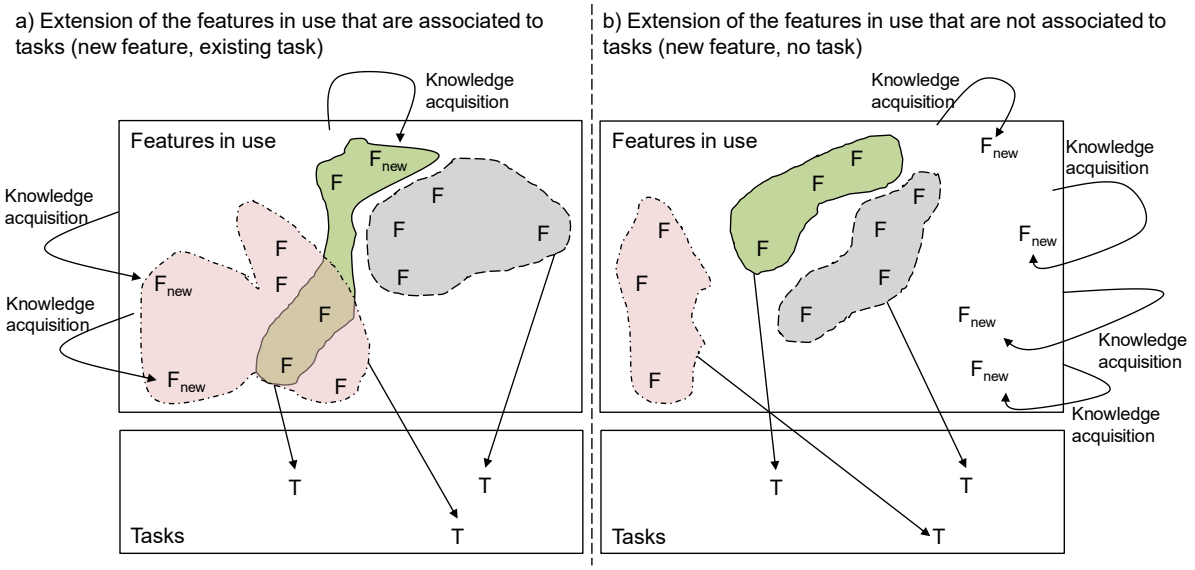


Figure 1. Configurations of feature set broadening.

Feature set deepening involves the transformation of task-solving-related knowledge (i.e. the creation and modification of "how-to knowledge") and requires the user to connect different features to current and potential tasks [6]. The higher the feature set depth, the better will be the solutions produced to solve tasks. We have summarized different conceptualizations of feature set deepening in Table 2. The table shows that the deepening process is often linked to individual abilities in mastering an IT. Hence, it involves a transformation of *existing and new feature set* by recombining, substituting, or removing included features. Technically, an increase in feature set depth has also been linked to the number of feature combinations [18], [28], the number and length of sequences of features [29], or the number of feature set (i.e. the number of tasks that can be solved) [8], [28].

Table 2. Conceptualizations of feature set deepening.

[28]	Different combinations of features to find a solution for solving a task.
[10]	Understanding of how to use IT features and how these features complement other features.
[18]	Feature combining or repurposing as an adaptation of how the features are used, separately or together, in an existing or new way.
[29]	Learn how to apply known IT features in entirely new sequences and contexts.
[8]	Increase the mastery of already known features and functionalities. Fully grasp the features' affordances, effects, and their associations with already-known IT features.
[19]	Extent to which features in the system that relate to the core aspects of the task are used.

Two manifestations of feature set deepening exist (Figure 2). The deepening may lead to a modification of existing task-related feature sets. In this regard, feature sets may be recombined or adjusted by repurposing, adding or removing features [18], [31]. The know-ledge transformation may also lead to a compilation of new feature sets for new tasks [8], [32]. In sum, a theoretical postulate in this paper is that the only way to include *new features* into feature set is by feature set broadening, and the only process by which *new tasks* can be addressed is by feature set deepening. Using this understanding as research framework, we will next show how feature set broadening and deepening lead to innovative outcomes in a real-world IT adoption case.

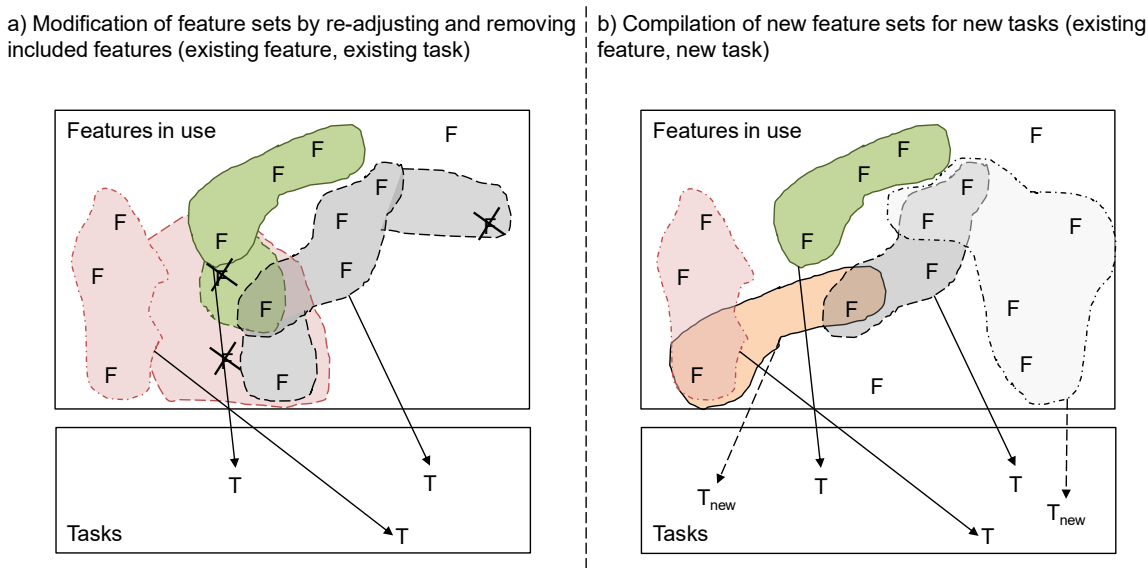


Figure 2. Configurations of feature set deepening.

Research method

We decided for a qualitative research setup [33], [34], [35] to capture the multifaceted phenomena related to innovative IT use. We chose a case of an innovative self-tracker, John, to which we had access via the professional network of one of the co-authors and who, in a private effort to improve his meditation, found a way to become more productive at work using the same IS that he used for meditation. Table 3 presents a brief profile of the case.

Table 3. Case profile.

Context	Stress tracking to increase work productivity
Origin	John works as a professional data analyst and, as a hobby, organizes a large group of self-trackers from the USA
IT	emWave2, a tool for tracking and visualizing heart rate variability
# of feature set recombinations	Feature set broadening: 4 Feature set deepening: 3
Affected tasks	Stress reduction, investigating the occurrences of stress
Impacted outcomes	Performative and creative outcomes of innovative IT use

We employed multiple data sources for our analysis. We reverted to three different detailed presentations on John's stress tracking he gave on his experiences between 2012 and 2015. Contrasting John's different presentations as part of the data analysis allowed us not only to construct his course of adoption longitudinally. We were also able to triangulate our findings by analyzing John's varying emphasis of certain aspects of his adoption over time. This way, we could differentiate routine use from innovation processes, construct a logical chain of evidence, and rule out rival explanations [35], [36]. To further enrich and triangulate our database, we systematically browsed and analyzed John's tweets between January 2011 (beginning of stress tracking) and May 2016. We further analyzed two interviews he gave to self-tracking magazines, browsed and collected relevant posts from all his forum activities in the self-tracking forum of his group. We also compiled a case diary and detailed case write-up.

For our data analysis, we first developed a detailed code list, in which we operationalized our major concepts using established definitions and operationalizations [8], [18], [19], [27]. We came up with altogether twenty codes representing the different feature set recombinations, the deep structure usage, outcomes, features, and tasks. One of the co-authors selectively coded the collected empirical materials [35] using the qualitative data analysis software Atlas.ti (version 8). Another co-author performed a detailed review and revision of the coding. Different interpretations were discussed by these two co-authors and resolved by rechecking the different data sources to increase coding reliability and reduce coding bias. Ad-

ditionally, one further co-author was asked to provide input regarding the coding of randomly chosen excerpts or those whose interpretation was difficult, thus, helping to exclude rival explanations [34]. The pattern coding was performed jointly by the same two co-authors [35]. The goal of this step was to detect re- and co-occurring patterns of feature set modifications and their relationship to performative and creative outcomes during the IT use process. To that end, the initial codes were grouped together into synthesizing categories [33]. High-level code categories and the relationships between them were organized with help of network displays [35], which also fostered the building of logical chains of evidence.

Case description

Having started with tracking his sleep in 2005, John has gone on to engage with tracking stress. He tracks his stress levels using emWave2, a tool for tracking and visualizing heart rate variability (HRV). John has experimented with hooking up his HRV tool to his computer while working and installed software to alert him when it detects periods of stress. John understood how particular behaviors acted as triggers and how he could manage those triggers in order to reduce and defuse stress, and ultimately, through a number of learning cycles, increase his work productivity by using the HRV tool in an innovative way. John bought his HRV tool initially for meditation to relax. The way the HRV tool is intended to be used is to sit down every day twice for at least 10 minutes each time and meditate by putting a thumb on a sensor that senses the heartbeat. If the heartbeat is mild and regular, the tool will play a sound and show a green light. If the heartbeat is moderately regular, the sound will be different and a blue light is shown. If the heartbeat is hectically, thus indicating restlessness and stress, no sound is played and a red light is shown. Beyond using a thumb on the sensor, the user can also employ an ear clip to track the heartbeat. The tool also shows an oscillating light moving in the speed of the heartbeat. There is an app available that can support meditation by visualizing the heart rate, logging and scheduling the meditation times, configuring the sensitivity of three meditation stages (red, blue, green) and configuring the sounds.

John started using the tool with the thumb sensor and by listening to the sounds, trying to adjust his breathing this way. After a while, he noticed the ear clip device as reported during his first presentation in 2012: “And so I found out there is this ear clip accessory and so you put that on attached to the device and you use it with hands free.” Meditating this way allowed him to improve his meditation by putting the HRV tool on a chair and focus on the oscillating light of the tool: “So, in a way it was actually kind of cheating the meditation, I was trying to abandon stalls by looking at the device and trying to stay in green – but I felt a lot better afterwards [he laughs].” Reflecting on his initial use, John subsumed “So, I tried doing that [way of meditation], but it didn’t work for me. The reason why is that during [meditation], I feel okay, but then I would get stressed later in the day… And I also found out that any kind of habit that requires me to carve out time, even if it’s just 10 minutes, would be very hard to adopt for me.” Despite these downsides, John kept meditating irregularly using the tool: “Using emWave, I became a lot better at meditating. […] So, actually spending some time with the device helped me – it just ain’t doing it [the stress reducing] quick enough for me to really appreciate it.” Realizing that the meditation would only moderately help with his stress problem, John started exploring the app available for the tool. He was interested in the “challenge levels” – an app feature that could be used to configure the sensitivity of the tool to determine a green, blue, or red state. By setting the challenge level to hard (i.e. for a green state an extremely constant heartrate was required) he tried to become more balanced and hoped the state would last longer. But he struggled to reach the green state then and reset it to the previous setting. He kept exploring the app and came across a feature to configure the sounds, for which he had immediately no idea about which task he could meaningfully use it for: not only the sound itself could be changed, it was also possible to activate or deactivate the red, blue, or green sounds. John remembered: “And then [after a while] I was thinking: well if the whole point of this is to be in green as much as possible and to reduce the amount of time that you are actually in red during the day, I need some sort of alert system, some sort of waver to tell me when I’m feeling stressed and through some creative configuration, I was

actually able to setup such a system using emWave." This "creative configuration" involved using the tool with the ear clip all day during work and tracking his HRV as indicator of stress constantly: "I turned off the tones for the green and the blue and just left on the red one. So, that's how it alerted me when I was in stress and the protocol that whenever I would here that tone, I would stop what I was doing, turn to it, and do, yeah, deep breathing until I got into a blue state or into that green state and then I go back to work."

Tracking his stress during the day eventually helped John recognizing some productivity affecting negative behaviors: "I started detecting behavioral patterns I wasn't aware of before. So, I checked the news a lot during the day I work. And I noticed that whenever I opened up Google News, the alert would go off. Well, it wasn't the news that was stressing me out. What it was: the stress was actually a trigger and checking the news was the behavior to kind of get away from the stress […]. As you can imagine that cycle would repeat many times throughout the day." As a result, John installed a website blocker to block such stress reduction sites, and instead took a deep breathe. This was not just less time-consuming, it also helped reducing the stress feeling more effectively than checking news websites: "I am not going to make some statement like 'I am not feeling stressed anymore.' I still do. I mean, there are still some patterns …, but I am making progress. And I think it's because now I can better understand where the problem is and have a good way of marking that progress."

Results

Our data analysis reveals three findings. First, successful feature set broadening and deepening alternate constantly in cycles. Second, a cycle is always initiated by a broadening. Third, different configurations of feature set broadening and deepening impact different outcomes, and to come up with creative outcomes, it is instrumental to pass through a broadening that does not immediately relate to a task. We summarized the broadening and deepening phases in Table 4 and illustrate the development of stress reduction in Figure 3.

Table 4. Summary of case analysis: cycles of broadening and deepening.

Feature set modification	Description	Impacted outcomes
#0 Original use	Meditate using the thumb sensor, with closed eyes, sitting on the floor, infrequently check oscillating light to see heartrate Features included in feature set: thumb sensor, sounds, lights	Meditation and stress reduction considered inefficient and ineffective.
#1 Feature set broadening (new feature, existing task)	*Include* ear clip into used features for meditation Features included in feature set: *ear clip*, thumb sensor, sounds, lights Task: stress reduction	Having hands free and having not to constantly push thumb on sensor led to higher level of meditation, thus increasing the effectiveness of meditation.
#2 Feature set deepening (existing feature, existing task)	*Replace* thumb sensor constantly by ear clip, place HRV tool on chair, focus on oscillating light Features included in feature set: ear clip, sounds, lights Task: stress reduction	Focus on oscillating light got John quicker into green state, thus increasing efficiency of meditation. He was also able to stay a lot longer in green state, thus indicating increased effectiveness.
#3 Feature set broadening (new feature, existing task)	*Explore app* offered for the tool to learn challenge levels Features included in feature set: ear clip, sounds, lights, *app/challenge levels* Task: stress reduction	Unsuccessful, no impact on outcomes – challenge levels were too difficult. Reset to previous feature set configuration.
#4 Feature set broadening (new feature, no task)	*Discover* sound configuration option in tool Features included in feature set: ear clip, sounds, lights, *app/sound configuration* Task: no task at that point	The discovering of the sound configuration left John wondering what it would be good for, but no immediate impact happened.

#5 Feature set deepening (existing feature, new task)	Understand that the sound configuration could be used to *configure a stress alert system* Features included in *new feature set*: ear clip, sounds, app/sound configuration Task: *investigating the occurrences of stress*	No impact on efficiency and effectiveness of meditation or work, but creative outcome.
#6 Feature set broadening (new feature, existing task)	Install a website blocker Features included in feature set: ear clip, sounds, app/sound configuration, *website blocker* Task: stress reduction	Since the website blocker prevented prejudicial behavior, he was able to stay more focused on tasks, thus increasing effectiveness of reducing stress.
#7 Feature set deepening (existing feature, existing task)	Do not just block websites. To *better overcome stress* overreaction, perform conscious breathing during work (around 1-3 minutes) Features included in feature set: ear clip, sounds, app/sound configuration, website blocker Task: stress reduction	The breathing helped in reducing stress states quicker, thus reducing stress levels more efficiently, and – since he actively and sustainably reduces his stress level – also was more effectively in reducing stress.

As can be seen from Figure 3, with four out of seven feature set modifications John was able to improve his stress level. In essence, the stress level at the end of iteration 7 was much less during and the end of the day than in iteration 0. However, only after iteration 5, when John started realizing his actual stress curve, he was able in his future feature set modifications to improve the effectiveness and efficiency of stress reduction. The relative development of the change in stress reduction effectiveness and efficiency is displayed in Figure 4. As can be seen from this figure, there is a more frequent increase in effectiveness than in efficiency. There were also periods of stagnation, which was terminated by a "no task – new task" pattern. Furthermore, there is the same increase pattern in effectiveness and efficiency at the beginning and the end of the observed feature set modifications. Reflecting these observations jointly with the summary presented in Table 4 indicates that successful feature set recombinations always alternate. The only exception from this reoccurring pattern was #3, which was reset due to unsuccessful recombination. Furthermore, the starting point to a new cycle is always the broadening. We derive:

P1a: Feature set recombination will always alternate between feature set broadening and deepening.

P1b: A feature set recombination cycle will always start with a broadening.

P1c: Unsuccessful recombination processes are reset, and a new cycle begins regardless of the last recombination process.

Backing for our propositions comes from recent literature on IT use. For instance, [37] show that all forms of IT use are the result of two different types of cognition processes that interact in cyclical patterns but leave the details of associated use or outcomes open for future research. Several studies have documented a nonlinear relationship between experience and creativity or innovation: increased experience contributes to creativity and innovation up to a certain point, with diminishing returns at high levels of experience [38], [39], [40]. For instance, [3] indicate that this complex relationship might result from the actual context of routines and practices in which creativity occurs, since some routines seem more or less favorable to creativity or innovation. However, the actual relationship between these concepts and innovativeness remains mostly unexplored. As a result, the generation of creative outcomes of innovativeness has consistently been pointed out as the more complex, more uncertain and hazardously type of innovativeness, but also the most promising in terms of larger impacts such as inventions, new products, or processes [3], [4], [6], [14], [22].

Our data analysis reveals an effect that may help shedding further light on the generation of creative outcomes, namely the relevance of a feature set broadening *without* a clear task association (#4), i.e., an individual explores a feature for which no immediate use is evident (to that individual). Our case indicates that such a feature set broadening leaves an individual

in a state of curiosity and increased cognitive involvement – an important trigger for subsequent creative action [2], [6].

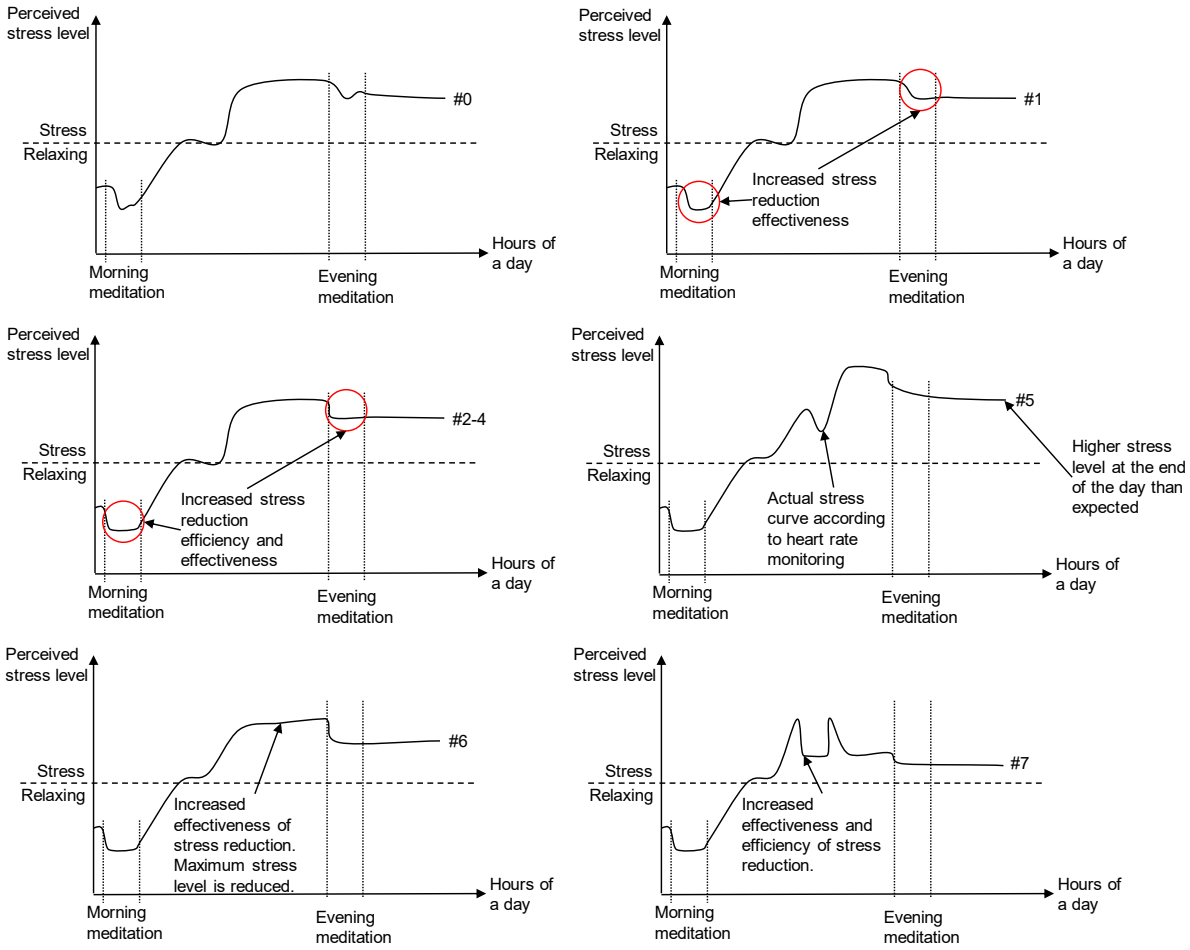


Figure 3. Development of John's stress level during the seven feature set modifications

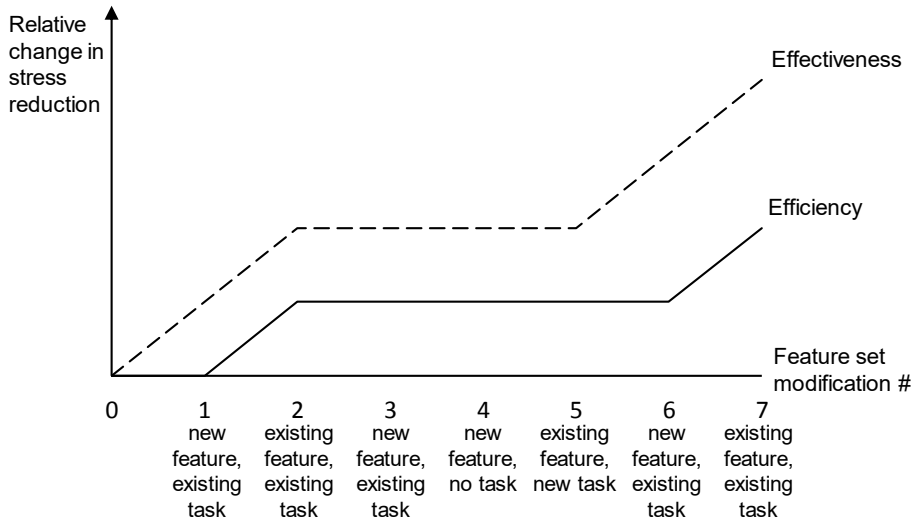


Figure 4. Relative development of change in stress reduction effectiveness and efficiency

To sum up, a strong interrelation between the recombination processes associated to no task and a new task is evident. This situation must not be confused with an unsuccessful recombination process (see P1c). While in the case of an unsuccessful recombination, the affected feature set is reset, in the no-task configuration, the discovered feature remains in the features in use, indicating that an individual is convinced that there might be value of a feature, which that person has just not uncovered. Thus:

P2a: The attainment of creative outcomes of innovative IT use operates via a feature set broadening with no task-related outcomes.

P2b: Only when the feature set broadening does not lead to later reset of the features in use, a subsequent feature set deepening related to new tasks is probable.

From the perspective of performative outcomes of innovative use, we also detected different impacts on task efficiency and effectiveness from the different recombination processes. In the two occurrences of a feature set broadening associated with new features and an existing task (#1, #6), we detected an impact on effectiveness. While previous literature frequently postulated a more general impact of feature set broadening on performance, [8], [10], [19], a few authors have also speculated on that this type of broadening might be rather related to effectiveness [32]. Furthermore, the impact of this feature set broadening on effectiveness seems lower than that of a feature set deepening (#2, #7) related to existing tasks. With respect to task-related efficiency, we were, in our case, only able to detect an impact for feature set deepening related to existing tasks. Evidence for this finding comes from [24], who show that knowledge transformation processes such as feature set deepening will also "streamline" the feature sets and, thus, improve task-related efficiency. We derive:

P3a: Feature set broadening involving new features and only existing tasks has a positive impact on task-related effectiveness.

P3b: Feature set deepening related to existing tasks has a higher positive impact on task-related effectiveness than a feature set broadening related to existing tasks.

P3c: Feature set deepening related to existing tasks has a positive impact on task-related efficiency.

Discussion and Conclusions

Given that one, conveniently sampled case, regardless of its richness and longitudinal character, can only provide limited insights on such a complex phenomenon such as innovating with IT, we see the work presented here as a starting point for future investigations. Case research is, per se, never generalizable. The purpose of case research is to shed light on relevant and complex phenomena that cannot easily be investigated using large samples, but require scrutiny and a high level of granularity to extract the relevant mechanisms. Our case started in a specific private setting. However, the way the person used the IT was completely different (work productivity), and, to some degree, unpredictable. We hold that this situation is prototypical for most innovative use scenarios. Beyond that, we believe that our unique longitudinal research setup, the way we analyze the case, and the way we extract the relevant mechanisms of innovative use allow for the derivation of conclusions that go beyond our case. First, we demonstrate that findings from existing organizational research on IT use also apply to our case. Second, we offer an analytical explanation of the attainment of innovative outcomes. In particular, we show that innovation only occurs through repeating cycles of recombination processes, and propose that it is instrumental to run through a phase that does not benefit any task-related outcomes to trigger true creative outcomes. This insight sheds completely new light on the process of innovating with IT and contributes to the recent and lively discussions in IT use research, on why it is so challenging to thoroughly explain innovative IT use. Our insights are thus relevant both for researchers that investigate innovation with IT, but also for practitioners who want to stimulate more innovative behavior with their employees. In presenting our findings, we hope to stimulate further discussions on the nature of the processes underlying individual innovative IT use. We are confident that our work contributes to both recombination research and research on innovative IT use by shedding light on the nonlinearity of innovating with IT. In particular, we showed that innovating with IT operates in constant cycles of feature set broadening and deepening, with broadening preceding the deepening. By linking feature set broadening and deepening to existing tasks as well as to new deliverables, we clarify the relationships and transitions between different configurations of innovative use and show which patterns of innovative use occur over time.

Consequently, we can also extend recombination research by showing that the fundamental processes of recombination, broadening and deepening, occur in cycles and show that processes related to individual innovation behavior follow specific patterns. Furthermore, we shed light on the relevance of individuals on innovativeness – a facet, typically not considered in recombination research. On a higher abstraction level, we highlight the role of learning and feature set modifications as the key mechanisms to the attainment of outcomes. Further, we show that users cannot directly move from efficiency or effectiveness-optimizing behavior to innovative behavior, but that at least one exploration cycle that leads to no immediate outcomes needs to be successfully passed. Future researchers may draw on this insight and rather consider how this insight reflects concrete types of post-adoption use and their mutual relationships. Our paper is also of interest to IT practitioners who wish to better understand why the users in their organization arrive at dedicated outcomes. We shed light on the relevance of learning processes and show which behaviors may have to be incentivized to end up with specific task-related results. In addition, the feature set typology and understanding we present provides guidance to set up training programs (e.g. by not only introducing features, but explicitly linking them to multiple tasks and inspiring creative uses) and helps better structuring communication with users in problem situations.

References

1. L. Argote, E. Miron-Spektor, “Organizational Learning: From Experience to Knowledge.” Org Sci, vol.22, no.5, pp. 1123–1137, 2011, https://doi.org/10.1287/orsc.1100.0621
2. C. M. Flath, S. Friesike, M. Wirth, F. Thiesse, “Copy, transform, combine: Exploring the remix as a form of innovation.” JIT, vol.32, no.4, pp. 306–325, 2017, https://doi.org/10.1057/s41265-017-0043-9
3. L. Fleming, S. Mingo, D. Chen, “Collaborative Brokerage, Generative Creativity, and Creative Success.” Admin Sci Quart, vol.52, no.3, pp. 443–475, 2007, https://doi.org/10.2189/asqu.52.3.443
4. M. Gruber, D. Harhoff, K. Hoisl, “Knowledge Recombination Across Technological Boundaries: Scientists vs. Engineers.” Mgmt Sci, vol.59, no.4, pp. 837–851, 2013, https://doi.org/10.1287/mnsc.1120.1572
5. A. Majchrzak, L. P. Cooper, O. E. Neece, “Knowledge Reuse for Innovation.” Mgmt Sci, vol.50, no.2, pp. 174–188, 2004, https://doi.org/10.1287/mnsc.1030.0116
6. S. Nambisan, R. Agarwal, M. Tanniru, “Organizational mechanisms for enhancing user innovation in information technology.” MISQ, vol.23, no.3, pp. 365–395, 1999, https://doi.org/10.2307/249468
7. J. Carlo, K. Lyytinen, G. Rose, “A Knowledge-Based Model of Radical Innovation in Small Software Firms.” MISQ, vol.36, no.3, pp. 865–895, 2012.
8. A. Benlian “IT feature use over time and its impact on individual task performance.” JAIS, vol.16, no.3, pp. 144–173, 2015, 10.17705/1jais.00391
9. M. K. Ahuja, and J. B. Thatcher, “Moving beyond intentions and toward the theory of trying: Effects of work environment and gender on post-adoption information technology use.” MISQ, vol.29, no.3, pp. 427–459, 2005.
10. J. Jasperson, P. E. Carter, R. Zmud, “A Comprehensive Conceptualization of Post-Adoptive Behaviors Associated with Information Technology Enabled Work Systems.” MISQ, vol.29, no.3, pp. 525–557, 2005, https://doi.org/10.2307/25148694
11. R. Agarwal, “Individual acceptance of information technologies.” In Framing the domains of IT management research: Glimpsing the future through the past (Zmud, R. W. Ed.), pp. 85–104, Cincinnati, Ohio: Pinnaflex, 2000.
12. S. Nambisan, “Information Technology and Product/Service Innovation: A Brief Assessment and Some Suggestions for Future Research.” JAIS, vol.14, no.4, 2013, https://doi.org/10.17705/1jais.00327

13. Y. Rahrovani, A. Pinsonneault, “User’s Perceived IS Slack Resources and their Effects on Innovating with IT.” ICIS 2014, Auckland, New Zealand, 2014, https://doi.org/10.1109/HICSS.2015.500.
14. A. Ortiz de Guinea, J. Webster, “An investigation of information systems use patterns: Technological events as triggers, the effect of time, and consequences for performance.” MISQ, vol.37, no.4, pp. 1165–1188, 2013, https://doi.org/10.25300/misq/2013/37.4.08
15. F. F. Bagayogo, L. Lapointe, G. Bassellier, “Enhanced use of IT: A new perspective on post-adoption.” JAIS, vol.15, no.7, pp. 361–387, 2014, 10.17705/1jais.00367
16. J. P. A. Hsieh, A. Rai, S. X. Xu, “Extracting business value from IT: A sensemaking perspective of post-adoptive use.” Mgmt Sci, vol.57, no.11, pp. 2018–2039, 2011, https://doi.org/10.1287/mnsc.1110.1398
17. J. P. A. Hsieh, R. Zmud, “Understanding Post-Adoptive Usage Behaviors: A Two-Dimensional View.” Comp Info Sys Fac Publ, 2006, http://aisel.aisnet.org/digit2006/3
18. H. Sun “Understanding user revisions when using information system features: Adaptive system use and triggers.” MISQ, vol.36, no.2, pp. 453–478, 2012, https://doi.org/10.2307/41703463
19. A. Burton-Jones, A., D. W. Straub, “Reconceptualizing system usage: An approach and empirical test.” ISR, vol.17, no.3, pp. 228–246, 2006, https://doi.org/10.1287/isre.1120.0444
20. X. Li, J. P. A. Hsieh, A. Rai, “Motivational Differences Across Post-Acceptance Information System Usage Behaviors: An Investigation in the Business Intelligence Systems Context.” ISR, vol.24, no.3, pp. 659–682, 2013, https://doi.org/10.1287/isre.1120.0456
21. D. J. Beal, R. R. Cohen, M. J. Burke, C. L. McLendon, “Cohesion and performance in groups: A meta-analytic clarification of construct relations.” J Appl Psy, vol.88, no.6, pp. 989–1004, 2003, https://doi.org/10.1037/0021-9010.88.6.989
22. A. Burton-Jones, A., C. Grange, “From Use to Effective Use: A Representation Theory Perspective.” ISR, vol.24, no.3, pp. 632–658, 2013, https://doi.org/10.1287/isre.1120.0444
23. K. S. Lassila, J. C. Brancheau, “Adoption and utilization of commercial software packages: Exploring utilization equilibria, transitions, triggers, and tracks.” JMIS, vol.16, no.2, pp. 63–90, 1999, https://doi.org/10.1080/07421222.1999.11518246
24. S. Sundaram, A. Schwarz, E. Jones, W. Chin, “Technology use on the front line: How information technology enhances individual performance.” J Acad Mark Sci, vol.35, no.1, pp. 101–112, 2007. https://doi.org/10.1007/s11747-006-0010-4
25. L. P. Robert, T. A. Sykes, “Extending the Concept of Control Beliefs: Integrating the Role of Advice Networks.” ISR, vol.28, no.1, pp. 84–96, 2017, https://doi.org/10.1287/isre.2016.0666
26. C. K. W. de Dreu, M. A. West, “Minority dissent and team innovation: The importance of participation in decision making.” J Appl Psy, vol.86, no.6, pp. 1191–1201, 2001, https://doi.org/10.1037/0021-9010.86.6.1191
27. T. L. Griffith, “Technology Features as Triggers for Sensemaking.” Acad Mgmt Rev, vol.24, no.3, pp. p. 472, 1999, https://doi.org/10.5465/amr.1999.2202132
28. G. DeSanctis, M. S. Poole, “Capturing the complexity in advanced technology use: Adaptive structuration theory.” Org Sci, vol.5, no.2, pp. 121–147, 1994, https://doi.org/10.1287/orsc.5.2.121
29. A. Ortiz de Guinea, M. L. Markus, “Why break the habit of a lifetime? Rethinking the roles of intention, habit, and emotion in continuing information technology use.” MISQ, vol.33, no.3, pp. 433–444, 2009, https://doi.org/10.2307/20650303
30. M. Limayem, M., S. G. Hirt, C. M. K. Cheung, “How habit limits the predictive power of intention.” MISQ, vol.31, no.4, pp. 705–737, 2007, https://doi.org/10.2307/25148817
31. H. Barki, G. Paré, C. Sicotte, “Linking IT implementation and acceptance via the construct of psychological ownership of information technology.” JIT, vol.23, no.4, pp. 269–280, 2008, https://doi.org/10.1057/jit.2008.12

32. K. A. Saeed, S. Abdinnour, “Understanding post-adoption IS usage stages: An empirical assessment of self-service IS.” ISJ, vol.23, no.3, pp. 219-244, 2013, https://doi.org/10.1111/j.1365-2575.2011.00389.x
33. I. Benbasat, D. K. Goldstein, M. Mead, “The Case Research Strategy in Studies of Information Systems.” MISQ, vol.11, no.3, pp. p. 369, 1987, https://doi.org/10.2307/248684
34. L. Dubé, G. Paré, “Rigor in Information Systems Positivist Case Research: Current Practices, Trends, and Recommendations.” MISQ, vol.27, no.4, p. 597, 2003.
35. R. K. Yin, Case study research. 4th ed. Thousand Oaks, CA: Sage, 2009.
36. M. B. Miles, A. M. Huberman, j. Saldana, Qualitative Data Analysis. 3rd ed. Thousand Oaks, CA: Sage, 2015.
37. T. W. Ferratt, J. Prasad, E. Dunne, “Fast and Slow Processes Underlying Theories of Information Technology Use.” JAIS, vol.19, no.1, 2018.
38. P. G. Audia, J. A. Goncalo, “Past Success and Creativity over Time: A Study of Inventors in the Hard Disk Drive Industry.” Mgmt Sci, 53, no.1, pp. 1–15, 2007, https://doi.org/10.1287/mnsc.1060.0593
39. G. Hirst, D. van Knippenberg, J. Zhou, “A Cross-Level Perspective on Employee Creativity: Goal Orientation, Team Learning Behavior, and Individual Creativity.” AMJ, vol.52, no.2, pp. 280–293, 2009, https://doi.org/10.5465/AMJ.2009.37308035
40. R. Katila, R., G. Ahuja, “Something old, something new: a longitudinal study of search behavior and new product introduction.” AMJ, vol.45, no.6, pp. 1183–1194, 2002., https://doi.org/10.5465/3069433

24th International Conference on Business Information Systems (BIS 2021)

Applications

https://doi.org/10.52825/bis.v1i.52

Published: 02 July 2021

An integrated group decision-making approach considering uncertainty conditions

Daniela Borissova[1,2][https://orcid.org/0000-0002-4457-3703], Zornitsa Dimitrova[1][https://orcid.org/0000-0003-2688-9840]

[1] Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences, Sofia, Bulgaria

[2] Laboratory of Telematics at the Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract. The management of business information processes needs effective decision-making models. That means to involve different methods, techniques, and principles to improve competitiveness and to achieve the planned business results. In this context, the article deals with the problem of group decision-making under uncertain conditions. To cope with such problems some well-known optimization strategies of Wald, Laplace, Hurwitz, and Savage are modified to take into account the experts' opinions with different importance when forming the final group decision. Numerical testing is based on a case study for CRM software selection. The results are discussed based on the proposed models under two different cases derived from the case study. The conducted numerical testing of the proposed models demonstrates their applicability to cope simultaneously with multiple experts' evaluations and uncertainty conditions.

Keywords: Group decision-making, uncertainty conditions, cost-benefit estimation, Wald's, Laplace's, Hurwitz's, Savage's criteria

Introduction

Standing dynamics in the global world is the premise with which managers face every day and need to make decisions to solve different problems. The COVID-19 pandemic forced companies to move quickly to shift to remote working and to build new business models to reflect the new technology capabilities and users' needs. This could be done with the important role of chief information officer [1]. Now, the companies need to determine how to keep and to improve the new business models considering the changing customer demands and ongoing economic uncertainty. To cope with the complex decision-making process, it is necessary to involve multi-disciplinary teams with different qualification area [2]. Involving experts with different expertise will contribute to the credibility of the group decision-making considering different points of view [3]. On the other hand, considering the problems in real life and human judgment are in most cases unclear and cannot be represented by fixed values. Therefore, it is necessary to use different approaches to overcome such conditions. The Bellman-Zadeh approach with some of the well-known optimization strategies proposed by Wald, Laplace, Hurwitz, and Savage could be used to cope with problems under information uncertainty [4]. A systematic review of expert judgement for dependence in probabilistic modelling is presented [5].

The management of operational processes require proper modelling of planned analytical applications to support business in specifics tasks processing [6]. To be successful every company from the large organizations to family businesses needs well-planned information strategy [7]. Therefore, many different approaches are proposed to cope with different practical problems such as effective management in streetlight modernization [8], decision making in publishing sector [9] and selection of supplier under public procurement

[10], for doing business [11] and business management [12], different real-life applications with time series forecasting [13], etc. The development of business information systems should be based on innovative models and computational intelligence methods. In addition, the development of intelligent software needs to meet the principles of agile software development [14]. Depending on particular domain area different specifics and some common business processes could be recognized from portfolio risk optimisation to IoT systems [15, 16]. In order for a better understanding of the business process maturity models, a systematic literature review is proposed [17]. This is also in line with the policy of the European Data Strategy aiming at the development of an attractive, secure and dynamic data economy. All of these should be an underling of the principles of findability, accessibility, interoperability, and reusability. These principles could be applied also to other digital objects, e.g. algorithms, tools, and workflows, that led to that data, as all these elements must be available to ensure transparency, reproducibility and reusability [18, 19].

In the context of digital transformation, a new perspective of digital entrepreneurship driven by the concepts of digital transformation and entrepreneurship is proposed. The authors identify several significant relationships between three categories of digital transformation as technology readiness (ICT investments), digital technology exploration (research and development) and digital technology exploitation (patents and trademarks) that are essential key for business management [20]. Here should be noted also the important role of government to enhance digital transformation in a small service business [21]. The authors show interesting links between business models and business performance. They found correlations between innovation and sustainability, revealing that digital transformation tools contribute over the long-term to the value creation process [22]. Management of business process relies on different methods, techniques and principles aiming to achieve higher business results and competitiveness. The proper methodology for market structures analysis with possibilities for ranking could contribute to business process improvement too [23, 24]. The success of both business process management and digital innovation could be done by the active role of chief information officers and information technology executives [25].

Business process management requires the use of optimization in order to improve operational efficiency [26]. In addition, in the digital age, businesses need to be not only flexible but also responsive to market conditions. Therefore, it is necessary to provide new levels of optimization of business processes. However, this cannot be achieved only through classical planning techniques and it is necessary to develop new approaches that cover various aspects of the business environment, including the conditions of uncertainty in strategic planning. That is why contemporary business process systems need new methods for thinking about intelligent processes in ways that build on the concepts of collaboration, addictiveness, and awareness [27].

The main paradigm of decision-making is the subjective human factor that cannot be avoided as the decisions are made by humans. Therefore, any attempt to overcome such subjectivism is to be encouraged to reduce bigger and wrong decisions with subsequent accidents [28, 29]. The usage of group decision-making aims to reduce the subjectivism by involving more experts from different domain area. To get the group decision, it needs to involve proper mathematical models with the ability to aggregate individual preferences of DMs into the final group decision. This means that some quantitative evaluations of the alternatives are to be done concerning predefined criteria. Besides these evaluation criteria need to be further evaluated regarding their importance by each expert.

Integrated Group Decision Making Approach for Evaluation and Choice under Uncertainty Conditions

During the decision-making process, a set of different alternatives with very different likely consequences must be analysed, i.e. presence of uncertainty in decision making. When making decisions in conditions of uncertainty, it is assumed that the DM has an idea of the

goals to be achieved, but the information about the alternatives and future events is incomplete. In the decision-making process, the preferences of DM can be represented by a utility function $f(a)$ over a set of alternatives $A = \{a_1, a_2, \dots, a_m\}$ in different states $S = \{s_1, s_2, \dots, s_n\}$.

Algorithm for Evaluation Considering Different Criteria to Cope with Uncertainty Conditions

The proposed decision-making algorithm for evaluation considering different criteria to cope with uncertainty conditions is shown in Fig. 1.

Figure 1. Algorithm for group decision-making under uncertainty conditions.

The first stage for the implementation of group decision-making in the conditions of uncertainty is related to the description of the specific problem for choosing an alternative. Stage 2 is related to identifying the acceptable alternatives appropriate to the problem at hand. At Stage 3, the possible states of the environment in which the identified alternatives are implemented are determined. Depending on the specifics of the identified problem, a group of experts capable of evaluating the alternatives is selected on stage 4. Once the group of experts are determined, the respective weighting coefficients are to be determined in accordance with the expertise and the degree of importance of each DM in the group. Stage 6 is related to the process of alternatives assessment toward identified conditions and to the point of view of each expert in the group. To cope with uncertainty at stage 7, one of the well-known Wald's, Laplace's, Hurwitz's or Savage's criteria is to be chosen to determine the optimal strategy. Next, a corresponding optimization task for choosing an alternative in the conditions of group decision making is to be formulated and solved on stage 8. On the last stage 9, the most preferable group alternative is determined.

To realize the group decision-making under uncertainty conditions it is needed to formulate mathematical models able to aggregate experts' evaluation with respect to the selected criterion to overcome uncertainty.

Mathematical Models for Group Decision-Making Considering Uncertainty

As a quantitative analytical tool, revenue and expenditure analysis can be used to support decision making [30]. In this case, a utility function is formulated based on the estimation of expected revenues and expenses. In the general case, the utility function is:

$$CBE = \frac{Costs}{Benefits} \tag{1}$$

The overall $Costs$ could be expressed as a sum of different aspects from purchasing to installation and maintenance. The $Benefits$ can be expressed by revenue related to some activities that contribute to the reducing labour costs, automated processes, etc.

Using the generalized utility function (1) and the mention above criteria to cope with uncertainty, four modified group decision-making models are formulated as follows:

- Group decision-making model based on Wald's criterion (strategy):

$$maximin \sum_{i=1}^{M} \sum_{k=3}^{K} \lambda^k CBE_{ij}^k \tag{2}$$

subject to

$$\sum_{k=1}^{K} \lambda^k = 1 \tag{3}$$

where CBE_{ij}^k expresses evaluation of the i-th alternative toward j-state from k-expert, λ^k represent the importance of particular expert by corresponding weighted coefficient for the expertise.

Wald's principle is based on a pessimistic view and conservative moderation. It consists of the fact that for any of the chosen strategies, the objective circumstances will always represent the most unfavourable situation. For each alternative solution, the worst outcome can be determined.

- Group decision-making model based on Laplace's criterion (strategy):

$$max \left(\sum_{i=1}^{M} \sum_{k=1}^{K} \frac{\lambda^k CBE_{ij}^k}{M} \right) \tag{4}$$

subject to relation (3).

According to the Laplace's criterion, when is no other additional information all possible states are considered equally probable.

- Group decision-making model based on Hurwitz's criterion (strategy):

$$max\{\alpha max \sum_{i=1}^{M} \sum_{k=1}^{K} \lambda^k CBE_{ij}^k + (1-\alpha) min \sum_{i=1}^{M} \sum_{k=1}^{K} \lambda^k CBE_{ij}^k\} \tag{5}$$

subject to relation (3).

where α is the coefficient of optimism ($0 < \alpha < 1$).

The value of the coefficient $\alpha = 0$ corresponds to environment considered to be completely antagonistic, $\alpha = 0.5$ corresponds to an equivalent environment (neither antagonistic nor friendly) and at a value of $\alpha = 1$, the medium is the most favorable. In essence, the Hurwitz's criterion is a simplified version of the Laplace principle, namely – with certain probabilities of the individual states, the arithmetic mean of the results of the best decisions is taken.

- Group decision-making model based on Savage's criterion (strategy):

$$minimax \sum_{i=1}^{M} \sum_{k=1}^{K} \lambda^k R_{ij}^k \tag{6}$$

subject to relation (3) and one more relation

$$\forall i = 1,2, \dots M: (\forall k = 1,2, \dots M: R_{ij}^k = \left| CBE_{ij}^k - maxCBE_{ij}^k \right|) \tag{7}$$

where R_{ij}^{k} represent the regret as a result of opportunity loss if A_i is chosen and state S_j happens in accordance to k-th expert point of view.

It is possible to use different ranges for the coefficients that express the importance of DMs opinions according to the particular problem, background and expertise, for example between 0 and 10 or range between 1 and 100. In this situation, a proper normalization has to be done to get compatibility between dimensionless weighted coefficients (λ^k) and the CBE.

Numerical Testing

Numerical testing of the problem of group decision-making in conditions of uncertainty is made for a specific example of choosing specialized software. Among a variety of specialized systems for customer relationship management (CRM) software, a subset of suitable alternatives with similar characteristics are identified. Due to the dynamism of the economic environment, the companies' prospects can be generally represented by three possible situations: increasing revenues, reducing revenues, or maintaining the current state. Therefore, the subject of the decisions is in the conditions of uncertainty, as the possible situations are known, but it is not known which of them will come true and there is no information about the probabilities for the realization of these situations. The selection of the most appropriate software system could be made through costs benefit analysis as a tool to determine whether the investment decision is good taking into account the mention above possible situations. For the goal, the following representation of costs-benefit relation is formulated:

$$CBE = \frac{Costs}{Benefit} = \frac{C_{ac}+C_{cust}+C_{inst}+C_{test}+C_{staff}+C_{proc}+C_{file}+C_{un}+C_{lc}}{R_{lcr}+R_{stock}+R_{lc}+R_{ap}} \quad (8)$$

The overall $Costs$ are expressed as a sum of the costs of product acquisition (C_{ac}), customization (C_{cust}), installation (C_{inst}), post-installation testing (C_{test}), staff training (C_{staff}), file conversion (C_{file}), uninstallation of the old system (C_{un}), and loyalty policy (C_{lc}). Revenue is related to labor costs reducing (R_{lcr}), reducing the need to maintain the stock (R_{stock}), improved reliability through a new policy for loyal customers (R_{lc}), the use of more automated processes (R_{ap}).

Thus, the use of cost-benefit analysis makes it possible to determine the value of CBE as a result of the realization of certain situations.

For the purpose of a small company, it is needed need to make a decision to implement new CRM software. The higher management together with the chief information officer determined 3 suitable alternatives among those the selection should be done. In addition, three experts are also determined to conduct the evaluation. These experts are as follows: an expert in the database (E-2), one end-user of CRM (E-1), one expert from the IT support team (E-3).

The CBEs of the three CRM software systems (alternatives) are done by using the above relation and taking into account the possible three situations for increase, decrease, and maintain the current state of the company as shown in Table 1.

Table 1. Modified matrix for group decision making under the conditions of uncertainty.

Experts (DMs)	Weighted coefficient for DMs' expertise	Alternatives	Conditions		
			Increase	Reduction	Unchanged
E-1	0.25	A-1	0.52	0.83	0.75
		A-2	0.53	0.86	0.72
		A-3	0.54	0.84	0.71
E-2	0.35	A-1	0.56	0.87	0.78
		A-2	0.55	0.80	0.74
		A-3	0.58	0.81	0.72
E-3	0.40	A-1	0.62	0.85	0.72
		A-2	0.60	0.88	0.76
		A-3	0.61	0.82	0.70

The first two columns of Table 1 contain the weighted coefficients about the importance of each expert (DM) from the formed group.

Using the data from Table 1 and defined group decision-making models based on the Wald, Laplace, and Hurwitz criteria, corresponding tasks are formulated and solved. Savage's criterion requires drawing up a regret matrix. The values of the elements of this matrix represent the losses due to missed opportunities.

The calculation is performed according to the point of view of the DMs from the group according to the following formula:

$$R_{ij}^{k} = \left| CBE_{ij}^{k} - maxCBE_{ij}^{k} \right| \tag{9}$$

The values of regret resulting from lost profits are shown in Table 2.

Table 2. Regret matrix as result of opportunity loss.

Experts (DMs)	Alternatives	Regret as result of opportunity loss		
E-1	A-1	0.02	0.03	0
	A-2	0.01	0	0.03
	A-3	0	0.02	0.04
E-2	A-1	0.02	0	0
	A-2	0.03	0.07	0.04
	A-3	0	0.06	0.06
E-3	A-1	0	0.03	0.04
	A-2	0.02	0	0
	A-3	0.01	0.06	0.06

The described above problem is numerically tested for two cases that express different combination between DMs opinions importance as shown in Table 3.

Table 3. Two cases for the expertise of group members.

Cases	Weighted coefficient for the experts' expertise		
	λ^1 for E-1	λ^2 for E-2	λ^3 for E-3
Case-1	0.25	0.35	0.40
Case-2	0.34	0.46	0.20

The Case-1 illustrates the scenario where as the most important is taken on the expert E-1 with weight of 0.4, closely followed by expert E-2 with a weight of 0.35 and less important of

the expert E-1 by a weight of 0.25. The Case-2 represents the experts' importance in the following order E-2, followed by E-1 and finally E-3 (see Table 3).

Results and Discussion

The formulated four models for group decision-making are based on criteria behind a certain strategy for dealing with uncertainty. Therefore, each of the models can be considered as a model of the respective specific strategy. The results obtained from solving the respective optimization tasks based on the proposed models for group decision-making, taking into account two scenarios for the importance of DM's opinions (Table 3) are illustrated in Fig. 2.

When the Wald's criterion is used, the obtained solutions identified the alternative A-3 as the most preferred in both cases of weights for the experts in the group. Choosing this strategy, the worst possible consequences of each strategy are considered, choosing the least bad or the best of them, according to the point of view of the experts in the group. The choice of this strategy determines the worst possible consequences of each condition and choosing the least bad or the best of them, according to the point of view of the experts in the group.

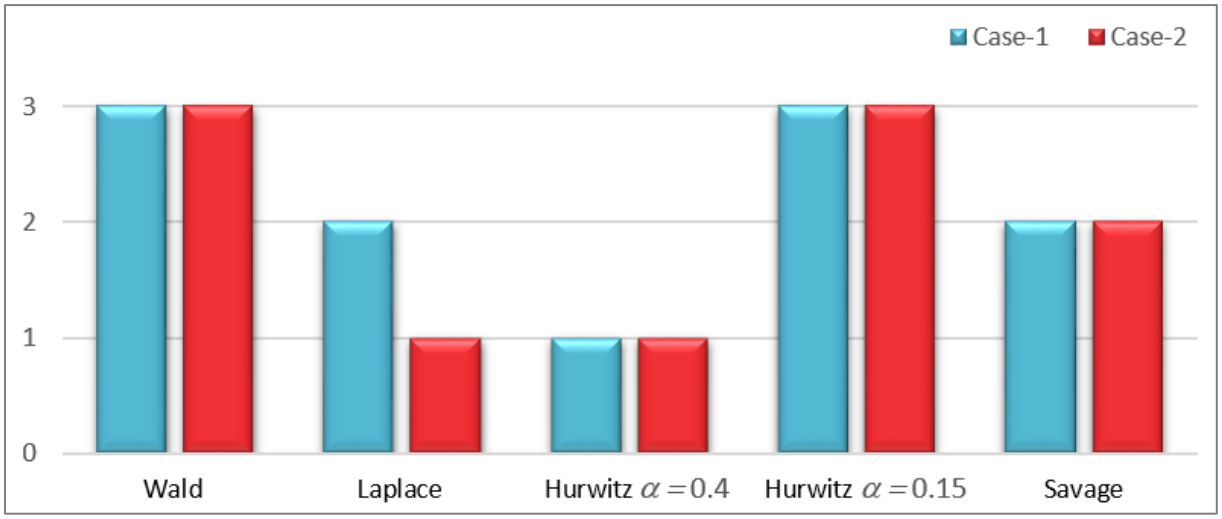


Figure 2. Selected alternative when using different criteria to overcome the uncertainty conditions and different weights for DMs' importance.

According to Laplace's criterion, it is assumed that all objective conditions have the same probability of occurrence due to lack of other grounds. The result is obtained by calculating the arithmetic mean for each strategy, according to the views of the experts in the group and the determined weighting factors, choosing the strategy with the highest score. When using the first combination of weights for experts by Case-1 ($\lambda^1 = 0.25$; $\lambda^2 = 0.35$; $\lambda^3 = 0.40$) the solution determines as the most preferred alternative A-2, while in the second combination of weights noted asCase-2 ($\lambda^1 = 0.34$; $\lambda^2 = 0.46$; $\lambda^3 = 0.20$) the most preferred alternative is A-1.

Unlike Wald's and Laplace's criteria, Hurwitz's criterion requires the determination of the so-called coefficient of optimism, respectively the coefficient of pessimism, which are applied to each strategy, and the choice is determined by the highest result obtained. Two cases corresponding to two different values for the coefficient of optimism ($\alpha = 0.40$ and $\alpha = 0.15$) is considered. When using both combinations of weights for experts, for the specific example, and a value of the coefficient of optimism $\alpha = 0.40$, the choice for the most preferred is the alternative A-1, and when using $\alpha = 0.15$, the solution identifies alternative A-3 as the most preferred alternative.

The use of Savage's criterion requires the compilation of a regret matrix as a result of lost profits, choosing this alternative whose maximum losses are minimal. In the particular

numerical experiment for both uses combinations (Case-1 and Case-2) of weights for the opinions of experts, the most preferred alternative is A-2.

According to the Wild and Hurwitz ($\alpha = 0.15$) strategies and particular input data, the final group decision is to chosen alternative A-3 (Fig. 2) in both cases for DMs' opinions importance. If the strategy behind the Savage principle is chosen, the final group decision is to select the alternative A-2 for both cases for DMs' opinions importance, but according to the strategy behind Laplace principle, the importance of DMs' opinions determines alternative A-2 for Case-1 and alternative A-1 for Case-2 (Fig. 2). The alternative A-1 is also preferable group decision in accordance to the Hurwitz strategy and coefficient of optimism $\alpha = 0.40$.

As a result of the analysis, it is found that the use of different strategies according to the principles of Wald, Laplace, Hurwitz, Savage, and in combination with the experts' point of views (Case-1 and Case-2), lead to a different selection of group alternative. It is therefore important to choose in advance the most appropriate decision-making strategy in the face of uncertainty. In summary, it can be concluded that the results of numerical testing of the proposed models for group decision-making in conditions of uncertainty show the applicability of the described modifications based on the optimization criteria of Wald, Laplace, Hurwitz, and Savage.

Conclusions

The article deals with problems related to group decision-making under uncertain conditions. Such problems are in the focus of many different organizations from SME, including small family businesses, universities, and other non-profit organizations. Due to nowadays dynamic in different aspects of business processes, organizations are facing a variety of decision-making problems with different degrees of uncertain conditions. For the goal, the well-known criteria able to cope with uncertainty like Wald, Laplace, Hurwitz, and Savage are modified to be able to integrate the points of view of experts with different importance when forming final group decision. This is realized by introducing weighted coefficients assigned to each expert. The advantage of such an approach is the possibility to take into account the experts' opinions accordingly to their expertise, background, and closeness to the particular decision-making problem. A drawback of the proposed algorithm could be considered the subjective factor of DM who determines the weighted coefficients of the experts within the group.

The major contributions of the article are related to modified optimization strategies of Wald, Laplace, Hurwitz, and Savage to consider experts' opinions with different importance. These opinions are based on the usage of the calculated value for costs benefits ratio for each particular problem. The components that make up the costs and benefits in case of implementation of a software system are identified, involved in determining the value of the cost-benefit ratio. The proposed approach is applied in the selection of CRM software. It is shown that the final group decision depends not only on the used strategy according to the principles of Wald, Laplace, Hurwitz, and Savage, but is influenced also by the introduced weighed coefficients expressing the experts' opinion importance when aggregating the final group decision.

Future developments concern the development of different models for more precise and objective estimations of the weights for experts' opinions in an aggregation of a final group decision. Another perspective direction is related to the use of a different scale for the alternatives estimation that differs from the used cost-benefit estimation ratio.

References

1. V. Shalamanov, V. Sabinski, T. Georgiev, “Optimization of the chief information officer function in large organizations”, *Information & Security*, vol. 46, no. 1, pp. 13-26, 2020, doi: https://doi.org/10.11610/isij.4601.
2. D. Borissova, Z. Dimitrova, V. Dimitrov, “How to support teams to be remote and productive: Group decision-making for distance collaboration software tools”, *Information & Security*, vol. 46, no. 1, pp. 36-52, 2020, doi: https://doi.org/10.11610/isij.4603.
3. D. Borissova, “A group decision making model considering experts competency: An application in personnel selections”, *Comptes rendus de l’Academie Bulgare des Sciences*, vol. 71, no. 11, pp. 1520-1527, 2018.
4. P. Ekel, J.S.C. Martini, R.M. Palhares, “Multicriteria analysis in decision making under information uncertainty”, *APPL MATH COMPUT*, vol. 200, no. 2, pp. 501-516, 2008, doi: https://doi.org/10.1016/j.amc.2007.11.024.
5. C. Werner, T. Bedford, R.M. Cooke, A.M. Hanea, O. Morales-Napoles, “Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions,” *EUR J OPER RES*, vol. 258, no. 3, pp. 801-819, 2017, doi: https://doi.org/10.1016/j.ejor.2016.10.018.
6. C. Hrach, R. Alt, “Configuration approach for analytical service models – development and evaluation”, in: *2020 IEEE 22nd CONF BUS INFORM*, Antwerp, Belgium, 2020, pp. 260-269. https://doi.org/10.1109/CBI49978.2020.00035.
7. M. Kamariotou, F. Kitsios, “Information Systems Planning and Success in SMEs: Strategizing for IS”, in: *BIS 2019*, LNBIP, vol. 353, 2019, pp. 397-406, https://doi.org/10.1007/978-3-030-20485-3
8. D. Borissova, P. Cvetkova, I. Garvanov, M. Garvanova, “A framework of business intelligence system for decision making in efficiency management”, in: *CISIM’2020*, LNCS, vol. 12133, 2020, pp. 111-121. https://doi.org/10.1007/978-3-030-47679-3_10.
9. D. Borissova, N. Keremedchieva, D. Keremedchiev, “Business intelligence approach to support decision making in publishing sector”, *MIPRO*, pp. 1532-1537, 2020, doi: https://doi.org/10.23919/MIPRO48935.2020.9245424.
10. D. Korsemov, D. Borissova, I. Mustakerov, “Group decision making for selection of supplier under public procurement”, in: *ICT Innovations 2018*, COMM COM INF SC, vol. 940, 2018, pp. 51-58. https://doi.org/10.1007/978-3-030-00825-3_5.
11. D. Borissova, D. Korsemov, I. Mustakerov, “Multi-criteria decision making problem for doing business: Comparison between approaches of individual and group decision making”, in: *CISIM’2019*, LNCS, vol. 11703, 2019, pp. 385-396. https://doi.org/10.1007/978-3-030-28957-7_32.
12. I. Stankov, G. Tsochev, “Vulnerability and protection of business management systems: Threats and challenges. *Problems of Engineering Cybernetics and Robotics*, vol. 72, pp. 29-40, 2020, https://doi.org/10.7546/PECR.72.20.04.
13. R. Ketipov, G. Kostadinov, P. Petrov, I. Zankinski, T. Balabanov, “Genetic algorithm based formula generation for curve fitting in time series forecasting implemented as mobile distributed computing”, in: *ADV HIGH PERF COM 2019*, STUD COMP INTELL, vol. 902, 2021, pp. 40-47. https://doi.org/10.1007/978-3-030-55347-0_4.
14. M. Perkusich, L. Chaves e Silva, A. Costa, F. Ramos, R. Saraiva, A. Freire, E. Dilorenzo, E. Dantas, D. Santos, K. Gorgonio, H. Almeida, A. Perkusich, “Intelligent software engineering in the context of agile software development: A systematic literature review”, *INFORM SOFTWARE TECH*, vol. 119, no. 106241, 2020, doi: https://doi.org/10.1016/j.infsof.2019.106241.
15. K. Stoyanova, V. Guliashki, “Two-stage portfolio risk optimisation based on MVO model”, *International Journal of Reasoning-based Intelligent Systems*, vol. 12, no. 1, pp. 70-79, 2020, doi: https://dx.doi.org/10.1504/IJRIS.2020.105011.
16. A. Vodyaho, R. Yoshinov, N. Zhukova, A.M. Thaw, A. Saddam Ahmed, “Fog oriented model for data collection in the networks of mobile devices”, in: *2020 IEEE 10th INT*

CONF INTELL SYST, Varna, Bulgaria, 2020, pp. 421-425. https://doi.org/10.1109/IS48319.2020.9200138.
17. A. Tarhan, O. Turetken, H.A. Reijers, "Business process maturity models: A systematic literature review", *INFORM SOFTWARE TECH,* vol. 75, pp. 122-134, 2016, doi: https://doi.org/10.1016/j.infsof.2016.01.010.
18. L. Liu, W. Li, N.R. Aljohani, M.D. Lytras, S.-Ul Hassan, R. Nawaz, "A framework to evaluate the interoperability of information systems – Measuring the maturity of the business process alignment", *INT J INFORM MANAGE*, vol. 54, no. 102153, 2020, doi: https://doi.org/10.1016/j.ijinfomgt.2020.102153.
19. A.-L. Lamprecht, et al., "Towards FAIR principles for research software," *Data Science*, vol. 3, no.1, pp. 37-59, 2020, doi: https://doi.org/10.3233/DS-190026.
20. V. Jafari-Sadeghi, A. Garcia-Perez, E. Candelo, J. Couturier, "Exploring the impact of digital transformation on technology entrepreneurship and technological market expansion: The role of technology readiness, exploration and exploitation", *J BUS RES*, vol. 124, pp. 100-111, 2021, doi: https://doi.org/10.1016/j.jbusres.2020.11.020.
21. C.-L. Chen, Y.-C. Lin, W.-H. Chen, C.-F. Chao, H. Pandia, "Role of government to enhance digital transformation in small service business", *Sustainability*, vol. 13, no. 3, 2021, doi: https://doi.org/10.3390/su13031028.
22. A. Di Vaio, R. Palladino, A. Pezzi, D.E. Kalisz, "The role of digital innovation in knowledge management systems: A systematic literature review", *J BUS RES*, vol. 123, pp. 220-231, 2021, doi: https://doi.org/10.1016/j.jbusres.2020.09.042.
23. I. Petrov, "Improving the methodology of market structures analysis with innovative concepts for phase-structure states and set concentration index", *Economic Alternatives*, vol. 1, pp. 5-15, 2016.
24. I. Mustakerov, D. Borissova, "Investments attractiveness via combinatorial optimization ranking", *International Journal of Management Science and Engineerin*g vol. 7, no. 10, pp. 230-235, 2013, doi: https://doi.org/10.5281/zenodo.1088218.
25. A. Van Looy, "A quantitative and qualitative study of the link between business process management and digital innovation", *INFORM MANAGE*, vol. 58, no. 2, 103413 (2021). https://doi.org/10.1016/j.im.2020.103413.
26. M. Camargo, M. Dumas, O. Gonzalez-Rojas, "Automated discovery of business process simulation models from event logs", *DECIS SUPPORT SYST*, vol. 134, 113284, 2020, doi: https://doi.org/10.1016/j.dss.2020.113284.
27. H. Kir, N. Erdogan, "A knowledge-intensive adaptive business process management framework", *Information Systems*, vol. 95, 101639, 2021, doi: https://doi.org/10.1016/j.is.2020.101639.
28. F. Klapproth, "Time and decision making in humans", *Cognitive, Affective, & Behavioral Neuroscience*, vol. 8, pp. 509-524, 2008, doi: https://doi.org/10.3758/CABN.8.4.509.
29. R. Moura, C. Morais, E. Patelli, J. Lewis, M. Beer, "Human factors influencing decision-making: tendencies from first-line management decisions and implications to reduce major accidents", in: *ESREL'2017*, Proc. of International Conference on Engineering Sciences and Technologies, 2017, pp. 69-34. doi: 10.1201/9781315210469-34
30. A. Boardman, D. Greenberg, A. Vining, D. Weimer, "Cost benefit analysis: Concepts and practice", *The Pearson Series in Economics*, 4th edn., Pearson, 2010.

24th International Conference on Business Information Systems (BIS 2021)

Applications

https://doi.org/10.52825/bis.v1i.53

Published: 02 July 2021

Interoperability of Health Digitalization:

Case Study on Use of Information Technology for Maternal and Child Health Services in Indonesia

Lutfan Lazuardi[1] [https://orcid.org/0000-0001-5146-8162], Guardian Yoki Sanjaya [2], Pungkas Bahjuri Ali[2][https://orcid.org/0000-0002-0650-0925], Renova Glorya Montesori Siahaan[2] , Lia Achmad[1], Hanifah Wulandari[1]

[1] Universitas Gadjah Mada, Bulaksumur, Yogyakarta 55281, Indonesia

[2] the Ministry of National Development Planning/National Development Planning Agency. Taman Suropati 2 Jakarta Pusat 10310 DKI Jakarta, Indonesia

lutfan.lazuardi@ugm.ac.id

Abstract. Introduction : Introduction : Maternal and child health (MCH) is a global priority as health care innovation continues to evolve, including the use of information and communication technology. Studies showed that interoperable information systems can improve the quality of health services and at the same time facilitate the integration of data for the purpose of monitoring and evaluating the performance of health services, especially MCH. Aims : This study aims to identify various maternal and child health information systems used in Indonesia and opportunity of interoperability between systems to support continuum of care services. Methodology: Qualitative descriptive research was conducted in Yogyakarta Province from November to December 2020. This study assessed MCH applications that have been used in public and private primary health care, hospitals, health offices and in the community by identifying their functions and mapped data elements used by each application to assess potential interoperability between systems. The online focus group discussions with various application providers was conducted to explore the challenges of interoperability between digital systems. Results and Discussion : There were 18 maternal and child health information systems have been developed by the government (central and local), health facilities and private sector. The initiation of interoperability between systems has not yet occurred, except to support regular reporting at the health office and Ministry of Health level. Interoperability between information systems required efforts to improve information technology facilities and infrastructure, development of health data standards, strengthening governance and regulation and utilization of data as an effort to monitor, evaluation and continuity of interoperability between systems to support the digitalization of services and routine reporting. Conclusions and Recommendations : Digitalization of MCH services in Indonesia has the opportunity to support the continuum of care through an interoperable system. However, several enabler factors need to be prepared to support interoperability between information systems.

Keywords: health digitalization, interoperability, maternal and child health.

Introduction

Mother and child health program (MCH) is one of the basic health care programs in Indonesia and a global priority. It is an early phase of the continuum of care where mothers, newborns, and children are inseparably linked in life and in health care needs. Unfortunately, there are differences in MCH outcomes between regions that are caused by the uneven capacity of health care systems such as human resources, facilities and infrastructure,

pharmaceuticals and medical devices, and financing. In addition, the COVID-19 pandemic has also led to a decrease in maternal and child health care reporting, including immunizations, nutrition services [1], as well as reproductive health services. In addition, the regular monthly reports of nutrition programs in the health office are also under utilized for decision making. Such low data utilization is indicated due to difficulties in data accessibility [2].

Utilization of information technology in the health sector can support the availability of accurate and timely data and information. Time-consuming and less efficient manual logging should be switched to digital-based logging that can improve access to the data and better on the managing information system. Digital-based information systems provide ease of access to all stakeholders. However, a number of digital-based health information systems that have been developed need to be integrated to support better data and information management.

Policy innovation is also needed to accelerate the progress of MCH's outcomes throughout the region. With the challenge of disparity in the capacity of health care systems between regions, the regulation on digitalization of services is one solution to overcome them. In addition, the integrated health care information system, user friendliness, and accessibility require strong policy in the implementation of healthcare information systems to be able to provide the latest data and information, accurate, valid, fast, and transparent. Therefore, mapping the condition of health care digitalization, strategies and policies and alternative solutions to solve the problem need to be formulated appropriately.

Research Method

This study used a qualitative descriptive study approach in Yogyakarta Special Region which was conducted from November to December 2020. The method was appropriate to describe more comprehensive phenomena related to the current situation of digitalization of health services [3].

The study consisted of several stages: stakeholders and application identification, online focus group discussions by involving resource persons from different organizations, observation of the application by installing and exploring the features that are displayed, data analysis, and dissemination of preliminary results to reconfirm findings.

Table 1. FGD Participants

Time	**Activities**	**FGD Participants**	
		Health Sectors	**Total participants**
Wednesday, 11 November 2020	FGD with District Health Office and Health Facilities from west and central regions	District Health Office Public Health Care Regional Public Hospital Mother and Child Hospital	6 3 3 1
Thursday, 12 November 2020	FGD with District Health Office and Health Facilities from eastern regions and private sectors	District Health Office Public Health Care Regional Public Hospital Mother and Child Hospital Private sectors	6 3 3 1 3

Result and Discussion

Condition of MCH Digitalization in Indonesia

There were many applications used for data capture and reporting MCH services. We mapped about 18 applications that have been identified based on the results of discussions with various stakeholders with a continuum of care approach (Figure 1).

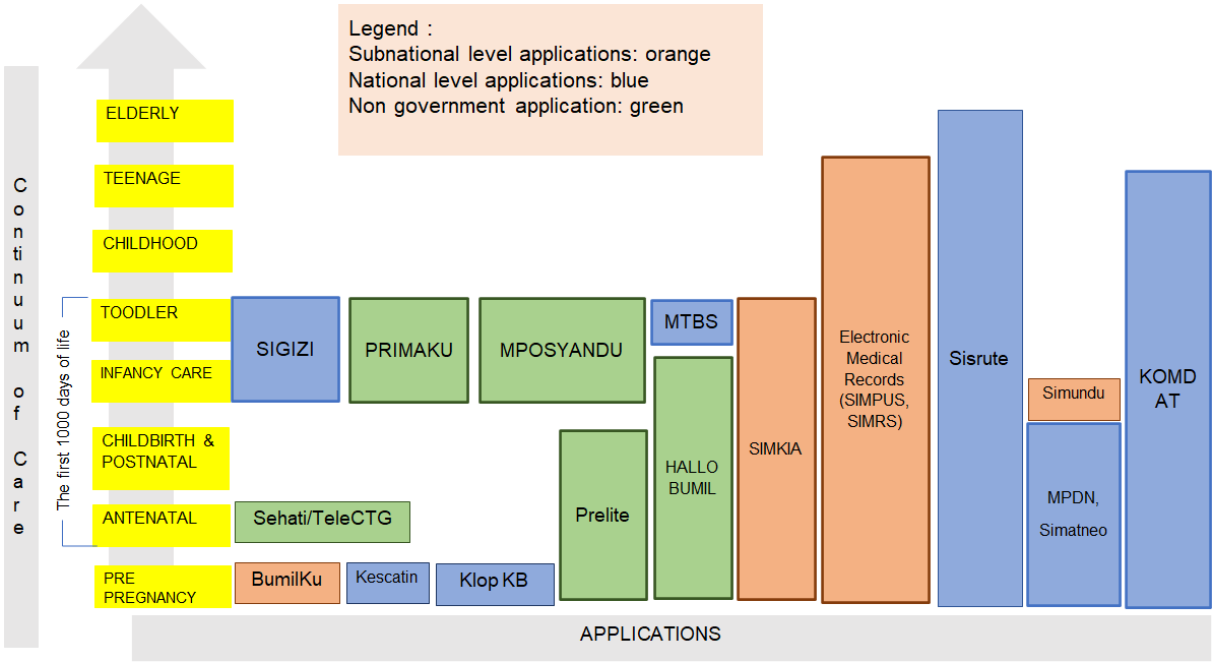


Fig. 1. Mapping of various applications related to MCH with continuum of care approach

The continuum of care approach suited to describe the number of applications that has interlink services from pre-pregnancy stage, antenal, childbirth and postnatal, infancy, toddler, childhood, teenage children, and the elderly. This mapping illustrated that there were already applications that served as a container for data collection in every journey of life. The application has two main purposes, 1). to facilitate health services (digitalization of services) and 2. to enable ease of reporting. The digitization was initiated both at the government (national and sub-national level) and private sector and health facilities. The reporting purpose was mostly initiated by the government, while health services and individual monitoring applications mostly initiated by health facilities and the private sector.

Based on our finding, the digital health intervention for MCH can be divided based on WHO classification of digital health interventions [4] for tracking patient and health information management. In addition, the digital interventions were distinguished by the main users of the system consisting of 1). clients or the community who used health care facilities, health promotion and self monitoring, 2). health workers who conduct data capture from healthcare services they provided, and 3). health data managers who responsible for routine health management information systems which includes data collection, validation, analysis, visualization, data use and dissemination.

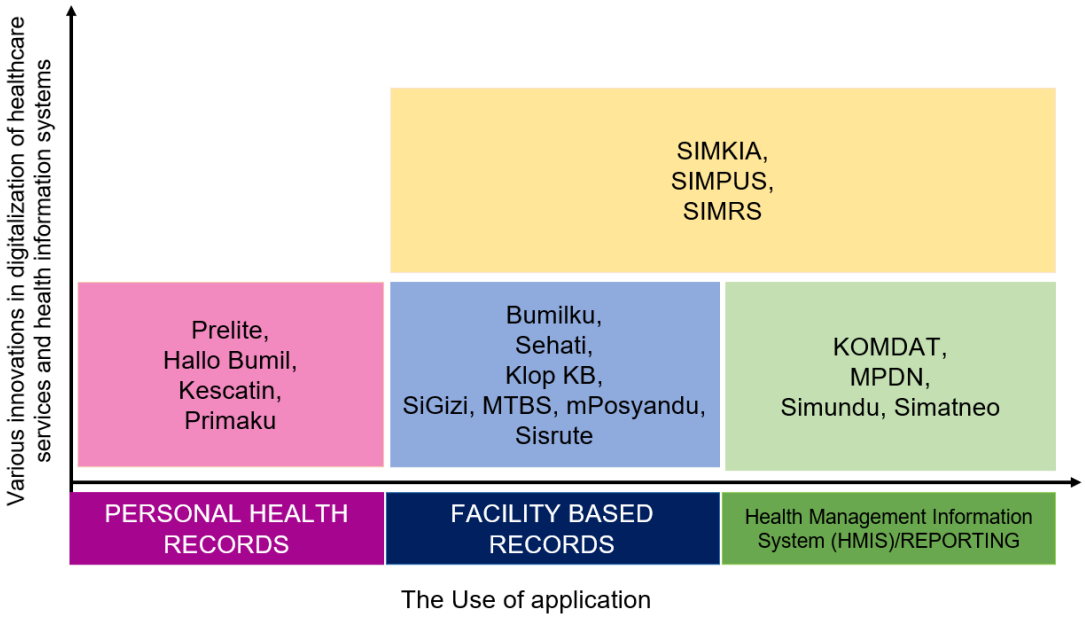


Fig. 2. Classification of Digital Health Intervention in applications related to MCH

MCH-related applications that exist today can be categorized into personal health records used by individuals / communities directly, facility based records filled by officers in health facilities and HMIS for the purposes of recording and reporting. Among these applications, there are applications that have facilitated facility-based service activities as well as for reporting purposes both comprehensively and specifically MCH.

One of the interesting things about mapping is the application has not facilitated the principle of Continuum of Care because each application is still fragmented, so that efforts are needed to make it sustainable between applications. Operational applications for service needs can also be used to fill reporting needs. Constraints due to the number of fragmented applications is the need to enter data repeatedly (double entry) by officers. For example, if there are patients visiting Primary Health Care, when going to the hospital supporting data has not been connected between health care facilities so it needs to be re-entry and may even need to be re-examined. It is certainly inefficient and increases the risk of medical and medication error.

Differences in Application Usage between Regions

In this study, we explored the use of MCH applications in different regional zones in Indonesia by grouping into several categories as in the following table.

Table 2. Comparison of the use of MCH digitalization between different regions in Indonesia [5]

Category	Regions	
	West and Central	**Eastern**
Implementation of MCH digitalization (healthcare service and information system management)	There were already applications for healthcare services and health information systems	Focus on reporting, for services were still few. Still relying on WhatsApp and paper based or not digitized.
Standard interoperability of MCH	Some subnational levels have developed a variety of	- Haven't developed many subnational level

digitalization (national and sub-national level)	applications that overlapped and there was no interoperability with national level applications. In this region the national level application was less used	applications - Already used the application from the national level but not running optimally due to limited infrastructure. Thus, paper based reporting was still used.
IT infrastructure and internet connection	Minimum internet constraints in some areas	Apart from internet access, there were still limited hardware availability

MCH-related applications that exist could be categorized into 1). personal health records used by individuals or communities that in some extent link to health facility, 2). facility based records that are used by healthcare workers in health facilities and 3). routine health management information systems (HMIS) for the purposes of data capture and reporting that are mostly used by data managers in primary health centers and health offices. Among these applications, there were applications that have facilitated both facility-based service and for reporting purposes for the MCH services.

We found that the application had not facilitated the principle of Continuum of Care because each application was still fragmented. There was no interoperability in place to ensure sustainable data sharing between applications. Some of the applications were actually collecting data that can generate necessary routine reports. However these applications did not contribute to the routine reporting. Thus, inefficient data collection appeared due to the number of fragmented applications required repeatedly data capture by health care workers (double entry). For example, if there were patients visiting Primary Health Care, when going to the hospital supporting data had not been connected between health care facilities so it needed to be re-entry and may even need to be re-examined. It was certainly inefficient and increased the risk of medical and medication error.

The difference in application usage between regions was inseparable from the resource capabilities of each region. In areas that have adequate resources to innovate technology utilization would certainly differ the level of application usage compared to areas where the condition of resources was less supportive. The western and central regions already started to develop applications independently to meet the needs of their respective regions related to the digitalization of maternal and child health services. While in the eastern zone there were still many who have not implemented the digitalization of maternal and child services so that in the process of reporting data was still done manually.

Differences between National and Sub National Initiated Applications

The national government in initiating application development generally aimed to facilitate MCH reporting systems, such as Komdat, Simatneo, MPDN, and reporting in general that support MCH such as Sisrute. Some applications aimed to facilitate services such as SiGizi and MTBS. Sub national and private governments in developing applications tend to aim to support services, such as SIMPUS, SIMRS, mPosyandu, Sehati TeleCTG, Klop KB, Primaku, Hallo Bumil, and Prelite.

The National Government played a role in facilitating tiered reporting, initiating applications to be further developed by regions, including facilitating areas that did not yet have a system. While the Sub National Government played a role in continuing the interoperable system with the national level. This indicated the implementation of a system that combines a centralized and decentralized system.

The implementation of the centralized system was developed by the national government and run to all regions under the coordination of the center. While the implementation of decentralized systems was developed by the sub national level to meet the needs of specific health services in the region. Decentralized systems must meet the

standards of interaction with other systems, but allow them to be developed without dependence from other systems. Implementation measures that combine centralization and decentralization efforts can meet the needs of health services in each region as well as needs related to reporting to the center [6].

Differences between Public and Private Initiated Applications

Currently applications developed by private parties mostly aimed to facilitate the needs of individuals or communities directly (personal health records), for example Prelite, Hallo Bumil and Primaku. This was in line with the use of personal applications that began to develop, especially among people who have good health knowledge. Some applications have been intended to facilitate services by health workers or health facilities such as Klop KB, Sehati TeleCTG and mPosyandu. While the application developed by the public more broadly covers personal needs, health facilities to recording and reporting.

The development and implementation of digital health services for the government had the main goal to provide access to fair, affordable and useful services for the entire community [7]. In addition, in developing applications private parties would have a specific purpose either referring to non-profit organizations, the commercial sector that generates profits for their owners, as well as academic institutions involved in the dissemination of knowledge through research, education and training. Collaboration between private parties and the government would have a significant impact on public health while fulfilling the principles and objectives of each private party [8].

MCH Digitalization Resources

To support the digitalization of MCH, efforts were made to fulfill the necessary resources such as facilities and infrastructure in various regions. Indonesia has carried out various infrastructure developments as a manifestation of these efforts, such as the following:

- provided internet access in the underdeveloped, outermost, isolated areas (3T) managed by the Ministry of Communication and Informatics through the Telecommunications and Information Accessibility Agency (BAKTI)
- provided Base Transceiver Station in the blank spot area managed by BAKTI. This effort has reached several underdeveloped, outermost, isolated areas in Indonesia.
- telecommunication network development projects to all districts / cities using Sea Cable Communication System (SKKL) and Fiber Optic Communication System (SKSO). The project was known as Palapa Ring.
- the planned launch of the Satellite of the Republic of Indonesia (SATRIA) in 2022 for the provision of internet

Regulatory Analysis and Technical Guidance

Various applications that have been developed for health care services and health information systems were still fragmented both from health facilities and sub national level so that not thoroughly MCH data in each developer of MCH service applications, health care facilities and regional health agencies delivered to the center. On the other hand, the national government was also developing applications for MCH reporting needs that were poorly adhered to by some areas in its charging because the sub national level has developed applications as needed, applications were less user friendly, infrastructure constraints, and human resources. Based on these conditions, regulations related to the interoperability of digitalization of healthcare services and the integration of health data in the digitalization of health information systems was necessary especially for MCH data.

MCH health services that support the Continuum of Care required collaboration between health facilities by exchanging patient data from one system to another, including personal health records used by patients or patients' families in order to monitor health independently and contribute to medical records. MCH services are a combination of institutional and community-based services that complement each other. While MCH service standards are already available and used evenly in Indonesia (MCH Books), MCH health service

digitalization technical standards that cover various needs of recording and exchanging MCH health service data have not been widely developed. For example, the interoperability of digitalization of MCH services could be [9]:

- Scheduling and appointment of MCH services at health facilities
- Recording of health services ranging from registration, anamnesis, vital signs, pregnancy status, physical examination, medical support examination, medical action, treatment, birth, neonatal registration, outcome and medical resume
- Clinical decision support system
- Public access (patients and families) to MCH service data through web portals or personal health records that are widely developed by the private sector.
- Codification standard or unique ID for various actors of MCH health services such as patient identity, codification of health facilities and codification of health workers.
- Standard codification and dictionary of maternal and child health care data that includes codification and standard terminology for vital signs, anthropometry, medical symptoms and signs, pregnancy status, diagnosis, medical support examination, medication, medical procedures, clinical outcomes and others.
- The protocol and format of electronic data exchange, for example, adopts the HL7 FHIR standard which can be xml and JSON documents [10].

The above three aspects must be managed professionally either by specialized institutions or consortiums, documented, accessible and developed continuously in accordance with the necessary changes. In Indonesia there were no institutions that formally and had a concrete responsibility to develop health data standards, develop protocols and formats of electronic data exchange and manage interoperability infrastructure between digital systems of healthcare services.

In the concept of OpenHIE interoperability (Open Health Information Exchange), in addition to standards, health data codification and electronic data exchange protocols, electronic data exchange was recommended using a centralized model (shared health records), where individual data was collected in one database system that were managed safely and reliably [11]. Australia through the concept of PCEHR (personally controlled electronic health record) is one of the countries that has a central data exchange model played by the federal government (central government) [12]. A centralized system makes it easier to connect data from different healthcare digital systems. There needed to be regulation for centralized database system managers who were not health care providers but participated in collecting individual data and were responsible for facilitating the exchange of such individual data in order to support sustainable health services (Continuum of Care). In various models of electronic data exchange, the central government had a dominant role in facilitating interoperability.

Integrated MCH Digitalization

Integration Initiatives

Integration is required so that different applications are interconnected and form a single entity. Integration should involve standard specifications and communication media that enable interoperability between health services [9]. Various initiatives to support system integration in Indonesia has been carried out including:

Kamus Data Kesehatan Indonesia (Kata Hat-I)

Kata Hat-I (is a list of information about data standards in clinical/health terminology in all Information Systems and Health Services in Indonesia in order to create common meaning and improve the validity and reliability of health data to improve communication of exchange, collection, and use of integrated data. The word Hat-I (https://idn-hdd.kemkes.go.id) is an initiative that aims to make local applications developed independently can be easily integrated, read and exchanged between systems.

SPBE Integration

Electronic-Based Governance System (SPBE) was regulated in Presidential Regulation No. 95 of 2018 concerning Electronic-Based Governance System. SPBE aimed to realize clean, effective, efficient, transparent, and accountable governance; realizing quality and reliable public services; and realize an integrated electronic-based governance system. SPBE services include government services(e-office, e-planning, e-budgeting, e-monev), civil servant (e-staffing, e-pension), community (e-complaints, e-health, e-education) and businesses (e-procurement, e-licensing). SPBE principles include effectiveness, integration of supporting resources, sustainability, efficiency, accountability and interoperability, and security.

The Need to Support Integration and Interoperability in Indonesia

Although there have been initiatives to support data integration and interoperability between digital health systems, the results of the analysis showed that Indonesia still needs strengthening several resources. The following figure summarizes the aspects found in the study based on the framework from Mansoor (2010) [13].

Table 3. The Need to Support Integration and Interoperability in Indonesia [13]

No	Aspects that need to be supported	Description and examples
1	Regulatory framework	Regulations and policy such as electronic data exchange for meaningful health data, the national guideline for data integration and interoperability, the use of electronic health records.
2	Informative structure	Standard health terminology and unique ID for different domains such as health facility registry, client registry, health workers registry.
3	Technical infrastructure	Communication network, electronic directory and data warehouse, identification, authentication, information structure: Service Program by the Ministry of Communication and Information Technology
4	Adequate and interoperable ICT system	Electronic health records, national patient discharge summary, administrative support systems, community-based health program information systems (immunization, nutrition)
5	Data Accessibility	● Access data between health and non health organization: sharing policy through compliance with data standards, metadata, interoperability ● Accessibility for the community

It was identified that several interoperability and integration requirements have been in place that require additional arrangement. Registration of health facilities with a unique identification is carried out by 3 different units in the MoH, the Directorate General of Health Services responsible for hospitals and private clinics, the Center for Data and Information of the Ministry of Health for primary health centers and the Directorate of Pharmaceutical Services for pharmaceutical facility registration. Unique ID for resident or client registry can be used Resident Number (NIK) from the Ministry of Home Affairs or membership number of the social health insurance (BPJS Kesehatan). Health Management Information System has

been managed by the Center for Data and Information of Ministry of Health through One Health Data application as the national health data warehouse for routine reporting. Registration of Health Workers have been attempted by different institutions such as the Human Resources Development Agency (BPSDM) of the MoH, the Indonesian Medical Council (KKI), as well as various professional organizations such as midwives, nurses, and others. Currently also being proposed is the Integrated Medical Record as a shared health records that is led by the Directorate of Health Services Ministry of Health that covers electronic medical record and medical discharge summary standards.

Civil registration is a legal identity for individuals to access public services. Unique National IDs can link important information collected from a variety of sectors including health. Multiple countries assign one unique civil registry ID at a time to health data. There are also those who add an identification number for health in addition to a unique civil registry ID, but the two data are interconnected. Connecting the health system with civil registration and national identity management systems will provide effectiveness and efficiency in universal health services [14].

Leadership and Governance

Leadership and governance are critical to ensure the process of integration and interoperability of health information systems [15,16]. It is a cross-sectoral collaboration that involves different ministries such as the Ministry of Health, Ministry of Communication and Information and the Ministry of Home Affairs. Our study found that The National Development Planning Agency has a strategic role to coordinate policy formulation for data sharing amongst different ministries.

Expectations of Interoperable and Integrated Information Technology

Based on identification and analysis of conditions related to MCH digitalization in Indonesia, the following information system architecture is proposed (Figure 3).

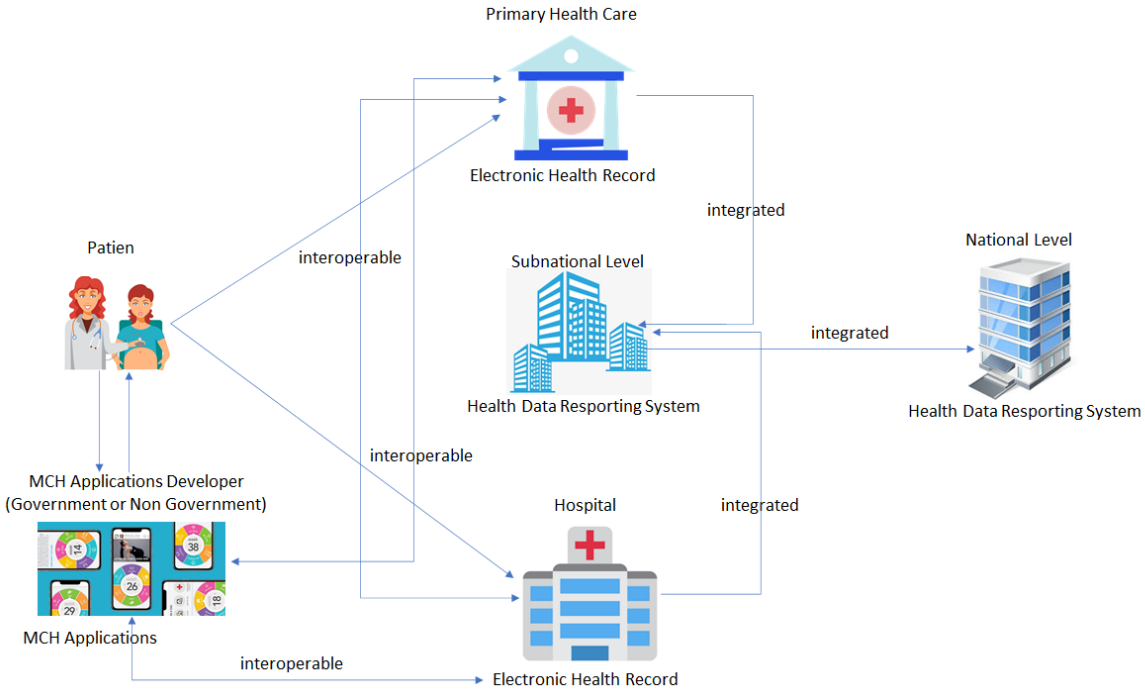


Fig 3. Proposed an integrated and interoperable health information system architecture for MCH services [17]

The proposed health information system architecture includes interoperability of individual patient data through the existing electronic health records (EHR) in health care facilities (Primary Health Care and hospitals). The MCH services should be incorporated into

the EHR system that has capability to be interoperable with other EHR systems and any MCH applications that are used by a patient or community. These applications can be developed by the government and the private sector. Any healthcare service application such as an EHR or MCH application should have capability to produce monthly reporting based on a standard that is required for the sub-national level up to national level.

Interoperability between systems in the context of healthcare services to support continuity of care and capability to generate routine reports as an output of such systems is a comprehensive function that must be considered in any digital health intervention. With such a health system architecture, it is expected that quality MCH services as part of the continuum of care will improve, efficient, sustainable and promote data use for decision making.

Conclusion and Recommendation

A number of digital health innovations for the healthcare service and routine program reporting of Maternal and Child Health (MCH) have been developed either at the National, Sub-national (provincial and district) and health facility levels. Unfortunately, the digital health initiatives were fragmented which has an impact on redundancy of data collection and fragmentation of efforts to digitize healthcare services and health information systems.

Interoperability of digital healthcare services and integration of digital MCH information systems under the framework of continuum of care (continuity between services) is indispensable, both for individual healthcare services and monitoring of maternal and child health programs. More technical regulations or guidelines are required for electronic data exchange between digital systems providers and integration of routine report (digitalization of management information systems). An institution or consortium needs to be established to develop terminology and vocabulary standards, unique ID for different registration systems, and provide infrastructure, facilitate maintenance and electronic data exchange amongst digital MCH applications.

Acknowledgment

We appreciate all key informants that have contributed to this study from Primary Health Center, Maternal and Child Hospitals, District Public Hospitals, District and Provincial Health Offices as well as the private sector as digital innovation developers. In addition, we are thank to stakeholders at the national level include the Ministry of Health, The Ministry of Home Affairs, the National Population and Family Planning Agency (BKKBN), the Ministry of Communication and Informatics as well as experts or consultants who have provided directions related to the digitalization of MCH services in Indonesia in the future. There is no conflict of interest to disclose.

References

1. Smeru. COVID-19 Impacts on Nutrition and Maternal Child Health Services: Case Study in Five Regions in Indonesia. Catatan Penelitian Smeru No.5.(2020).
2. Yusni Zainal, Guardian Yoki Sanjaya, Mubasysyir Hasanbasri. The Importance of Information System for Managing Routine Data for Maternal and Child Health Monitoring. (2013).
3. Sofaer, Shoshanna. Qualitative Methods: What Are They and Why Use Them?. Health Services Research 34 : 5 Part II (1999).
4. World Health Organization. Classification of Digital Health Interventions. (2018)
5. World Health Organization and International Telecommunication Union. 2020. Digital Health Platform: Building a Digital Information Infrastructure (Infostructure) for Health. Geneva. Licence: CC BY-NC-SA 3.0 IGO

6. Hugoson, MA. Centralized versus Decentralized Information Systems: A Historical Flashback. International Federation for Information Processing, AICT 303, pp. 106–115. (2009), doi: https://doi.org/10.1007/978-3-642-03757-3_11
7. World Health Organization. Draft Global Strategy on Digital Health 2020–2025. (2020)
8. World Health Organization. 2016. Sixty-Ninth World Health Assembly: Framework of Engagement with Non-State Actors. (2016).
9. Ismail, Nanang, et al. "Interoperability and Reliability of Multiplatform MPLS VPN: Comparison of Traffic Engineering with RSVP-TE Protocol and LDP Protocol." Communication and Information Technology Journal, vol. 11, no. 2, 2017, pp. 57-65, doi:10.21512/commit.v11i2.2105. (2017)
10. Binus. 2019. FHIR: Fast Healthcare Interoperability Resources. https://mti.binus.ac.id/2019/08/09/fhir-fast-healthcare-interoperability-resources/.
11. Integrating the Healthcare Enterprise (IHE) Europe. 2018. The IHE Connectathon: What is it? How is it done? Version 006. www.ihe-europe.net
12. Vimalachandran, P., Wang, H., Zhang, Y., Zhuo, G. The Australian PCEHR system: Ensuring Privacy and Security through an Improved Access Control Mechanism. European Alliance for Innovation. (2016) doi: 10.4108/eai.9-8-2016.151633
13. Mansoor, Muhammad Ehsan., Majeed, Rashid. Achieving Interoperability among Health Care Organizations. Sweden. (2010).
14. Mills, Samuel, et al. Unique health identifiers for universal health coverage. Journal of Health, Population and Nutrition. 38 (1):22. (2019), doi: 10.1186/s41043-019-0180-6
15. Adler, Julia, et al. Benchmarking health IT among OECD Countries: Better Data for Better Policy. J Am Med Inform Assoc. 2014;21: 111–16. doi: 10.1136/amiajnl-2013-001710. (2013), doi: 10.1136/amiajnl-2013-001710
16. Michelsen, H. Brand, Achterberg P., Wilkinson J. Health Evidence Synthesis Report. Promoting Better Integration of Health Information System: Best Practices and Challenges. World Health Organization. (2015).
17. Data and Information Center, The Ministry of Health. 2020. One Health Data Policy.

24th International Conference on Business Information Systems (BIS 2021)

Applications

https://doi.org/10.52825/bis.v1i.54

Published: 02 July 2021

The Digitalization of Local Owner-Operated Retail Outlets: How environmental and organizational factors drive the use of digital tools and applications

Lars Bollweg[1], Sören Bärsch[2], Richard Lackes[2], Markus Siepermann[3], and Peter Weber[1]

[1] Fachhochschule Südwestfalen, Soest, Germany
{bollweg.lars, weber.peter}@fh-swf.de

[2] Technische Universität Dortmund, Germany
{soeren.baersch, richard.lackes}@tu-dortmund.de

[3] Technische Hochschule Mittelhessen Giessen, Germany
{markus.siepermann}@mni-thm.de

Abstract. The digitalization of the retail industry is a disruptive innovation process which endangers the very existence of Local Owner Operated Retail Outlets (LOOROs). Despite the manifold digital options to regain competitive power, LOOROs struggle in their digital transformation and persist often in their traditional business behaviour. As their customers get more and more used to buying via digital channels, they more and more expect the provision of digital services. This paper and the presented survey among 223 LOORO owners from 26 cities in Germany aim to understand why the LOOROs are so hesitant. Our findings show high insecurity among LOOROs about what to do and where to begin the digitalization route. The owners of LOOROs are often decoupled from their near and far business environment. This leads to a wrong self-assessment and implies the risk that the services provided do neither match the competitive environment nor customer expectations.

Keywords: digitalization, innovation, business transformation, retail outlets.

Introduction

The digital transformation of the retail industry creates enormous challenges for local owner-operated retail outlets (LOOROs), which are characterized by a small-sized store area, a restricted number of employees and a high degree of owner-involvement in the business operations [1]. This kind of “local retail market” enables a personal relationship between the shop owners and their customers and provides along with that a lot of advantages to sustain this relationship compared to online shops.

However, despite these possible advantages LOOROs seem not to be able to make use of it. LOOROs are pressured by the digital development of all their value chain partners (customers and suppliers), as well as by the competitive environment (Big-Box retail outlets, multichannel chain stores and pure online trade). Furthermore, LOOROs have to realize that their most important value chain partner – the customer, is no longer satisfied with the current digital approach in traditional small shops [2]. Customers have already changed their shopping habits and do use more and more digital sales channels and services. For shopping the customers expect the high level of convenience they are used to online also in local shops like those from LOOROs.

Anyway, LOOROs are still hesitant to offer digital services. On the one side, they fail to adopt emerging technology like digital systems due to the high complexity [3]. On the other side, the reason may lie in their limited resources (e.g. lack of time or Know-how regarding

new technology etc.). Being typical micro-enterprises (MEs) [3], their internal structure does not give them much room for manoeuvre. Therefore, implementing digital structures and processes in the daily business operations is hardly possible without external aid. However, as mentioned, LOOROs are not without opportunities in this situation. Digital tools and applications like for example digital inventory management systems, additional online shopping channels, customer relationship management systems (CRM), or also marketing automation tools exist and could help LOOROs to overcome their inherent restrictions [4] and to regain competitive power. Despite the importance of LOOROs for the local economy or the attractiveness of the city centers, research with a clear focus on the technology adoption of LOOROs and small retails is still scarce. A reason could be the high diversity of the retail sector that hinders the study of a sufficient number of retailers to obtain significant results [5].

Therefore, this paper attempts to examine the reasons why LOOROs are hesitant to digitalize their infrastructure and business processes. We are aiming to provide insights to enable LOOROs, municipal leaders and city governments to identify opportunities for action on how to help local retail to grow digitally and transform into multi-channel local commerce. Accordingly, the following research question shall be answered:

RQ: How do environmental factors influence the adoption of digital tools and applications by owners of LOOROs?

The remainder of this paper is organized as follows: Section 2 discusses the theoretical background. The research model for the survey is developed in section 3 and analyzed in section 4. The paper closes with a discussion of the results, the managerial implications and future research in section 5.

Theoretical Background

SME Retail in Research

To discuss the theoretical background a structured literature review was conducted. The reviewed papers and studies in the literature review where mainly identified through a keyword search with focus on the term “SME retail” as research on “ME” retailers is scarce. Most of the identified studies have classified “SME retailers” along the number of employees as a size indicator, like, e.g. a range from three to 80 employees [6], [7], [8], as part of SME retail chains [5] and others had a focus on single-location outlets [8], [9].

The reviewed studies had one feature in common: the unique role of owners / managers of the SME retailers. SMEs like LOOROs are mainly owned and operated in personal union. Subsequently, in SMEs a strategic decision is highly dependent on the owners. A positive attitude of the owners towards change creates an organizational environment that is open to innovation [9]. The structural lack of internal and external resources is another hallmark of SME retailers like LOOROs [7]. Reluctant implementation of new retail technologies also relies on scarce financial capital and the lack of technical know-how [3]. Moreover, many non-adopter SMEs do not have the requisite infrastructure and procedures to implement new technologies [5], [8].

Internal and external influence factors

Innovation and technology acceptance processes' driving factors are mainly divided into two types: 1) internal and 2) external factors. The decision of a company to implement emerging innovations is greatly affected by internal and external variables based on innovation attributes: perceived benefits, organizational readiness and external pressure [6].

Previous studies examined further internal effects like the risk perception, advantages of IT use, the owner’s perspective, the attitude and internal demand of the retailer but also external effects like competition, government or the society as factor for the adoption of new technologies [2], [3], [5].

Research Framework and Conceptual Model

Unlike big corporations, the owner is the primary decision-maker in ME like LOOROs, who decides on strategic issues alone. Hence, organizational factors can be seen – to a certain extent – as external factors. As a result, this paper focuses on an owner-centric examination on the individual level [10]. Therefore, this study research design will be based on the Stimulus-Organism-Response Model (S-O-R Model) [5], [8]. Mehrabian and Russell's (1974) [11] Stimulus-Organism-Response Model comes from the field of environmental psychology [12]].The S-O-R model shows how environmental processes and changes, called stimuli (S), are perceived by an organism (O) and instigate (emotional) reactions of the organism called behavioural response (R) (see Figure 1).

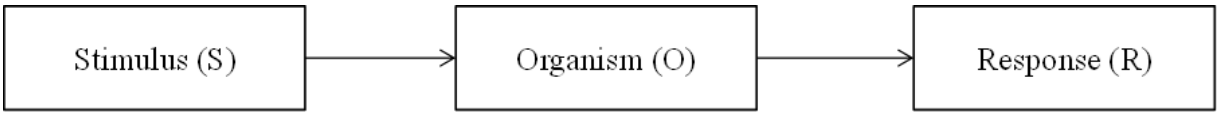


Figure 1. S-O-R Model

Based on the environmental psychology, three central aspects of emotional responses to the perception of the encountered environments are used: pleasure, arousal and dominance (the PAD-Scale). Thereby, pleasure is described purely in terms of positive or negative feelings, arousal as a feeling state that concerns mental activity, and dominance as a feeling of control and behaviour restrictions caused by physical or social barriers [11].

However, the S-O-R framework is often criticized for its bipolar measurement when using the PAD-Scale [13], as it allows the joint occurrence of pleasant and unpleasant states [14]. Thus, the current study uses a unipolar view linking the three dimensions to one joint model that is more suitable [14], [15]. Pleasure, arousal, and dominance can be seen as affective (feeling), cognitive (thinking), and conative (acting) responses. Then, these responses can be unified as one joint measure for the organism [15].

Conceptual Model

In the Literature there seems to be two streams that can be distinguished: technology-centered theories focus and on the characteristics of the technology itself and the diffusion of technology through different channels [16]. These theories are manly used for understanding the technology adoption on an organizational level. In contrast, decision maker centered theories concentrate on the individual level to analyze human behaviour as well as its impact on the decision-making process regarding technology adoption and use [17], [18].

Looking deeper into the decision maker centered theories, the Theory of Reasoned Action (TRA) [19] and its successor, the Theory of Planned Behavior (TPB) [17] state that attitudes, control beliefs, and subjective norms influence behavioural intention, which in turn influences actual behaviour. Davis et al. (1989) [18] applied TRA/TPB to the individual level of technology adoption behaviour in the well-known Technology Adoption Model (TAM).

The organism, namely the owner as the decision maker in LOOROs, is thus captured as the attitude towards a technology by the TRA/TBP concept and influences the intention to use it [15]. This thought process is triggered by internal and external stimuli. We assume that the perception of organizational resource availability and the perception of external pressures can both be seen as such environmental stimuli leading to the organism’s emotional reactions [20]. Finally, the usage of the technology is the stimulated organism’s emotional response.

For a better understanding, we interpret digitalization as the use of digital tools and applications in one of the two following areas:

(1) Front-end sales channels: all digitalization efforts with direct customer touch points [21].

(2) Administrative back-end: all digitalization efforts are invisible to the customer [21], [22].

Figure 2 shows how the digital tools and applications integrate into customer service delivery. Digital services and digital sales are front-office activities and exceed to the line of interaction, while digital marketing exceeds to the line of visibility. Digital administration comprises back office activities and exceeds to the line of internal interaction.

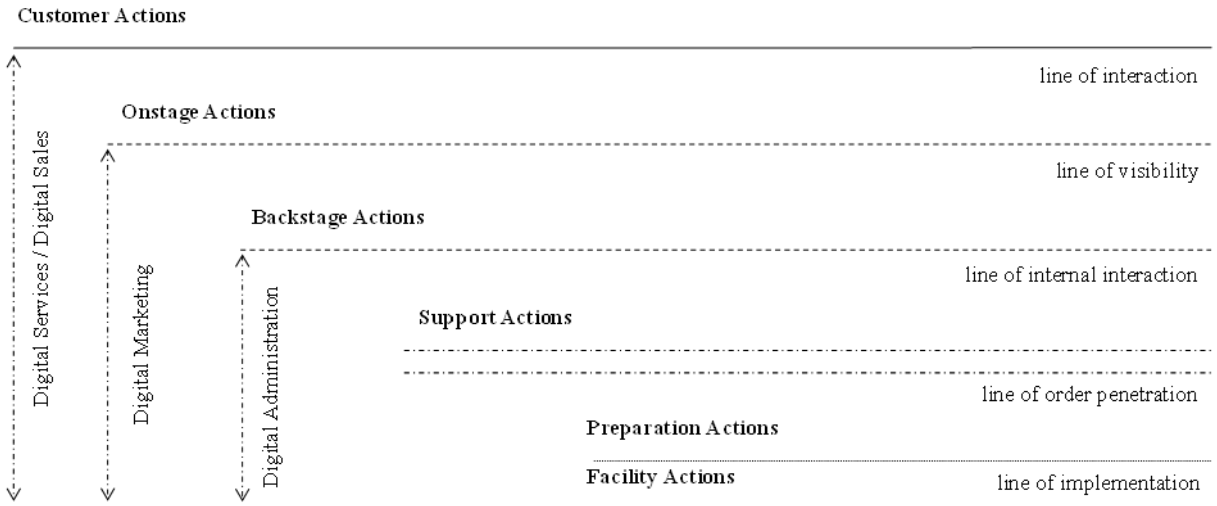


Figure 2. Service Blueprint including digital tools and applications [21]

Hypotheses Development

Stimulus (S) to Organism (O): companies with restricted access to capital and inadequate infrastructure are hesitant to invest in digital tools and applications that could have a competitive edge [23]. Resources can be categorized into tangible and intangible resources [23]. The availability of tangible organizational infrastructure is embodied in the availability of general resources, required capability and the IT infrastructure. Without the first two resources, emerging innovations are becoming increasingly difficult for companies like LOOROs to implement. [24].

This is particularly relevant for the IT infrastructure when digital tools and applications are adopted. We hypothesize the effect of the available infrastructure on the emotional reactions of an organism (O) (attitudes towards digitalization) as follow:

H1a: The availability of infrastructure has a positive influence on the attitude towards the digitalization.

The availability of intangible organizational human capital is reflected by employee expertise and motivation, which have been found to be the most influential success factors [24]. Also, the innovative strength of employees plays an important role [9]. Hence, we hypothesize:

H1b: The availability of human resources has a positive influence on the attitude towards the digitalization.

Previous studies have shown that external environmental pressures have an impact on the adaption of technology among companies [5], [9]. Correspondingly, external pressures comprise influences from the near and far environment. The near (specific) environment is formed by influences of competitors and customers that exert a direct impact on the examined organization. The perceived pressure of the competitors is demonstrated by the perception of own development relative to the development of the competitors, the perception of the need for own development to remain competitive and the perception of external pressure to remain competitive in digitalization. [25]. Hence, we hypothesize:

H2a: Perceived pressure from competitors towards digitalization has a positive influence on the attitude towards digitalization.

The perceived pressure of the customers for LOOROs is represented by the perception of customer actions, the perception of customer pressure, the perception of customer expectations [25]. We hypothesize:

H2b: Perceived pressure from customers towards digitalization has a positive influence on the attitude towards digitalization.

Government and socio-political situations characterize a far environment [26]. The perceived pressure of society is thus reflected by the perception of the general relevance of digitalization, political pressure, and social expectations. [25]. We hypothesize:

H2c: Perceived pressure from politics and society towards digitalization has a positive influence on the attitude towards digitalization.

Organism (O) to Response (R): Attitudes as well as control beliefs and subjective norms do not directly influence actual behavior, but rather influence behavioral intention (intention to use), which in turn influences actual behavior (current use) [17], [18].

We then use "Digitalization Attitude" and "Intention to use Digitalization." In accordance with the TRA/TPB/TAM theory, the Digitalization Assessment, the ease of learning and the expected effectiveness of digitalization [17] are considered for the measurement of the construct.

H3: A positive attitude towards digitalization has a positive influence on the intention to use digitalization.

Behavioural intentions are said to influence actual behaviour and therefore to have direct impact on the current use of digital tools and applications [17], [18]. Hence, we hypothesize:

H4: A high intention to use digitalization has a positive influence on its current use.

We distinguish the back-end from the front-end operations to frame the umbrella term digitalization into an operational interpretation [22]. All activities without consumer contact points reflect the back-end operations of retailers. For the consumer, these activities are unseen. We focus on front-end operations for customer contact points since the retail industry's digitalisation is very consumer-oriented.

These activities are noticeable to consumers and differ only in terms of their level of customer interaction [21], [22]. In detail, the following four areas are investigated [22]:

1. Digital administration includes all back-end operations, such as inbound and out-bound distribution or human resource management, without customer contact points and engagement.

2. Digital marketing covers all front-end marketing activities with customer touchpoints but without direct customer interaction.

3. Digital sales channels cover all front-end sales activities with customer touchpoints and low customer interaction.

4. Digital services cover all digital front-end services with customer touch points and high customer interaction.

We then divide the (behavioural) intentions (“Intention to Use”) and the actual behaviour (“Current Use”) towards digitalization into the four dimensions administration, marketing, sales, and services. Thus, we extend the above stated hypotheses 3 and 4 as follows:

H3a: A positive attitude towards digitalization has a positive influence on the intention to use digital administration.

H3b: A positive attitude towards digitalization has a positive influence on the intention to use digital marketing.

H3c: A positive attitude towards digitalization has a positive influence on the intention to use digital sales channels and provide them to customers

H3d: A positive attitude towards digitalization has a positive influence on the intention to use digital services

H4a: A high intention to use digital administration has a positive influence on the current use of digital administration tools and applications.

H4b: A high intention to use digital marketing has a positive influence on the current use of digital marketing.

H4c: A high intention to use digital sales channels has a positive influence on the current use of digital sales channels and their provision to customers.

H4d: A high intention to use digital services has a positive influence on the current use of digital services.

The concentration on attitude, behavioural intention, and use alone is inadequate because it does not fully capture the mechanism of adoption in organizations. Prior knowledge and inexperience are major influences affecting the use of technology [20], [27]. This is consistent with several IS studies on technology adoption which show that perceived benefits of already used technologies are a main factor for the implementation of new technologies [3], [28].

We therefore postulate that prior experience with digital administration in the back-end will have a positive impact on use of digital marketing and on the front-end areas of digital sales channels and digital services. Subsequently, we state the following hypotheses:

H5a: A high prior use of digital administration has a positive influence on current use of digital marketing.

H5b: A high prior use of digital marketing has a positive influence on current use of digital sales channels.

H5c: A high prior use of digital sales channels has a positive influence on current use of digital services.

H6a: A high prior use of digital administration has a positive influence on current use of digital services.

H6b: A high prior use of digital administration has a positive influence on current use of digital sales channels.

H6c: A high prior use of digital marketing has a positive influence on current use of digital services.

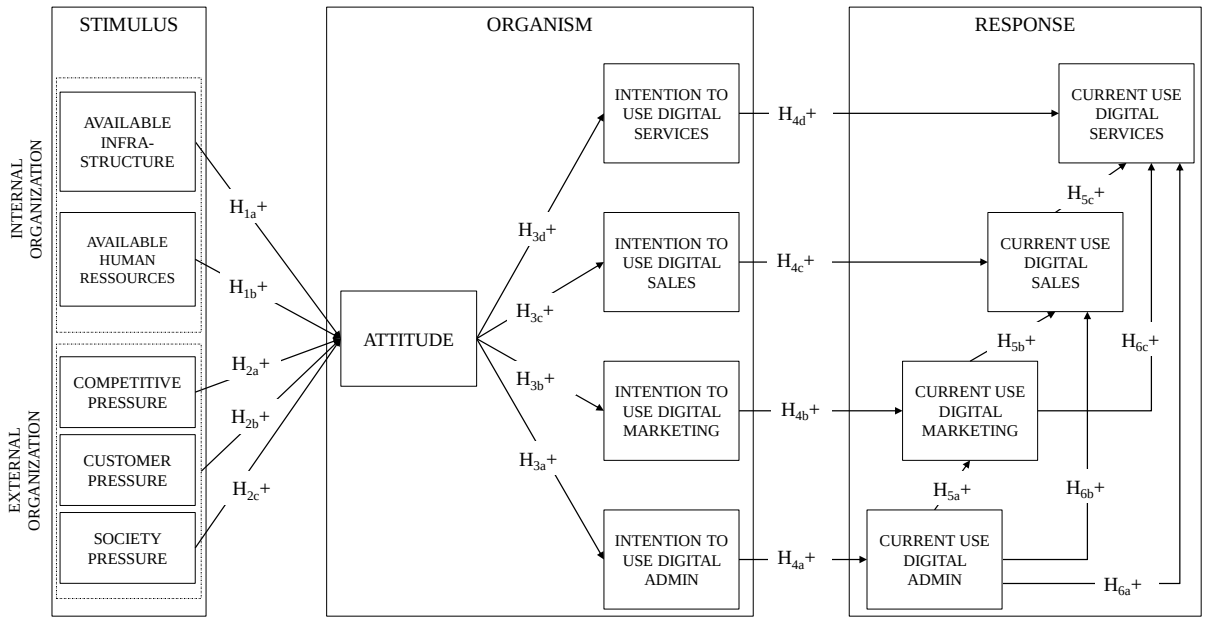


Figure 3. Conceptual model

Analysis

Data Collection

We surveyed shop owners of LOOROs in 26 cities in the area of South Westphalia in Germany between May and July 2016. Two opening questions and 34 individual questions were included in the questionnaire. Via an online form, 124 participants replied and 119 participants replied on paper. In total, the questionnaires of 243 companies were received, including 223 questionnaires with full data sets.

The descriptive analysis shows that 25% of the respondents sell clothing, fashion and shoes. Other important groups of retailers in this study are jewellers, stationery and office suppliers, each with a share of 9%. Drugstores, electronic shops, toys and art shops, curtains and photographic supply shops with each around 5%. Finally, the remaining 16% of the examined retailers that do not belong to any of the above-mentioned categories can be summarized as “other”. For the analysis of the collected data and the evaluation of the research model, we used SmartPLS. Bootstrapping was done with 5,000 samples and 223 cases, determining the significance of weights, loadings and path coefficients.

Measurement Model

The research model has one reflective construct (‘Attitude towards Digitalization’). The other thirteen constructs are formative, so that different analyses are needed [29]. The significance of the constructs’ indicators is assessed by their loadings (reflective constructs) or weights (formative constructs) and their t-values. Concerning the reflective construct, all indicators are significant [30]. For the convergence criterion, the model fits to the convergence criteria AVE (Average Variance Extracted) is 0.576 (minimum > 0.5), the composite reliability is 0.844 (min. 0.7) and Cronbach’s alpha is 0.751 (min. 0.7) [31-34].

The prediction validity Q^2 is with 0.381 higher than the minimum of 0 [34]. For the formative constructs, the discriminant validity must be verified. The highest correlation between the latent variables with a value of 0.85 still matches the maximum of 0.9, so that the criterion is met [34]. In addition, multicollinearity between indicators of formative constructs is not permitted [35]. The variance inflation factor (VIF) for all indicators $_i$, with $VIF_i=1/(1\ R_i^2)$ is lower than five so that there is no sign for multicollinearity [34].

Structural Model

The variance inflation factor of constructs with two or more influencing factors (here: Attitude, VIF=1.00) is lower than the required level, which shows that there is no multicollinearity [35]. The value of R2 indicates a substantial (moderate, weak) influence if the value exceeds 0.67 (0.33; 0.19) [37]. Since endogenous and exogenous variables are collected together using one questionnaire [37], the survey is prone to common method bias (CMB). However, our VIF values indicate that the model is free from CMB [34].

In sum, only two hypotheses are not significant. We could confirm 17 of 19 hypotheses of our research model, two of which could be confirmed at the 10%, one at the 5%, and 14 at the 1% level (see Figure 4). The explanatory power of the model (R^2) is on a medium (Current Use of Digital Administration and Digital Marketing) to high level (Current Use of Digital Sales and Digital Services).

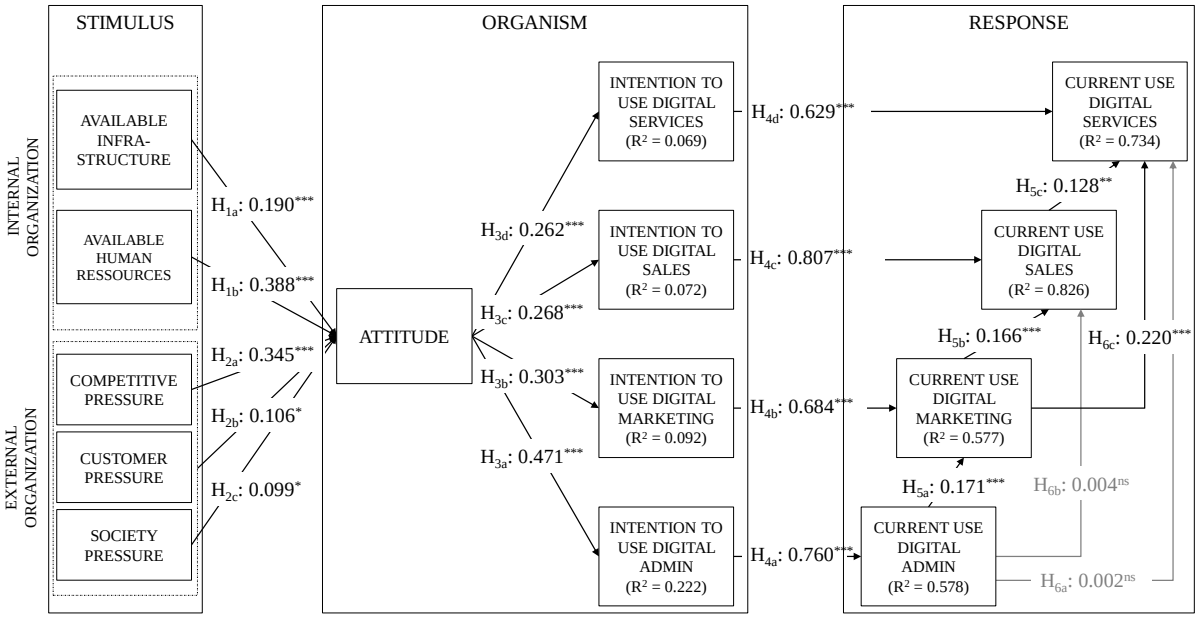


Figure 4. Conceptual model results

Conclusion

Results

The "Available Organizational Infrastructures" and the "Available Human Resources" impact attitudes towards digitalization with respect to internal organizational impact factors. Both hypotheses (H1a, H1b) were verified and proved to be extremely relevant. Specifically, this refers to the availability of human resources among all factors that have by far the most significant effect. This holds in particular for the availability of human resources which have the highest impact by far among all factors. That means that employees drive innovation processes of LOOROs and influence the shop owners if they have enough competencies. Digital competencies are a prerequisite for the adoption of technological innovations.

However, the descriptive results show only a medium availability. While only 11% of the respondents attribute innovativeness to their employees, at least 44% found their available human resources to have enough “competencies” and to be “motivated” to handle digitalization (58%). The level of “infrastructural readiness” is even lower. Only about 30% of the respondents agreed or strongly agreed to have sufficient “infrastructural resources” to face the digitalization challenge, to have sufficient “capacities”, or to have a sufficient “IT-Infrastructure” for the challenges of the digitalization. Obviously, LOOROs suffer from lack of internal resources.

External pressure is commonly found to be an adoption driver [5], [6] which is confirmed by our study. All examined factors show an impact on the attitudes towards digitalization, particularly competitive pressure. Around 40% of respondents recognize their own digitalization and feel a desire to keep up with competitors. In addition, the influence of customer pressure is small and of society pressure nearly negligible. The reason lies in the perception of customer demand. About 70% of all LOORO owners cannot report that their customers ask for digital offers and services. Many LOOROs seem to live in an “offline bubble”. They only have customers who prefer the offline offer of LOOROs and the owners do not get in contact with other consumers who prefer online offers. Obviously, LOOROs seem to be decoupled from their near and far environment [38], [39]. Interestingly, if LOOROs perceive direct pressure from customers or society, this decreases their positive attitude towards digitalization. On the one hand, this is not surprising. Pressure from policy usually comes from regulations that are often regarded as impositions. On the other hand,

LOOROs should orientate to the wishes of their customers. If they directly ask for digital services, this should enhance and not hinder LOOROs' wish to digitalize.

However, the general attitude towards digitalization, which concerns the organism of the model, is positive. Nearly 60% perceive digitalization as being good and "easy to learn". While a positive attitude fosters all intentions to use digital tools and services, the highest effect is exerted on the intention to use digital administration tools. Because the intention to digitalize an activity significantly influences the use of digital tools (H4a-H4d), this in turn activates a domino effect from back office to front office activities to the line of customer interaction.

The use of digital administration tools encourages LOOROs to use digital marketing tools (H5a) which consecutively fosters the use of digital sales (H5b) and the provision of digital services (H5c, H6c). That means that in case the prerequisites for the operation of digital tools are given, this not only facilitates but fosters the usage of these tools. In more detail, the use of digital administration tools is prerequisite for the use of digital marketing tools and so on.

However, the digitalization of the administrative backend only influences its direct successors (H6a and H6b are not significant). That means that the digitalization process of LOOROs seems to evolve from backend to frontend step by step and is not customer driven. If we have a deeper look at the intentions of LOOROs to digitalize, we can observe a medium level for the administrative stage (52% to 62%). Digital marketing activities are intended to use by 23% to 45%, digital sales by 8% to 28%, and digital services by 21% to 39%. Digital administration in the backend is used by more than half (54%) of the participants, while digital sales (8%), digital marketing and digital services (both mean: 25%) are rarely used.

Our findings indicate that LOOROs are facing a lack of available human resources and infrastructure and that they are facing a situation of insecurity. LOOROs seem to be holding and waiting for their digitalization decision, not understanding whether or not their own usable technology is appropriate and in which technologies they should invest [40]. Surprisingly, they do not experience pressure from changing consumer demands and thus do not see a need to respond to competitors' digitalization efforts.

Managerial Implications for LOOROs

First of all, LOOROs need to be reconnected to their potential customers with regard to changing habits and the growing competition from Internet and chain stores. Hence, for LOOROs to recover competitive strength there is a need for an external (public or governmental) push to help the requisite internal turnaround. The owners/managers need to focus mainly on their understanding of the present and future consumer demands and preferences to reconnect LOOROs with environmental developments [41]. Talking with and involving their employees might help with this as their competencies and motivation is one of the main drivers.

Therefore, employees should secondly be bolstered with digital knowledge. The lack of competence among employees is one of the major resource problems. Chambers of commerce and government can offer trainings tailored to the needs of LOOROs to overcome this digitalization barrier. LOOROs as well as their employees are inexperienced with the according tools and applications and therefore neglect opportunities of digital sales channels. Appropriate trainings can introduce them to this new digital world so that LOOROs can start using online sales and marketing channels with low entry barriers, like third-party platforms (also local shopping platforms), to keep in touch with existing customers, explore new markets and to get started in the e-commerce arena. Providing help in this manner seems to be a more promising approach than exerting direct pressure on them. The direct digitalization pressure already exerted by policy through for example electronic cashier systems reduces the positive attitude towards digitalization and subsequently the LOOROs' intention to digitalize.

However, even if it reduces LOOROs positive perception of digitalization, forcing them to digitalize their back office is nonetheless and thirdly a suitable approach. The administrative backend is the area with the highest use intentions and with the highest current use.

Moreover, for subsequent digitalization areas, it is the starting point. Since legal regulations will control the administrative backend to some degree, policy can use the openness of LOOROs as a door opener for digital support of their administrative backend. This could trigger a promising impulse and launch a chain reaction in all subsequent areas towards the use of digitalization tools and applications.

Research Implications

Firstly, we contribute to the technology adoption research by means of an examination of the internal and external influence factors of the technology adoption process of Micro Enterprises (like LOOROs with an adapted and improved S-O-R Model. The new model includes an improved organism (O) section (by integration of the TRA/TPB core constructs) as well as an extended response (R) section and a usage-related examination. It offers a toolbox for future research on micro enterprises of all kinds.

Secondly, the subdivision of the analysis model into four digital business areas (Digital Administration, Digital Marketing, Digital Sales Channels and Digital Services) offers a systemized approach to frame the ambiguity of the umbrella term digitalization into an operational understanding. Previous research usually neglected that companies already have adopted different digital tools which are used to support parts of their business processes. Yet, the degree to which digital tools are already used determines the readiness of a company to adopt other technologies [7], [27].

Limitation and Future Research

The very limited sample size, first of all, restricts the explanatory power of our results. Second, this analysis is focused on the German retail industry context. The findings should, however, not simply be generalized to other countries with their unique retail cultures.

Thirdly, only owners of LOOROs, but not their customers, have been studied. Although several recent surveys have had a look at the customers' view in the cities we investigated, the connection between retailers and customers is only indirect. This could be improved in further studies by distributing questionnaires to owners and their customers at the same time.

Lastly, the technologies (tools and applications for services, sales, marketing and administration) considered when measuring the "intention to use" and the "current use" are just one possible selection. The inclusion of other technologies could lead to different results.

Future research would be valuable on at least the following aspects: (1) Technology: Systematic research is needed to identify promising technologies and digital tools and applications that can help LOOROs improve their businesses and win back competitive power. (2) Technology adoption under uncertainty: Further studies should investigate what other factors may impact the technology adoption process. Additionally, more research on how to overcome the high uncertainty of local shop owners is needed, as this uncertainty currently clearly hinders the technology adoption of LOOROs. (3) Public and governmental support: Research is needed on how the public can trigger the digital development of LOOROs.

References

1. L. Bollweg, R. Lackes, M. Siepermann, P. Weber, "Mind the Gap! Are local retailers misinterpreting customer expectations regarding digital services," *In Proceedings of the MCCSIS, pp. 111–117,* Las Palmas de Gran Canaria (2015), doi: https://doi.org/10.1016/j.jretconser.2013.06.007

2. E. Pantano, M. Viassone, “Demand Pull and Technology Push Perspective in Technology-Based Innovations for the Points of Sale: The Retailers Evaluation,” *Journal of Retailing and Consumer Services*, 21(1), 43–47 (2014), doi:10.1016/j.jretconser.2013.06.007
3. V.E Erosa, “Technology Illiteracy in retail SMEs: Exploring late adopters characteristics,” *In PICMET’09-2009 Portland International Conference on Management of Engineering & Technology*, pp. 2623–2630. IEEE, Portland (2009), doi:10.1109/picmet.2009.5261816
4. Statista. “Veränderung der Besucherfrequenz im Einzelhandel”. Statista. https://de.statista.com/statistik/daten/studie/291581/umfrage/besucherfrequenzim-%0Aeinzelhandel-in-deutschland-ggue-dem-vorjahr (last accessed 2020/12/18.)
5. S. Kurnia, J. Choudrie, R.M. Mahbubur, B. Alzougool: “E-commerce technology adoption: A Malaysian grocery SME retail sector study,” *Journal of Business Research*, 68(9), 1906–1918 (2015), doi:10.1016/j.jbusres.2014.12.010
6. J. Mehrtens, P.B. Cragg, A.M. Mills, “A model of Internet adoption by SMEs,” *Information & Management*, 39(3), 165–176 (2001), doi:10.1016/s0378-7206(01)00086-6
7. R. Rahayu, J. Day, “Determinant factors of e-commerce adoption by SMEs in developing country: evidence from Indonesia,” *Procedia-Social and Behavioral Sciences*, 195, 142–150 (2015), doi:10.1016/j.sbspro.2015.06.423
8. S. Kabanda, I. Brown, “A structuration analysis of Small and Medium Enterprise (SME) adoption of E-Commerce: The case of Tanzania,” *Telematics and Informatics*, 34(4), 118–132 (2017), doi:10.1016/j.tele.2017.01.002
9. M.R. Amin, H. Hussin, “E-commerce adoption in SME retail sector: A conceptual model,” *The 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, pp. 1–6. IEEE, Kuching (2014), doi:10.1109/ict4m.2014.7020677
10. A. Marcati, G. Guido, A.M. Peluso, “The role of SME entrepreneurs’ innovativeness and personality in the adoption of innovations,” *Research Policy*, 37(9), 1579–1590 (2008), doi:10.1016/j.respol.2008.06.004
11. A. Mehrabian, J.A. Russell, “An approach to environmental psychology,” *The MIT Press*, Cambridge (1974).
12. R.S. Woodworth, “Psychology: A study of mental life,” *Henry Holt and Co*, New York (1921)
13. S. Kim, G. Park, Y. Lee, S. Choi, “Customer emotions and their triggers in luxury retail: Understanding the effects of customer emotions before and after entering a luxury shop,” *Journal of Business Research*, 69(12), 5809–5818 (2016), doi:10.1016/j.jbusres.2016.04.178
14. R.A. Westbrook, “Product/consumption-based affective responses and postpurchase processes,” *Journal of Marketing Research*, 24(3), 258–270 (1987), doi: https://doi.org/10.2307/3151636
15. I. Bakker, T. Van der Voordt, P. Vink, J. de Boon, “Pleasure, arousal, dominance: Mehrabian and Russell revisited,” *Current Psychology*, 33(3), 405–421 (2014), doi:10.1007/s12144-014-9219-4
16. E. M. Rogers, “Diffusion of innovations”, 5th edn. *Simon and Schuster*, New York (2010).
17. I. Ajzen, “The theory of planned behavior,” *Organizational Behavior and Human Decision Processes*, 50(2), 179–211 (1991), doi:10.1016/0749-5978(91)90020-T
18. F.D. Davis, R. P Bagozzi, P.R. Warshaw, “User acceptance of computer technology: a comparison of two theoretical models,” *Management Science*, 35(8), 982–1003 (1989), doi:10.1287/mnsc.35.8.982
19. I. Ajzen, M. Fishbein, “Belief, attitude, intention and behavior: *An introduction to theory and research”* (Vol. 27). Addison-Wesley, Boston (1975).
20. R. Vize, J. Coughlan, A. Kennedy, F Ellis-Chadwick, “Technology readiness in a B2B online retail context: An examination of antecedents and outcomes,” *Industrial Marketing Management*, 42(6), 909–918(2013),doi:10.1016/j.indmarman.2013.05.020

21. S. Fließ, M. Kleinaltenkamp, “Blueprinting the service company: Managing service processes efficiently,” *Journal of Business Research*, 57(4), 392–404 (2004), doi:10.1016/s0148-2963(02)00273-4
22. A. Enders, T. Jelassi, “The converging business models of Internet and bricks-and-mortar retailers,” *European Management Journal*, 18(5), 542–550 (2000), doi:10.1016/s0263-2373(00)00043-8
23. J. Barney, “Firm resources and sustained competitive advantage,” *Journal of Management*, 17(1), 99–120 (1991), doi:10.1177/014920639101700108
24. Y.J. Wang, M.S. Minor, J. Wei, “Aesthetics and the online shopping environment: Understanding consumer responses,” *Journal of Retailing*, 87(1), 46–58 (2011), doi:10.1016/j.jretai.2010.09.002
25. T. Stapleton, “Complexity and the External Environment,” *Milton Keynes* GB: The Open University (2000).
26. N. Melville, K. Kraemer, V. Gurbaxani, “Information technology and organizational performance: An integrative model of IT business value,” *MIS Quarterly*, 28(2), 283–322 (2004), doi:10.2307/25148636
27. T. Oliveira, M.F. Martins, “Understanding e-business adoption across industries in European countries,” *Industrial Management & Data Systems* (2010), doi:10.1108/02635571011087428
28. B. Ramdani, P. Kawalek, “SME adoption of enterprise systems in the Northwest of England,” *IFIP International Working Conference on Organizational Dynamics of Technology-Based Innovation*, 409–429 (2007), doi:10.1007/978-0-387-72804-9_27
29. C. Fornell, F.L. Bookstein, “Two structural equation models: LISREL and PLS applied to consumer exit-voice theory,” *Journal of Marketing Research*, 19(4), 440–452 (1982), doi:10.1177/002224378201900406
30. C.B. Jarvis, S.B. MacKenzie, P.M. Podsakoff, “A critical review of construct indicators and measurement model misspecification in marketing and consumer research,” *Journal of Consumer Research*, 30(2), 199–218 (2003), doi:10.1086/376806
31. W.W. Chin, “Commentary: Issues and opinion on structural equation modeling,” *MIS Quarterly,* 22(1), vii-xvi (1998).
32. L.J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, 16(3), 297–334 (1951), doi:10.1007/bf02310555
33. C. Fornell, D.F. Larcker, “Structural equation models with unobservable variables and measurement error: Algebra and statistics,” *SAGE Publications* CA: Los Angeles (1981), doi:10.2307/3150980
34. J.F. Hair, G.T.M Hult, C. Ringle, M. Sarstedt, “A primer on partial least squares structural equation modeling,” *SAGE Publications* CA: Los Angeles (2016).
35. A. Diamantopoulos, P. Riefler, K.P. Roth, “Advancing formative measurement models,” *Journal of Business Research*, 61(12), 1203–1218 (2008), doi:10.1016/j.jbusres.2008.01.009
36. A. Diamantopoulos, J.A. Siguaw, “Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration,” *British Journal of Management*, 17(4), 263–282 (2006), doi:10.1111/j.1467-8551.2006.00500.x
37. E.T. Tornikoski, H. Rannikko, T.P. Heimonen, “Technology-based competitive advantages of young entrepreneurial firms: Conceptual development and empirical exploration,” *Journal of Small Business Management*, 55(2), 200–215 (2017), doi:10.1111/jsbm.12315
38. E. Pantano, “Innovation drivers in retail industry,” *International Journal of Information Management*, 34(3), 344–350 (2014), doi:10.1016/j.ijinfomgt.2014.03.002.
39. A. Parasuraman, V.A. Zeithaml, L.L. Berry, “Servqual: A Multiple-Item Scale for Measuring Consumer Perc,” *Journal of Retailing* 64, 64(1), 12 (1998), doi:10.1037/t09264-000
40. A. Purvis, W.G. Boggess, C.B. Moss, J. Holt, “Technology adoption decisions under irreversibility and uncertainty: an ex ante approach,” *American Journal of Agricultural Economics*, 77(3), 541–551 (1995), doi:10.2307/1243223

41. D. Grewal, A.L. Roggeveen, J. Nordfält, “The Future of Retailing,” *Journal of Retailing*, 93(1), 1–6 (2017), doi:10.1016/j.jretai.2016.12.008
42. P.B. Lowry, J. Gaskin, “Partial Least Squares (PLS) Structural Equation Modeling (SEM) for Building and Testing Behavioral Causal Theory: When to Choose It and How to Use It,” *IEEE Transactions on Professional Communication* (57:2), 123–146 (2014), doi:10.1109/tpc.2014.2312452

24th International Conference on Business Information Systems (BIS 2021)

Applications

https://doi.org/10.52825/bis.v1i.61

Published: 23 June 2021

A Novel Example-Dependent Cost-Sensitive Stacking Classifier to Identify Tax Return Defaulters

Sanat Bhargava[1], M. Ravi Kumar[2], Priya Mehta[2], Jithin Mathews[2], Sandeep Kumar[2], and Ch. Sobhan Babu[2]

[1]Indian Institute of Technology Roorkee, Roorkee, India

[2]Indian Institute of Technology Hyderabad, Hyderabad, India

Abstract. Tax evasion refers to an entity indulging in illegal activities to avoid paying their actual tax liability. A tax return statement is a periodic report comprising information about income, expenditure, etc. One of the most basic tax evasion methods is failing to file tax returns or delay filing tax return statements. The taxpayers who do not file their returns, or fail to do so within the stipulated period are called tax return defaulters. As a result, the Government has to bear the financial losses due to a taxpayer defaulting, which varies for each taxpayer. Therefore, while designing any statistical model to predict potential return defaulters, we have to consider the real financial loss associated with the misclassification of each individual. This paper proposes a framework for an example-dependent cost-sensitive stacking classifier that uses cost-insensitive classifiers as base generalizers to make predictions on the input space. These predictions are used to train an example-dependent cost-sensitive meta generalizer. Based on the meta-generalizer choice, we propose four variant models used to predict potential return defaulters for the upcoming tax-filing period. These models have been developed for the Commercial Taxes Department, Government of Telangana, India. Applying our proposed variant models to GST data, we observe a significant increase in savings compared to conventional classifiers. Additionally, we develop an empirical study showing that our approach is more adept at identifying potential tax return defaulters than existing example-dependent cost-sensitive classification algorithms.

Keywords: goods and services tax, tax evasion, example-dependent cost-sensitive stacking classifier, example-dependent cost-sensitive ANNs,Benford's analysis, social network analysis, cosine similarity.

1 Introduction

Taxes can be classified into direct taxes, which are payable directly to the government (Eg. Income tax). These taxes cannot be transferred to any other third party, and indirect taxes, which can be shifted to a third party by the entity that is levied the tax (Eg. VAT, excise duty). The Goods and Services Tax (GST) system is an indirect taxation system introduced in India in July 2017. This paper proposes a methodology to predict potential tax return defaulters for the GST system [1].

1.1 Working of the GST system

For demonstration purposes, we take a fictitious ornament manufacturer as an example, and 10% as the GST rate levied at every step (See Figure 1). Note that throughout the paper, we

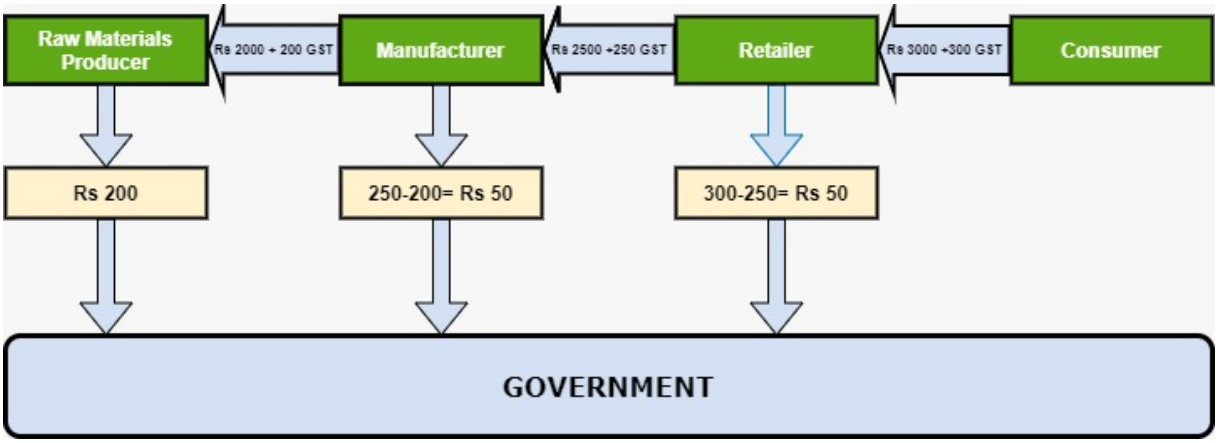


Figure 1. Flow of Tax in GST.

represent currencies in "Indian National *Rupees (INR)*," denoted henceforth as "Rs.". Assume the manufacturer purchases raw material worth Rs. 2000 and is hence levied a GST of Rs. 200 (10% of 2000). Suppose, the manufacturing process adds a value of Rs. 500 to the ornament. Hence, the value of the ornament is now Rs. 2500. Now, the total tax levied on the sales of this ornament to the retailer is Rs. 250 (10% of 2500). By setting off the tax he had already paid at the time of purchasing the raw material, the GST payable to the manufacturer is Rs. 50 (tax collected-tax already paid), $i.e.$, Rs. 50 (250-200). The retailer adds his margin of Rs. 500 increasing the total value to Rs. 3000 and sells it to the consumer for Rs. 3000, and the consumer is levied Rs. 300 as tax for the purchase. Similarly, the retailer is liable to pay a GST of Rs. 50 (tax collected - tax already paid), $i.e.$, Rs. 50 (300-250) at the time of purchasing the ornament from the manufacturer. Finally, the GST received by the Government is Rs. 300, which is completely borne by the end consumer.

1.2 Motivation for this work

In the GST system, taxpayers are required to file their tax returns once a month. Defaulting on filing tax returns has the following consequences: First, defaulters have enough time at their disposal to manipulate their records; second, the penalty imposed by the Government is negligible compared to the going interest rates of the market, and therefore not an effective deterrent. Lastly, having movable assets is always beneficial to businesses involving large monetary transactions. This work's motivation is to construct a classification model to identify potential return defaulters and implement preventive measures such as sending emails, SMS, or physically visiting their business premises. We are working with the Government of Telangana, India, and are using their data for analysis and building models to increase tax returns compliance.

To attack a classification problem such as the one presented here, one would be inclined to use conventional cost-insensitive classification algorithms such as Logistic Regression, K-Nearest Neighbors, $etc.$, to design the classifier. However, this presents us with a significant problem. A conventional classifier assigns equal misclassification costs for every example. In practice, however, the misclassification cost associated with classifying a genuine taxpayer as a return defaulter might vary significantly from classifying a return defaulter as an honest taxpayer. Similarly, the misclassification cost associated with misclassifying a return defaulter as a genuine taxpayer would vary for individual taxpayers based on their respective turnovers. Hence, there is a trade-off between choosing a model with better cost savings and choosing a model with better performance. To deal with this trade-off better, we propose four variants of an *example dependent cost-sensitive stacking classifier*. In a later section, we show that our Proposed Approach (PA) is adept at identifying potential tax return defaulters for the upcoming month with high accuracy. The approach that we have adopted in this paper can be generalized

to any indirect taxation system used globally.

Our paper is structured as follows. In Section 2, we brief on existing works that are related to ours. Section 3 describes the data set and the feature extraction techniques used for designing the model. In Section 4, we describe the framework for the proposed variants of the example-dependent cost-sensitive stacking classifier. Section 5 discusses the performance of the PA on the data set and compares it with some example-dependent cost-sensitive and conventional cost-insensitive classifiers commonly in use. Finally, in Section 6, we provide concluding remarks for our work.

2 Related Work

In [2], Jasmien Lismont et al. used social network analysis concepts to develop a model to predict tax avoidance by including a wider variety of network features. In [3], Bianchi et al. use network measures of centrality to show that the taxpayers who collaborate with better-connected auditors are likely to have lower effective tax rates. In [4], Veronique Van Vlasselaer et al. worked on identifying entities that indulge in social security fraud by assigning a time-dependent exposure score to each business entity based on its involvement with known fraud business entities in the social network. In [5], Yusuf Sahin and Ekrem Duman have built classification models for detecting credit card fraud using Logistic Regression and Artificial Neural Networks, one of the first studies to compare the performance of Logistic Regression and ANNs for this use case. In [6], Charles X. Ling and Victor S. Sheng showed that cost-sensitive learning is a common approach to solve data imbalance problems. In [7], A. C. Bahnsen et al. proposed an example dependent cost matrix for credit scoring. They proposed a cost function that introduces the example dependent costs into logistic regression. In [8], A.C Bahnsen et al. propose a framework for cost-sensitive classifiers, including Cost-Sensitive Decision Trees, Cost-Sensitive Random Forests, and ensembles of cost-sensitive models based on techniques such as majority voting and stacking Cost-Sensitive Logistic Regression generalizers. In [9], David H. Wolpert introduces Stacking or Stacked Generalization, an ensemble learning technique that aims to deduce generalizers' biases for the training set provided. In [10], Matjaz Kukar and Igor Kononenko designed a cost-sensitive analog for ANNs, with their study being the first to do so.

3 Data Description and Feature Extraction

3.1 Benford's law

Benford's law is a mathematical method for identifying fraud [11], [12],[13] in naturally-occurring numbers, considering that these numbers are neither highly constrained nor purely random. This law posits that the percentage of numbers with the first digit as $k \in \{1, 2, ..., 9\}$ follows the formula $log_{10}(1 + 1/k)$.

3.2 Description of the data set

We now proceed to briefly describe the data used to design our models. We were provided two types of data sets to prepare our models, namely: GSTR-1 data and month-wise GST returns data.

3.2.1 GSTR-1 Data

GSTR-1 is a financial statement that every taxpayer is required to submit monthly. This statement consists of details of all outward supplies, $i.e.$, all sales done during the month corresponding to this statement. A fictitious sample of this statement is given in Table 1. Every row in Table 1 corresponds to one transaction. The data set contains several millions of such rows. The actual statement contains more information, such as the tax rate, the number of goods sold

etc.

S.No.	Month	Seller	Buyer	Invoice Number	Amount (Rs)
1	Jul 2017	A	D	AB323	13000
2	Aug 2017	B	C	ZX362	16000
3	Sep 2017	B	A	BC9414	14490

Table 1. GSTR-1 Data

3.2.2 Monthly GST Returns Data

Table 2 contains a few select fields of GST returns data. Each row in this table corresponds to the monthly returns filed by a taxpayer. ITC (Input tax credit) is the amount of tax paid by the taxpayer during purchases of services and goods. The output tax is the amount of tax collected by the taxpayer during the sales of services and products. The taxpayer has to pay the Government the difference between the output tax and ITC, $i.e.$ (output tax – ITC). The actual dataset consists of much more information like tax payment method, return filing data, international exports, exempted sales, and sales on RCM (Reverse Charge Mechanism).

S.No.	Firm	Month	Purchases	Sales	ITC	Output Tax
1	D	Jul 2017	170000	250000	17000	25000
2	C	Oct 2017	230000	300000	11500	30000
3	F	Dec 2017	350000	450000	17500	45000

Table 2. Monthly GST Returns Data

3.3 Creation of Network of taxpayers

In this model, we have attempted to quantify the amount of interaction between taxpayers. To compute this independent variable, we created a weighted, directed graph (social network) in which each vertex (node) corresponds to a taxpayer. The weight assigned to the vertices is the average tax paid per month [$ATPM$] associated with the corresponding taxpayer from July 2017 to November 2019. Vertex weights have been normalized using min-max normalization. We have utilized the month-wise GST Returns Data explained in Table 2 to compute each taxpayer's vertex weights. We have placed a weighted, directed edge from taxpayer a to taxpayer b, where the weight of the edge is the amount of sales done by taxpayer a to taxpayer b during the period July 2017 to November 2019. Similar to the vertex weights, the edge weights have been normalized using min-max normalization. For the same, we have used the GSTR-1 data explained in Table 1. This graph captures the scale of interaction between taxpayers.

3.4 Feature Extraction

3.4.1 Ratio

This is the variable extracted from the weighted, directed graph defined in subsection 3.3. This graph captures the degree of interaction and the monetary transactions between taxpayer b and other taxpayers. This variable captures the influence of other taxpayers on b. If b has close ties with taxpayers who are known tax return defaulters, they will influence b not to file GST returns and vice-versa [2]. Let B be the set of all vertices corresponding to defaulters (who have filed at most $1/4^{th}$ of their returns) and Y be the set of all vertices corresponding to taxpayers who have filed their returns in time (who have filed more than $3/4^{th}$ of their returns in time).

- $b_{11} = \sum_{v \in B} \frac{\omega(v) * \omega(vb)}{\omega(v) + \omega(vb)}$, where $\omega(v)$ is the weight of vertex v and $\omega(vb)$ is the weight of directed edge vb
- $b_{12} = \sum_{v \in B} \frac{\omega(v) * \omega(bv)}{\omega(v) + \omega(bv)}$.
- $b_{21} = \sum_{v \in Y} \frac{\omega(v) * \omega(vb)}{\omega(v) + \omega(vb)}$, where $\omega(v)$ is the weight of vertex v and $\omega(vb)$ is the weight of

directed edge vb

- $b_{22} = \sum_{v \in Y} \frac{\omega(v) * \omega(bv)}{\omega(v) + \omega(bv)}$.
- $Ratio = \frac{b11+b12}{b21+b22}$.

3.4.2 Filed

This is the *dependent variable* in the model with a binary outcome. This variable gives the GST return filing status (whether the taxpayer has filed returns in-time or not) of the taxpayer b for December 2019. $Zero$ denotes returns were filed in-time (negative class) and, one denotes returns were not filed in-time (positive class).

3.4.3 Not Filed Count

This is the number of GST returns not filed in-time before the due date of the corresponding month by b from July 2017 to November 2019.

3.4.4 Division-Name

The state of Telangana is divided into 12 geographic divisions for simplification of administration works. This independent variable gives the geographic location in which b is located.

3.4.5 ATPM

This is the average tax per month paid by b. We included square, cube, log and the square root terms of the $ATPM$ in the model as the relation between $ATPM$ and *Log of Odds* of the Filed variable is a polynomial.

3.4.6 MAD Value

It is the Mean absolute deviation value of the first digit Benford's Law (Section 3.1) on sales transactions of b.

3.4.7 Seasonality

Case A: Retailers selling a single commodity:
In an actual market, the annual revenue of some businesses may show a seasonal trend. For example, for a taxpayer involved in the yogurt business, one might observe higher revenues in the peak summer (May-June in India) and lower revenues in the winter (November-February in India). To quantify this seasonality, we have calculated the cosine similarity between the *output tax* of each taxpayer selling a particular type of commodity and the mean of the *output tax* of all taxpayers selling that commodity.

$$similarity(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

Here A denotes a vector of the *output tax* for every month for each taxpayer selling a particular type of commodity, and B denotes the vector of the mean of the *output tax* of all taxpayers selling that commodity for every month.

Case B: Retailers selling multiple commodities:
In a more general case, a retailer might generate revenue by selling multiple commodities, and each commodity might have its own seasonal trend. Consider a retailer sells commodities from a set $I = A, B, C, D$. For this retailer, we calculate the seasonality parameter as follows

$$Seasonality = \sum \omega_i s_i, \forall i \in I$$

Here , ω_i weight associated with each commodity i, defined as
$\omega_i = \frac{\textit{Total revenue generated by sales of commodity i (for that retailer)}}{\textit{Total revenue generated by that retailer}}$,
$s_i =$ *Similarity of commodity i for that retailer*. Here, similarity is calculated as in case A.

3.5 Class Imbalance

The ratio of genuine taxpayers to defaulters was noted to be 0.22, hence, the use of sampling techniques was not deemed necessary.

4 Framework for Example Dependent Stacking Classifier

In problems such as the identification of tax return defaulters, it is of paramount importance to minimize the government's losses on account of the defaulters. For this task, an *example-dependent cost-sensitive classifier* would be the most prudent choice, as opposed to cost-insensitive classifiers [7]. Intuitively, one can deduce that a cost-sensitive classifier would minimize the total cost (or increase the total savings), compromising overall model performance. On the other hand, a cost-insensitive classifier would aim for optimal model performance while leading to higher losses to the government. It follows that there is a trade-off between higher cost savings and better model performance. To alleviate this problem, we propose a novel framework for example-dependent cost-sensitive stacked classifiers that give a competitive model performance and increased savings compared to cost-insensitive classifiers.

4.1 Stacked Generalizers and General Framework

Stacked Generalization or stacking is an ensemble learning technique that aims to improve upon the performance of its constituent generalizers by deducing the biases of each of the individual generalizers. However, stacked generalizers do not always perform better than individual generalizers, and their efficacy depends on the choice of generalizers. While there is no defined architecture for a stacked generalizer, it is observed that stacking is most effective when the choice of individual generalizers is as diverse as possible.
We propose a framework for a two-level stacked generalizer constructed as follows: The first level G_1 consists of conventional cost-insensitive classifiers to deduce the biases of classifiers on the input space, such that,

G_1 = {K-Nearest Neighbors Classifier, XGBoost Classifier, Random Forest Classifier, Logistic Regression, Artificial Neural Network, AdaBoostClassifier}.

The second level G_2, consists of the meta-generalizer, which generalizes on the second space, consisting of the predictions of G_1. We consider four choices of generalizers for G_2, which gives rise to the following four variants:

- **Variant A** (G_2=Cost Sensitive Decision Tree Classifier[8]),
- **Variant B** (G_2=Cost-Sensitive Bagging Classifier[8]),
- **Variant C** (G_2=Cost-Sensitive Random Forest Classifier[8]),
- **Variant D** (G_2=Cost-Sensitive ANN [10]).

The choice of the meta-generalizers was dictated by the savings score and the AUC-ROC score (Section 5.3.1). The models with the highest savings score and highest AUC-ROC score were chosen as meta-generalizers.

4.2 Meta Learners

4.2.1 Cost function

Let S be a set of N examples $\mathbf{x}_i$, where each example is represented by the augmented feature vector with associated costs $\mathbf{x}_i^* = [\mathbf{x}_i, C_{TP_i}, C_{FP_i}, C_{FN_i}, C_{TN_i}]$ and labelled using the class label y_i. A classifier f, which generates the predicted label c_i for each example i is trained using the set S. Then the cost of using f on $\mathbf{x}_i^*$ is calculated by

$$\begin{aligned} Cost(f(\mathbf{x}_i^*)) =& y_i(c_i C_{TP_i} + (1-c_i)C_{FN_i}) \\ &+ (1-y_i)(c_i C_{FP_i} + (1-c_i)C_{TN_i}), \end{aligned} \tag{1}$$

and the total cost defined as

$$Cost(f(S)) = \sum_{i=1}^{N} Cost(f(\mathbf{x}_i^*)). \tag{2}$$

C_{TP_i}, C_{FN_i}, C_{FP_i}, C_{TN_i} are defined in Table 3.

	Actual Positive $y_i = 1$	Actual Negative $y_i = 0$
Predicted Positive $c_i = 1$	C_{TP_i}	C_{FP_i}
Predicted Negative $c_i = 0$	C_{FN_i}	C_{TN_i}

Table 3. Cost Matrix

4.2.2 Variant A

For variant A, we have G_1 as defined above, and we use G_2= Cost-Sensitive Decision Tree Classifier (CSDT) [8]. In CSDTs, instead of using traditional splitting criteria such as Gini, entropy, or misclassification, the cost as defined in (1) is calculated for each node, and the gain of using each split is evaluated as the decrease in the total cost of the algorithm. The cost-based impurity measure is defined by comparing the costs when all the examples in a leaf are classified as negative and as positive,

$$I_c(S) = min\left\{Cost\left(f_0\left(S\right)\right), Cost\left(f_1\left(S\right)\right)\right\}.$$

Then, using the cost-based impurity, the gain of using the splitting rule $(\mathbf{x}^j, l^j)$, that is the rule defined as splitting the set S on feature $\mathbf{x}^j$ on value l^j, is calculated as

$$Gain(\mathbf{x}^j, l^j) = I_c(S) - \frac{|S^l|}{|S|} I_c(S^l) - \frac{|S^r|}{|S|} I_c(S^r),$$

where $S^l = \{\mathbf{x}_i^* | \mathbf{x}_i^* \in S \wedge x_i^j \leq l^j\}$,$S^r = \{\mathbf{x}_i^* | \mathbf{x}_i^* \in S \wedge x_i^j > l^j\}$, and $|\cdot|$ denotes the cardinality. Afterward, a decision tree is grown using the cost-based gain measure until no further splits can be made. After the tree is constructed, it is pruned by using a cost-based pruning criteria

$$PC_c = Cost(f(S)) - Cost(f^*(S)),$$

where f^* is the classifier of the tree without the pruned node.

4.2.3 Variant B

For variant B, we have G_1 as defined above, and we use G_2= Cost-Sensitive Bagging Classifier (CSB). Bagging or Bootstrap Aggregating is an ensemble technique that involves fitting base estimator(s) to random samples of the data set. The individual predictions are then aggregated using *majority voting* or *weighted average* to form a final prediction. To build our CSB, we have used the CSDTs mentioned above as base estimators and aggregated the individual predictions using majority voting [8].

4.2.4 Variant C

For variant C, we have G_1 as defined above, and we use G_2= Cost-Sensitive Random Forest Classifier (CSRF). Cost-Sensitive Random Forest Classifiers are ensemble classifiers that work by creating multiple CSDTs and outputting the mode of the predictions made by the CSDTs as the final prediction of the ensemble classifier [8].

4.2.5 Variant D

Finally, we have implemented a Cost-Sensitive analog for an Artificial Neural Network [10]. To design our Cost-Sensitive ANN Classifier (CSANN), we have used the ReLU function as the activation function for the hidden layers and the logistic (sigmoid) function for the output layer. We have used equation (1) as the loss function for the neural network to incorporate the example-dependent cost-sensitive losses.

5 Experimental Results

5.1 Software Used

All the models in this work have been designed using **Python** as it is a high-level, open-source language with an extensive library ecosystem. Python can also handle large amounts of data very well.

5.2 Cost Matrix

Table 3 gives different miss-classification costs of a given taxpayer.

- **True-negative cost ($\mathbf{C_{TN}}$)** is zero. We would not incur any cost for classifying an in-time return filer (actual class zero) as an in-time return filer (predicted class zero).
- **True-positive cost ($\mathbf{C_{TP}}$)** is the expenditure towards sending SMS, calling the taxpayer and other preventive measures and the cost of associated manpower. This cost is the same for all taxpayers whose actual class is one and predicted class is one. This cost is Rs. 150.
- **False-positive cost ($\mathbf{C_{FP}}$)** is the expenditure towards sending SMS, calling the taxpayer and other preventive measures and the cost of associated manpower. This cost is the same for all taxpayers whose actual class is zero and predicted class is one. This cost is also Rs. 150.
- **False-negative cost ($\mathbf{C_{FN}}$)** depends on the $ATPM$ of each taxpayer and the expected number of days of delay in filing return by a taxpayer. This is given by $\frac{ATPM * \textit{expected number of days of delay} * 18}{36500} * 3 + 100$.

Here $\frac{ATPM * \textit{expected number of days of delay} * 18}{36500}$ is the loss incurred due to late filing of return, where interest rate is 18%. This cost is different for every taxpayer as the $ATPM$ and expected number of days of delay in filing return may vary for each individual taxpayer. We have multiplied this loss by three times and added 100 to it, in order to deter a defaulter from becoming a chronic defaulter.

5.3 Performance of Proposed Variants

In this section, we have compared the four proposed variants (variants A, B, C, and D) on tax return data vis-à-vis each other. The models have been compared on the following metrics:

5.3.1 Savings score

The savings score is defined as the relative improvement in cost using a classifier $f(S)$, compared to the cost of classifying all entries as class one or class zero, whichever is lesser [7].

$$Savings(f(S)) = \frac{Cost(f(S)) - Cost_l(S)}{Cost_l(S)}.$$

5.3.2 Balanced Accuracy Score

The balanced accuracy score is a metric for models trained on imbalanced data sets, which avoids inflated performance metrics due to the abundance of one class (in a binary classification

problem). It is defined as follows:

$$\textit{Balanced accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right).$$

5.3.3 Recall

Recall or Recall score refers to the fraction of relevant records correctly classified by the models. It is defined as follows:

$$Recall = \frac{TP}{TP+FN}.$$

In the context of this paper, the recall score is the fraction of tax return defaulters correctly identified by the model.

5.3.4 F1-Score

The F1-Score is defined as the harmonic mean of the precision and recall of a model. Thus,

$$\textit{F1-Score} = 2 * \left(\frac{precision * recall}{precision + recall}\right).$$

The comparative performance of the four variants is summarized in the Table 4. From the four variants, we propose Variant D (G_2=Cost-sensitive ANN) to be our proposed approach (PA) for this data set as it is the most adept at correctly predicting tax return defaulters, with the highest savings score and the highest AUC-ROC predicted on the train and test data set among the four variants.

Proposed Models	Savings Score		Balanced Accuracy		F1-Score		Recall		AUC-ROC	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Variant A	0.536	0.174	83.36%	82.23%	0.76	0.75	0.95	0.94	0.92	0.92
Variant B	0.535	0.530	83.90%	83.82%	0.77	0.77	0.91	0.90	0.93	0.94
Variant C	0.583	0.572	85.58%	85.94%	0.83	0.83	0.90	0.91	0.94	0.94
Variant D	0.520	0.582	85.14%	85.21%	0.79	0.79	0.94	0.95	0.94	0.93

Table 4. Performance of variants

Proposed Models	Savings Score		Balanced Accuracy		F1-Score		Recall		AUC-ROC	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
ANN	0.272	0.242	84.11%	83.40%	0.69	0.66	0.76	0.74	0.93	0.93
CSANN	0.393	0.440	84.31%	84.29%	0.84	0.84	0.84	0.84	0.93	0.93
CSDT	0.610	0.600	81.52%	79.00%	0.84	0.84	0.78	0.71	0.93	0.94
CSB	0.557	0.633	82.00%	78.50%	0.83	0.83	0.83	0.83	0.94	0.94
CSRF	0.495	0.517	79.15%	83.34%	0.52	0.62	0.93	0.91	0.91	0.92
Proposed Approach	0.520	0.582	85.14%	85.21%	0.79	0.79	0.94	0.95	0.94	0.93

Table 5. Performance of PA compared to existing algorithms

5.4 Performance of Proposed Approach (PA) with existing algorithms

In this section, we have compared our proposed approach's performance with some cost-sensitive algorithms mentioned in [8]. Additionally, we compare the performance of the PA with a cost-sensitive ANN [10]. We have also compared the performance of our PA with a cost-insensitive ANN. We have chosen a cost-insensitive ANN as it gave the most promising results

among various cost-insensitive algorithms we experimented with, including, KNNs, Random Forests, XGBoost Classifier, AdaBoostClassifier, and Logistic Regression. The performance has been compared using the same metrics described in section 5.3. The results have been summarized in Table 5.

5.5 Model Validation for PA

5.5.1 Confusion and Cost Matrices

Tables 6 and 7 are the training and the testing confusion matrices for the PA. Tables 8 and 9 are the training matrix and the testing cost matrix for the PA. These give the true-positive cost, false-negative cost, true- negative cost, and false-positive cost of both the training and testing data sets.

	Predicted 0	Predicted 1
Actual 0	9842	2946
Actual 1	196	2742

Table 6. PA Train Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	3320	1006
Actual 1	48	930

Table 7. PA Test Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	0	441900
Actual 1	276796	411300

Table 8. PA Train Cost Matrix

	Predicted 0	Predicted 1
Actual 0	0	150900
Actual 1	68217	130200

Table 9. PA Test Cost Matrix

5.5.2 Training and Testing ROC Curves

Training and testing ROC curves for the PA are given in Figure 2 and 3. The AUC value of training ROC curve is 0.94 and AUC value of testing ROC curves is also 0.93. From these values, one can conclude that the model is neither under-fitting nor over-fitting.

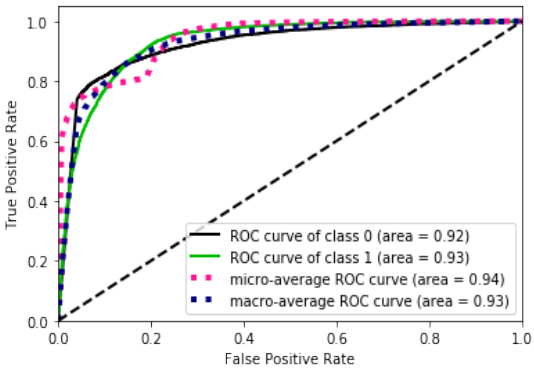


Figure 2. PA ROC on Train.

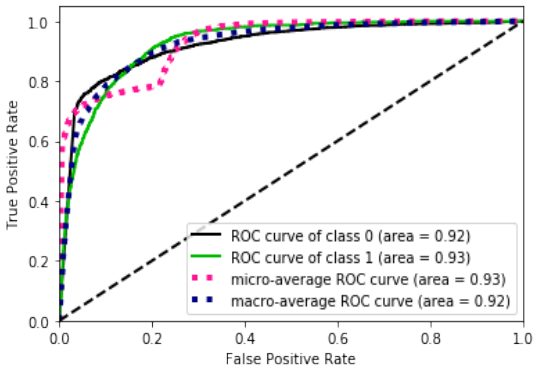


Figure 3. PA ROC on Test.

5.5.3 Savings score

To measure an example-dependent cost-sensitive algorithm's performance, we use the savings score (Section 4.3). As observed in Table 5, the savings score for the PA is 0.520 and 0.582 for the training and test sets, respectively. Since the values of the savings score for the training and testing set are reasonably high and almost similar, we can conclude that our PA is performing well.

6 Conclusion

In this paper, We propose a framework for example-dependent cost-sensitive stacked generalization comprising four variant models. We show that our Proposed Approach (PA) outperforms commonly used example-dependent cost-sensitive classifiers. We use our PA to predict whether a given taxpayer is a potential tax return defaulter or not for the upcoming month. While this framework was designed on the GST returns data for Telangana, it can be generalized to predict potential tax return defaulters using any of the four proposed variants depending on their performance, for any indirect taxation system around the world.

References

[1] S. Dani, "A research paper on an impact of goods and service tax (gst) on indian economy," *Business and Economics Journal*, vol. 07, Jan. 2016. DOI: 10.4172/2151-6219.1000264.

[2] J. Lismont, E. Cardinaels, L. Bruynseels, S. D. Groote, W. Lemahieu, and J. Vanthienen, "Predicting tax avoidance by means of social network analytics," *Decision Support Systems*, vol. 108, pp. 13–24, 2018.

[3] P. A. Bianchi and M. Minutti-Meza, "Professional networks and client tax avoidance: Evidence from the italian statutory audit regime," *SSRN Electronic Journal*, Jan. 2016. DOI: 10.2139/ssrn.2601570.

[4] V. V. Vlasselaer, L. Akoglu, T. Eliassi-Rad, M. Snoeck, and B. Baesens, "Guilt-by-constellation: Fraud detection by suspicious clique memberships," in *2015 48th Hawaii International Conference on System Sciences*, 2015, pp. 918–927.

[5] Y. Sahin and E. Duman, "Detecting credit card fraud by ann and logistic regression," in *2011 International Symposium on Innovations in Intelligent Systems and Applications*, 2011, pp. 315–319.

[6] C. Ling and V. Sheng, "Cost-sensitive learning and the class imbalance problem," *Encyclopedia of Machine Learning*, Jan. 2010.

[7] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive logistic regression for credit scoring," in *2014 13th International Conference on Machine Learning and Applications*, 2014, pp. 263–269.

[8] A. C. Bahnsen, D. Aouada, and B. Ottersten, *Ensemble of example-dependent cost-sensitive decision trees*, 2015. arXiv: 1505.04637 [cs.LG].

[9] D. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, Dec. 1992. DOI: 10.1016/S0893-6080(05)80023-1.

[10] M. Kukar and I. Kononenko, "Cost-sensitive learning with neural networks," in *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI-98*, John Wiley & Sons, 1998, pp. 445–449.

[11] M. L. J. Nigrini Mark J., "The Use of Benford's Law as an Aid in Analytical Procedures," *Auditing: A journal of practice & theory*, vol. 41, p. 52, 1997.

[12] A. Asllani and M. Naco, "Using Benford¡¯s Law for Fraud Detection in Accounting Practices," *Journal of Social Science Studies*, vol. 2, no. 1, pp. 129–143, Jan. 2015. [Online]. Available: https://ideas.repec.org/a/mth/jsss88/v2y2015i1p129-143.html.

[13] C. Durtschi, W. Hillison, and C. Pacini, "The effective use of benford's law to assist in detecting fraud in accounting data," *J. Forensic Account*, vol. 5, Jan. 2004.

24th International Conference on Business Information Systems (BIS 2021)

ICT Project

https://doi.org/10.52825/bis.v1i.43

Published: 02 July 2021

Development of the Information Security Management System Standard for Public Sector Organisations in Estonia

Mari Seeba[1, 2] [https://orcid.org/0000-0002-9066-2467],Raimudas Matulevičius[1, [https://orcid.org/0000-0002-1829-4794], and Ilmar Toom[2]

[1]Institute of Computer Science, University Of Tartu, Estonia

[2]Estonian Information System Authority, Tallinn Estonia

Abstract. Standardisation gives us a common understanding or processes to do something in a commonly accepted way. In information security management, it means to achieve the appropriate security level in the context of known and unknown risks. Each government's goal should be to provide digital services to its citizens with the acceptable level of confidentiality, integrity and availability. This study elicits the EU countries' requirements for information security management system (ISMS) standards and provides the standards' comparison requirements. The Estonian case is an example to illustrate the method when choosing or developing the appropriate ISMS standard to public sector organisations.

Keywords: Information Security Management System, ISMS, Public Sector, Requirements of Security Standards, Estonia

Introduction

Standardisation aims to optimise the process management, compare defined objects with each other, enable integration and interoperability of systems, cost optimisation and preparedness to adapt to new situations [1]. There are standards designed for information security management systems (ISMS) as well (few examples are [20, 21, 22]). In private organisation the management decides which ISMS standard to follow based on organisation requirements. At the national level the stakeholders' objectives and the national characteristics (e.g. unique technologies such as the X-tee [2] or electronic identity solutions [3]), and cultural and linguistic peculiarities should be considered independently of each organisation requirements. There is also a need for the standard long-term central maintenance and reduction of administrative costs, or compliance with the regulations (e.g. EU GDPR [5]). At the national level the ISMS standard must ensure a comprehensive national defence and systems interoperability carried out by each organisation. EU regulation (NIS Directive [6]) defines t h e c r oss-union incident management and information sharing rules, but it does not provide the information security management framework for public sector organisations.

There is no standardised method or requirements on how to compare and show different approaches of the ISMS standards for public sector organisations at the national level. This method should consider the standards substantive comparison, the national security strategic objectives, and external interested parties' requirements or abilities. On the national strategic level this method can support decision makers, and also security specialists to find relevant

arguments when choosing or planning to create an ISMS standard. This study aims to investigate *what are the requirements to develop information security management standards for public sector organisations at the national level.*

The paper is motivated by the development of the national ISMS standard for the Estonian public sector organisations. In this study we identify and structure the requirements for national ISMS using 12 EU national cybersecurity strategies. Then we share the example of how Estonian ISMS requirements can be structured using our study approach. Using the elicited requirements we compare three ISMS standards and illustrate how the assessment of the ISMS standards with the elicited requirements can be done based on the Estonian case. Our experience shows that the comparison of the elicited and sorted requirements and ISMS standards is a possible way that can be followed by the other countries that are looking for ISMS standards or framework for public sector organisations.

The paper is structured as follows: Sect. 1 gives an overview of the Estonian case and related work. Sect. 2 describes the research method. Sect. 3.1 guides the ISMS standards requirements elicitation and structuring and Sect. 3.2 illustrates the use of requirements in comparison of standards and presents the results with the Estonian case. Finally, Sect. 4 concludes the paper with the results and limitations.

1 Background

1.1 Case Description

Estonia is an EU country with 1.33 million inhabitants. Estonia is known for its digital society imago and with the successful response to the first large-scale cyberattack against the entire state [15]. Estonian citizens, e-residents and organisations can use or provide more than 2860 digital services via eGovernment supported Data Exchange Layer X-tee (Estonian instance of the X-Road). More than 150 million requests per month are made via X-tee [16]. Majority of the transactions are made between public sector organisations. This context requires a clear understanding and mutual recognition of information security from the data exchange partners and data processors. The Estonian first version of information security management baseline standard called ISKE was developed and published in 2004 [27]. Now Estonia is developing its new national ISMS standard. In this paper we use Estonian case to illustrate how the elicited requirements for national ISMS can be used.

1.2 Related Works

We investigated the studies dealing with requirements to the ISMS standards and standards comparison.
European Union Agency for Cybersecurity (ENISA) certification standards review report [12] is indispensable to understand the origin and functioning of standardisation organisations. The report is focused on certification, and provides assessment guidance on the certification schemes, but it does not provide direct input to the comparison of standards.

EU SPARTA project includes the overview of the security-related certification initiatives and the related standards at the national and international level, as one of it's deliverable [10]. It's aim is to inform project partners about available standards that the project partners can consider certifying their project deliverable against. The report does not follow any exact requirement or comparison requirement.

Pertinent collection of security standards are systematised by standardisation bodies authority, jurisdiction, applicability, document type and standards examples in [8]. This overview did not describe the requirements to follow or which characteristics of the standards to compare.

Overviews and summaries of standards can be found from security blogs or websites of the

consulting companies. A similar descriptive approach can be found in [9]. The paper covers ISO security-related standards and mentions the Information Security Forum (ISF) Standard of Good Practice for Information Security, COBIT (ISACA framework) and BSI IT-Grundschutz (IT baseline protection). This work only describes the standards, not focusing on the requirements or comparison.

A systematic approach to the content analysis of the standards can be found in [7], where the authors have created the conceptual model for security standards and provide the template for the standards content comparison. Their approach can help organisations, but do not help at the national strategic level.

By standards web-sites, the content comparison is provided for standards compliance confirmation. Usually, there are tables where each row represents similar control of comparable standards [24, 28]. These comparisons provide the sentence-by-sentence compliance confirmation on standard contents, but do not deal with other properties of the standard.

Finnish report [11] compares the cybersecurity situation of eight countries on the state level. The report provides a comparison of economical, educational, legal and social aspects of cybersecurity, and names the approaches of these eight countries. The report helped us to consider the relevant areas of the countries cybersecurity strategies.

The Estonian case can be illustrated with studies conducted in 1998 and 2003, which analysed the national security needs and security specialists ability to manage ISMS standards. The studies concluded, that Estonia needs baseline security with granular security measures catalog. [4] The same statements apply in today's Estonia [19]. ENISA report [13] compares 28 EU state cybersecurity strategies and has identified that one common strategic objective is to establish baseline security measures to harmonize the security practices in the public and private sector. Report did not create requirements for that.

Related works showed several approaches on how to compare the security standards and gave some overview of the standards. The related works did not give any suggestions or requirements on how to choose ISMS standards for public sector organisations on the national strategic level. Also, we revealed that national cybersecurity strategies could be an appropriate source of requirements elicitation for ISMS standards.

2 Research Approach

The research demonstrates the requirements elicitation when developing the ISMS standard, illustrated using the Estonian case presented in Sect.1.1. The paper's goal is to answer the question **RQ:** *what are the requirements to develop information security management standards for public sector organisations on the national level?* The research question can be divided into two subquestions: **RQ1:** how to find and what are the countries requirements to the ISMS standard? **RQ2:** how to use these requirements when developing the national ISMS standard?

Our research process is case-oriented and is illustrated in Fig. 1. We conducted two parallel processes. Firstly, theoretical approach is used to elicit requirements for ISMS standard (activity 1.1) at the national level. It is based on the National Cybersecurity Index (NCSI) [14] database (input 1.1.a) to answer RQ1. The structured result of ISMS standard requirements (artefact 1.1.b) were used to elicit the Estonian ISMS standard requirements (activity 2.1 and artifact 2.1.a). Secondly, activity 1.2 uses the output of 1.1.b to compare ISMS standards to answer the RQ2. Activity 2.2 illustrates the ISMS standards' comparison (1.2.a) in the case process and results in 2.2.a.

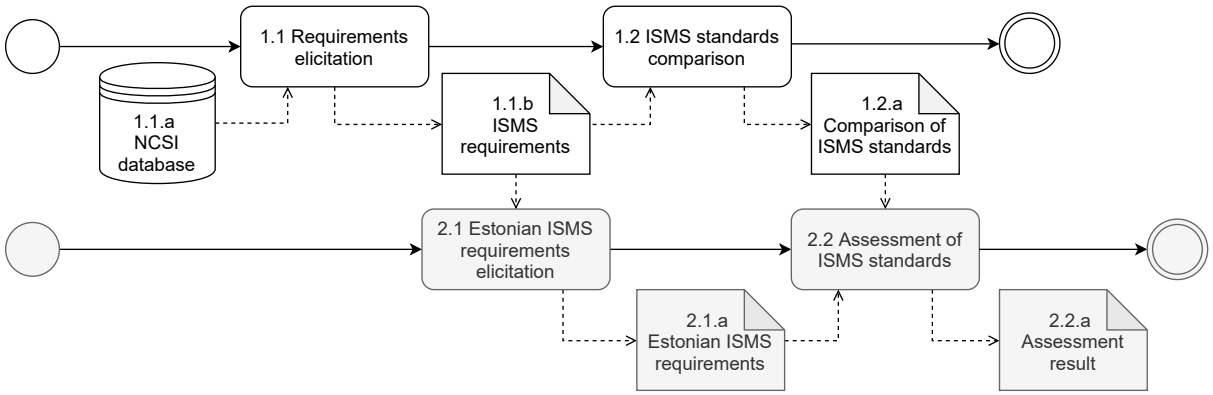


Figure 1. Study approach

3 Elaboration of ISMS Standards Requirements

3.1 Requirements Elicitation

NIS Directive [6] requires the EU member states to create and maintain national cybersecurity strategy and its implementation plan. National cybersecurity strategy is the fundamental source document for acceptable requirements of ISMS standards of the country among other strategic objectives.

For the security standard requirements elicitation we used the NCSI [14] database developed by the eGovernance Academy. eGovernance Academy collects links with publicly available evidence material of each country's cybersecurity documents [14]. We wrote out the ISMS standard's required properties of NCSI TOP 12 EU country's cybersecurity strategy and implementation plan. Then we collected similar requirements under one requirement. We generalised the elicited requirements to cover different countries' needs simultaneously. The requirements pass on the nature of the requirement, not the exact initial wording. Each requirement received the characteristic keywords. Finally, we got 15 requirements. We grouped the elicited requirements into three modules (see Table 1):

- **National security module** determines the national security aspects like compliance with jurisdiction regulations and the national authority right to make or influence to make changes into the content of the standard. This module allows assessing the possible future cost related to adoption and maintaining the ISMS standard. The target group of these requirements are the organisations responsible for ISMS standard development and maintenance on the national level.
- **Content module** helps to get to know the standard usability and adaptability issues related to implementation barriers and complexity. *Basic Controls* and *Levelled Controls* help to understand the implementation possibilities depending on the security needs. *Technology Dependence* and *Adaptability with National Needs* describe the flexibility of the standard controls. *Risk Management Approach* shows how risk management is included in the standard or requires separate management. The target group of these requirements are the organisations who have to implement the standard.
- **Assessment module** requires the monitoring and auditing capabilities to assess the organisation's information security. The module characterises the needs and requirements outside the public sector. It is necessary to consider the availability and cost of resources like external certified auditors and audit bodies (target group of the module).

Each requirement received a unique ID (Nx, Cx or Ax) where x is a requirement sequence number and letter corresponds to the module the requirement belongs to. The county code in Table 1 shows the origin owners(s) of the requirement.

Table 1. National cybersecurity strategy requirements for ISMS standards

Req ID	Requirement	Requirement description	Country Code
National security module			
N1	**Developer and jurisdiction**	Standard should take into account EU and NATO regulations.	FI, GR, LT, HR
N2	**Development financing**	It should be possible to influence the development of the standard by national authority.	FI, GR, LT
N3	**Licence conditions**	Standard should be freely available to all national implementer.	FI, LT
N4	**Language**	Standard should be available in national language.	BE, GR, LT, LV
N5	**Update cycle**	Standards should be improved continuously/regularly.	BE, ES, GR, HR
Content module			
C1	**Scope**	Standard should be usable by public/private sector organisations information systems / processes / assets / critical infrastructure.	BE, CZ, ES, FI, GR, LT, LV, PL, SK, HR
C2	**ISMS compliance**	Standard should be compliant with internationally recognised standards / frameworks / best practices.	BE, CZ, ES, FI, FR, GR, LT, HR
C3	**Basic controls**	Standard should include basic/minimum security controls/measures.	BE, CZ, ES, FI, GR, LT, LV, PL, NL, HR
C4	**Leveled controls**	It should be possible to implement the standard controls/measures depending on the security level.	CZ, ES, FI, GR, LT, LV, PL, HR
C5	**Risk management approach**	Standard should include risk management.	BE, CZ, ES, GR, LT, LV, SK, HR
C6	**Technology dependence**	Standard should be technology-independent.	PL
C7	**Integrability of local needs**	It should be possible to adapt the standard with the national technological needs.	GR, PL
C8	**Controls approach**	It should be possible to change the content of the standard by national authority.	FI, GR, LT, PL
Assessment module			
A1	**Auditability**	Standard implementations should be auditable/assessable.	BE, CZ, ES, FI, GR, LV, PL, SK, HR, NL
A2	**Certification Schema**	Standard should be certifiable for being in compliance with recognized standards.	GR, PL, HR, NL

Estonian case ISMS requirements. We elicited Estonian requirements to ISMS standard from the Estonian Cybersecurity Strategy for 2019-2022 [17], the long-term Information Society Development Plan of Estonia (IÜAK) [18], and new ISMS standard procurement document [19]. These sources take into account the requirements of information security regulations.

Identified Estonian requirements are sorted according to Table 1. The result is given in the Table 3 columns *ReqID* and *Estonian Requirements*. Some of the identified Estonian requirements have been collected under same requirement ID, as their final objective is similar (e.g. N4, C1, C2). Also, some are mentioned more than once under several requirements, because they serve several goals (e.g., N2, C8 - one of them requires the possibility to make changes in the standard, the other requires controls approach and flexibility to add national aspects).

3.2 ISMS Standards Comparison Example

By following the requirements in Table 1, we compared the three following ISMS standards:

- **ISO27001** *ISO/IEC 27001:2013 Information technology — Security techniques — Information security management systems — Requirements* [20], developed by international level standardisation body and recognised globally.
- **CIS20** *CIS Controls v 7.1* [21], developed by industrial body, focuses only on information security. CIS20 provides TOP 20 security measures for organisations.
- **BSI ITG** *BSI IT-Grundschutz Kompendium* [22], differs from previous standards by its included threats, requirements and security controls catalogues. BSI ITG is known as a baseline security framework which is developed by an EU member state national standardisation body.

Standards content comparison CIS20 has published separate web articles of *CIS Mapping and compliance* to provide the control-by-control mappings to ISO27001, GDPR, and some industry specific frameworks [24]. BSI has published the analysis of BSI Standards and Kompendium compliance from the ISO27001 perspective [28]. These compliance confirmation publications assert that through ISO27001 perspective, three comparable standards contents cover the same security areas and are compliant to each other's security objectives.

Standards comparison based on elicited requirements ISMS standards comparison results are presented in Table 2. The table gives a one-page overview of the similarities and differences of standards.

As the standard-setting similarities, we point out that the ISO27001 requirements and security objectives are reflected in other standards (C2). The standards are, thus, consistent with the security areas content. All three standards are intended to be used by a wide user community and do not impose restrictions to organisations by size, sectorality or industry field (C1). The introduction of risk management is required by all standards (C5). For all standards, there is one basic document supported by additional documents. It must be taken into account that the implementer must have all documents available (to take into account the cost to translation, maintenance, license fees) (C8, N4, N3, N5). None of the standard imposes restrictions on technologies directly (C6). An auditing and certification approach based on ISO27001 is suitable for all standards (A1, A2).

When deciding about standards, however, differences between standards become critical. For example, the chosen standards are part of different legal jurisdictions (N1) and there are also different funding schemes (at the moment: global, US, EU) (N2). Often, just through financing, it is possible to influence the content of the standards. This is important for national security considerations. The financing schemes of those three given standards differ by financier (national bodies, donations or state government) (N2). From the public sector's perspective, it could be a problem if the standard has a license fee (ISO27001) and is not freely available (N3). To assess the standard dynamics or statics we can compare the update cycle

Table 2. ISMS Standards Comparison

Req ID	ISO27001	CIS20	BSI ITG
National security module			
N1	International Standardisation Organisation (Switzerland), globally recognised	Centre for Internet Security (US based non-profit organisation), US industrial, wide adoption	Federal Office for Information Security (BSI) (Germany), German national EU jurisdiction
N2	National bodies participate in development and finance ISO. Sale of standards. [25]	Contributors: US agencies, commercial partners. Financing: donations, grants, paid programs, product sales [26]	Publicly reviewed contributions. Financing: German Gov.
N3	User based fee (also to translated versions)	Free for registered users, Creative Commons	Free download
N4	20+ languages	English, Spanish, Italian, Japanese, Lithuanian, Estonian	German, English
N5	5 year cycle	No exact rule, expectation is yearly update	Every February 1st
Content module			
C1	No limitations	No limitations	No limitations
C2	Officially compliant with ISO/IEC Management system standards, Management system standards adopted from Annex SL of ISO/IEC Directives, Consolidated ISO Supplement.	ISO 27001, NIST Framework [23]	ISO 27001
C3	Requirements mandatory, objectives with justified exclusions	User profile Implementation Group (IG) based basic requirements	Basic protection
C4	No	Three IG based levels	Standard and High level
C5	Mandatory. Guidelines: ISO/IEC 27005, ISO 31000	Guidelines: CIS RAM, ISO 27005, NIST SP 800-39, RISK IT (ISACA)	Embedded. Extension: BSI Standard 200-3: Risk Management
C6	No	No	User profile based technology modules
C7	Through risk management, local implementation	Through risk management, local implementation	Through risk management, central new technical modules development. Process modules are compliant to German regulations
C8	Control objectives (14) and controls (114). Related: ISO27000 series (50+ standards). Important: ISO/IEC 27000, ISO/IEC 27002, ISO/IEC 27003, ISO/IEC 27004, ISO/IEC 27005	Security mode: 3 Implementation Groups. Controls (20), subcontrols (171). Related: CIS Controls TM, CIS RAM	Security mode: Basic, Standard, Core. Security catalogue: process and technical modules(5+5), Submodules (94), 1680+ requirements and measures in modules. Related: IT-Grundschutz Compendium; standards BSI 200-1, 200-2,200-3; BSI 100-4
Assessment module			
A1	External audit based on ISO 27007	Self-assessment or auditing based on ISO27001 or other standards	External audit
A2	Based on ISO 27006, ISO 27007, ISO 27008	No	Based on ISO27001 requirements and BSI methodology

of standards (N5).

Organisations have different security needs and they are looking for matching security levels to optimise the security cost. So the organisations with lower security needs do not have to implement all the high-level measures. CIS20 and BSI ITG provide leveled approach (C3, C4). The volume of the guidance material can drive the usability of the standard (C8).

If ISO27001 and CIS20 are technology-free, then BSI ITG offers security measures suitable for the most common technologies (C6). Everyone can propose suitable profiles for the BSI, and if there exists a general approval, they will be integrated within a year into the composition of standard catalogues (C7, N5).

To summarize our comparison, the decision-maker should understand the differences and similarities of the standards, consider separately national security aspects (first module) and standards' content aspect (second module), and to weigh, how the auditing and certification schemes (third module) could work, and which resources are needed.

Estonian case standards assessment. From the perspective of the Estonian ISMS standard development it is important to compare the Estonian requirements with ISMS standards. In Table 3 we align the Estonian ISMS standard requirements with the compliance assessment to the three previously described ISMS standards (see Table 2). The qualitative sequence method has been used for the assessment: the most suitable standard in compliance with concrete Estonian requirement(s) is marked as "++", suitable with some exclusions is marked as "v" and not suitable is marked as "0", N/A is marked as "-". We used the assessment mark "+" for interim cases of "++" and "v". The result shows the differences between the standards in the National security module in Table 3. In the Content module the BSI ITG stands out with its positive results. In Estonian case the Assessment Module probably could not influence too much the decision making. The case shows that for Estonian public sector organisations, the most suitable standard to use is BSI ITG based standard.

4 Limitation and Conclusion

We investigated the national cybersecurity strategies and their implementation plans for requirements elicitation in their original languages using the Google Translate application (when needed). We avoided the progressing of the errors caused by machine translation by including the requirement in case ambiguity only if it appeared in both sources.

Second aspect to mention is that the national cybersecurity strategies are written in different detail and maturity levels. For example, the Greek documents covered 14 requirements out of 15, while we found only one requirement for the French public sector security. In order to bring the elicited requirements to the same maturity level, we ruled out very specific requirements for security measures and generalised them under Requirement ID C7. Also, the requirements are not with equal importance to national states. We suggest to assess them in the context of national objectives.

In the study, we elicited the ISMS requirements for public sector organisations in a form that supports reuse of the structured requirements. We used the structure of elicited requirements to compare three ISMS standards. In the example of the Estonian case, we showed how to compare requirements and standards. The result could be useful for small national states which wish to use the experiences and existing ISMS standards of other countries to develop their information security measures.

During the study, we perceived that all EU countries are simultaneously developing their standards or frameworks. Our working group came to the same conclusion with the ENISA report [13]. Hence the ENISA or other EU organisation could develop a central framework or baseline for public sector organisations security management, and each country could adapt

Table 3. ISO27001, CIS20 and BSI ITG standards assessment based on requirements to Estonian ISMS standard

(Notation: “++” - most suitable; “v” - suitable with some exclusions; “0” - not suitable; “-” - N/A; “+” - interim cases of “++” and “v”)

Req ID	Estonian Requirements	ISO27001	CIS20	BSI ITG
National security module				
N1	Standard should enable the baseline security to fulfil requirements of national and international regulations like GDPR, NIS-directive, etc. [17].	v	0	+
N2	Standard should be flexible enough to add national content, measures or modules [19].	v	v	+
N3	Standard should be available free of charge [19].	0	+	++
N4	The standards must transfer Estonian language and culture, i.e. be in correct language, terminologically validated and compiled for Estonians [17]. Correct language and consistent terminology should be used and validated [19].	++	v	0
N5	Standard should be updated regularly/yearly [19, 17].	v	v	++
Content module				
C1	Information security should be integrated widely in all type of organisations and their processes [17]. Standard should be extendable for all public administration and industry organisations [17]. Standard should support public sector business processes [19].	++	++	++
C2	Standard should be based on an European or internationally recognised standards and practices [17, 6]. In case of a translation adoption, the standard should retain the connections with original document sets [19].	+	v	++
C3	Standard should help optimising risk management by providing predefined measures for typical solutions [19].	0	v	++
C4	Implementation process should enable levels of implementations - the base implementation and advanced levels based on security requirements [19].	0	+	++
C5	Standard should use and adopt risk based approach for information and network security management [17].	++	++	++
C6	All technologies should be given equal opportunities regardless of the platform [17].	++	++	+
C7	The obligation to use Estonian based technological solutions. Therefore, the standards must enable and propagate the use of X-tee and Estonian public key infrastructure (PKI) solutions. [19]	+	+	v
C8	Standard should be flexible enough to add national content, measures or modules [19].	0	0	+
Assessment module				
A1	Standard should allow audit-ability [19].	++	v	++
A2	-	-	-	-

them to their national regulations.

Acknowledgement. This paper is supported in part by EU Horizon 2020 research and innovation programme under grant agreement No 830892, project SPARTA.

References

[1] Purser, S., Standards for Cyber Security. In: Best Practices in Computer Network Defence: Incident Detection and Response, pp. 97–107. IOS Press, (2014), 10.3233/978-1-61499-372-8-97

[2] Oja, T., X-Road Trust Model and Technology Threat Analysis. (2020), Master Thesis, Tallinn University of Technology

[3] Mets, T., Parsovs, A., Time of Signing in the Estonian Digital Signature Scheme, In: Digital Evidence and Electronic Signature Law Review,16(2019), pp.40–50, https://doi.org/10.14296/deeslr.v16i0.5076

[4] Seeba, M., A Specification of Layer-Based Information Security Management System for the Issue Tracking System (2019), Master Thesis, Institute of Computer Science University of Tartu

[5] European Union, General Data Protection Regulation.(2018), `http://eur-lex.europa.eu/`. Last accessed 28 Jan 2021

[6] European Union, Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union, (2016), `http://data.europa.eu/`. Last accessed 22 Jan 2021

[7] Beckers, K., Côté, I., Fenz, S., Hatebur, D., Heisel, M., A Structured Comparison of Security Standards, (2014), 10.1007/978-3-319-07452-8_1

[8] Nabi, S., I., Al-Ghmlas, G., S., Alghathbar, K., Enterprise Information Security Policies, Standards, and Procedures: A Survey of Available Standards and Guidelines, In: Strategic and Practical Approaches for Information Security Governance: Technologies and Applied Solutions, pp.67–89, (2012), IGI Global, ISBN 978-1-4666-0197-0

[9] Tofan, D., Information Security Standards. In: Journal of Mobile, Embedded and Distributed Systems (3). (2011), ISSN 2067 – 4074

[10] Grandclaudon, J. (Ed.), D11.1 International and national cybersecurity certification initiatives. Report of SPARTA project. (2020), `https://www.sparta.eu/`. Last accessed 10 Jan 2021

[11] KPMG OY Ab, Digitaalisen turvallisuuden kansainvälinen vertailu Valtiovarainministeriö. (2020) `https://vm.fi/documents/10623/307681/Digitaalisen+turvallisuuden+kansainv%C3%A4linen+vertailu/7aafe82e-86e7-7450-358c-f1adfeecb3e5/Digitaalisen+turvallisuuden+kansainv%C3%A4linen+vertailu.pdf`. Last accessed 10 Jan 2021

[12] ENISA, Standardisation in support of the Cybersecurity Certification, (2020), 10.2824/481787

[13] ENISA, Good practices in innovation on cybersecurity under the NCSS, (2021), 10.2824/01007

[14] e-Governance Academy (eGA), NCSI National Cyber Security Index, (2021), https://ncsi.ega.ee. Last accessed 10 Jan 2021

[15] Ottis,R., Analysis of the 2007 Cyber Attacks Against Estonia from the Information Warfare Perspective, Cooperative Cyber Defence Centre of Excellence, Tallinn, Estonia, https://ccdcoe.org/uploads/2018/10/Ottis2008_AnalysisOf2007FromTheInformationWarfarePerspective.pdf. Last accessed 10 Jan 2021

[16] Estonian Information Authority (RIA), X-tee factsheet, https://www.x-tee.ee/factsheets/EE/#eng. Last accessed 01 Nov 2020

[17] The Ministry of Economic Affairs and Communications of Estonian Republic, Cybersecurity Strategy Republic of Estonia 2019–2022, (2018). https://www.mkm.ee/sites/default/files/kyberturvalisuse_strateegia_2022_eng.pdf. Last accessed 10 Jan 2021

[18] The Ministry of Economic Affairs and Communications of Estonian Republic, Infoühiskonna arengukava 2020, (2013) https://www.mkm.ee/sites/default/files/elfinder/article_files/eesti_infouhiskonna_arengukava.pdf. Last accessed 10 Jan 2021

[19] Estonian Information System Authority Public Procurement No. 203534. Development of the Estonian information security standard. Description of works. (2019) https://riigihanked.riik.ee/. Last accessed 1 Nov 2020

[20] International Standardisation Organisation (ISO), ISO/IEC 27001:2013 Information technology — Security techniques — Information security management systems — Requirements, (2013).https://www.iso.org/standard/54534.html. Last accessed 1 Nov 2020

[21] Center of Internet Security (CIS), CIS Controls, 2020 hhttps://www.cisecurity.org/controls/cis-controls-list/. Last accessed 20 Nov 2020

[22] German Federal Office for Information Security (BSI), BSI IT-Grundschutz Kompendium, 1-02-2020, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Kompendium/IT_Grundschutz_Kompendium_Edition2020.html. Last accessed 10 Jan 2021

[23] German Federal Office for Information Security (BSI), BSI Standard 200-3: Risk Analysis based on IT-Grundschutz,(2017), https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Grundschutz/International/bsi-standard-2003_en_pdf.html. Last accessed 10 Jan 2021

[24] Center of Internet Security (CIS), Mapping and Compliance. Collaboration Enhances Cybersecurity Compliance, https://www.cisecurity.org/cybersecurity-tools/mapping-compliance/. Last accessed 10 Jan 2021

[25] International Standardisation Organisation (ISO), Frequently Asked Questions (FAQS), https://www.iso.org/footer-links/frequently-asked-questions-faqs/general-faqs.html. Last accessed 20 Nov 2020

[26] Pro Publica Inc., Center for Internet Security Inc., Full text of "Full Filing" for fiscal year ending Dec. 2019, https://projects.propublica.org/nonprofits/organizations/522278213/202041959349302934/full. Last accessed 10 Jan 2021

[27] Estonian Information System Authority (RIA), Three Level IT Baseline Security System ISKE, (2020), https://www.ria.ee/en/cyber-security/it-baseline-security-system-iske.html. Last accessed 10 Jan 2021

[28] German Federal Office for Information Security (BSI), Zuordnungstabelle. Zuordnung ISO/IEC 27001 sowie ISO/IEC 27002 zum modernisierten IT-Grundschutz, (2018) https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Kompendium/Zuordnung_ISO_und_modernisierter_IT_Grundschutz.pdf?__blob=publicationFile&v=1. Last accessed 10 Jan 2021

24th International Conference on Business Information Systems (BIS 2021)

ICT Project

https://doi.org/10.52825/bis.v1i.69

Published: 02 July 2021

Potentials and Barriers of Agility in Small and Medium Sized Enterprises

Insights from qualitative research in Germany

Joerg Bueechl[1], Ralf Haerting[1], Michael Pressl[1] and Raphael Kaim[1]

[1] Aalen University, Beethovenstr. 1, 73430 Aalen, Germany.

Abstract. Our explorative, qualitative study uncovers the dynamic processes of agility unleashing or inhibiting potentials within German small and medium sized enterprises through the lens of digitization. Based on an analysis of 22 interviews, we propose a conceptual model, which illuminates antecedents and external determinants of agility and their impact on potentials and performance in organizational settings. In this process we determine digitization both as an external driver and as an internal provider of agility. Resource constraints and traditional leadership styles are SME-specific barriers to agility. How extensively the potentials of agility can be utilized dependents on the moderating factors firm size and department.

Keywords: Agile Management, Small and Medium Sized Enterprises, Qualitative Research, Potential and Barriers.

1 Introduction

In recent years, markets and industries have been impacted by uncertainty and volatility as a result of digital innovations and disruptions on a global scale. This phenomenon is reflected in the emergence of the by now widely known acronym VUCA, which describes a world characterized by volatility, uncertainty, complexity and ambiguity [1]. An exponentially increasing amount of external economic or environmental disruptions, such as the United States presidential election 2016, the Brexit referendum or just recently Covid-19 forces organizations to cope instantly [2, 3].

Historic data is not necessarily applicable anymore to derive answers to current and future challenges. As a consequence, organizational agility emerged as a widely applied response to guide organizations strategically through these disruptive times to thus ensure their competitiveness. The greater body of research investigates agility in the context of large corporations. However, given the relevance of SMEs as they account for 66% of the workforce in the European Union [4] and 48% of the workforce in the USA [5], we want to address this gap by investigating the dynamics of agility in an SME context. Furthermore, there are distinctive characteristics that set SMEs extensively apart from large corporations, not only in quantitative, but also in qualitative terms, which propose promising results when investigating agility in SME contexts [6]. While SMEs appear to be more agile than large corporations [7], they face numerous limitations in the sense of capabilities, resources and external financial access which in turn makes it difficult to conquer organizational inertia along their way towards digital transformation [8, 9]. We consider the investigation of German SMEs, also known as German Mittelstand or Hidden Champions, as particularly promising as they are not only known as the ”backbone of the German economy”, but also as a leading factor of Germany’s overall competitiveness on a broad spectrum [10]. Given the limited amount of extant research on SME’s agility in digitization contexts [11], we apply a qualitative research approach, which is well-suited when engaging in inductive theory building [12].

Our data collection leads to 22 interviews of employees working in 9 different industries, transcribed on over 170 pages. Based on our data set we developed a model on potentials and barriers of agility in SMEs. Our paper contributes to the literature by identifying the role of digitization both as an external driver and as an internal provider of agility. Other internal drivers in SMEs are consistent with the literature on large companies. The barriers, on the other hand, specifically relate to the context of SMEs in its qualitative and quantitative manifestation and are rooted in resource constraints and traditional leadership structures. The opportunity to exploit the potentials delivered by agility in SMEs is moderated by the department and the firm size.

2 Theoretical Framework

2.1 Agility

Agility in its various forms, such as strategic agility and organizational agility, is often described as a way to cope with dynamic challenges and thereby not losing one's competitive edge in the VUCA world [13]. The concept of agility dates back to the 1960s and centers primarily around adaptability of the manufacturing industry and its processes [14]. Yet, it was in the 1990s when agility transcended into the business research context to find answers to the eroding competitiveness of the US-American market as a result of production overcapacities. During this time, the term agile described organizations which respond instantly to changing consumer and market expectations with product and process innovations [15]. In the early 2000s agility has experienced a renaissance with the emergence of agile project management methods, such as SCRUM, KANBAN or Design Sprints, especially in the field of software development [16, 3].

Ever since, an abundance of different approaches and concepts have surfaced in organizational and business research. However, over the course of the last twenty years, there appeared a consensus among scholars that agility serves as an expression of the interdependence and the interplay of sensing and responding abilities [13]. The sensing capability enables an organization to detect environmental change in forms of competitive market opportunities and evolving conditions, whereas the responding capability helps an organization to rapidly seize these sensed opportunities by efficient and effective reactions. [17]. Nevertheless, there is still a need for structuring the heterogenic landscape of agility to receive a holistic understanding of the concept across various disciplines [18].

The VUCA environment, in which companies find themselves, requires companies to sense and obtain a profound understanding of these disruptions in order to derive effective responses to meet market and customer expectations [11]. Information systems as a key prerequisite for data management are fundamental enablers for companies to match changing external conditions with internal resources and capabilities to successfully navigate through the challenges which have been imposed by the disruptive environment. In this context the situation around the COVID-19 pandemic may foster a mindset in which learning from change is leading into an era of agility, as Batra (2020) notices [21]. Contemporary agility research however often centers around the information system context, particularly around software development, as agility is often regarded both a key prerequisite as well as a consequence of digitization [19]. In this vein, Sambamurthy et al. (2003) argue that investments in IT competence enhances an organization's agility, digital options and entrepreneurial alertness, thus allowing competitive actions and improving financial performance [20]. In a next step, Leonhardt et al. (2016) distinguish between entrepreneurial and adaptive agility and investigate the positive influence of IT on these two constructs. Entrepreneurial agility aims at proactively sensing environmental changes and responding with the development of customized processes, services or products, while entrepreneurial agility takes a reactive approach and focuses on keeping pace with anticipated and upcoming innovations [22].

Although agility can have a positive impact on company performance, there is an ongoing controversy on whether or not it has an exclusively beneficial effect [13]. In this vein, a

lack of either element (sensing or responding) can lead to an impaired equilibrium which again results in a damaging effect on the organization and its performance [17]. While some authors call for unconditional agility and radical transformation, others take a more cautious approach and warn about costs and efficiency to be sacrificed for the sake of agility [23]. An increasing number of scholars are dedicated to resolve this quandary by introducing the concept of organizational ambidexterity. Ambidexterity refers to an organization's ability to efficiently manage current processes and business fields by exploiting existing competencies, while proactively exploring new opportunities by addressing constantly changing demands of the future at the same time [24]. These opposing pairs of concepts, namely exploring vs. exploiting and respectively sensing vs. responding, are highly interdependent in their nature. While the exploitation component mainly comprises scaling daily business activities and therefore represents the continuity-based end of the ambidextrous spectrum, the exploration component encompasses both elements sensing and responding and thus represents the disruptive and agile-based end of the spectrum. Agility can therefore be regarded as a dynamic capability for resolving the ambidextrous quandary [25, 24].

2.2 Agility among Small and Medium Sized Enterprises (SMEs)

The concept of agility and ambidexterity evolved from an idea to a complete area of research over the past 20 years: Overall 135 articles have been published in this domain in the year 2000, while in 2011 already nearly 600 articles have been published on this topic. This number continued to increase steadily, and last year a total of 1546 articles dealing with agility or ambidexterity have been published. However, only very few of these publications address SMEs and even less go beyond investigating agility in the software development context. Nonetheless, agility is not only a relevant concept for large corporations and IT related companies, but for SMEs across different industries as well. It is particularly the ongoing debate of whether or not SMEs are more agile than large corporations which requires further research. Compared to large corporations, SMEs are lean, more informal and less hierarchically structured, which are important features of an agile way of working [11]. Furthermore, in order to compensate for their lower degree of leverage due to their size, SMEs often heavily invest in building external relationship networks, which help them to tap on new opportunities together with other partners [27, 28, 29]. On the other hand, the limitation of resources which SMEs are confined to, such as financial and human resources as well as the capability to transform rigid processes into agile ones, are key liabilities of SMEs on their way to agility [11, 30]. Despite that, in order to ensure long-term success, every medium-sized enterprise and family business need to facilitate exploration and exploitation [31]. As a consequence, more research is needed to shed light on the context of agility in SME contexts.

From a cultural perspective, there exists only sporadic research on agility and ambidexterity that examine German companies. The majority of research investigates the cultural contexts such as USA and China [25, 32, 33, 34, 35, 34, 35]. We chose Germany as the country of investigation for various practical and conceptual reasons. German SMEs are of great significance as they account for 99.6% of all companies in Germany and for 58.5% of the German workforce [38]. Furthermore, Germany is known for the success stories of Hidden Champions, a specific subgroup of SMEs which are highly specialized world-market leaders shaped by international dominance in various product niches [10, 39]. Moreover, in German SMEs ownership, management and liability are typically interdependent. From a demographic perspective, in 2016 more than one third of all company owners were at least 55 years old [40]. This phenomenon goes along with a traditional, hierarchical style of leadership, which commonly is opposed to environments, in which agile mindsets can thrive as top-down communication and excessive control leaves no space for flexibility [23]. Next, Germany conceptually represents a risk avoidant culture, scoring relatively high on Hofstede's uncertainty avoidance scale [41]. As uncertainty avoidance goes along with tolerance for ambiguity, a beneficial factor for agility [40], Germany can be seen as culturally disadvantaged in terms of agile management. Lastly, Germany is worth investigating regarding its world share in GDP as well as its extensive trade activities [43].

The aim of our study is to examine specific boundary conditions, challenges, and benefits of organizational agility as a result of digitization for SMEs in Germany. Previous research calls for more studies to unveil boundary conditions on the restrictions and limitations of agility, namely to assess at which point the benefits of exploration endeavors outweigh the associated costs [26]. Second, the question of whether it is preferable to develop selective agile processes that are critical for organizational success or to develop a holistic agile landscape across the organization and its functions instead, also needs to be answered [13]. Moreover, Leonhard (2017) calls for more research to unveil the natural boundary of agile transformation, namely to identify the point including its boundary conditions at which organizational agility reaches its limit and cannot be enhanced anymore [32]. In order to address the aforementioned calls for future research and research gaps, we deduced following research question, which we try to find answers on with the help of our qualitative data:

What are the opportunities and barriers of agility among SMEs?

3 Methodology

3.1 Research Design

As the assessment of opportunities and barriers of agility on SMEs in a cross-cultural setting is still a nascent research area, lacking established theory, we regard an inductive and explorative research design as suitable. Such a research approach is appropriate when investigating dynamic and complex phenomena and engaging in inductive theory building [12] in form of mid-range theory [44]. Our research design comprises 22 in-depth, semi-structure interviews, which provide a more profound understanding of the overall dynamics than quantitative research, which applies a more deductive approach [45].

3.2 Research Setting

To ensure comparability of the data collected, we conducted all interviews in a single country for the above-mentioned reasons. We have further selected SMEs which account for 50-1000 employees [46]. Furthermore, we focused on organizations in the production sector, as service, software or IS companies in many cases already employ agility techniques and methods. We also consider agility in German SMEs worthwhile investigating due the high degree of trade activities of Germany resulting in extensive internationalization endeavors and growing complexity [3]. In order to guarantee ecological validity [47] and to obtain a holistic understanding of the dynamic processes under investigation, we interviewed individuals on a broad spectrum, working in different functions and different industries [48]. Our interviewees were individuals who are exposed to agility, ambidexterity and/or other topics relevant for this study in scope of their work. This approach ensures comparability across functions and companies to tease out individual phenomena and their specific context.

3.3 Data Collection

The data set of the research project consists of 22 semi-structured interviews conducted 2019 and 2020 in 20 different departments from twelve different organizations in nine different industries, with an emphasis on the manufacturing sector. Participants from as many different functional areas and hierarchical levels as possible were to be interviewed in order to obtain a broad view of the situation and to minimize the potential for bias of a single individual. 60% of the interviewees are in management positions. A total of almost 13 hours of interview material was recorded and transcribed to over 170 pages. A list of the individual interview partners can be found in table 1. Semi-structured interviews guarantee a reasonable degree of consistency in the questions and thus establish comparability between the interviews without preventing the discovery of unknown and unexpected phenomena [49]. We opted for a narrative interview design which helped us to collect rich data from people in various roles and situations [49] to access interviewees' motivations and thoughts.

Table 1. Overview of Interviewees

Firm	Interviewee	Gender	Department/ Position	Industry	Employees
1	Data 1	female	Human Resources	Telecom Services	75
2	Data 2	female	Project Management	Data/Media	320
3	Electro 1	male	Head of Logistics	Electrical Industry	350
3	Electro 2	male	Product Management	Electrical Industry	350
3	Electro 3	male	IT-Project Management	Electrical Industry	350
4	Food 1	male	Head of Logistics	Food Production	900
5	Food 2	female	Head of Production	Food Production	140
5	Food 3	male	CEO	Food Production	140
5	Food 4	female	Head of Logistics	Food Production	140
6	Machine 1	male	CEO	Mechanical Engineering	270
7	Machine 2	male	Head of Production	Mechanical Engineering	600
7	Machine 3	male	Production	Mechanical Engineering	600
7	Machine 4	male	Production	Mechanical Engineering	600
7	Machine 5	male	Logistics	Mechanical Engineering	600
8	Textile 1	female	Head of HR	Textile Industry	800
9	Textile 2	male	Finance/Sales	Textile Industry	650
10	Tool 1	male	HR	Tool Manufacturing	530
11	Trade 1	male	Head of IT	Production/Mail Order	800
12	Trade 2	female	Head of HR	Paper Industry/Mail Order	270
12	Trade 3	male	CEO	Paper Industry/Mail Order	270
12	Trade 4	male	Head of Sales	Paper Industry/Mail Order	270
12	Trade 5	male	CFO	Paper Industry/Mail Order	270

The interview guide had three sections. The first part dealt with basic information about the interviewee such as demographics, tenure, duration of the current position, current tasks, the company's industry or company culture. The second and main part of the questionnaire dealt with agility in the company. Questions covered topics such as experiences with agility, initial successes or problems and the connection to digitization and management style. The questions were first asked very openly to reduce biases. In the further course, probes were used in a more specific way to learn about specific examples and descriptions of relevant situations. In order not to disrupt the interview flow, the sequence of the questions was held in a flexible manner. At the end of the interview the interviewees were given the opportunity to reflect on issues which they considered to be relevant, but which have yet remained unaddressed. The interviewees were also given the opportunity to emphasize any open aspects or issue of their choice. An interview lasted on average 35 minutes, was digitally recorded and then transcribed verbatim in the original language German. We stopped conducting interviewing once we reached the point of data saturation, the point when no information emerged [50].

3.4 Analysis

The data was analysed using the Grounded Theory [51]. Therefore, first open coding, then axial coding and lastly selective coding was applied. In the open coding phase, all material was read and basic, general codes were assigned. These codes can originate either from the direct formulation of the interviews or from constant comparison with the literature. In a next step, the resulting large number of codes were then grouped into categories to transition from a descriptive to a conceptual level using the constant comparative method [51]. We have constantly compared and juxtaposed different parts within and across interviews to ensure consistency. In this way connections between our coded appeared and categories could be formed. For example, the codes ”Digital Technologies” and ”Data Security” have been merged into first order category ”Digitization”, and then combined with the first order category ”Increasing Speed” into the higher order category ”External Driving Factors of Agility”. In order to retrace our method, we provide an overview of exemplary codes and categories in figure 1. The quotes related to the exemplary codes are integrated in the results section.

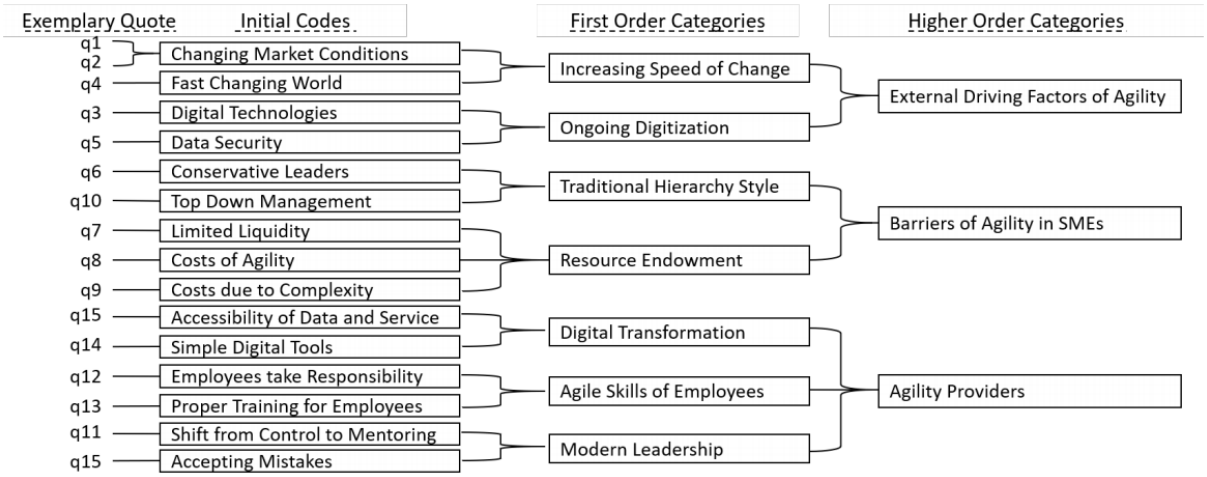


Fig. 1. Exemplary Procedure of the Data Analysis.

Subsequently, in the selective coding phase, the relationships and connections between the categories have been analysed, summarized and visualized in figure 2.

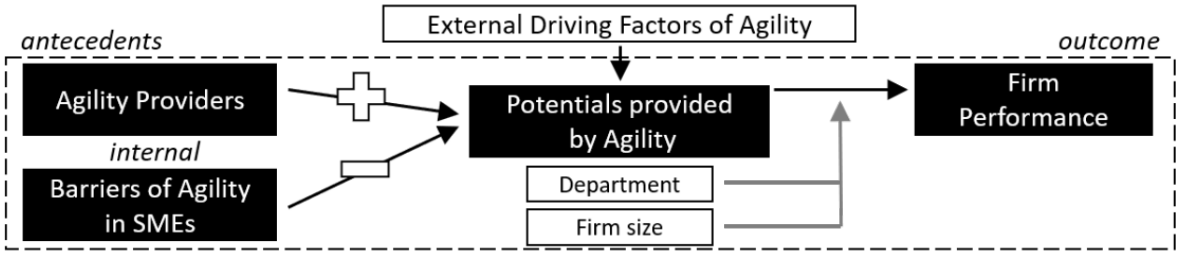


Fig. 2. Model of Potentials and Barriers of Agility in SMEs.

During the analysis, theoretical memos were written, which helped to cycle back and forth between literature and collected data to ensure validity. The data analysis was mainly carried out by the first author and the interim results and findings were regularly discussed intensively with the other authors. The transcribed interviews were analysed using the software for qualitative research MAXQDA.

4 Empirical Findings and Model Conceptualization

Based on our data we will introduce our findings, juxtapose these with existing research, and ultimately derive a comprehensive model which represents the process of agility facilitation and aligning it with firm performance. This chapter is organized according to the different

categories which emerged from the analysis of our data-set. First, we will begin by introducing the external driving factors of agility, followed by the internal antecedents of agility, namely facilitating agility providers and inhibiting barriers of agility among SMEs. Second, the resulting potentials delivered or provided by agility are presented, and third, agility-induced potentials, including moderating factors.

4.1 External Driving Factors of Agility

Digitization has proven to be a key driver to impact social and economic life as it helps to increase efficiency with the help of information technologies, particularly in the areas of production, administration and ext. communication. As a result, digitization and its implementation lead to disruptive changes in both – internal and external processes - which organizations need to flexibly adapt to. The interviewed experts mentioned numerous times how digitization-induced disruptive environments perpetually forces them to quickly adapt to these changes to retain their competitiveness. Electro 2 (q1) and Media 2 (q2) emphasized that changing market conditions forces organizations to be agile to respond faster and more flexibly to customer needs.

Furthermore, digital transformation processes differ extensively from traditional forms of strategic change. Interviewee Electro 2 (q3), a product manager for electrical components, concluded that digital technologies have accelerated the speed and shape in which enterprises compete with one another. Furthermore, agile working environments even impact the way in which companies create value and interact with their customers and partners [52]. In this fast-evolving context it has become vital to respond quickly to new situations and agility has emerged as a promising approach in mastering the rising degree in complexity. In a similar vein, as interviewee Trade 1 (q4) stated: *The world is changing faster and faster and I can only deal with a fast change by working in an agile way*. Another example for fast-paced changing opportunities, which agility offers is being raised by Trade 2. The HR Business Partner of a paper wholesaler claimed that the market and especially the sector they are doing business in is declining. Therefore, the interviewee sees the ability to adapt quickly and flexibly as one key to identify opportunities to diversify the company portfolio and to build up new pillars for the enterprise. His colleague Trade 4, who works in the sales department, concurs completely. Only those enterprises which are open to new ventures in this disruptive market are able to act successfully in the future.

There are also external factors which impede agility and limit a transformation as described above. Externally imposed norms and regulations inhibit quick and efficient execution of the jobs to be done. Interviewee Data 1, responsible for connecting cities to fibreglass lines, emphasized the relevance of strict regulations when executing roadworks. He outlined the constraints associated with complying with the mandatory construction steps. Opportunities for efficiency improvements are therefore strictly limited due to public regulations.

4.2 Agility Providers

After providing external driving factors, this section focuses on internal antecedents. The analysis has identified three basic types of agility providers: leaders, employees and digital transformation. For employees to be empowered to assume responsibility within agile and self-organized teams [16], in a first step it requires the management to abandon the mindset of an overall controlling instance. As the product manager Electro 2 (q15) mentioned, a leader in charge should abandon parts of his responsibility, accept mistakes and empower employees to act self-organized and independently. CEO Machine 1 added that this inner shift takes a lot of effort, time and patience to live up to these expectations. Also Trade 1 (q11) stated that this can be understood as an ongoing process for leaders and goes along with a transition in leadership style shifting from control to mentoring and coaching.

At the same time the organization relies on the employees' willingness to take an active part by taking over responsibility as the HR Business Partner Data 1 (q12) emphasized. But according to Project Manager Data 2 (q13), it is the company that is required to educate

employees in how to use agile working methods. The Head of IT continued that a central provider of agility is empowering employees to achieve the common goal in their own way. To unlock the full potential within the exchange between leaders and employees, transparency and open communication is key according to Trade 1. Data 1 further concluded that all employees, who are involved in the process, must be aware of the expected targets as well as possible challenges in order to succeed together in a creative way.

Interviewee Machine 1 stated that the processes and agile methods must be tailored to the specific context of the organization, or the organization needs to adapt own techniques. Either way - to design a perfect fit between the organization, the business model and the employees, transparent communication is needed. To achieve a sustainable shift in working behavior the capabilities for reconfiguration and process concurrency are further vital elements. One key provider thereof is once again digitization. Agile working methods often amplify the requirements for digitized processes and support the overall digital transformation. As Trade 3 and Trade 5 stressed, enhancing process speed is a key argument in favor of striving for a higher degree of agility. Interviewee Textile 1 (q14) considers the simplification in sophisticated digital tools as a major milestone. In this regard, Textile 2 (q15), who works in sales and finance, sees the main impact in the accessibility of services during mobile working. The opportunity of a flexible and agile working environment, sustained by digital technologies, is generally appreciated by several interviewees.

The identified agility providers are also proposed as the agility enablers of Conforto et al. (2014) as leaders show similar features to the described organization category, collaborators to project team category and digital transformation to the process category [53].

4.3 Barriers of Agility in SMEs

Besides the providers that enable agility in SMEs, there are also barriers that inhibit agility. Leaders and employees are agility providers, but in some cases, they can also be the barriers of agility in SMEs. Especially leaders that exert control can impede agility. Expert Electro 1 (q6) stated: *Especially as far as managers are concerned, there are still some really old-fashioned conservatives who dictate everything*. This attitude is difficult to reconcile with an agile way of working. Especially when it comes to the extraordinary situation of SMEs based in Germany, the resource endowment is worth to be taken a further look at [54]. As described above, in many German SMEs the ownership and management responsibility are integrated in the same person or circle of persons and SMEs are often externally financed by bank loans and equity capital. Expert Food 3 (q7) emphasized in this context the limited availability of liquidity compared to large corporations. In this regard, Machine 4 (q8) urged not to forget about time and resources needed to implement agile working methods like SCRUM and Kanban in the first place, especially as SMEs deal with fixed lead times which again restricts process flexibility. Expert Food 1 (q9) pointed out that a key challenge to respond to increasing complexity is represented by resource limitations, which makes it more difficult to compete with larger corporations.

As already mentioned, hierarchy is often regarded as one of the biggest obstacles when implementing agile working mindsets. Most interviewees also reported in this respect that their SMEs are shaped by a flat hierarchical landscape as well as by supportive and appreciative working relationships. However, a closer analysis of the management and communication style revealed discrepancies. Data 2 stated for example *We work collegially together, but we still have the hierarchies with the boss and you don't call him by his first name*. Also, Tool 1 (q10) says *it's very collegial, we also have rather few hierarchical levels*, while at the same time he says about the management style: *Since I've been here at the company* [2 years] *it's top down*.

This classical hierarchy often depends on the leadership style of the owner. With regards to agility, the relationship between managers and employees is an essential contextual factor. Agile working methods can only be successfully implemented if managers understand leadership not as a means of control, but as a concept which facilitates a coaching process

based on a mutual trusting relationship [55]. But even a change of management does not imply immediate change and may create additional barriers. As expert Textile 1 stated, the transition process, which is induced by a manager and directed at the shift from a strictly owner-driven enterprise to a more open one, takes a significant amount of time. Although the employees are encouraged to take responsibility, it takes time to open up into the new cooperative ways of working. Often the attempt to achieve more organizational agility fails because of organizational inertia. In this vein, a mature organization tends to continue on its current trajectory [13]. The CFO Trade 5 (q26) emphasizes an intense conflict between the two extreme positions never change a running system and a completely agile working methodology. Especially in SMEs, the mentality of the ever-lasting principles and the resistance to give up strictly structured methodologies tends to stick with the former position.

4.4 Potentials provided by Agility

A variety of potentials which agility provides could be derived from the data. The examples can be divided into the following categories responsiveness, speed, flexibility and competency [56]. With the help of agility, companies can involve the customer in the development process of new products. As Expert Machine 1 (q16) pointed out, customer feedback can be implemented at a faster rate. He explained this by an agile product development process, which the company he works for is currently establishing: *It is different in the agile environment, because at the beginning we bring techniques like design thinking into the process, where we try out what the customer needs, what kind of problems he actually has, and how can we solve them*. As a consequence, this leads to improved responsiveness. In addition, agility offers the fundamental potential to act and respond faster. Trade 1 (q17) mainly strives for agility to increase speed. Machine 3 (q18) emphasized: *You´re able to reduce complexity in modern production facilities which improves the process speed enormously*. Being agile enables organizations to be flexible in uncertain times. Machine 1 (q19) pointed out: *I need agility when I do not know exactly what I should or can or must do the day after tomorrow*. [...] *That means I have to be flexible*. Further potentials provided by agility can be clustered in the category competencies. One of them represents increased inspiration and creativity when seeking and including employees' opinions and feedback. This is a means to improve team spirit and intrinsic motivation as Textile 1, Food 4 and Textile 2 stated. According to Data 2, agility also creates transparency and thus helps to understand and set a common goal in a project. Machine 1 (q23) concludes that agility helps to optimize processes to resolve what he calls *monster processes*.

4.5 Moderating factors

Additional factors were identified which influence the ability to exploit the potential of agility described above. The task orientation of the department in which agility is used and the firm size in general moderate the influence of the firm's potential on its performance. The analysis of the data provided evidence that some departments are suited for agility processes while in other departments an agile way of working is impeding the working progress. The IT department was mentioned by numerous interviewees like Machine 1 and Electro 3 as a department where agility adds value. Further, in project management, research and development and strategic driven fields, there exists a basis for agility facilitation. Sales and marketing were also emphasized, as the opportunities of executing creative and customized ways of working were highly valued by experts Data 1 and Food 1. Also, logistical functions need to adopt agile working methods, not to endanger the supply chain performance as stated by Machine 5 (q20). In contrast, Trade 4 (q23) pointed out, there are some departments in which the adoption of agility does not achieve the expected potential. Especially operations and administrative functions, which are shaped by continuity-based routines, need to function reliably in an efficiently and effective manner, and therefore leave little for creativity and flexibility [23]. Or as Trade 2 (q24) stated it out: *Once a process is defined and running the going concern continuity is competing with agility and must be weigh up with each other.* Further, departments, such as accounting or controlling are bound by a multitude of norms and regulations. Experts therefore underlined efficiency losses rather than improved company perfor-

mance attributed to agile working methods. Production was also frequently mentioned as an area, where the implementation of agility does not attribute to an added value. This was rather surprising, as agile manufacturing or production is widely accepted and implemented, but apparently in the German research context only to a limited degree.

A further aspect with regards to departments as a moderating factor is the working-relationship between agile and non-agile working teams within organizations. As the experts summarized, digital transformation simplifies the interaction, but the interface still needs to be monitored very closely. A mutual comprehension as the CEO Food 3 (q21) called it, needs to be established as a foundation, although friction losses can occur. Trade 1 (q22) mentioned that in line with their transformation process it was crucial to completely shift project related work towards agile two-week iterative planning processes. In their experience it was necessary to align the entire value-chain to the new methodology. Otherwise, project steps would not have been met. However, a challenge emerges in inter-team collaborations in form of bottlenecks if agile teams fail to deliver as the descriptions of Data 1 indicate.

The other factor which moderates the influence of the potentials on firm performance, is the firm size itself. According to many experts, small companies close up to 50 employees are more applying agile methodologies as for their size. The underlying reason is that in small companies, every employee tends to flexibly fill gaps and acts on a day-to-day basis without consciously dealing with the construct agility itself. In SMEs, a standardized process landscape is often missing, as the Head of HR Textile 1 (q24) described, while larger corporations' function less efficient as for their cumbersome structures. The decision-making processes in larger organizations often is more time consuming and reaction speed also suffers along with growing hierarchy, as Machine 3 (q23) added. Therefore, especially large companies, which are commonly impeded by organizational inertia, can capitalize on the assets of agility. We therefore propose that firm size might positively moderate the influence of potentials provided by agility on firm performance.

5 Discussion

Based on the analysis of the empirical qualitative data we collected on agility in German SMEs, we developed a conceptual model, which is depicted in Figure 2. Both, similarities and differences between SMEs and larger organizations were identified. Both company types share similarities in the field of agility providers as well as with regards to potentials provided by agility, which concurs with previous research [51, 53]. In contrast, the barriers of agility appear to be specific to the SME context. These specific barriers can be mainly attributed to the above-mentioned resource constraints of SMEs, along with hierarchical structures and traditional leadership styles, which appear to be a German cultural phenomenon. Therefore, we see great value in agility research in SME contexts, also from a cultural perspective.

Despite these barriers, it was evident that German SMEs already demonstrate agile working practices in many cases. The examined service companies are already considerably further advanced in this respect than those companies with a manufacturing background, which might not be overly surprising. However, they often do so in an unstructured and even unconscious way, without explicit associations with agility. SMEs in many cases are therefore less methodologically agile than larger companies, which results in a call for action.

5.1 Implications for Practice

Our model helps SME managers to facilitate the process of agile transformation by utilizing their limited resources in a more targeted and thus more efficient way. During this process our model will reveal how many factors constitute an agile mindset and which steps need to be taken: First, managers need to differentiate internal agility providers as well as barriers. By juxtaposing these factors, managers can identify specific potentials of agility. Our study shows that agility is not a concept to be applied in every context, namely in all departments and business functions. However, if SME managers identify suitable departments and busi-

ness functions, they will succeed by being able to implement agility in a targeted and customized way. Furthermore, it is helpful for managers to be aware that they may encounter many internal barriers in such a transformation process. They need to understand that the involvement and empowerment of employees is key, which again often requires a reassessment of a leader's own leadership style.

5.2 Implications for Research

Our paper aims to integrate the key findings based on the conceptual model of agility of Sharifi and Zhang (1999) into an SME context [53]. By doing so, we provide boundary conditions for the application of agility in SME contexts. Specifically, our paper provides deeper insights into the agility providers and external factors driving agility as well as regarding barriers of agility. Additionally, our study indicates that cultural context impacts the benefits and challenges of agility on firm performance as well. In future studies, we propose to investigate in more detail the extent to which the concept of ambidexterity provides a possible solution to bridge the gap between agile and non-agile departments and business units.

Nonetheless, we acknowledge the limitations of our study. First of all, our results stem from qualitative empirical research and thus need to be validated by quantitative studies. Furthermore, our cross-sectional study would profit from a longitudinal study to tease out the dynamics associated with antecedents and consequences of agility in a SME context. Furthermore, our study highlights that particularly barriers of SMEs provide promising avenues for further research as they are clearly distinct from the barriers that apply to a larger corporation context. Last, the terms agility and ambidexterity are largely unknown in German SME's beyond IT business function. Therefore, agile working approaches in our targeted companies were rather tacit. Despite these limitations, we are confident that our research adds value to theory development in the fields of agility and SMEs in specific cultural settings.

Acknowledgement: This work is based on several projects and theses at Aalen University. We would like to especially thank Christiane Weiler, Markus Buck, Nils Urban and Christian Neldert who contributed within a project in master class. Also, Diana Schmidt and Alexander Schmidt supported with her raw data generation collection for their Master respective Bachelor theses. We would like to thank all contributors for their great support.

References

[1] Bennett N, Lemoine J. What vuca really means for you. Harvard Business Review, 2014.

[2] Economic Policy Uncertainty Index. Global economic policy uncertainty index. http://www.policyuncertainty.com/global_monthly.html. Accessed 2021 May 5.

[3] Kaim R, Härting RC, Reichstein C. Benefits of Agile Project Management in an Environment of Increasing Complexity—A Transaction Cost Analysis. Smart Innovation, Systems and Technologies, vol 143. Intelligent Decision Technologies 2019. 2019 Jun. https://doi.org/https://doi.org/10.1007/978-981-13-8303-8_17

[4] Dyfed L. Viele Kleinstunternehmen in der EU. https://de.statista.com/infografik/9755/unternehmenslandschaft-in-der-eu/. Accessed 2020 July 10.

[5] United States Census Bureau. Business Dynamics Statistics. https://www.census.gov/programs-surveys/bds.html. Accessed 2014.

[6] The Leader. In: Hidden Champions of the Twenty-First Century.. New York, NY: Springer; 2009. https://doi.org/https://doi.org/10.1007/978-0-387-98147-5_10

[7] Dibrell C, Davis PS, Craig J. Fueling Innovation through Information Technology in SMEs. Journal of Small Business Management. 2008 04;46(2):203-218. https://doi.org/10.1111/j.1540-627x.2008.00240.x

[8] Li L, Su F, Zhang W, Mao J. Digital transformation by SME entrepreneurs: A capability perspective. Information Systems Journal. 2017 06 20;28(6):1129-1157. https://doi.org/10.1111/isj.12153

[9] Argote L, Ingram P. Knowledge Transfer: A Basis for Competitive Advantage in Firms. Organizational Behavior and Human Decision Processes. 2000 05;82(1):150-169. https://doi.org/10.1006/obhd.2000.2893

[10] Audretsch DB, Lehmann EE, Schenkenhofer J. Internationalization strategies of hidden champions: lessons from Germany. Multinational Business Review. 2018 04 16;26(1):2-24. https://doi.org/10.1108/mbr-01-2018-0006

[11] Chan CM, Teoh SY, Yeow A, Pan G. Agility in responding to disruptive digital innovation: Case study of an SME. Information Systems Journal. 2018 08 09;29(2):436-455. https://doi.org/10.1111/isj.12215

[12] Eriksson P, Kovalainen A. Qualitative Methods in Business Research . Second Edition. 2015.

[13] Tallon PP, Queiroz M, Coltman T, Sharma R. Information technology and the search for organizational agility: A systematic review with future research possibilities. The Journal of Strategic Information Systems. 2019 06;28(2):218-237. https://doi.org/10.1016/j.jsis.2018.12.002

[14] Burns T, Stalker M. The management of innovation. London: Tavistock Publications; 1961.

[15] Nagel RN. 21ST Century Manufacturing Enterprise Strategy Report. Defense Technical Information Center; 1992 01. https://doi.org/10.21236/ada257032

[16] Manifesto for Agile Software Development. https://agilemanifesto.org/. Accessed 2001.

[17] Overby E, Bharadwaj A, Sambamurthy V. Enterprise agility and the enabling role of information technology. European Journal of Information Systems. 2006 04;15(2):120-131. https://doi.org/10.1057/palgrave.ejis.3000600

[18] Conboy K. Agility from First Principles: Reconstructing the Concept of Agility in Information Systems Development. Information Systems Research. 2009 09;20(3):329-354. https://doi.org/10.1287/isre.1090.0236

[19] Batra D. The Impact of the COVID-19 on Organizational and Information Systems Agility. Information Systems Management. 2020 Oct 01;37(4):361-365. https://doi.org/10.1080/10580530.2020.1821843

[20] Lindner D, Leyh C. Organizations in Transformation: Agility as Consequence or Prerequisite of Digitization?. Business Information Systems. Lecture Notes in Business Information Processing, vol 320. BIS 2018.. https://doi.org/https://doi.org/10.1007/978-3-319-93931-5_7

[21] Sambamurthy, Bharadwaj, Grover. Shaping Agility through Digital Options: Reconceptualizing the Role of Information Technology in Contemporary Firms. MIS Quarterly. 2003;27(2):237. https://doi.org/10.2307/30036530

[22] Leonhardt D, Mandrella M, Kolbe LM. Diving into the Relationship of Information Technology and Organizational Agility: A Meta-Analysis. International Conference on Information Systems (ICIS). 2016.

[23] Teece D, Peteraf M, Leih S. Dynamic Capabilities and Organizational Agility: Risk, Uncertainty, and Strategy in the Innovation Economy. California Management Review. 2016 08;58(4):13-35. https://doi.org/10.1525/cmr.2016.58.4.13

[24] Raisch S, Birkinshaw J, Probst G, Tushman ML. Organizational Ambidexterity: Balancing Exploitation and Exploration for Sustained Performance. Organization Science. 2009 08;20(4):685-695. https://doi.org/10.1287/orsc.1090.0428

[25] Lu, K. (Ram) Ramamurthy. Understanding the Link Between Information Technology Capability and Organizational Agility: An Empirical Examination. MIS Quarterly. 2011;35(4):931. https://doi.org/10.2307/41409967

[26] O'Reilly CA, Tushman ML. Organizational Ambidexterity: Past, Present, and Future. Academy of Management Perspectives. 2013 Nov;27(4):324-338. https://doi.org/10.5465/amp.2013.0025

[27] Rehm S, Goel L. Using information systems to achieve complementarity in SME innovation networks. Information & Management. 2017 06;54(4):438-451. https://doi.org/10.1016/j.im.2016.10.003

[28] Hite JM. Evolutionary Processes and Paths of Relationally Embedded Network Ties in Emerging Entrepreneurial Firms. Entrepreneurship Theory and Practice. 2005 01;29(1):113-144. https://doi.org/10.1111/j.1540-6520.2005.00072.x

[29] Smith-Doerr L, Powell W. Networks and Economic Life. In: Smelser NJ, Swedberg WW, eds. The Handbook of Economic Sociology. Princeton: Princeton University Press; 1994.

[30] Neirotti P, Raguseo E. On the contingent value of IT-based capabilities for the competitive advantage of SMEs: Mechanisms and empirical evidence. Information & Management. 2017 03;54(2):139-153. https://doi.org/10.1016/j.im.2016.05.004

[31] Goel S, Jones RJ. Entrepreneurial Exploration and Exploitation in Family Business. Family Business Review. 2016 01 28;29(1):94-120. https://doi.org/10.1177/0894486515625541

[32] Leonhardt D, Haffke I, Kranz J, Benlian A. Reinventing the it function: the role of it agility and it ambidexterity in supporting digital business transformation. European Conference on Information Systems (ECIS). 2017, 968.

[33] Lee O(, Sambamurthy V, Lim KH, Wei KK. How Does IT Ambidexterity Impact Organizational Agility?. Information Systems Research. 2015 06;26(2):398-417. https://doi.org/10.1287/isre.2015.0577

[34] Zhou J, Bi G, Liu H, Fang Y, Hua Z. Understanding employee competence, operational IS alignment, and organizational agility – An ambidexterity perspective. Information & Management. 2018 09;55(6):695-708. https://doi.org/10.1016/j.im.2018.02.002

[35] Lubatkin MH, Simsek Z, Ling Y, Veiga JF. Ambidexterity and Performance in Small-to Medium-Sized Firms: The Pivotal Role of Top Management Team Behavioral Integration. Journal of Management. 2006 Oct;32(5):646-672. https://doi.org/10.1177/0149206306290712

[36] Hoyer V, Stanoevska-Slabeva K. IT impacts on operation-level agility in service industries. European Conference on Information Systems. 2009, 111.

[37] Cao Q, Gedajlovic E, Zhang H. Unpacking Organizational Ambidexterity: Dimensions, Contingencies, and Synergistic Effects. Organization Science. 2009 08;20(4):781-796. https://doi.org/10.1287/orsc.1090.0426

[38] Held H. KMU- und Start-up-Management. Kohlhammer; 2019.

[39] Venohr B, Meyer KE. The German Miracle Keeps Running: How Germany's Hidden Champions Stay Ahead in the Global Economy. SSRN Electronic Journal. 2007;. https://doi.org/10.2139/ssrn.991964

[40] KFW. Mittelstand ist der Motor der deutschen Wirtschaft. www.kfw.de/KfW-Konzern/KfW-Research/Mittelstand.html. Accessed 2020 June 10.

[41] Hofstede Insights. What about Germany?. www.hofstede-insights.com/ country-comparison/germany. Accessed 2020 July 10.

[42] Nemkova E. The impact of agility on the market performance of born-global firms: An exploratory study of the ‘Tech City’ innovation cluster. Journal of Business Research. 2017 Nov;80:257-265. https://doi.org/10.1016/j.jbusres.2017.04.017

[43] Clark D. Gross domestic product at current market prices of selected European countries in 2018. https://www.statista.com/statistics/685925/gdp-of-european-countries/. Accessed 2020 June 15.

[44] Eisenhardt KM. Building Theories from Case Study Research. Academy of Management Review. 1989 Oct;14(4):532-550. https://doi.org/10.5465/amr.1989.4308385

[45] Rubin HJ, Rubin IS. Qualitative Interviewing: The Art of Hearing Data. SAGE Publications, Inc; 2011.

[46] Gnyawali DR, Park B(. Co-opetition and Technological Innovation in Small and Medium-Sized Enterprises: A Multilevel Conceptual Model. Journal of Small Business Management. 2009 07;47(3):308-330. https://doi.org/10.1111/j.1540-627x.2009.00273.x

[47] Lee TW. Using qualitative methods in organizational research. Thousand Oaks, Calif.: Sage Publications; 1999.

[48] Hollensbe EC, Khazanchi S, Masterson SS. How Do I Assess If My Supervisor and Organization are Fair? Identifying The Rules Underlying Entity-Based Justice Perceptions. Academy of Management Journal. 2008 Dec;51(6):1099-1116. https://doi.org/10.5465/amj.2008.35732600

[49] Myers MD. Qualitative Research in Business and Management. 3rd edition. SAGE Publications Ltd; 2019.

[50] Locke K. Grounded Theory in Management Research. SAGE Publications Ltd; (SAGE series in Management Research) 2001.

[51] Glaser BG, Strauss AL, Strutzel E. The Discovery of Grounded Theory; Strategies for Qualitative Research. Nursing Research. 1968 07;17(4):364. https://doi.org/10.1097/00006199-196807000-00014

[52] Warner KS, Wäger M. Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal. Long Range Planning. 2019 06;52(3):326-349. https://doi.org/10.1016/j.lrp.2018.12.001

[53] Conforto EC, Salum F, Amaral DC, da Silva SL, de Almeida LFM. Can Agile Project Management be Adopted by Industries Other than Software Development?. Project Management Journal. 2014 06;45(3):21-34. https://doi.org/10.1002/pmj.21410

[54] Arbussa A, Bikfalvi A, Marquès P. Strategic agility-driven business model renewal: the case of an SME. Management Decision. 2017 03 20;55(2):271-293. http://doi.org/10.1108/md-05-2016-0355

[55] Nold H, Michel L. The performance triangle: a model for corporate agility. Leadership & Organization Development Journal. 2016 05 03;37(3):341-356. https://doi.org/10.1108/lodj-07-2014-0123

[56] Sharifi H, Zhang Z. A methodology for achieving agility in manufacturing organisations: An introduction. International Journal of Production Economics. 1999 05;62(1-2):7-22. https://doi.org/10.1016/s0925-5273(98)00217-5