

A Min Tjoa · Li-Rong Zheng
Zhuo Zou · Maria Raffai
Li Da Xu · Niina Maarit Novak (Eds.)

LNBIP 310

Research and Practical Issues of Enterprise Information Systems

11th IFIP WG 8.9 Working Conference, CONFENIS 2017
Shanghai, China, October 18–20, 2017
Revised Selected Papers



ifip

 Springer

Lecture Notes in Business Information Processing

310

Series Editors

Wil M. P. van der Aalst

RWTH Aachen University, Aachen, Germany

John Mylopoulos

University of Trento, Trento, Italy

Michael Rosemann

Queensland University of Technology, Brisbane, QLD, Australia

Michael J. Shaw

University of Illinois, Urbana-Champaign, IL, USA

Clemens Szyperski

Microsoft Research, Redmond, WA, USA

More information about this series at <http://www.springer.com/series/7911>

A Min Tjoa · Li-Rong Zheng
Zhuo Zou · Maria Raffai
Li Da Xu · Niina Maarit Novak (Eds.)

Research and Practical Issues of Enterprise Information Systems

11th IFIP WG 8.9 Working Conference, CONFENIS 2017
Shanghai, China, October 18–20, 2017
Revised Selected Papers

Editors

A Min Tjoa
Vienna University of Technology
Vienna
Austria

Li-Rong Zheng
Fudan University
Shanghai
China

Zhuo Zou
Fudan University
Shanghai
China

Maria Raffai
Szechenyi Istvan University
Gyor
Hungary

Li Da Xu
Old Dominion University
Norfolk, VA
USA

Niina Maarit Novak
Vienna University of Technology
Vienna
Austria

ISSN 1865-1348

ISSN 1865-1356 (electronic)

Lecture Notes in Business Information Processing

ISBN 978-3-319-94844-7

ISBN 978-3-319-94845-4 (eBook)

<https://doi.org/10.1007/978-3-319-94845-4>

Library of Congress Control Number: 2018947443

© IFIP International Federation for Information Processing 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

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

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

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

International Conference on Research and Practical Issues of Enterprise Information Systems

11th IFIP WG 8.9 Working Conference, CONFENIS 2017

**Crowne Plaza Fudan, Shanghai, China
October 18–20, 2017**

Preface

The 11th edition of the IFIP WG 8.9 Working Conference, CONFENIS 2017, was held at Crowne Plaza Fudan, in Shanghai, China, during October 18–20, 2017, marking the return of the CONFENIS conference series to China after 10 years (previously CONFENIS was held in Beijing, China, in 2007).

This year’s conference theme – “Industrial Internet of Things and Made in China 2025” – brought together researchers as well as practitioners and representatives of industry, academia, and public authorities to present and discuss latest research ideas and findings with the aim of facilitating the exchange of ideas and developments in all aspects of EIS. In addition, a specialized workshop on the topic of “Smart Electronics and Systems for Industrial IoT” was jointly held with the conference.

The 2017 edition of CONFENIS focused mainly on aspects of EIS concepts, theory and methods, IoT and emerging paradigm, EIS for Industry 4.0, big data analytics, intelligent electronics and systems for industrial IoT. A total of 39 high-quality papers from 19 countries and 42 organizations were received. After a rigorous peer-reviewing process, a total of 17 papers were accepted. We believe that the selected papers will trigger further EIS research and improvements. We express our thanks to the authors for their valuable work and to the Program Committee members for their advice and support. At the same time, we would like to acknowledge the great support by Fudan University and the organization team for their timely contribution and help that made this edition of the conference possible.

Finally, we hope that CONFENIS 2017, as a platform for both academia and industry representatives to discuss the current issues in enterprise information systems, triggered innovative approaches in the different EIS areas.

October 2017

A Min Tjoa
Li-Rong Zheng
Zhuo Zou
Maria Raffai
Li Da Xu
Niina Maarit Novak

Organization

International Conference on Research and Practical Issues of Enterprise Information Systems – CONFENIS 2017

Honorary General Chairs

A Min Tjoa	IFIP TC 8.9 Chair, Vienna University of Technology, Austria
Li Da Xu	IFIP TC 8.9 Founding Chair and Vice Chair, Shanghai Jiao Tong University, China

General Chairs

Li-Rong Zheng	Fudan University, China
Zhuo Zou	Fudan University, China

Program Chairs

Ling Xia Li	Old Dominion University, USA
Maria Raffai	Szechenyi University, Hungary

Program Co-chairs

Hannu Tenhunen (Smart System Track)	KTH Royal Institute of Technology, Sweden
Rose Hannes (Industry Track)	BOSCH, Germany
Yuan Li (EIS Track)	Shanghai Jiao Tong University, China
Tomi Westerlund (IoT Track)	University of Turku, Finland

Publication Chair

Niina Maarit Novak	Vienna University of Technology, Austria
--------------------	--

Program Committee

Liuliu Fu	Tsinghua University, China
Xiaojun Xu	Jilin University, China

Ao Zhang	Jilin University, China
Xiaolin Zhou	Fudan University, China
Yong Chen	Pennsylvania State University, USA
Yang Lu	University of Manchester, UK
Anjee Gorkhali	Tribhuvan University, Nepal
Yongwei Zhong	Fudan University, China

Contents

EIS concepts, Theory and Methods

Modeling of Service Time in Public Organization Based on Business Processes	3
<i>Larisa Bulysheva, Michael Kataev, and Natalia Loseva</i>	
A Behavior Analysis Method Towards Product Quality Management.	12
<i>Congcong Ye, Chun Li, Guoqiang Li, Lihong Jiang, and Hongming Cai</i>	
Method of Domain Specific Code Generation Based on Knowledge Graph for Quantitative Trading.	21
<i>Jianshui Bi, Hongming Cai, Bo Zhou, and Lihong Jiang</i>	
Image Database Management Architecture: Logical Structure and Indexing Methods	34
<i>Larisa Bulysheva, Alexander Bulyshev, and Michael Kataev</i>	

IoT and Emerging Paradigm

Internet of Things or Surveillance of Things?	45
<i>Petr Doucek, Antonin Pavlicek, and Ladislav Luc</i>	
The Economic Value of an Emergency Call System	56
<i>Tomas Lego, Andreas Mladenow, Niina Maarit Novak, and Christine Strauss</i>	
An IoT-Big Data Based Machine Learning Technique for Forecasting Water Requirement in Irrigation Field	67
<i>Fizar Ahmed</i>	

EIS for Industry 4.0

Penetration of Industry 4.0 Principles into ERP Vendors' Products and Services – A Central European Study.	81
<i>Josef Basl</i>	
Systematic Analysis of Future Competences Affected by Industry 4.0	91
<i>András Gábor, Ildikó Szabó, and Fizar Ahmed</i>	
Process-Based Analysis of Digitally Transforming Skills	104
<i>Ildikó Szabó and Katalin Ternai</i>	

Big Data Analytics

Big Data Analytics – Geolocation from the Perspective of Mobile Network Operator 119
Antonin Pavlicek, Petr Doucek, Richard Novák, and Vlasta Strizova

Pattern Discovery from Big Data of Food Sampling Inspections Based on Extreme Learning Machine. 132
Yi Liu, Xin Li, Jianxin Wang, Feng Chen, Junyu Wang, Yiwei Shi, and Lirong Zheng

Big Data Analytics Using SQL: Quo Vadis? 143
K. T. Sridhar

Intelligent Electronics and Systems for Industrial IoT

Rethinking ‘Things’ - Fog Layer Interplay in IoT: A Mobile Code Approach. 159
Behailu Negash, Tomi Westerlund, Pasi Liljeberg, and Hannu Tenhunen

A Security Framework for Fog Networks Based on Role-Based Access Control and Trust Models 168
Farhoud Hosseinpour, Ali Shuja Siddiqui, Juha Plosila, and Hannu Tenhunen

IoT Platform for Real-Time Multichannel ECG Monitoring and Classification with Neural Networks 181
Jose Granados, Tomi Westerlund, Lirong Zheng, and Zhuo Zou

Deep Ensemble Effectively and Efficiently for Vehicle Instance Retrieval . . . 192
Zhengyan Ding, Xiaoteng Zhang, Shaoxi Xu, Lei Song, and Na Duan

Author Index 203

EIS concepts, Theory and Methods



Modeling of Service Time in Public Organization Based on Business Processes

Larisa Bulysheva^{1(✉)}, Michael Kataev², and Natalia Loseva³

¹ Old Dominion University, Norfolk, USA
lbulyshe@odu.edu

² Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia
kataev.m@sibmail.com

³ Regional Branch of the Social Insurance Foundation, Tomsk, Russia
lonat@bk.ru

Abstract. Time is essence in all processes involved while providing customer service. Very often the duration of a specific service is regulated by the law. However, it is also affected by other external and internal factors. Quantitative approaches to model the service time and creating systems that support the workflow management process are vital. This paper proposes a model of business process analysis on the operational level for various service-oriented organizations (including governmental agencies). The model estimates time of rendering services for heavily regulated public organizations. The analysis of the obtained results is discussed, potential applications are identified, and future research directions are formulated.

Keywords: Process-oriented approach · Business process · Mathematical model
Time of service · Workflow management · Service · Public organizations
Business decision making · Operational level · Tactical and strategic

1 Introduction

The application of business processes in the description of any enterprise or institution allows to formalize and organize core business processes. It is the transparency of the description of workflows is a crucial factor that allows the institution to achieve specified improvements in its activities. The process of developing and formalizing business processes is quite laborious and complex task, which in the end allows us to achieve success in comparison with competitors. Business processes allow you to get some “portrait” of the main working routine functions that have certain parameters.

These parameters are related to the organizational structure, legislative and regulatory documents, cost and resource factors. All the factors can be divided into external and internal. Each of them, influence business processes and leads to their change. So, to maintain a given level of success, you must constantly modify business processes. This change is the only what is constant in modern business, and the change is an iterative process in nature. Extensive research is done in business

process modeling, business performance management, and Enterprise Information systems [1, 2, 7–11]. Number of effective algorithms based on Petri nets were developed to model service workflows [3–7].

The model of activity of the enterprises built with the help of the business process underlies the new direction of the “virtual enterprise” [2, 6, 12]. This approach allows us to build different variants of the model, depending on the possible modeling purposes. The objectives of the simulation associated with the control, management, forecasting, etc. For example, the control problem requires constructing a model that would enable a “big picture” of the institution and see the quality and quantity of the production processes in a temporary mode.

In general, the review processes of the institution through business processes (each of which has the specified parameters), allows us to build for the manager the equivalent of a “business game”. The rules of the “games” allow the supervisor to observe each business process separately, as well as their combined activities. The quantitative aspect of this “game” leads to an understanding of the current state and the analysis of previous states, to assess the causes of the situation, also there is the possibility of forecasting.

Combination of technological aspects such as hardware, software and accumulated data, business processes, and people defines the needs for development of automated information system of control for public enterprise. Note, that this aspect is now rare considered in the literature in relation to public institutions providing services to general public. However, business performance management and measurement via identification of key performance indicators, usage of balanced scorecards, dashboards, and Six Sigma are well known. This allows the authors to reveal the importance of our research area and emphasize the value of formal methods development for modeling complex business processes and workflow management.

In this paper, we describe a mathematical model of transient processes of the public organizations activities at the lowest level (operational level) of business processes. We will start with a context diagram of business processes involved. Then, we will define the problems and formulate a mathematical model. Finally, we will conclude with examples of problem solved and discuss directions for future research.

2 Mathematical Model for Service Delivery Processes

The functional model, which reflects the currently existing processes of the subject area (model “as is”), has been created. Figure 1 shows the context diagram, where the major entities and the flows of data involved are identified. The flows of data of the model processes of the subject area which reflects the boundaries of the model in width and describes most of the processes associated with the performance of services are included.

Management system is formed in accordance with the existing “Rules and regulations”, where the service providers are “ministries and departments”, “local government”, “State, municipal and commercial institutions.” The process of rendering services can be described as a sequence of phases implemented by key stakeholders. Structural diagram of this process is shown in Fig. 2.

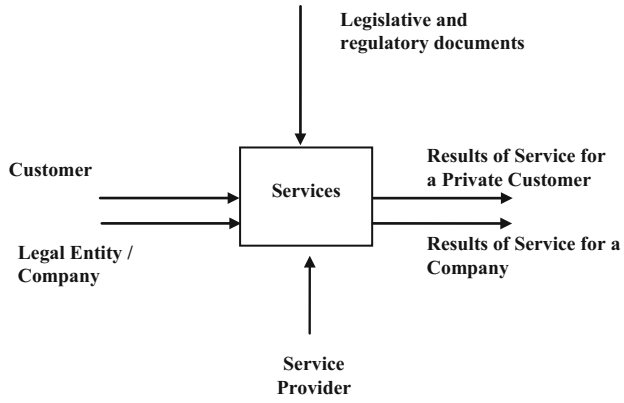


Fig. 1. Model of service execution.

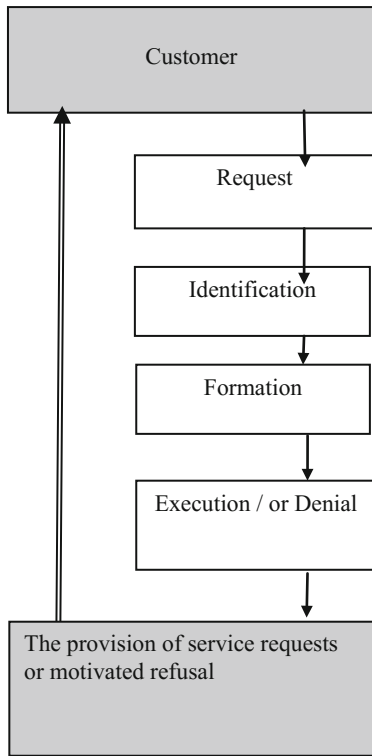


Fig. 2. Service rendering block diagram.

3 Problem Definition

Currently, there are many applications of business processes for solving problems of document management and business activities. However, there are no applications for management of the system mentioned above. There are only a few scientific papers devoted to this area of research. One reason for this is the fact that a qualitative approach is dominated in management activity of government institutions and quantitative is left out of researchers' attention.

Therefore, we propose to introduce the quantitative indicators, which are calculated based on business processes. Interdisciplinary integration of methods from operations research or management science including the linear programming models could be applied to formulate mathematical models for optimization of service provider operations in any business, including public.

Let's identify the main business-processes in providing services to public customers or legal entities. Typically, these activities are: "Service" or "Advice", or both. Figure 3 shows the partitioning of the service provider day associated with the execution of certain business processes per request of services. Every day is unique due to qualification of the employees, the influence of external and internal factors. In a separate time period can be performed a variety of sequence of business processes (BP), rest (To) or running errands (Tr), as shown in Fig. 3.

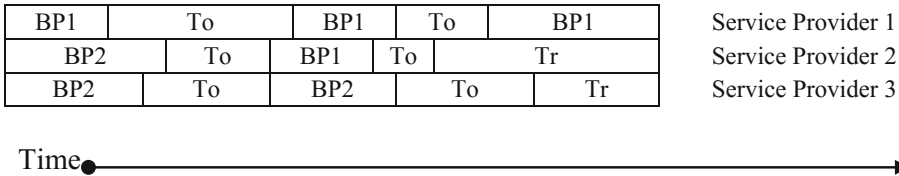


Fig. 3. The sequence of business processes execution in a workday.

During process of solving problems of service management many situations arise which could affect the quality of services. Figure 3 shows that the control over the activities of specialists is quite difficult without introduction of automation systems. One of the elements, which is needed to be considered, is time of service providing. Calculation of this time should be separated from the breaks time and time to implement other duties such as supervisor requests.

We propose a general mathematical formulation of the problem of organization management for service provision. The schematic of the model is shown in Fig. 4. This approach allows to see the process of providing services at different levels (operational, tactical and strategic) simultaneously.

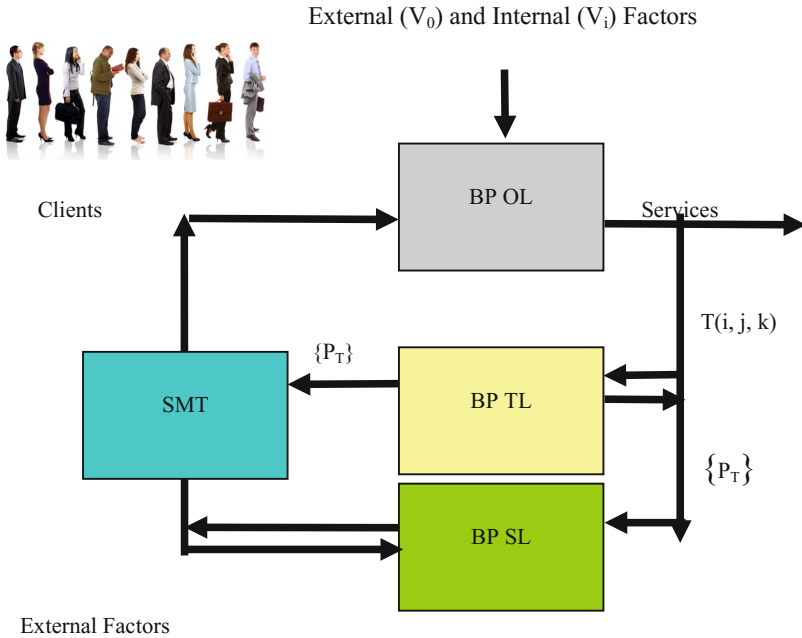


Fig. 4. Model of organization management for service provision.

Where BP OL – business processes on operational level,

BP TL – business processes on tactical level,

BP SL – business processes on strategic level,

SMT – system of generating the managerial action,

$\{P_T\}$ – the tactical level, calculated based on temporal measurements,

$T(i, j, k)$ the operational level.

The process of customer service, during the working time T_w (work = 8 h) can be represented at the operational level in the form of the following models:

$$T_w = T_r + T_0 + T_p \tag{1}$$

where T_r - is the total time used for completion of customer request via all business processes,

T_0 - time is not associated with the business processes, and

T_p - is the total time of performing supervisor requests.

Analyzing Fig. 3, we can notice many different sequences in business processes that could take place while processing a customer request in a public institution. Management of employees' performance as the service providers to the public is complicated and currently poorly automated. The decision to control through usage of the special devices (cards) is currently present, however, it is not yet universally implemented. How this information could be used in management activities is the topic that needs to be investigated further and presented in the research literature.

For the solution of managerial tasks, you need to have a model that adequately, at a quantitative level describes the daily activities of the organizations. Working day description on the average values of T_w, T_r, T_0, T_p is a well-known norm and enables the assessment activities at the tactical level (from weeks to months).

4 Model Description

Let's describe each element of the time model at the operational level of business processes in more details:

$$T_w = \sum_{i=1}^{i=N} \left[\sum_{j=1}^M (T_{i,j} + T_{0i,j}) + T_{pi} \right] \quad (2)$$

where N - is a number of employees ($i = 1, N$) and

M - a number of services ($j = 1, M$),

$T_{i,j}$ - execution time of the business process associated with the service j by employee i,

$T_{0i,j}$ - time between the processes, which are not directly associated with the implementation of business processes,

T_{pi} - time that the i-th employee spends on executing other duties.

The execution time of business processes can be written as:

$$T_{ri,j} = \sum_{k=1}^{k=K} \sum_{d=1}^{d=D} (t_{i,j,d,k,f} - t_{i,j,d,k,s}) \quad (3)$$

where K - total number of clients and D - number of days of service.

$t_{i,j,d,k,f}, t_{i,j,d,k,s}$ - time completion of f (finish) and the start (s) of the process s of j-th service by the i-th expert in day number d ($d = 1, D$).

The time between the business process services k and (k + 1) customer is written like this:

$$T_{0i,j} = \sum_{k=1}^{k=K-1} \sum_{d=1}^{d=D} (t_{i,j,d,k+1,f} - t_{i,j,d,k,s}) \quad (4)$$

where K - total number of clients and D - number of days of service provision for adoption of the report.

$t_{i,j,d,k,f}, t_{i,j,d,k,s}$ - time completion of f (finish) and the start (s) of the process s of j-th service by the i-th employee in day number d ($d = 1, D$).

The execution time of other duties of each employee i during the d-th day is calculated as follows:

$$T_{pi} = \sum_{d=1}^{d=D} t_{pi,d} \quad (5)$$

When calculating the total time of providing services, the next constraints must be considered:

$$(1) \quad (t_{i,j,d,k,f} - t_{i,j,d,k,s}) \leq 15 \text{ min (legally defined rule)} \quad (6)$$

(2) total working time that could be spent on providing services to population is specified as:

$$T_{\text{rab}} = N * D * 480 \quad (7)$$

(3) the waiting time in queue for customers must not exceed legally defined limits in minutes:

$$(t_{r,i,j} - t_{0,i,j}) / (KD) \leq 30 \text{ or} \quad (8)$$

$$(K \times t_{pi} - t_{0,i,j}) / (KD) \leq 30 \quad (9)$$

5 Solution of Managerial Problems Using Model

There are various situations that arise from influence of external and/or internal factors during the actual organization activities [9]. The reaction to these situations by the management team is to bring the project (business processes) into a desirable state. Managerial decisions presented as the following steps:

(Plan → Business process → Measurement → Analysis → Solution) [9] are cyclical in nature. The purpose of which is to detect deviations of the plan parameters (in our case, normative values) from the real values at this time. In the event of deviations from the plan there is a need to take certain administrative actions to return the situation to the plan or norm.

To make an accurate decision, management needs to know the problem which cause the situation. Empirically, it is possible to evaluate a set of circumstances that are related to the problem, and then to find the options that at the legislative level allow you to build managerial decisions. Thus, in the presence of some deviations from the planned values, the management team appears able to know the set of management decisions for direct decision making. This intellectual component of the decision-making process should contribute to the objectivity and efficiency of decision.

It is well known that the working activities of any organization linked to the business processes that are designed to provide services to general population or legal entities. Each business process is characterized by a set of input and output parameters, service providers, and managerial decisions that are based on legal and other regulatory requirements. Every day for each employee the basic elements of the expected work activities are well known, and therefore, it becomes possible to assess the main temporal elements of their activities during the day and during each work time period. The model that is presented above is based on this assumption.

Note, that it is also possible to build a model, given the fact that the working activities of each employee with various business processes are essentially sequential steps. The sequence of these steps might be recorded. For example, if we know the beginning and end of each process or if we know even just beginning time of each business process, then it gives us quantitative material for analysis. The results of the analysis offer an opportunity, despite various changes within the process, to identify the parameters to help in decision making process. Please note, that the classical ways of decision making, if we consider such approaches as scheduling, calendar planning and others [9–11], cannot be applied in this case because the business processes served by the staff coming in a random order with a random duration (see Fig. 3).

Considering that each public institution has its own specific personnel features, such as staff properties, age, qualification, culture, etc., which might be reflected in the style and types of managerial decision making, the use of the proposed model will allow to unify the decision making process, reduce the time and cost of rendering services, improve overall customer satisfaction, and to help in making higher quality and timely business decisions.

6 Potential Applications and Future Research

A specific feature of the temporary aspect of the organization of civil servants' activity is its intermittent nature, formed by the random arrival of different groups of clients, as well as by the large number of various orders coming from the management of different levels.

As a result, employees need to constantly adjust their plans and change time intervals in their activities:

- (1) the implementation of all business processes associated with customers in a given time limit;
- (2) performance (sometimes simultaneously with the main activity) of several orders with the same deadline and the same importance. These and other conditions for the temporary organization of civil servant activity are determined by the external and internal circumstances. The external frame of the temporal behavior is formed under the influence of external normatively given rhythms (in the activity of government bodies and civil servants) and the terms (execution of instructions). The internal frame of the temporary behavior of employees is related to solving the tasks of executing business processes for each of the clients according to the normatively defined timeframe.

When solving internal route tasks, important parts are cognitive, performing, communicative and emotional components for each employee (the components of the quality of the service). The timely completion of the assigned tasks of the state organization on rendering services to clients, in the presence of external and internal factors of the temporarily behavior, forces to involve new technologies. These technologies should make it possible to minimize the subjectivity in decision-making by the manager in the complex processes of daily routines allowing access to automated technologies.

Such technology may be the approach proposed in this paper, when the concept of a business process is introduced into the basis of the activity of a governmental organization. The system of business processes allows you to describe all the processes of labor when solving the main tasks of the organization, to evaluate objectively of the activities of the process executors and to perform forecasts. Our paper serves as the basis for the development of a formal approach that is considering uncertainties and varieties in the activities of the governmental organization presented above.

Future research will be focused on continuation of model development to make it even more adequate to the variety of business applications. It might be associated with considering additional parameters, such as qualifications of each service provider, the temporary features of the arrival of clients depending on various conditions, and other. As a result of modeling, we plan to provide recommendations for business process improvements, and to assist in developing of standardized protocols of employee actions with the goal to automate processing the main or routine operations.

Acknowledgments. The authors would like to thank the IFIP Confenis 2017, General Chairs: Dr. Zhou Zou and Dr. Li Rong Zheng, and IFIP WG 8.9 members for making this conference a great success.

References

1. Xu, L.: Enterprise systems: State-of-the-art and future trends. *IEEE Trans. Ind. Inform.* **7**(4), 630–640 (2011)
2. Kataev, M., Bulysheva, L., Emelyanenko, A., Emelyanenko, V.: Enterprise systems in Russia: 1992–2012. *Enterp. Inf. Syst.* **7**(2), 169–186 (2013)
3. Viriyasitavat, W., Xu, L., Viriyasitavat, W.: compliance checking for requirement-oriented service work flow interoperations. *IEEE Trans. Ind. Inform.* **10**(2), 1469–1477 (2014)
4. Viriyasitavat, W., Xu, L., Viriyasitavat, W.: A new approach for compliance checking in service workflows. *IEEE Trans. Ind. Inform.* **10**(2), 1452–1460 (2014)
5. Van der Aalst, W.M.P.: The application of Petri nets to workflow management. *J. Circ. Syst. Comput.* **8**(1), 21–66 (1998)
6. Zisman, M.: Representation, specification and automation of office procedures. Ph.D. dissertation, Wharton School of Business, Univ. Pennsylvania, Philadelphia, PA (1977)
7. Xu, L., Viriyasitavat, W.: A novel architecture for requirement-oriented participation decision in service workflows. *IEEE Trans. Ind. Inform.* **10**(2), 1478–1485 (2014)
8. Li, Y., Cao, L., Xu, L., Yin, J., Deng, S., Yin, Y., Wu, Z.: An efficient recommendation method for improving business process modeling. *IEEE Trans. Ind. Inform.* **10**(1), 502–513 (2014)
9. Tan, W., Shen, W., Xu, L., Zhou, B., Li, L.: A business process intelligence system for enterprise process performance management. *IEEE Trans. SMC Part C* **38**(6), 745–756 (2008)
10. Repin, V.: *Business Processes. Modeling, Implementation, Management*, M: Mann, Ivanov, Ferber (2013). (in Russian)
11. Kataev, M.: Process oriented approach to enterprise management. *Izvestia Tomskogo Polytechnicheskogo Universiteta* **313**(6), 20–23 (2012). (in Russian)
12. Kataev, M., Bulysheva, L., Xu, L.D., Loseva, N.: Influence of external and internal environment on the governmental institutions decision making processes. In: 2016 Proceedings of 22 International Scientific-Practical Conference, Tomsk, Russia, pp. 50–54 (2016). (in Russian)



A Behavior Analysis Method Towards Product Quality Management

Congcong Ye¹, Chun Li², Guoqiang Li¹, Lihong Jiang¹(✉), and Hongming Cai¹

¹ Shanghai Jiao Tong University, Shanghai, China

{yecongcong, li.g, jianglh, hmcai}@sjtu.edu.cn

² China Shanghai Institute of Precision Measurement and Testing, Shanghai, China

Abstract. Quality management is the basic activity in industrial production. Assuring the authenticity of testing datasets is extremely important for the quality of products. Many visual tools or association analysis methods are used to judge the authenticity of testing data, but it could not precisely capture behavior pattern and time consuming. In this paper, we propose a complete framework to excavate the features of testing datasets and analyze the testing behavior. This framework uses min-max normalization method to pre-process datasets and optimized k-means algorithm to label the training datasets, then SVM algorithm is applied to verify the accuracy of our framework. Using this framework, we can get the features of dataset and homologous behavior model to distinguish the quality of datasets. Some experiments are carried to measure the complete framework and we use various visual formats to show these results and to verify our method.

Keywords: Behavior analysis · Product quality management · Data mining

1 Introduction

Product quality testing data are the key factors of industrial production to discover inconsistencies and other anomalies. Ensuring the quality of data has been a continuing concern to industrial production. Testing the quality of electronic components is a tough and humdrum job. Testers are always not willing to do this work, and they sometimes directly write the testing records based on their own experience or others' data. The feature of recording result depends on the testers' psycho-behavior. Different people have different strategies and it is extremely difficult to distinguish real testing data from fictional data.

At present, there is not much researches on quality of testing data, which tends to be regarded as real and authentic data. Some engineers only use simple data visual methods like histogram and scatter diagram to show the feature of data and judge the reliability of testing result. It is hard to distinguish the real data from fictional data because testers write testing records within the allowable range of error. Illusory testing data will decline the quality of industrial products and damage the corporate image.

To address the aforementioned challenges, a complete behavior analysis framework is proposed to distinguish the reliability of electronic components testing behavior. We

use an optimized k-means clustering method to find the aberrant data and give every record a label. After that one third of the dataset are used as training data to optimize the kernel function of support vector machine(SVM) and another dataset are used to test the accuracy of our method. This optimized SVM can be used to distinguish the real testing data from illusory data.

Our main contributions are summarized as follows:

- A joint probability distribution method is used to prove that the relationships of attributes are not fixed. The illusory testing data can not be distinguished by merely association analyzing.
- An optimized clustering algorithm is proposed to mine the inner features of testing datasets. This cluster algorithm is the optimized k-means algorithm based on density and distance.
- A complete behavior analysis framework is constructed to associate the dataset with behavior. The data model and behavior model can be changed according to different features of dataset.

The rest of this paper is organized as follows. Section 2 discusses the related work in this field. Section 3 presents the detailed behavior analysis process including data pre-processing, data mining algorithm and behavior analysis. Section 4 uses various methods to show the result of our methods and analyzes the experimental result. Section 5 concludes the work in this paper and illustrates the future direction of this work.

2 Related Work

Data mining is the analysis of datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [1]. It contains four major branches: clustering or classification, association rules and sequence analysis [8]. Especially, clustering algorithm is very useful when the number of classification is unknown. The major clustering algorithms are k-means, hierarchical clustering and self-organizing map algorithm. The initial K-means clustering algorithm aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [9]. The implementation of k-means is simple and the distance function can be changed based on the feature of datasets. Through K-means is the most common clustering algorithm, there are still some disadvantages about K-means like the choice of centers, the number of centers and so on. Many methods have been proposed to improve the performance of the K-means clustering algorithm. [16] combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters.

As for hierarchical clustering, it yields a dendrogram representing the nested grouping of patterns and similarity levels at which groupings change and the clustering process is performed by merging the most similar patterns in the cluster set to form a bigger one [3]. Hierarchical clustering uses greedy algorithm, so the result of hierarchical clustering is only local optimum. Another common clustering algorithm is self-organizing map, which is a neural network algorithm to create topographically ordered spatial

representations of an input data set using unsupervised learning [10]. All these common clustering algorithms can be chosen based on the features of datasets.

Another important branch of data mining is classification. Decision trees [5] is one of the most popular methods for classification [7]. In decision tree algorithm, each internal node split the instance space into two or more parts with the objective of optimizing the performance of classifier and every path from the root node to the leaf node forms a decision rule to determine which class a new instance belongs to [4].

In general, various clustering and classification algorithms can be used to dig out the inner features of datasets. However, there are not many researches about human behavior analysis. Data quality analysis method is not enough to find the behavior pattern which is reflected by the datasets if it only focuses on data features.

3 Methodology

3.1 A Framework of Mining Behavior Model from Datasets

The features of datasets are the reflection of testers' behavior. In order to analyze the relationship between the data model and behavior model, we construct a behavior analysis framework as Fig. 1. This framework consists of four modules: data pre-processing, training data construction module, behavior analysis module and data visualization module. In data pre-processing module, we can delete the outliers and use min-max

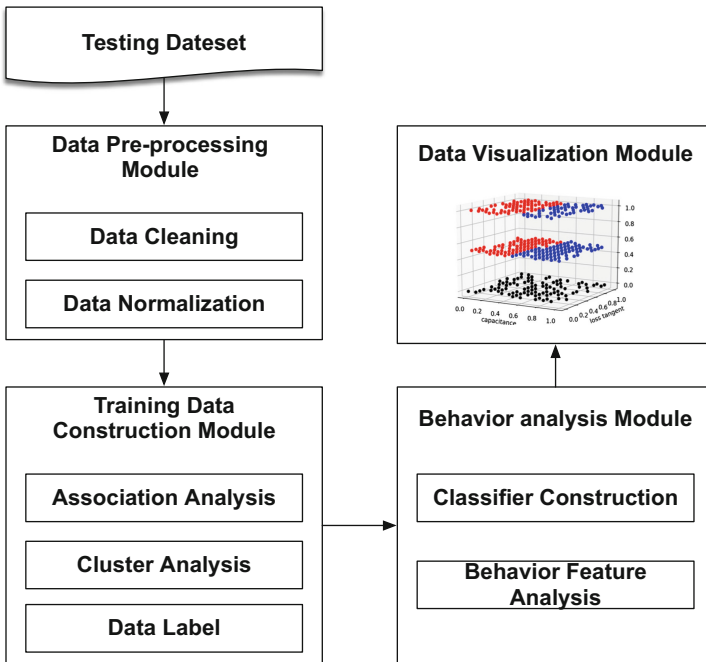


Fig. 1. A framework of mining behavior model from datasets

normalization method to process the datasets and improve the quality of datasets. In training data construction module, each dataset will be given a label through an optimized K-means algorithm and used to train the SVM classifier based on the association analysis and clustering result. At last, various visual methods are used to show the data mining results and behavior features.

Other data mining algorithms and behavior models, instead of k-means and SVM, can also be applied in case the datasets are replaced.

Data Pre-processing

Data pre-processing includes data preparation, compounded by integration, cleaning, normalization, transformation of data and data reduction tasks; such as feature selection, instance selection, discretization, etc. [6]. It will improve the quality of training data and increase reliability of data mining. First step of data pre-processing is data cleaning, which will dispose the noise and outliers. In order to ignore the effect of data dimension, the dataset will be normalized. There are many data normalization methods like min-max normalization, z-score normalization and decimal scaling normalization. The most common method is min-max normalization, which can transform every attribute into a value from zero to one. The calculation formula is defined as follow:

$$X^* = \frac{X - \min}{\max - \min} \quad (1)$$

After that we can use some attributes like mean value, variance, maximum, minimum and joint probability to excavate the features of data. All these features can reflect the fluctuation of every dataset and help engineers get an overview of dataset. However, only these methods are not enough for engineers to distinguish the authenticity of data and it needs more data mining methods to dig their inner features out.

Optimized Data Mining Algorithm

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems [11]. Clustering algorithms are usually used to process the datasets when their types of classification are not clear. In industrial production, the behaviors of testers depend on their different psychological activities. And we even don't know the number of different behavior types, which satisfies the conditions of clustering algorithms.

The most common clustering algorithm is k-means. Given a set of n points in d-dimension and the number of centers k, the problem is to determine k points in n as centers, so as to minimize the distance from each data point to its nearest center. The accuracy of k-means depends on the choice of k centers, which is also the main imperfection of k-means. In order to solve this problem and improve the accuracy of k-means, we optimize the original k-means algorithm and propose a k-means algorithm based on density and distance.

According to the relevant researches, the accuracy will be superior to others when the number of centers is set as the radication of the dataset number expect we have professional knowledge and ensure the clustering results. Center points are extremely crucial, but the original k-means algorithm chooses k centers randomly,

which is difficult to get the best cluster result. Then we optimize the strategy on choosing centers. Algorithm 1 represents the main framework of an optimized k-means algorithm based on density and distance. This algorithm has four parameters and the threshold values can be adjusted according to the features of datasets. This optimized method chooses the largest density point and other evenly distributed points as centers, which will avoid the situation that center points are excessive centralization. Using this framework, the noisy points or outliers will be ignored and the accuracy will be improved.

Algorithm 1: An optimized k-means algorithm based on density and distance

Input: Testing dataset d_n , the number of clusters k , threshold density m , threshold radius r

Output: Clustering result S_k

```

1: for  $i = 0$  to  $n$  do
2:   calculate the distance of any two points in dataset  $\text{dis}(i;j)$  ;
3: end for
4: delete the points when their density is lower than threshold density;
5: Collect the points in high density area  $D$ ;
6: find the largest density point as the first center, and put it in the center sets  $Z$ ,
   then delete it from  $D$ ;
7: find the furthest point from the first center and put it in the center sets  $Z$ , then
   delete it from  $D$ 
   repeat
8: calculate the distance product of every point in  $D$  with center sets;
9: find the furthest point and put it in  $Z$ ;
10: delete it from the set  $D$ 
11: Until the number of clusters do not reach  $k$ ;
12: put every point in  $D$  to the suitable center which is most similar with it;
13: return Clustering result  $S_k$ 

```

After clustering, we will give different clusters different labels and then use one third of these records about six hundred illusory and real records as training datasets to optimize the SVM classification. Finally, other datasets are used to judge the accuracy of our optimized algorithm and SVM. Not only this SVM can evaluate the accuracy of our labels, but also it can be used as an automated tool to distinguish illusory data quickly.

4 Experiments

The entire experiments were performed on a machine with 2.7 GHz Intel Core i5 cpu, equipped with 8G memory. The datasets come from an electronic component factory and they consist of four batches of product testing datasets. The dataset is defined as three dimensions, which are leakage current, minimum loss angle tangent and capacitance.

4.1 Association Analysis of Each Variate

Joint probability distribution [14] is an effective method to analyze the relationships between variates. In testing datasets, the crucial attributes are leakage current and minimum loss angle tangent. Hence we calculate their joint probability distribution and use 3-dimension figures to show the results.

Figure 2 contains four kinds of joint probability distributions on different batches testing datasets. Z-axis represents probability density and X or Y-axis are leakage current and minimum loss angle tangent. From these figures, we know the more points in front of the plane, the quality of datasets is higher. These four figures represent four types of the testing data: the multimodal state, the singlet state, the bowl state and the strip state. All these figures show the relationships between testing variates are not sure and it is not enough for us to distinguish the real datasets from others.

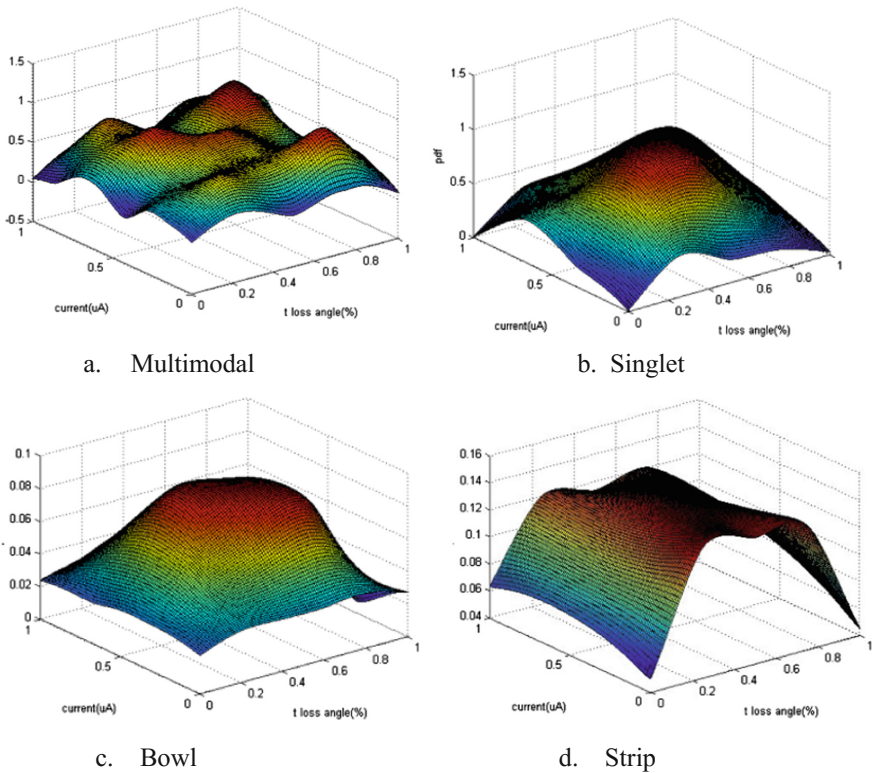


Fig. 2. Joint probability distribution of leakage current and minimum loss angle tangent

4.2 Data Clustering and Label Generation

According to the formed researches, only analyzing the relationships between variates are not enough to distinguish the illusory data, hence we adopt optimized k-means [13]

algorithm to dig out the inner features. Figure 3 is the clustering results on four batches datasets and these four batches all have six hundred records. The points in batch 1, 2 and 4 are very similar and they are equidistribution, but the points in batch 3 cluster in some fixed points, which illustrates the value of some attributes in batch 3 are same. The distribution in batch 3 is much similar to the crowd psychology phenomenon in social psychology. The crowd behavior is heavily influenced by the loss of responsibility of the individual and the impression of universality of behavior, both of which increase with the size of the crowd [2]. Hence we guess the third batch records are the illusory data. Then we use one third of real and illusory data as training datasets to optimize the kernel function and other datasets will be used to verify the accuracy of classifier.

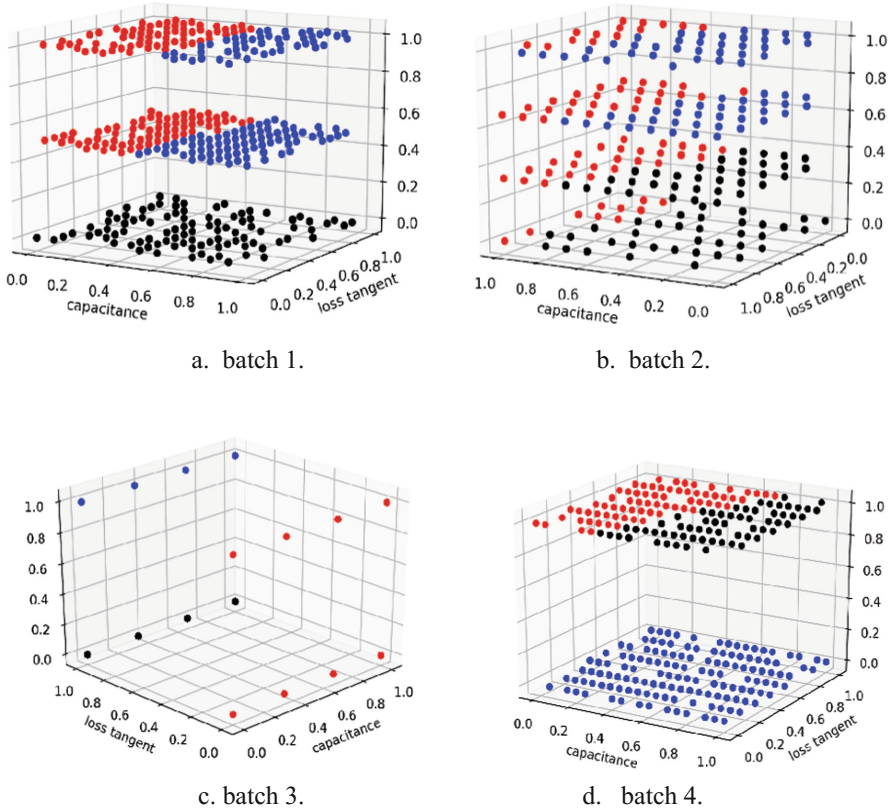


Fig. 3. The clustering result of optimized k-means algorithm

4.3 Data Classification and Testing Behavior Analysis

Support vector machine [15] is one of the classical machine learning techniques that can still help solve big data classification problems, especially, it can help the multi-domain applications in a big data environment [12]. According to the clustering results, the real datasets are not linearly separable and the points will spread over some special points

uniformly. The parameters of kernel function will be changed based on the training datasets and then it will find a support vector to separate the points into different clusters. The Fig. 4 is the classification result on other two third of datasets, which is t for our clustering result and speculation. Using 2400 records to verify the accuracy of classification, 2115 points are same as their labels and the accuracy is about 0.88125, which means most records can get a proper label. Hence this classification can be used to distinguish illusory data from real data quickly if the features of datasets are similar to ours.

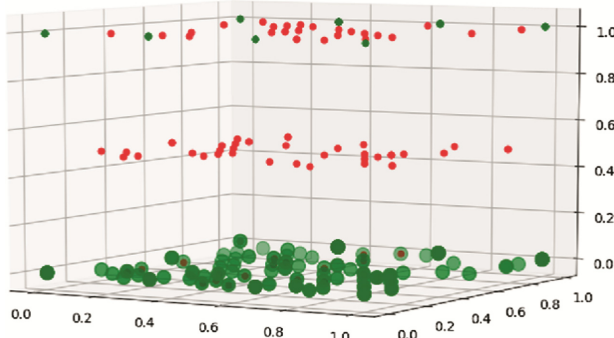


Fig. 4. The classification result of support vector machine.

From the data model, we can infer the testers' behavior. Most of the datasets are uniform distribution, only in batch 3, the points aggregate on some fixed points. This phenomenon is similar to the crowd psychology in social psychology. For example, if a tester is not willing to measure the variates of electronic components, they will see and copy others' records to decrease the risk of mistake. At first, they do not have a uniform criterion, so it will produce fluctuating results and be steady gradually. We still need more processes to confirm our conjecture.

5 Conclusion

In this paper, we propose a complete data mining framework to analyze the testing behavior and distinguish authenticity of testing records. This framework is generic and many datasets can use this framework to analyze their inner features and relevant behavior model. In order to increase the accuracy of clustering, we optimize the original k-means algorithm, which will make points spread uniformly. This classification can distinguish illusory datasets quickly, which will save much manpower and resource.

The proposed method still has some limitations. Firstly, the testing behaviors depend on different psychological activities, so every state of datasets are acceptable and we can only give a speculation which datasets may be illusory. Secondly the differences between real data and illusory are very tiny, and the common normalization method may decrease even eliminate these differences. In the future, we plan to implement more data mining algorithms in our framework to support other behavior analysis. And we can invite some professors to join the experiment to verify our clustering result. Meanwhile,

we will develop a data mining tool to detect the accuracy of testing datasets automatically, which will improve the quality of industry production.

Acknowledgement. This research is supported by the Shanghai Institute of Precision Measurement Project under Grand No. SAST2017-128 and the National Natural Science Foundation of China under Grant No. 61373030.

References

1. van der Aalst, W.M.P.: Process mining in the large: a tutorial. In: Zimányi, E. (ed.) eBISS 2013. LNBIP, vol. 172, pp. 33–76. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05461-2_2
2. Arain, S.M., Arain, A.M.: National Highways and Motorway Police in Pakistan: An Illuminative Study. Lulu.com, Morrisville (2016)
3. Bouguettaya, A., Yu, Q., Liu, X., et al.: Efficient agglomerative hierarchical clustering. *Expert Syst. Appl.* **42**(5), 2785–2797 (2015)
4. Dai, W., Ji, W.: A mapreduce implementation of C4.5 decision tree algorithm. *Int. J. Database Theory Appl.* **7**(1), 49–60 (2014)
5. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.* **40**(2), 139–157 (2000)
6. García, S., Luengo, J., Herrera, F.: Data Preprocessing in Data Mining. Springer, Cham (2015). <https://doi.org/10.1007/978-3-319-10247-4>
7. Hawkins, S.H., Korecki, J.N., Balagurunathan, Y., et al.: Predicting outcomes of nonsmall cell lung cancer using CT image features. *IEEE Access* **2**, 1418–1426 (2014)
8. Kesavaraj, G., Sukumaran, S.: A study on classification techniques in data mining. In: 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1–7. IEEE (2013)
9. Mundada, M., Gawali, B., Kayte, S.: Recognition and classification of speech and its related fluency disorders. *Int. J. Comput. Sci. Inf. Technol. (IJCSIT)* **5**, 6764–6767 (2014)
10. Rumbell, T., Denham, S.L., Wenckers, T.: A spiking self-organizing map combining stdp, oscillations, and continuous learning. *IEEE Trans. Neural Netw. Learn. Syst.* **25**(5), 894–907 (2014)
11. Satyanarayanan, K.S., Srikanth, B., Murugesan, M.: Tree dataset extraction using HAC based algorithm (2016)
12. Suthaharan, S.: Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. *ISIS*, vol. 36. Springer, Boston, MA (2016). <https://doi.org/10.1007/978-1-4899-7641-3>
13. Basu, S., Banerjee, A., Mooney, R.: Semi-supervised clustering by seeding. In: Proceedings of 19th International Conference on Machine Learning (ICML-2002) (2002)
14. Horn, J.T.H., Krokstad, J.R., Amdahl, J.: Joint probability distribution of environmental conditions for design of offshore wind turbines. In: ASME 2017 36th International Conference on Ocean, Offshore and Arctic Engineering. American Society of Mechanical Engineers, p. V010T09A068 (2017)
15. Kisi, O., Parmar, K.S.: Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *J. Hydrol.* **534**, 104–112 (2016)
16. Nazeer, K.A.A., Sebastian, M.P.: Improving the accuracy and efficiency of the k-means clustering algorithm. In: Proceedings of the World Congress on Engineering, vol. 1, p. 1–3 (2009)



Method of Domain Specific Code Generation Based on Knowledge Graph for Quantitative Trading

Jianshui Bi¹, Hongming Cai¹(✉), Bo Zhou², and Lihong Jiang¹

¹ Shanghai Jiao Tong University, Shanghai, China
{bijianshui, hmcai, jianglh}@sjtu.edu.cn

² Jiangsu Hoperun Software Co., Ltd., Nanjing, Jiangsu, China
zhou_bo1@hoperun.com

Abstract. Quantitative methods have been adopted by more and more individual investors for investment activities. Many third party platforms have been developed to help users complete the process of backtesting, which fills the gap between the trading strategy code and the trading strategy model. However, using a quantitative platform for backtesting has a high threshold for users who do not have programming experience. There is still a huge gap between the description and the code of trading strategy. Code generation allows developers to focus more on business related design and implementation, thereby increasing the efficiency of software development. The import of domain knowledge can improve the accuracy of requirement parsing to improve the quality of constructed code model. The general knowledge base is often incomplete in terms of domain specific terms and relationships, and the construction of domain knowledge graphs requires more domain related data. In this paper, encyclopedia data and the financial report data are used to extract domain terms and relations. And then a domain knowledge graph for quantitative trading is constructed to realize the automatic generation of quantitative trading strategy code.

Keywords: Code generation · Knowledge graph · Quantitative trading

1 Introduction

Quantitative trading takes the advantage of mathematical models instead of artificial subjective judgments, thus avoiding the adverse effects of human emotions in financial products trading. In addition to institutional investors, more and more individual investors have begun to invest through quantitative methods. The key step in establishing a quantitative trading model is to validate the effectiveness of the proposed model. Historical data are used to simulate the model performance in history. This process is often referred to the backtesting of a quantitative trading strategy.

There are many third party backtesting platform providers at present, which provides the historical market data and the transaction simulation program package. Users can write code of trading strategy on the platform, which will be processed based on historical market data using the trading simulated engine provided by the platform. Then an

evaluation report is generated to tell users the performance of the strategy in history, including the rate of return, the sharp ratio, etc., which is supposed to improve the strategy for users. These platforms enable users not need to obtain historical transaction data in advance, nor to deal with the logic of analog transactions, and to calculate the transaction results, greatly reducing the threshold for users to develop quantitative trading strategies.

However, the existing quantitative trading backtesting platform usually requires users to follow certain rules defined by the platform. Strategies are supposed to be written to code in a specific programming language. Users without engineering background usually have a high learning cost to be skilled in a programming language. And because different platforms are implemented based on different frameworks, the code of trading strategies between different platforms is often difficult to migration.

Code generation is to generate computer programs by some mechanisms, thus allowing software engineers to program at a higher level of abstraction, which can improve the quality of the code. Code generated quality only depends on the code template, code model and the data file. Developers can focus more on business related design and implementation, which improves software development efficiency. Code generation can also reduce the difficulty for code migration in different frameworks. It takes the business logic into the language independent model, and code for different frameworks can be generated by providing code templates for different frames based on the same business logic model.

Code generation is usually based on different forms of requirements descriptions, and requirements descriptions are highly relevant to application domains. Domain terms need to be identified in natural language requirement descriptions. Domain knowledge graphs, as a collection of domain terms and relationships, can be used for text requirement descriptions identification. This paper uses the domain specific data of open knowledge base and open data as the data source to build the knowledge graph in the field of financial analysis. By the analysis of the user's strategy description text based on the domain knowledge, the strategy code in backtesting platform can be generated. Therefore, users can focus more on the strategy, instead of the policy of written code about the specific platform. Users can develop quantitative trading strategies to guide investment activities without programming skills, which reduces the threshold for users to use the quantitative trading backtesting platform.

The paper is presented as follows: Sect. 2 describes the related work; Sect. 3 provides an overview of framework; Sect. 4 describes the proposed method for the domain knowledge graph construction for quantitative trading; Sect. 5 presents the method and the example of generate code for a specific backtesting platform; and Sect. 6 gives the conclusion of this paper.

2 Related Work

Code generation usually takes different forms of requirement description as input. Through the analysis and modeling of requirement description, combining code template or code generation rule, the program source code is generated. Requirement description

is usually based on the specific application domains, including a large number of terms in the field of text description. The identification of domain terms is a prerequisite for accurate identification of requirements in the domain text, as well the relationship between domain terms and program execution logic. As a collection of domain terms and relationships, domain knowledge graphs can play an important role in the identification of text requirement descriptions.

Alkhader et al. [1] used the natural language description of the requirements document as input, automatically generate the corresponding UML class diagram design. After processing with natural language related technology to generate requirements documents described in XML format, it is necessary for requirements engineers to use domain knowledge to manually eliminate redundancy and solve similarity problems. Popescu et al. [2] transformed the constraint grammar representation requirements specification statements into object oriented analysis model based on domain terms, to help identify the manual review of ambiguity and inconsistency. The grammar and part of speech rules are used to recognize the domain terms.

Bolloju et al. [3] introduced the knowledge based model quality and model ontology to evaluate the quality of object model, so that the model constructed has better semantic quality. Li et al. [4] proposed an engineering process that enables domain ontologies to guide the requirements elicitation process. Kong et al. [5] using the dom4j analytical framework and the Velocity template engine, automatic generation of database definition language, database manipulation language and the specific operation page code for information management system. Wei et al. [6] used predefined code templates to generate project code that meets the specific J2EE MVC framework based on Free-Marker. Lopata et al. [7] proposed that import enterprise model and enterprise meta model as knowledge data source in the model driven architecture development process.

As a collection of domain terms and relationships, domain knowledge graphs have important applications in the fields of natural language processing and other fields. The extraction of domain terms and relationships is the core work of domain knowledge graph construction. Hua et al. [8] used a rule-based approach to identify sentences containing technical terms from academic literature. Technical terms were extracted based on combining the rules and the vocabulary lists. Song et al. [9] used dependency parsing and automatic annotation of semantic roles to formulate extraction rules and weights, and extract knowledge units automatically from the term definition sentences.

The goal of relation extraction is to extract the fact between entities, and the relation extraction problem is usually abstracted into a two classification problem, that is, to determine whether the selected two entities have the relation. Perera et al. [10] put forward the characteristics of words, semantic features and other dimensions for the application of relation extraction. Chen et al. [11] proposed to use iterative methods to extract patterns of specific relationships based on predefined entity relationships, and then further extract relationships.

In this section, we present a brief overview of importing domain knowledge in code generation process, as well the term extraction and relation extraction in knowledge graph construction.

3 Overview of Framework

Figure 1 gives an overview of domain knowledge graph construction framework, while Fig. 2 gives the framework of strategy code generation based on the domain knowledge graph. The input in knowledge graph construction phase is the Wikipedia and Baidu encyclopedia data as the data source for term extraction and relation extraction, while XBRL (Extensible Business Reporting Language) format report data as the data source for relation extraction, and the output is the domain knowledge graph for quantitative trading, including the domain terms and the relations between domain terms. The input in strategy code generation phase is the strategy description written in natural language following some rules, and the output is the strategy code for a specific quantitative trading backtesting platform which is determined by the *code template* and the *comparison table* which is shown in Fig. 2.

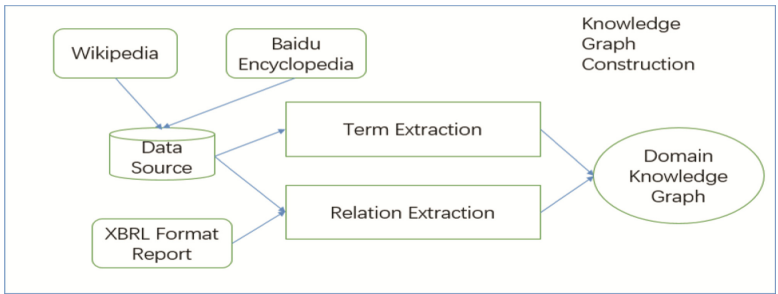


Fig. 1. Framework of domain knowledge graph construction.

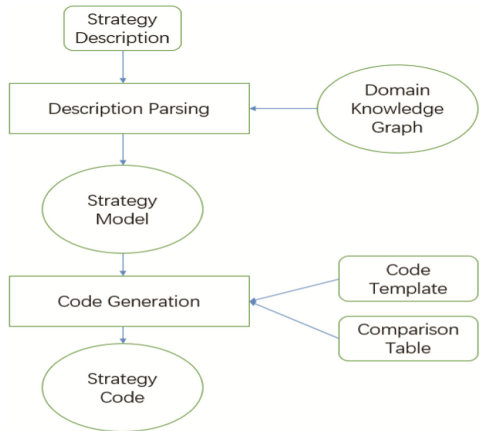


Fig. 2. Framework of strategy code generation.

The method of domain specific code generation based on domain knowledge graph for quantitative trading mainly has the following steps:

1. **Term Extraction:** In this step, the domain terms of financial analysis for quantitative trading need to be extraction for later relation extraction. Encyclopedia data are used as the data source to extract domain terms by defined rules according to the features different from common words.
2. **Relation Extraction:** In this step, the relation between domain terms needs to be extraction for domain knowledge graph construction. We use some domain specific text in open data in addition to encyclopedia data to establish the relation between domain terms extracted in above step.
3. **Description Parsing:** In this step, the quantitative trading strategy written in natural language need to be parsed. The domain terms are identified based on domain knowledge graph. And the domain knowledge in the domain knowledge graph is used to construct the strategy platform independent model.
4. **Code Generation:** In this step, the quantitative strategy code can be executed in a specific quantitative trading backtesting platform will be generated based on the strategy model and the code template related to the specific backtesting platform.

In this section, we give a brief overview of the framework we proposed in this paper. We use encyclopedia data and domain specific data in open data to extract domain terms and the relation between domain terms. And a domain knowledge graph for quantitative trading is constructed to provide domain knowledge in domain specific code generation.

4 Domain Knowledge Graph Construction

In this section, we describe the method of constructing a domain knowledge graph in financial analysis for quantitative trading, including the domain terms exaction and the term relation extraction.

4.1 Domain Terms Extraction

To construct a domain knowledge graph, we need to extract the terms in the domain firstly. This paper builds a knowledge graph that applies to financial analysis, so it is necessary to extract the terms in the field of financial analysis. There are some entries to summarize the relevant terms in the encyclopedic data and domain terms can be extracted from them. For example, *financial ratios* entry in Wikipedia and Baidu encyclopedia includes *inventory turnover rate* and *total asset turnover ratio*, and *financial analysis entry* in Baidu encyclopedia contains *price earnings ratio*, *price to book ratio*, etc. An example of financial ratio entry in Wikipedia is shown in Fig. 3.

Ratios [\[edit \]](#)

Profitability ratios [\[edit \]](#)

Profitability ratios measure the company's use of its assets and control of its expenses to generate an acceptable rate of return

Gross margin, Gross profit margin or Gross Profit Rate^{[7][8]}

$$\frac{\text{Gross Profit}}{\text{Net Sales}} \quad \text{OR} \quad \frac{\text{Net Sales} - \text{COGS}}{\text{Net Sales}}$$

Operating margin, Operating Income Margin, Operating profit margin or Return on sales (ROS)^{[9][9]}

$$\frac{\text{Operating Income}}{\text{Net Sales}}$$

Note: Operating income is the difference between operating revenues and operating expenses, but it is also sometimes used as a synonym for EBIT and operating profit.^[10] This is true if the firm has no non-operating income. (Earnings before interest and taxes / Sales^{[11][12]})

Profit margin, net margin or net profit margin^[13]

$$\frac{\text{Net Profit}}{\text{Net Sales}}$$

Return on equity (ROE)^[13]

$$\frac{\text{Net Income}}{\text{Average Shareholders Equity}}$$

Return on assets (ROA ratio or Du Pont Ratio)^[6]

$$\frac{\text{Net Income}}{\text{Average Total Assets}}$$

Fig. 3. An example of financial ratio entry in Wikipedia.

These domain terms in summary entries usually have common characteristics. For example, in the field of financial analysis, domain terms usually associated with formula. At the same time the related domain terms also has some domain independent common characteristics. For example, if the word contains hyperlinks to the other entry, the word is more likely to be a term. In this paper, these features are used to define rules for extracting domain terms of financial analysis from relevant summary entries. And These domain terms will be used for subsequent relation extraction and domain knowledge graph construction.

4.2 Term Relation Extraction

In the previous section, the extraction of domain specific terms has been completed. This section describes the method of extracting the relationships between domain terms based on the domain terms extracted from the previous section. The data source of relation extraction mainly includes two parts, one is the entry pages in the encyclopedia, and the other is the other open data related to the field as a supplement.

For the term entry page in the encyclopedia, we use pattern matching method to extract relations between domain terms. An example of same as relation in *price earnings ratio* entry in Wikipedia is shown in Fig. 4. For example, *price earnings ratio* entry has the following statement in Baidu encyclopedia, also known as *market price earnings ratio*, *pe ratio*, and *PER*. The *price earnings ratio*, *market price earnings ratio*, *pe ratio*, and *PER* are the same meaning in the field of financial analysis can be represented by three tuples are as follows:

- <price earnings ratio, same_as, market price earnings ratio>
- <price earnings ratio, same_as, pe ratio>
- <price earnings ratio, same_as, PER>

Price–earnings ratio

From Wikipedia, the free encyclopedia

The **price/earnings ratio** (often shortened to the **P/E ratio** or the **PER**) is the ratio of a company's stock price to the company's earnings per share. The ratio is used in valuing companies.

Fig. 4. An example of same as relation in price earnings ratio entry in Wikipedia.

For the extraction of relation between domain terms, the formula of field terms can be obtained by matching the patterns of specific computational symbols. For example, with the term *quick ratio*, the Baidu encyclopedia describes the entry as Eq. (1).

$$\text{Quick Ratio} = \text{Quick Assets} / \text{Current Liabilities} \quad (1)$$

The segmentation of the formula to a symbol of operation, we can get the relationship in three domain terms about *quick ratio*, *quick assets* and *current liabilities*, and the *quick ratio* is equal to the *quick assets* divided by *current liabilities*. This relationship can be represented by a set of tuples:

- <quick ratio, f1_dividend, quick assets>
- <quick ratio, f1_divider, current liabilities>

The *f1* in the above tuples means that the tuples are number 1 group statements for the *quick ratio* term. In this relation, *quick assets* is as the dividend in *quick ratio* calculation, and the *current liabilities* is as the divider in the *quick ratio* calculation. The relationship can be expressed as *quick ratio* equal to the *quick assets* divided by the *current liabilities*.

For the domain terms extracted from the above mentioned method, the relations are extracted from the corresponding entries in the encyclopedia. The relationship between domain terms can be obtained to establish a domain specific knowledge graph.

However, the encyclopedia data usually depends on the manual editing, which means the cover of the specific domain terms may not be very complete. This paper uses other open data related to this domain, in order to enhance the degree of coverage of the domain terms. XBRL (Extensible Business Reporting Language) format of standard financial report has been more and more accepted by the exchange and the listed company. The calculation link library including in the XBRL format financial report can be used as a supplement for encyclopedia of data to domain term relation extraction.

XML format is used in extensible business reporting language format of financial report to describe the calculation relation. We defined rules to transform XML format description to our three tuple format in domain knowledge graph.

The following is an example of an extensible business reporting language format of financial report:

- <arcrole = <http://www.xbrl.org/2003/arcrole/summation-item> from = “gross profit” to = “sales revenue” order = “1” weight = “1”/>
- <arcrole = <http://www.xbrl.org/2003/arcrole/summation-item> from = “gross profit” to = “cost of sales” order = “2” weight = “-1”/>

The <http://www.xbrl.org/2003/arcrole/summation-item> means the items described in the code are the total sum of the relationship. In this relationship, the *gross profit* is the summary item, while the *sales revenue* and the *cost of sales* are the add items. The order attribute means that in this relationship, *sales revenue* is the first and *cost of sales* is the second. The weight attribute means that in this relationship, the weight of *sales revenue* is 1 and the weight of *cost of sales* is -1. The code fragment represents the relationship between *gross profit*, *sales revenue* and *cost of sales*, follows the Eq. (2).

$$\text{Gross Profit} = \text{Sales Revenue} - \text{Cost of Sales} \tag{2}$$

Converting the above relation into a three tuple format expresses as follows:

- <gross profit, fl_minuend, sales revenue>
- <gross profit, fl_subtrahend, cost of sales>

Figure 5 shows a part of constructed domain knowledge graph for quantitative trading. The relation between domain terms in constructed domain knowledge graph can be represented by tuples. The *net operating profit rate* is equal to the *net profit* divided by *operating income* as an example, the following tuples can be used to be the representation of the relationship.

- <net operating profit rate, fl_dividend, net profit>
- <net operating profit rate, fl_divider, operating income>

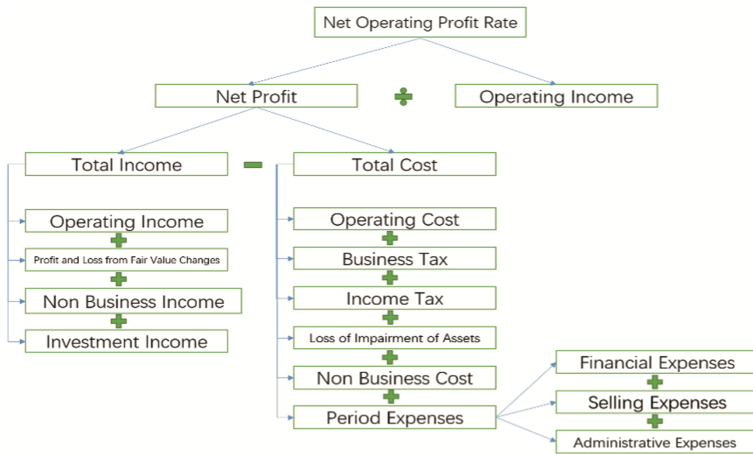


Fig. 5. A part of constructed domain knowledge graph for quantitative trading.

Two tuples above mean that the *net operating profit rate* can be obtained by the calculation of the *net profit* and the *operating income*, the *net profit* for the dividend, the *operating income* for the divider, namely the *net operating profit rate* is equal to the *operating income* divided by the *net profit*. And *fl* means it is the first groups related to the *operating net profit rate*. The same entity tends to have multiple sets of relational

representations, such as *price earnings ratio*, which can be calculated either by *stock prices* and *earnings per share*, or by *market capitalization* and *net profit*.

For the *total income* equals to the summary of *operating income*, *profit and loss from fair value changes*, *non-business income* and *investment income*, the following four tuples can be expressed:

- <total income, f1_addend, operating income>
- <total income, f1_addend, profit and loss from fair value changes>
- <total income, f1_addend, non-business income>
- <total income, f1_addend, investment income>

This section presents the method of constructing domain knowledge graph for quantitative trading using the encyclopedia data and the domain specific data, including the method of term extraction and the method of relation extraction.

5 Domain Specific Code Generation

In this section, we describe the method and the example of generating code for a specific backtesting platform, including the strategy description parsing and the backtesting platform code generation.

5.1 Strategy Description Parsing

The user's strategy description can be parsed based the domain terms in the domain knowledge graph constructed above. The strategy description needs to include the start date and the end date of the backtesting, trading target, and the execution logic of the strategy. A description of the user strategy logic condition needs to be followed by certain priority order rules.

In the following strategy description as an example, the trading target is ticker 000540, if the net operating profit rate is more than 20% and the current stock price is less than 10 yuan, or the moving average of the stock price in 5 days is more than the moving average in 10 days, open (buy); otherwise close (sell). The start date of backtesting is January 1, 2016, and the end date of backtesting is December 31, 2016.

The generated strategy model is illustrated as Fig. 6.

In the phase of data initialization, the backtesting start time (20160101), end time (20161231), and trading target (000540) are initialized. And the moving average of the stock price in 5 days, the moving average in 10 days and the *net operating profit rate* which required for the *net profit* and *operating income* (or further the *total cost*, *total income*, and *operating income*) need to be extracted in this phase as well. Then the data will be used in the strategy simulation stage to determine the conditions of strategy by calculation.

And in the phase of the strategy simulation, by modeling strategy logic, based on the extracted data in the initialization phase, the target backtesting platform uses historical market data to calculate whether the stock price or other data meets the strategy opening or closing conditions. Between the start date and the end date, the backtesting platform

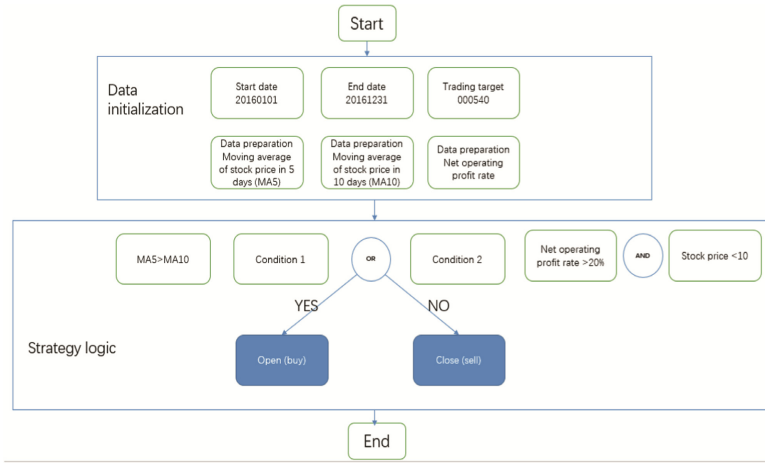


Fig. 6. An example of generated strategy model.

will simulate the trading according to the strategy conditions by calling the logic of the strategy. Then the backtesting platform will generate a report about the strategy performance in the period determined by the start date and the end date by processing the simulated trading record. The report usually includes indicators such as return rate, sharp ratio, etc., which provides a reference for the users to improve the strategy.

5.2 Backtesting Platform Code Generation

The section above describes the method to transform the user requirement descriptions of quantitative trading strategies to platform independent model based on the constructed domain knowledge graph for generating backtesting platform strategy code. This section will describe the method of generating strategy code which can be executed by the backtesting platform based on the platform independent model and the platform specific code template.

For specific quantitative trading backtesting platform, code templates are written using the FreeMarker template language defined by the FreeMarker template engine. The code to specify the trading target, start date, end date of backtesting and extract the data in the initialization phase should be included in the templates. And the templates should also include the code to generate the logic code of the strategy according to the strategy logic model. The platform should also give a comparison table to define the transform relation between the term in domain knowledge graph and the data field name in the platform. The FreeMarker engine is used to generate code that can be executed by a specific quantitative trading backtesting platform, based on the strategy model which is constructed by parsing strategy descriptions.

Figure 7 shows the generated code of data initialization, including the trading target, start date, end date and the moving average of the stock price. Figure 8 presents the generated strategy code and Fig. 9 gives an example of trading strategy backtesting report generated by a backtesting platform.

```

def initialize(self):
    ticker_id = '000540'
    from_time = time.strptime("2016-01-01 09:15:00", "%Y-%m-%d %H:%M:%S")
    to_time = time.strptime("2016-12-31 15:15:00", "%Y-%m-%d %H:%M:%S")
    self.add_data('Tick', ticker_id, 'data', ['TimeNum', 'LastPX'],
                 time.mktime(from_time), time.mktime(to_time))
    self.add_transform('MovingAverage', ['data'], ['LastPX'], 600)
    self.add_transform('MovingAverage', ['data'], ['LastPX'], 2400)

```

Fig. 7. An example of generated data initialization code.

```

def handle_data(self, data):
    ticker_id = '000540'
    now_time = data['data.TimeNum']
    price = data['data.LastPX']
    short_ma = data['data.LastPX.moving_average600']
    long_ma = data['data.LastPX.moving_average2400']

    if (self.compare == -1 and short_ma > long_ma):
        if (self.ordered == False):
            self.order(ticker_id, True, True, 'LimitOrder', price + 0.02, 100, now_time)
            print now_time
            self.trade_is_buy = 1
            self.ordered = True
        else:
            self.order(ticker_id, True, False, 'LimitOrder', price - 0.02, 100, now_time)
            print now_time
            self.trade_is_buy = 0
            self.ordered = False
    elif (self.compare == 1 and short_ma < long_ma):
        if (self.ordered == False):
            self.order(ticker_id, False, True, 'LimitOrder', price - 0.02, 100, now_time)
            print now_time
            self.trade_is_buy = -1
            self.ordered = True
        else:
            self.order(ticker_id, False, False, 'LimitOrder', price + 0.02, 100, now_time)
            print now_time
            self.trade_is_buy = 0
            self.ordered = False
    if (short_ma > long_ma):
        self.compare = 1
    elif (short_ma < long_ma):
        self.compare = -1

```

Fig. 8. An example of generated strategy logic code.

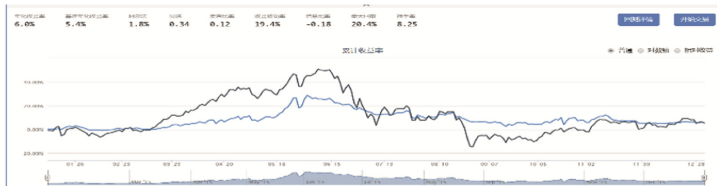


Fig. 9. An example of backtesting platform generated report.

This section we describe the method and the example of generating code for a specific backtesting platform. And the method of strategy description parsing and the backtesting platform code generation are presented in detail.

6 Conclusion

The third party backtesting platform providers make it is possible for individual investors to develop quantitative trading strategies without establishing a complete backtesting framework. However, it is difficult for users to migrate the strategy code because of the difference between these platforms. And there is also a huge gap between the description and the code of trading strategy especially for investors without programming skills.

Code generation is usually used to enhance the efficiency of application development. We use code generation related technologies in this paper to generate quantitative strategy code. In previous research, domain knowledge is rarely used in the code generation process. This paper uses the relevant data of open knowledge base and open data as the data source, and constructs a domain knowledge graph in the field of financial analysis to analyze user's strategy description text. Then quantitative trading strategy code in the backtesting platform is generated to reduce the threshold for users to use the backtesting platform. It allows users to focus more on research and development of quantitative trading strategies, while the development of the strategy can be easier migrated among different platforms. This paper has constructed the domain knowledge graph of financial analysis applying in Chinese listed companies. While the construction of knowledge graph depends on the financial statements of listed companies to disclose China criteria, because of different national accounting rules and legal differences, which may not be applied to the process of financial analysis of foreign companies directly.

In the future research, we will apply domain knowledge graph in more fields, such as intelligent agriculture, where knowledge graph could be used for agricultural product tracking.

Acknowledgement. This research is supported by Key R&D Project of Zhejiang Province under Grand No. 2017C02036.

References

1. Alkhader, Y., Hudaib, A., Hammo, B.: Experimenting with extracting software requirements using NLP approach. In: International Conference on Information and Automation, pp. 349–354. IEEE Xplore (2007)
2. Popescu, D., Rugaber, S., Medvidovic, N., et al.: Reducing ambiguities in requirements specifications via automatically created object-oriented models. In: Innovations for Requirement Analysis. From Stakeholders' Needs to Formal Designs, Monterey Workshop 2007, Monterey, CA, USA, 10–13 September 2007. Revised Selected Papers, pp. 103–124. DBLP (2007)
3. Bolloju, N., Sugumaran, V.: A knowledge-based object modeling advisor for developing quality object models. *Expert Syst. Appl.* **39**(3), 2893–2906 (2012)
4. Li, G., Jin, Z., Xu, Y., et al.: An engineerable ontology based approach for requirements elicitation in process centered problem domain. In: Knowledge Science, Engineering and Management - International Conference, KSEM 2011, Irvine, CA, USA, 12–14 December 2011, Proceedings, pp. 208–220. DBLP (2011)
5. Kong, D., Luo, F., Lin, W., et al.: Research on a velocity-based automatic code generation technology. *Computer Applications & Software* (2014)
6. Wei, Y.: Backstage management system code generator based on J2EE and Maven. *Comput. Modernization* **2**, 012 (2014)
7. Lopata, A., Ambraziunas, M.: Knowledge-based MDA approach. In: Business Information Systems Workshops - BIS 2011 International Workshops and BPSC International Conference, Poznań, Poland, 15–17 June 2011, Revised Papers, pp. 160–165. DBLP (2011)
8. Hua, B.: Extracting information method term from Chinese Academic Literature. *New Technol. Libr. Inf. Serv.* **6**, 68–75 (2013)

9. Song, P., Lu, Q., Liu, N.: A new method for knowledge unit automatic extraction using definitions of terms. *J. Intell.* **33**(4), 139–143 (2014)
10. Perera, R., Nand, P.: A multi-strategy approach for lexicalizing linked open data. In: Gelbukh, A. (ed.) *CICLing 2015*. LNCS, vol. 9042, pp. 348–363. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18117-2_26
11. Chen, C., He, L., Lin, X.: REV: extracting entity relations from world wide web. In: *International Conference on Ubiquitous Information Management and Communication*, pp. 1–5. ACM (2012)



Image Database Management Architecture: Logical Structure and Indexing Methods

Larisa Bulysheva^{1(✉)}, Alexander Bulyshev², and Michael Kataev³

¹ Old Dominion University, Norfolk, USA
lbulyshe@odu.edu

² NextVR, Inc., Newport Beach, USA
abulyshev@nextvr.com

³ Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia
kataev.m@sibmail.com

Abstract. Visual information is an important type of information in modern life. However, it is still not used by organizations in a full capacity. The major reason for that is the lack of internal structure of visual information. The existence of this structure in numerical data allows to build very effective tools for classification, storage, and retrieval of numerical information, such as a relational data management system. In case of visual information, each value of the picture is basically meaningless, but the set of pixels starts carry meaningful information. In this paper, we aim to classify different types of images based on the areas of origination and application. We also suggest the possible structure of the database management system with images as elements of it. Another objective is to propose the indexing methods, which allow to avoid the direct comparison of visual query consequently to entire database. We also introduce the idea of applying multi frame super-resolution method to development of store-retrieval procedures for a database with dynamical visual information.

Keywords: Image database management · Super-resolution · Visual data
Data indexing · Hash function · Content-based image retrieval (CBIR)
Industrial information integration systems · Video database

1 Introduction

Accumulations of enormous amount of visual information, such as industrial images, aerial and satellite images, medical images, and others require development of new approaches on how to store, pre-process and retrieve it. The challenge to make of use of Big Data or vast information generated by industrial information integration systems should be addressed [1–3]. In [4], one of the co-author of the current paper proposed a hybrid method of visual database organization and image retrieval. The main idea is to combine annotation approach with SBIR approach. In [5], authors investigated several types of visual data, and proposed indexing algorithms for each of those types.

In general, there are two major approaches to store and retrieve visual information: text-based approach and content-based image retrieval (SBIR) approach. First approach

operates not with image itself, but with a text which accompanies the image (annotation). In some sense the annotation serves as an index. When descriptive and reliable annotation exists, this method is very effective and efficient. However, it is rare the case. Since annotations are produced manually most of the time, the retrieval procedures suffer with low reliability.

Another approach extracts all necessary information from images themselves [6]. As a rule, intensity histograms and RGB proportions are major components. Both characteristics are important; however, they cannot be universal measures of image differences. There are some attempts (see for example [7–9]) to add the scene characteristics to the retrieval process, but they are very objects specific, and it is difficult to generalize those approaches.

In the current paper, we discuss extended version of the hybrid approach, and also discuss the possibility of database organization for storage and retrieval of data based on video information.

2 Different Types of Search

Talking about visual databases, authors very often mention entirely different types of storage and retrieval procedures. First, web-based storage is completely non-organized structure, where search is stochastic and produced “hits” or something similar on the query image. Second, database type storage where all elements are allocated into some logical structure, the search is deterministic, and the results are unique (including empty output).

In the current paper, we describe approach to image information organization, storage, and retrieval related to second type. Before an image is placed into database, it is pre-processed, annotated, and indexed [5]. The process of annotation and indexing depends on the nature of images. As was discussed in [5], the numerical indexing algorithms can be produced for different types of image collections: medical, satellite, aerial, virtual show room, forensic science. In all this cases annotation can be done based on meta information associated with the image itself. Simple but effective and efficient indexed method based on explicit hash functions are described in [5].

Much more complicated situation arises, when collection of images comes from variety of non-organized sources, such as the Internet, images collections, surveillance images and videos. In this case, we propose to use B+ tree indexing algorithm [10–12].

3 Extended Hybrid Method

3.1 Stored Procedures

As was described in [4, 5], the hybrid method was intended to combine strengths of both approaches mentioned above. Prior the storage procedure, all images should be annotated automatically. Image processing steps can be described as shown in Fig. 1.

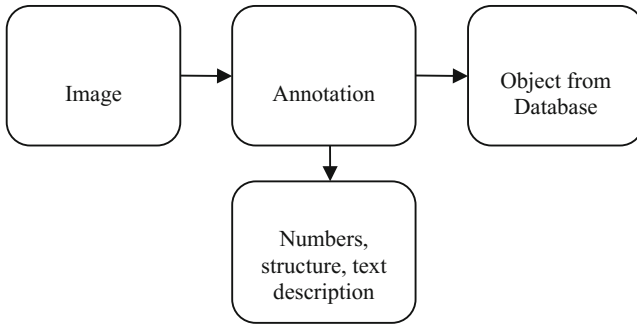


Fig. 1. Image processing steps.

The main idea of the method, that the image is associated with physical object(s). The image database stores formal object descriptions (classification) and detailed categorical or numerical attributes of the object.

Annotations needs to be stored in the database for annotations together with links to the images themselves. Annotations should be converted into indexes. For medical, satellite, aerial, industrial, and surveillance databases an explicit expression of hash functions can be formulated. Using explicit formula with just a few arithmetical operations as a hash function is the most effective and efficient choice. Details are presented in [5].

In case of general image collections or for collections with a priory unknown internal logical structures a B+ tree algorithms [10, 11] is a natural choice. Image annotation needs to include the following attributes shown in Fig. 2.

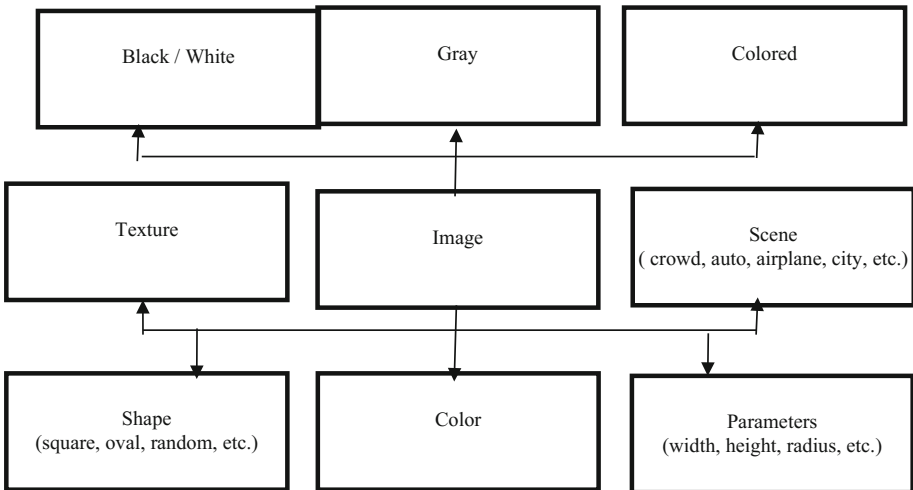


Fig. 2. Basic image attributes.

Basic image attributes should be extracted and stored on the highest level of description. Attribute “Scene” will be determined as the result of imaging analysis. Major steps of the process are shown in Fig. 3.

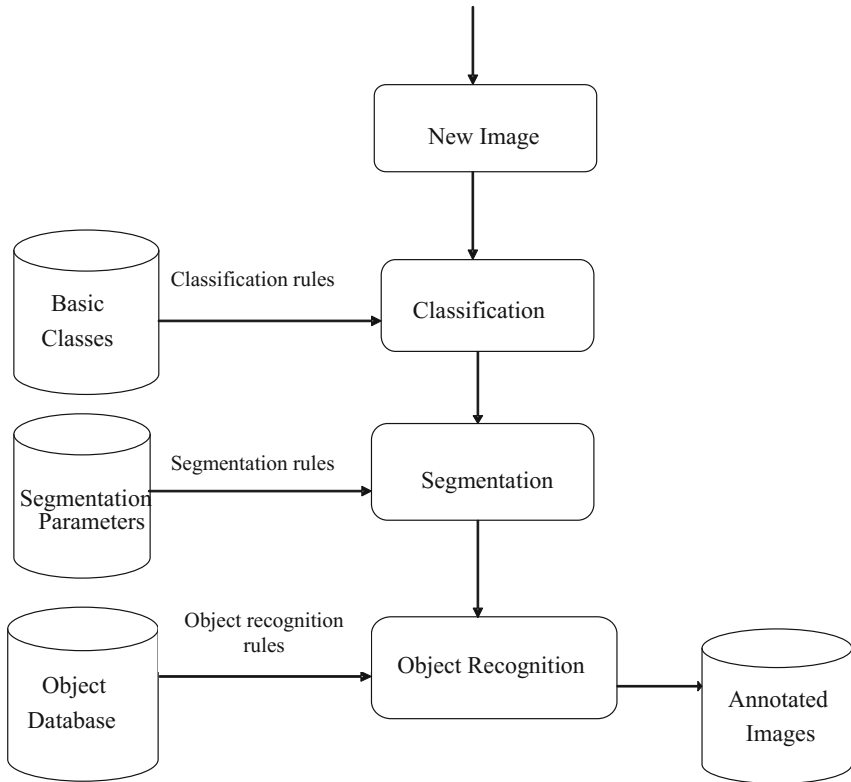


Fig. 3. Image processing block diagram of storage procedure.

The following steps need to be implemented to prepare annotation:

- Image Classification
- Image Segmentation
- Object Classification
- Annotation Generation

3.1.1 Image Classification

Images can be classified according to the image types. Following classes can used [5]:

- Photographic images
- Medical images
- Satellite images
- Aerial images

- Industrial images
- Surveillance images
- Unstructured and semi structured documents.

Images can be easily classified and automated by identifying the sources where these images are acquired.

3.1.2 Image Segmentation

The major purpose of the annotation process is to understand what kind of object is represented in the image. However, in general case many different objects can be presented simultaneously. It makes the segmentation of image the important step. Image segmentation is a decomposition process, when parts of image with different characteristics are separated from each other. Image segmentation provides data compression also. The techniques for data segmentation are very mature now [13–16]. However, the segmentation method which is capable of complete automation is still a challenging problem. We plan to use well known methods [13, 14] to segment images. The result of image segmentation will be a plurality of sub-images derived from an original image.

3.1.3 Object Classification

After image is segmented into set of smaller images, each of which contains individual object, these objects should be identified by comparison with examples from database. This step is critical for entire method. There were many attempts to solve this problem for different types of objects [17, 18]. We plan to use multi step hierarchical approach.

The first task is to determine an object class: building, road, person, animal, crowd, and so on. For this purpose, object database contains major features of the objects of a given class. The most popular example is a face recognition. One of the well-known



Fig. 4. The areas of human faces are underlined with white circles.

method in this area [19] was tested by us for the case of multiple faces. Some of results of this procedure are shown in Fig. 4.

To recognize other objects like mountains, buildings, animals, ships, cars and so on, should be used different algorithms. It is important that these image classification procedures can be automated completely.

3.1.4 Generating Annotations

The last step is to assign individual characteristics to the objects. It is logical to use the tree type structure to store object parameters. Typically, very general objects descriptions are stored on the top level of this structure, and individual categorical and numerical values are stored in the leaves level. The set of individual characteristics and the total number will depend on the types of image. Very often the total number of the characteristics relate to geometries. For example, the dimensions and proportions of an object. There are many different algorithms to determine categorical and numerical values of parameters: neural networks [20], regression [21], logistical regression [22], conditional probabilities approach [23], Markov chains [24], and others. Our experience shows that following methods are most effective in this case: correlation analysis [20], Lucas-Kanade algorithm [25], and conditional probability method [23].

The image description, which is based on the annotation method, can be presented as a set of attributes, such as the types of image or landscape, type of crowd. It could be rural landscape features, urban landscape features, human face or figure features, or other unique characteristics of objects.

3.2 Retrieval Procedures

The procedures described above are done over the query image. Then, the search is performed using two stages: first, search by index, second, search by image itself. This approach allows to reduce processing time dramatically. The choice of index search algorithm depends on the type of images collection. As it was stated above, explicit hash functions can be used for indexing for most cases of structured image data. However, for initially unstructured data, like a random pictures collection, B+ tree approach [11, 12] is the preferred method.

4 Video Database and Super-Resolution Method

Even more complicated case of visual information database is organization and retrieval information from movie collections. Since every individual movie is a set of mages (frames), the size of each individual element is large, and the size of all collection is extremely large. This circumstance requires using special approaches.

There are different kind of queries can be applied in case of a movie database. Queries based on textual information should be processed a traditional way. This option is widely used in our days. Visual based queries should be processed using approaches which are described above. However, there are two important differences are noted comparably with image databases.

First, there is dynamic information available in video databases. That means that some objects are moving and not just object itself but its velocity (or acceleration) could be subject of indexing [26].

Second, the fact that video is a sequence of scene which are slow changed from frame to frame opens new opportunities which were not available in case of static images. The set of frames can be used as input data for super-resolution algorithm [27]. It gives possibility to make query with depicted small object, which barely resolved in each individual frame, but becomes sufficiently resolved due to super-resolution method.

Combination of these two features can make possible to resolve and recognize a small moving object. It could be very useful for analysis of the surveillance camera or to find the image of unknown object, such as UAV (Unmanned Aerial Vehicle), for example.

5 Conclusion and Future Research

In the current paper, we investigate new possibility to construct databases which contain visual information. The description of a new indexing approach which covers the most complex case of unstructured data collection is presented. It was shown, that the data structure for image store - retrieval procedures can be built in general case. It was also shown, that object-based store-retrieval procedures can use already developed image processing functions. We provided examples of these functions.

We also propose a paradigm of the database which contains dynamical visual information. The idea is to apply multi frame super-resolution method to development of store-retrieval procedures with a higher spatial resolution. These procedures are not available on individual frame level and might be used to identify new moving objects. Using Super-Resolution Technique allows to create a database of digital movies with following advanced features comparable with static images:

- make queries not just about objects, but about objects velocity and acceleration;
- make query about small objects which are not resolved and recognizable on stand-alone images.

The future research will be focused on building a prototype of visual database applying principles described in the current paper. We plan to demonstrate that using super-resolution method can provide high resolution store - retrieval procedures capable to work in real time on the modern mass-produced hardware.

Acknowledgments. The authors would like to thanks the IFIP Confenis 2017, General Chairs: Dr. Zhou Zou and Dr. Li Rong Zheng, IFIP WG 8.9 members for making this conference a great success.

References

1. Xu, L.: Enterprise systems: state-of-the-art and future trends. *IEEE Trans. Ind. Inform.* **7**(4), 630–640 (2011)
2. Xu, L.: Engineering informatics: state of the art and future trends. *Front. Eng. Manage.* **1**(3), 270–282 (2014)
3. Xu, L.: *Enterprise Integration and Information Architectures*. CRC Press (2015). ISBN: 978-1-4398-5024-4
4. Bulysheva, L., Jones, J.: A hybrid model for image databases. In: *Proceedings - 2nd International Conference on Enterprise Systems, ES 2014* (2014). <https://doi.org/10.1109/es.2014.48>
5. Bulysheva, L., Jones, J., Bi, Z.: A new approach for image databases design. *Inf. Technol. Manage.* **18**(2), 97–105 (2015). <https://doi.org/10.1007/s10799-015-0224-6>
6. Liu, Y., Zhang, D., Lu, G., Ma, W.-Y.: A survey of content-based image retrieval with high-level semantics. *J. Pattern Recognit.* **40**(1), 262–282 (2007). <https://doi.org/10.1016/j.patcog.2006.04.045>
7. Li, Y.: *Object and Content Recognition for content-based Image Retrieval*. Ph.D Thesis, Washington University (2005)
8. Li, Y., Shapiro, G.: *Object Recognition for content-based Image Retrieval*. Lecture Notes in Computer Science, Washington University (2004)
9. Oberoi, A., Singh, M.: Content-based image retrieval system for medical databases (CBIR-MD) -lucratively tested on endoscopy, dental and skull images. *IJCSI Int. J. Comput. Sci.* **9** (2012). ISSN (Online): 1694–0814
10. Cormen, T., Leiserson, C., Rivest, R., Stein, C.: *Introduction to Algorithm*. MIT Press, Cambridge (1990)
11. Zhang, D., Lin, X., Jia, Y.: The volume cutting of three-dimensional image based on B + tree. In: *Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference* (2010). doi: <https://doi.org/10.1109/ICBBE.2010.5515679>
12. Navathe, R.E., Shamkant, B.: *Fundamentals of Database Systems*, 6th edn, pp. 652–660. Pearson Education, Upper Saddle River (2010)
13. Bulyshev, A., Bulysheva, L.: Modeling segmentation algorithm. In: *Proceedings of the 3rd World Congress on Software Engineering, WCSE 2012, Wuhan, China, 6–8 November*, pp. 5–9 (2012)
14. Bulysheva, L., Bulyshev, A.: Segmentation modeling algorithm: a novel algorithm in data mining. *Inf. Technol. Manage.* **13**(4), 263–271 (2012). <https://doi.org/10.1007/s10799-012-0136-7>
15. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation. *Ann. Rev. Biomed. Eng.* **2**, 315–337 (2000)
16. Florack, L., Kuijper, A.: The topological structure of scale-space images. *J. Math. Imaging Vis.* **12**(1), 65–79 (2000)
17. Kumar, A., Kannathan, N.: A survey on data mining and pattern recognition techniques for soil data mining. *IJCSI Int. J. Comput. Sci. Issues* **8**(3) (2011). ISSN (Online): 1694-0814
18. Zare, M.R., Mueen, Z., Seng, W.C.: Automatic medical X-ray image classification using annotation. *J. Digit. Imaging* **27**, 77–89 (2014)
19. Viola, P., Jones, M.: Robust real-time face detection. *Int. J. Comput. Vis.* **57**(2), 137–154 (2004)
20. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(1) (1998). <https://doi.org/10.1109/34.655647>
21. Dowdy, S., Wearden, S.: *Statistics for Research*. Wiley, New York (1983)

22. Dreiseitl, S., Ortho-Mochado, L.: Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inf.* **35**(5–6), 352–359 (2002)
23. Bishop, C.: *Pattern Recognition and Machine Learning*. Springer, Boston (2006). <https://doi.org/10.1007/978-1-4615-7566-5>
24. Chen, X., Yuille, A., Zhu, S.U.: Image parsing: unifying segmentation, detection, and recognition. *Int. J. Comput. Vis.* **63**(2), 113–140 (2005)
25. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: a unifying framework. *Int. J. Comput. Vis.* **56**(3), 221–255 (2004)
26. Kotov, A.: Indexing of video flow based of face recognition. Master Thesis, Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia (2008). (in Russian)
27. Bulyshev, A., Amzajerdian, F., Roback, V., Hines, G., Pierrottet, D., Reisse, R.: Three-dimensional super-resolution: theory, modeling, and field test results. *Appl. Opt.* **53**(12), 2583–2594 (2014)

IoT and Emerging Paradigm



Internet of Things or Surveillance of Things?

Petr Doucek , Antonin Pavlicek , and Ladislav Luc

Faculty of Informatics and Statistics, University of Economics, W. Churchill Sq. 4,
Prague, Czech Republic
{doucek, antonin.pavlicek, ladislav.luc}@vse.cz

Abstract. The paper deals with digital surveillance in the postmodern world. We define a new term ‘Surveillance of Things’ in the context of the study of the surveillance, and try to determine, whether and how the surveillance of people is connected with surveillance of things. We pay particular attention to the Internet of things and analyze in detail the principles of Sigfox network.

We work on the presumption that information about people obtained through surveillance of things are interpreted incorrectly and can have a direct impact on groups of people and also individuals.

Keywords: Internet of Things · Surveillance of Things · Sigfox

1 Introduction

Surveillance studies are not a young scientific discipline, yet they have been enjoying an unprecedented development lately. With the advent of digitization, miniaturization, and the Internet, substantial cultural and social changes have taken place and the surveillance has been adapted, changed its structure and organization.

Surveillance studies in a traditional concept are primarily focused on human resources – including the individuals, the family, the whole social entity, the interest groups, the states, etc. They study causes and consequences of supervision, mutual interaction of individuals and groups, their opinions, feelings, social inclusion, culture and countless other aspects across many disciplines.

Nowadays people are used to the fact that they are being watched from time to time and that they leave their digital footprint on cell phones, computers and computer networks. People are even actively involved in their own monitoring: they voluntarily give up some private information, and surveillance systems count on their cooperation (Facebook, Google etc.). A DIY-style surveillance was born, transferring the burden of monitoring and the associated responsibility for supervising to individuals themselves [1].

Perhaps the biggest novelty is that monitoring of people begins to be achieved through digital monitoring of things.

P. Doucek—Scopus Author ID: 55916686400, ResearcherID: A-7703-2015.

1.1 Internet of Things (IoT)

If we are looking for a platform where we can monitor things en masse, the Internet seems to be a great choice. The Internet itself is a great tool of supervision and, in essence, has made it possible for surveillance to become conspicuous. But the current Internet as we know it is primarily an ‘Internet of Users’. For monitoring things, we need a platform that can handle many more orders of entities than the technological limits of the current Internet network allow. It is not practically conceivable for all postal packages, water or gas meters, light switches, or merchandise items in the shops to be Internet-connected. Internet protocols do not allow to connect directly such a great number of devices that can move freely for a long time without fixed wires. But new networks are emerging to bridge the divide between the Internet of Users and Internet of Things [2, 3].

The Internet of Things (IoT) can be defined as a dynamic global network based on standardized protocols where physical and virtual things have their identity, physical attributes, and virtual personalities. It is expected that things integrated into IoT become active participants in trade, information, and social processes, allowing them to communicate with each other and with their surroundings. These integrated things can independently respond to events and trigger processes even without direct human intervention [4].

1.2 Surveillance of Things (SoT)

As we mentioned at the beginning of the paper, the surveillance studies in the traditional concept focus mainly on people. For the purposes of this work, we define a new concept that will express the principles of monitoring (in the context of Surveillance Studies), in which surveilled or surveilling subject is a non-human thing. We also include combined cases where a supervisor is a person watching things and vice versa.

Surveillance of Things (SoT) means a focused, systematic and routine attention to data about things – collected and analyzed in order to influence, manage, protect and monitor them.

“Things” in the context of supervising things can be understood as physical and virtual entities that exist in space and time and can be identified.

1.3 Ownership of Things

Ownership of things has been one of the key aspects of the development of SoT. If things are to be possessed (and keep in mind, that possession is considered to be nine-tenths of the law), it means we have to ensure their inviolability. We must have some power over our things, have them under the sovereign control, we must have the ability to prove and defend our property when disputed or stolen. In order to achieve this, we need to be aware of their whereabouts, keep them under surveillance. Even though the right to own things is guaranteed by the modern state, in practice we still have to deal with security individually. One of the security features is monitoring or surveillance. We also began to use SoT techniques to track things that do not belong to us, but for some reason, we are interested in them.

There have recently been new reasons and needs for tracking things through new technical possibilities and social changes within society.

1.4 The Degree of Tracking Things

The degree of tracking of things could be seen on two levels. First, quantitatively, the number of things and the number parameters we can monitor. Secondly, in terms of quality, the depth and complexity of the tracking we perform (whether the tracking is done with all relevant aspects).

Since we are dealing with digital surveillance, we also need to take into account the aspects of digitization. Let us first mention the first Manovich's [5] principle of numerical representation, from which we can infer that the digital tracking is discrete. For example, digital video-recording may seem to be continuous, but it is not. In fact, it is just a stream of individual pictures taken at the rate of 25 shots per second. For normal human activities, it is unlikely that a person would do anything significant in a time span shorter than 33 ms.

However, the situation is different for monitoring things. Things can be very fast in both movement and change (for example a fired bullet). With virtual things, the situation is even more difficult. In the virtual environment, we can even reach the physical boundary of Planck's time. (What happens in a shorter time span than Planck's time (t_P) is not physically apparent).

$$t_p \equiv \sqrt{\frac{\hbar G}{c^5}} \approx 5.39124(27) \times 10^{-44} \text{s} \quad (1)$$

For our needs, however, it is sufficient to conclude that digital surveillance is discrete and, for SoT, the sampling frequency becomes an essential variable.

2 Technical Aspects of SoT

When compared to humans, things behave highly predictable, since they lack their own intelligent thinking. Things can be easily parameterized, described and identified. A special category is the existence of virtual things; items without their physical representation in the real world. Things are similar to people in having a life cycle. Two technical aspects (problems) are essential for the purposes of this work.

The first technical problem is the fact that there are many times more things than humans. There would be nothing special about that, but the amount of the things being watched is exponentially growing. Buyya and Dastjerdi [6], in their book *Internet of Things*, in accordance with McKinsey's estimate that around one trillion devices will be connected to the network by 2025. And here comes the trouble - traditional mobile networks, such as 2G, 3G, LTE, are not capable of handling such huge number of connected things.

The second technical problem is that things can move freely. It is quite easy to monitor things in a fixed perimeter, but the problem occurs when they are geographically

unstable or they are moving all the time. Mobile networks cover a large part of the mainland, but the terminal devices require a relatively large amount of electricity to operate.

The answer to these technical problems is building communication platforms dedicated to IoT.

2.1 IoT Trends

Let's see how IoT is perceived by leading technological companies:

- Samsung predicts further growth in IoT technologies, with the IoT market volume reaching USD 1.7 trillion by 2020. Samsung provides a solution in IoT and supports NB LTE and LoRa technologies [7].
- Computer giant HPE (Hewlett-Packard Enterprise, formerly HP) is developing a product called the “HPE Universal IoT Platform” to connect a variety of different IoT technologies and manage their lifecycle. HPE forecasts that in the year 2020, the amount spent on IoT hardware alone will reach USD 3 trillion [8].
- Another world technological leader, IBM predicts the value of the economic potential of IoT at USD 11 trillion in 2025 [9].
- Microsoft has been working in the IoT area for a long time and provides a full range of IoT products. From software and solutions for end-to-end devices to cloud-based platforms that provide comprehensive IoT operation [10].
- From Czech companies, we can name ČD - Telematika, which offers its own concrete solutions in the field of IoT [11].
- Similarly, other companies like Google, Amazon, Dell, or Cisco are actively engaged in the world of IoT.

2.2 IoT for SoT

The cornerstone of the Internet of Things is secure two-way communication of a large number of things anywhere, anytime. Contrary to that, one-way communication thing → supervisor is sufficient for SoT, provided that the subsequent stimulus from supervisor will be communicated through another channel.

Practical implementation of IoT must, therefore, ensure ubiquitous connectivity. The supervisor's goal is to be able to keep track of the things as continuously as possible – by systematic and continuous monitoring without failures (with respect to the discrete nature of digital surveillance). Any gap in tracking leaves room for speculation and the possibility of questioning the “pedigree” of monitored things. And let us remind, that for complete global coverage, there is a quite substantial space (the world as we know it and commonly use - the land, underground, water areas, and airspace within the reach of conventional commercial aircraft) to be covered.

In order to monitor things, a considerable degree of autonomy of the monitored things must be ensured. Having things constantly connected to the network (both electricity and data) is impractical and expensive, making it impossible for them to move easily. A properly dimensioned wireless power supply and data connectivity need to be solved.

However, in addition to the above, the viability of SoT depends also on the following criteria: economic aspects, miniaturization, restrictive legislation, increasing complexity - increasing number of connected elements in the network.

3 Sigfox – IoT Solution for SoT

One possible solution for IoT is the French project Sigfox. Sigfox builds a brand new network dedicated to the Internet of things. As of May 30, 2016, the Sigfox network was available in 18 European countries and covered an area of 1.2 million km² [12]. In the Czech Republic, Sigfox has established a partnership with SimpleCell Networks, which has contracted T-Mobile Czech Republic (TMCZ). TMCZ is actually building the Sigfox Network for the Internet of Things [13]. By the end of 2016, more than 350 base stations were connected to the network, covering roughly 3 × 95% of the Czech Republic's territory, and further expansion is planned to increase reliability. The situation illustrate the coverage maps of the Czech Republic and Europe of May 2017, see Fig. 1.

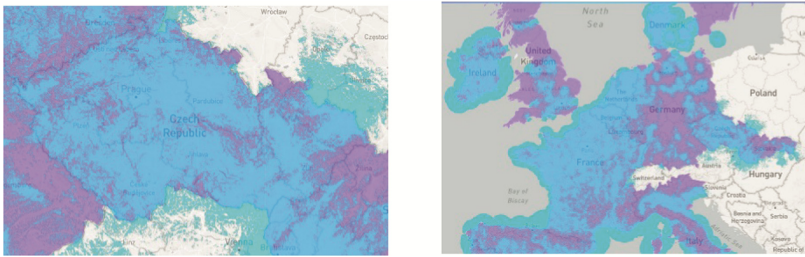


Fig. 1. A coverage of Sigfox network in the Czech Republic and Europe – as of May 2017. (<http://www.sigfox.com/en/coverage>)

3.1 Sigfox Connectivity

In this chapter, we will get closer look how Sigfox works to meet the requirements for SoT. The aim is not to describe the technical details, but to introduce this specific IoT solution.

The Sigfox wireless network operates at 868 MHz, an unlicensed but regulated frequency. Thanks to that no licensing neither authorization is needed, but the country-specific arrangements must be respected. The Decree R/10/05.2014 enforces the limits within the Czech Republic. The Decree states inter alia that the maximum transmitting power should be only 25 mW [14]. This is the same bandwidth and power we commonly use in remote controllers to open garage doors. However, the Sigfox transmission signal is directed to a very narrow beam, providing connection within a direct visibility of up to 200 km (a practical test performed by TMCZ repeatedly succeeded in connection between the highest mountain of the Czech Republic, Snezka and Prague – which are 120 km apart), declared reach in the populated area is 5–30 km.

A maximum of 140 messages per day can be sent from the Sigfox device (in other words - the device can send a message every ten minutes). The device can receive up to 4 messages per day. These parameters are strictly legislatively enforced, see the CTU Decree, technology itself allows to broadcast almost continuously.

However, the size of one message is quite tiny - only 12 bytes of user data plus some technological information (for example device battery status, etc.) as well as headers with message identifiers. The entire message is encrypted with an AES-128 certificate before sending.

The device sends the message repeatedly three times in a row and does not monitor, whether a base station received it. After a broadcast, the device powers off to wake up just before the next message is ready to be broadcasted. When the device is in the sleep mode, it is not possible to contact it remotely.

The sent message is received by the base stations. Sigfox assumes that each geographic location should be covered at least three times in order to increase the likelihood of receiving the message.

The mobile operator places Sigfox base units at standard locations (masts) next to the traditional mobile phone technology. From mobile operator’s point of view, the system is complete “black box” - the device works completely autonomously, sending data through encrypted VPN tunnels to the Sigfox headquarters in France. The Sigfox network headquarters is a cloud platform. Data is preprocessed and ready to be sent to users (from the IT architecture point of view it is a back end). Users can connect to the cloud platform via either a web interface or they can use third party applications that connect to the Cloud Sigfox through the API (Fig. 2).

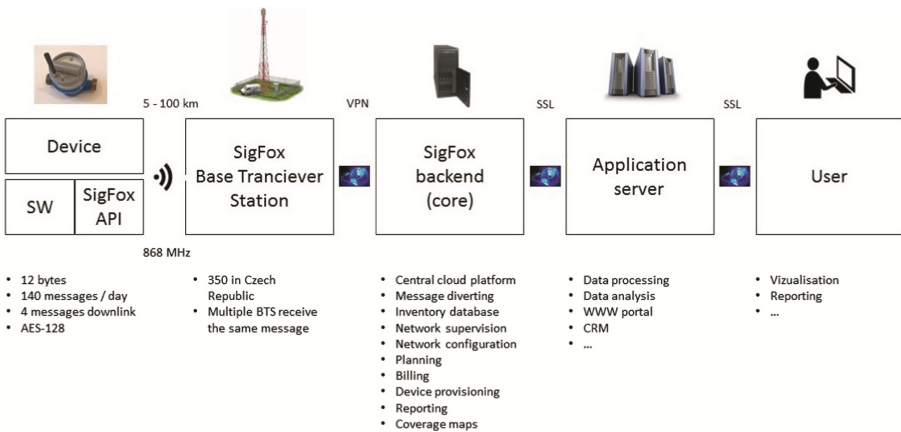


Fig. 2. Message transmission over Sigfox network (authors)

3.2 Sigfox Devices

Sigfox terminal devices consist of three main parts. The first part is Sigfox modem with the antenna. This module (usually a printed circuit or a chip of several square centimeters in size, with a unique ID - similar to a SIM card) is responsible for sending and receiving

messages. In the second part are the sensors – can be practically any digital detector – e.g. sensors of temperature, humidity, air pressure, sound and vibration, flowmeter, sensor of chemical states and gas states, pressure gauge and weight sensor, fluid detection and fluid level measurement, magnetic sensor, acceleration sensor, light and optical sensor, GPS, motion sensors, position change sensor, ... The third major part is the battery.

The Sigfox modem is designed to have extremely low power consumption. The whole device is designed to last for a long time without a battery replacement. For example, to measure daily the temperature around the device, i.e. send one message a day with the user's temperature information, the device can last for more than fifteen years. But energy-intensive user device (such as a GPS module) or high message frequency decrease the battery life.

The Sigfox is a half-open concept. The chip is available to any regular user to make amateur devices with just the right equipment. Professional modules can be made by any company without need of Sigfox license. Examples of these devices are shown in Fig. 3.

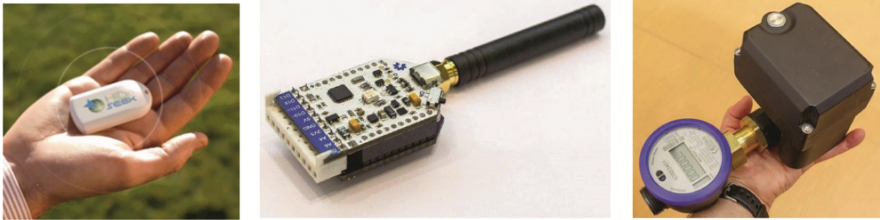


Fig. 3. Sigfox devices (<https://www.cnet.com/devices-for-sigfoxs-super-cheap-network>)

3.3 Examples of Using Sigfox Technology

Sigfox is best suited for use in situations where we need a long-running, autonomous system that provides regular or event generated information of low data volume. The technology has an excellent reach and can be deployed both in natural and urban agglomerations. Moreover, Sigfox's cloud concept is not limited by state borders. Devices purchased and paid in the Czech Republic will work equally well in France or the UK. There are no extra roaming fees. The system is quite robust, it has been designed to handle up to hundreds of millions of devices.

Sigfox solution does not provide information in real time – there is slight delay within one minute, so the system is “near online”. The system is not designed as highly reliable. In practice, some messages are not received and there is no feedback (with the device). Sigfox does not suit even fast movement due to Doppler effect. However, the practice has shown that movement in motorway speeds will not affect functionality, a flying plane is problematic.

Specific use cases are still being devised. For some solutions, there is already a pilot study in operation. A feasibility study is being developed. Of the existing ones, let us mention at least some of them:

- Remote readings of water meters, gas meters, heat sensors on radiators and the like (smart metering);
- Electronic seals;
- Flood sensors;
- Environmental measurements (temperature, pressure, humidity, air quality, river level);
- Waste management (dustbin alerts the collection agency being full);
- Sensors of parked cars (placed directly under the asphalt);
- Shock sensors on bridge structures;
- Availability of meeting rooms (“digital company” concept);
- Soil quality (can be buried in soil and monitor soil moisture, composition, pH, etc.);
- Tracking of consignments, containers or livestock and the like (GPS tracking technology);
- Remote health monitoring.

4 Case Studies: IoT in the Czech Republic

IoT is not a newcomer with the arrival of Sigfox in TMCZ. Previously, there were also systems called M2M (machine to machine) built on 2G, 3G, LTE, and the like. They serve primarily for corporate clients and were usually built as tailor-made customer solutions. We can include, for example, connection of ATMs and payment terminals, MeR (measurement and control) systems, or corporate fleet management systems. With Sigfox, however, it is a dedicated network for IoT, and TMCZ will have to form another strategy, invent new products and run traffic.

IoT still does not have a defined world standard. There are other technologies (i.e. LoRa) on which IoT can be built. From this perspective, Sigfox can be considered as an uncertain project.

4.1 Alternatives to the Sigfox Network

As mentioned, the global IoT standard does not yet exist and Sigfox is not the only platform for the Internet of Things. In the Czech Republic, CRA (České Radiokomunikace a.s.) is also trying to build a network for the Internet of Things.

LoRa technology is technically and strategically different from Sig-Fox technology. The LoRa operator can be anyone. The city may decide to build its IoT on its territory fully under its management. With LoRa it is possible. But there is only one specific company, which can produce a proprietary chip. LoRa has its pros and cons, but the intention and purpose are the same. The ability to connect millions of devices to a network of things. Identical are therefore cases of use.

Another system, which can be used for IoT, is a system based on LTE technology - LTE M.

4.2 Case Studies - Sigfox Technology from the SoT Perspective

Sigfox meets the basic definition of surveillance because many uses (such as Sigfox in combination with GPS) are concentrated, systematic and routine attention to data about things to influence, manage, protect and route them. By providing supervision, we also provide factual power over things and consolidate the principle of ownership.

Sigfox is a highly centralized database on a pan-European scale that associates information from all connected devices in one place under the supervision of a single organization. Sigfox allows you to convey information about a particular thing and to make whole pedigree of things, for example in agriculture [15] or energy industry [16].

Imagine an apple orchard. The farmer has installed sensors of humidity, Ph level, sunshine, and the like. The nearby apiculture company has honeybee sensors installed in its hives. Growing fruits have been monitored since birth, and even though the farmer cannot see the beekeeping company's data (and vice versa); this information may come together at the Sigfox headquarters. After harvest, a logistics company is using the GPS tracking device, the apple is once again monitored – as it goes to the vendor's warehouse – where its environment is again monitored. When someone orders the delivery of apples on the internet, Sigfox logistics company can transport them again. None of these companies see each other in the data but in Sigfox they know that the apple has grown under the specific conditions when it was harvested, knowing how much it weighted, where it went, how long and under what conditions it was stored and eventually where it was taken by the consumer - including address). Such pedigree of things of unprecedented proportions is feasible today.

The practical applicability of Sigfox tracking technology is evidenced, for example, by the pilot operation in the TMCZ implementation demanded by Václav Havel Airports in Prague. The airport considered Sigfox tracking of luggage trolleys – passengers leave carriages in car parks and do not return them to the marked places. Because of the vastness of the airport, carriages pose even a security threat. Sigfox can provide both indoor and outdoor protection and solve the problem.

5 Conclusion

The Sigfox network bypasses the need for a large number of IP addresses, thus avoiding IPv6 issues. Terminal devices are not connected directly to the Internet and do not have an IP address. Sigfox has its “last mile”, or the connection to the last section between the network and the end device by a separate network. It is only from the base station that the data is moving over the IP network. Terminal devices are not available for IPv4 or IPv6.

Sigfox's top authority is Sigfox, which has access to all transferred data. Due to the fact that only 12 bytes are sent from a single device, user encryption is impracticable; the data are open and readable for Sigfox. The message is signed with a digital certificate, to forge a system of fake law is difficult. Sigfox can thus act as a guardian over the entire global network. In theory, however, it is possible that regional Internet sovereignty (for example, the state) could order its internet service providers to disallow data streams to Sigfox servers. In that case, the Sigfox would cease to function in that country.

Geographical boundaries are significant in the Sigfox network so far, but in general, it should not be a problem. Although the network is still covered only by part of Europe, the coverage is planned to be complete. Additionally, in the Sigfox network, automatic roaming works across borders. The communication protocol is the same everywhere. Perhaps the Internet of Things is going to be the Internet without borders, as Internet creators introduced it because things do not have their own language, culture or thinking. For the time being, however, it would seem that if Sigfox copied the geographic Internet scenario, it would be a restriction from the top (meaning state regulations) rather than from users.

In conclusion, the Sigfox standard meets the surveillance studies theory and, at least in Europe, has the best prospects of becoming state-of-the-art Surveillance of Things tool.

References

1. Li, B., Yu, J.: Research and application on the smart home based on component technologies and Internet of Things. In: Ran, C., Yang, G. (eds.) *Ceis 2011*. Elsevier Science Bv, Amsterdam (2011)
2. Haiyan, S., Xiaobin, L.: Research on practical teaching system of the Internet of Things technologies and application. In: Zhang, H.M. (ed.) *Proceedings of the 2014 International Conference on Education, Management and Computing Technology*, Atlantis Press, Paris, pp. 536–538 (2014)
3. Sun, E., et al.: The Internet of Things (IOT) and Cloud Computing (CC) based tailings dam monitoring and pre-alarm system in mines. *Saf. Sci* **50**(4), 811–815 (2012)
4. Sundmaeker, H., et al.: *Vision and Challenges for Realising the Internet of Things*. Publications Office of the European Union, Luxembourg (2010)
5. Manovich, L.: *The Language of New Media*. The MIT Press, Cambridge, Mass (2002)
6. Buyya, R., Dastjerdi, A.V. (eds.): *Internet of Things: Principles and Paradigms*. Morgan Kaufmann, Amsterdam Boston Heidelberg (2016)
7. SAMSUNG: IoT Solution Samsung Business. <http://www.samsung.com/global/business/networks/solutions/solutions/iot-solution>. Accessed 21 Dec 2017
8. Hewlett Packard Enterprise: Delivering on the IoT customer experience The HPE Universal IoT Platform 1.4. <http://h20195.www2.hp.com/V2/getpdf.aspx/4AA6-5353ENW.pdf?ver=3.0>. Accessed 21 Dec 2017
9. IBM: IBM Watson Internet of Things (IoT). <https://www.ibm.com/internet-of-things/>. Accessed 21 Dec 2017
10. Microsoft: Internet of Things (IoT) Microsoft, <https://www.microsoft.com/en-us/internet-of-things/>, last accessed 2017/15/21
11. ČDT: Internet věcí ČD-Telematika a.s. <http://www.cdt.cz/cz/internet-veci-1272/>. Accessed 21 Dec 2017
12. Sigfox: Sigfox - The Global Communications Service Provider for the Internet of Things (IoT). <https://www.sigfox.com/en>. Accessed 21 Dec 2017
13. SimpleCell: simplecell.eu – Connecting Things (2017). <https://simplecell.eu/>
14. CTU: Všeobecné oprávnění č. VO-R/10/05.2014-3 k využívání rádiových kmitočtů a k provozování zařízení krátkého dosahu (2014)

15. Ma, J., et al.: Connecting agriculture to the Internet of Things through sensor networks. In: Presented at the Proceedings - 2011 IEEE International Conferences on Internet of Things and Cyber, Physical and Social Computing, iThings/CPSCoM (2011)
16. Sladek, P., Maryska, M.: Internet of Things in energy industry. In: IDIMT 2017 - Digitalization in Management, Society and Economy, pp. 411–418. Trauner Verlag Universitet, Linz (2017)



The Economic Value of an Emergency Call System

Tomas Lego¹, Andreas Mladenow¹ (✉), Niina Maarit Novak², and Christine Strauss¹

¹ Department of e-Business, Faculty of Business, Economics and Statistics, University of Vienna,
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

{tomas.lego, andreas.mladenow, christine.strauss}@univie.ac.at

² Institute of Information Systems Engineering, Vienna University of Technology,
Favoritenstr. 9-11, 1040 Vienna, Austria
niina.novak@tuwien.ac.at

Abstract. eCall is a complex solution, aimed at supporting drivers and car passengers in the event of an accident in Europe. This automatic emergency call system for motor vehicles, planned by the European Union, is installed in all new models of passenger cars and light commercial vehicles. In this context, the contribution analyses the monetary value of the eCall system implementation.

Keywords: eCall · Emergency call · eCall implementation · Monetary value
Business value · Value creation · Economic value

1 Introduction

Information technologies (IT) and innovative solutions influence humanity both positively and negatively [1–6]. In this context, modern technologies offer new solutions to well-established problems, e.g. automobile accidents. A good example of such an innovative solution is the automatic emergency call system (eCall) which will be implemented in the European Union beginning in March 2018 [7].

This IT solution is also important from the psychological point of view, since the need for safety is one of the basic needs of Maslow's well-known hierarchy of needs [8]. The development and search for an EU-wide implementation of the eCall system is only one evidence that this statement is still valid.

Still, such a system brings more than one solution to a specific problem and encompasses numerous aspects when considered thoroughly. Some of the effects of this safety system can be expressed in monetary values that will be analysed in this paper, which is structured as follows. Section 2 provides a brief overview of the development, the functions, and the necessary infrastructure of the technology. Section 3 focuses on positive and negative monetary and non-monetary aspects of the economic impacts of its implementation, which are further analysed from two different perspectives. Finally, Sect. 4 of the paper summarizes the initial findings and presents possible future developments for eCall.

2 Emergency Call System – Functions and Infrastructure

The eCall system is part of the “Road Safety Strategy” which was launched by the European Union in 2010. The system is meant to work throughout the entire European Union based on the existing platform of a uniform emergency call number 112 [9], with the main goal to manage emergency calls. The general aim of the system’s implementation can be summarized as “automating the notifications of traffic accidents from anywhere in the European Union and associated countries” [9].

In the event of an emergency, most likely a car accident, the system is either automatically or manually triggered and an emergency call is established [10]. The purpose of such a system is to facilitate the making of the call. It is especially helpful in situations when the passengers of the vehicle are not able to make the call themselves [11], e.g. in situations when an individual does not have a mobile phone to use for an emergency call. Furthermore, it provides an additional advantage as the system is not only meant to establish a voice connection, but also establishes data transfer. To ensure the operability of the system, the Public Safety Answering Points (PSAP), serving as call-centres receiving the emergency calls, must be upgraded and vehicles have to be equipped with in-vehicle devices, which will act as triggering mechanisms [10].

The implementation of an automatic notification system for emergency situations on roads has several goals. One of the main goals among these is to cut the number of road casualties in half by the year 2020. This is a significant challenge given that alone in the year 2014, there have been approximately 26,000 people who died as a consequence of a car accident in the European Union. Yet, the system does not only aim at lowering the number of human casualties on roads; but it aims as well at reducing the number and the severity of injuries in general [9]. Moreover, the introduction of the eCall system is also an attempt to unify the in-vehicle emergency systems within the European Union. As a matter of fact, many similar systems such as the “Volvo On Call” are operational throughout Europe [9]. However, these systems are voluntarily purchased by car-owners as optional add-on service and are not standard. Considering their benefits, this should change with the implementation of the eCall system.

eCall was introduced by the European Union in 2001 and was originally meant to be fully operational by 2009 [12]. However, in 2006 the European Commission postponed the desired year of implementation to 2010 [13], and again to October 2015 [14]. As of today, the system is to be launched in all EU-member states by 31 March 2018, making it mandatory that all new types of vehicles introduced after this date will have to be equipped with eCall technologies [15].

To ensure operability of the system, each EU-member country must equip and ensure that the infrastructure of all national Public Safety Answering Points (PSAP) is capable of receiving eCalls by 1 October 2017 at the latest. It must be guaranteed that all PSAPs work on the basis of the 112-emergency number [7], which was launched in 1991 [10] and is still active as a pan-European emergency call number. It should be highlighted that the eCall system is not meant to substitute this number, but is rather based on this well-established emergency call platform and should enable the triggering of an automatic emergency call from the inside of a vehicle.

As indicated, the road towards an EU-wide implementation of the system was interrupted or postponed several times. However, the introduction of the EU “Road Safety Strategy” in 2010 marked the beginning of a more dynamic development [9]. One of the largest steps in ensuring a required level of preparedness of the European market was taken in 2011 when a pilot project co-founded by the European Union, referred to as the Harmonised eCall European Pilot (HeERO), was launched. For three years, the eCall system was tested in nine European countries. The number of these testing countries was increased with an additional six countries in January 2013, and the second stage of testing began. Furthermore, several other countries desired to join this pilot project, but their request was not successful. The goals which were followed by HeERO included: expressing a necessity of upgrading infrastructures to support the uniform eCall system, boosting investments into these, and generally preparing the entire area for the compulsory introduction of the system [9] by October 2017 [7].

To ensure full functional capacity of the eCall system, both vehicles and the infrastructure must be prepared accordingly. Vehicles have to be equipped with eCall technology consisting of an In-Vehicle System (IVS) that is capable of establishing an emergency call using a mobile network [16]. Further, this IVS is attached to collision detection sensors such as airbags or any other triggering sensors [17] that are integrated in the vehicle and is also connected to the audio-communication mechanisms of the car [9].

Once these devices are installed and operational within the vehicle, the system is able to work and can be activated in two ways. It can be triggered either automatically or manually. The automatic triggering is meant for serious accidents when the in-built sensors mentioned above detect an impact and independently initiate an emergency call [9]. The second method for activating the system is meant for cases when an individual witnesses another accident, when the car crash has not been so impactful that the system would have been triggered automatically [18], or for any other emergency case which has not evoked an automatic response by the vehicle.

The PSAP have to be prepared for handling eCalls to ensure a state of readiness within the infrastructure. This is a crucial premise since the PSAP do not only have to be able to establish a voice connection with the vehicle but are also required to be capable of decoding a bundle of digital information that is sent to them by the IVS. This information is often referred to as the minimum set of incident data (MSD) [10]. One of the main pieces of data which is transferred from the crashed car to the PSAP is a so called eCall Flag - a piece of information that is decodable by the answering points which clearly states that the incoming emergency call is an eCall [9]. The eCall Flag is crucial and especially valuable in situations when the passengers of the vehicle are not able to speak. If such a call would be not routed to an eCall-supported PSAP, the emergency call [19] would be defined as a silent call and would thus most probably be terminated by the operator.

However, knowing that it is not a silent call and given the MSD, the emergency call operator still has information to work with. Based on this, the operator is facilitated in making better, well-informed decisions and can arrange for the needed rescue services. In this regard, the MSD contains information such as the car’s location, the direction in which it had been moving, the vehicle’s description, the time [9], and whether the system

was triggered automatically or manually [17]. Depending on whether the used PSAP is privately or publicly owned, additional information can be shared useful for road assistance or the Global Positioning System (GPS navigation) [9].

Once a call is triggered and the MSD is transferred, the system works in such a way that all installed microphones and loudspeakers in the vehicle are solely dedicated to the system and used as means of communication with the mobile operator. Interestingly enough, this connection can only be terminated by the PSAP and cannot be discontinued from inside the vehicle. Still, the audio systems within the car remain designated to the system in the event of a call back [9]. Generally, the establishment of a voice connection is primarily meant to be a source of additional information for the operator of the emergency call line [17]. The determination of the exact location of the car is complemented with the international GPS location [20]. Notably, these positioning services can also be provided by the European Galileo system [16].

3 Analysis of Non-monetary and Monetary Value

3.1 Non-monetary Value

More than two thirds of car accident-related emergency calls are not made by the people directly involved in the accidents [21] but are made by witnesses. This could be due to the fact that the victims of an accident are incapable of making these calls because of their injuries, but it can also be that those involved do not even realize the severity of the situation. This is often the case when occupants of a vehicle are intoxicated, under the influence of other addictive substances, or have suffered a severe shock or trauma during the crash itself. Thus, a system, which is able to make such a call automatically, would be of major importance. It is expected to reduce the notification delay in approximately one third of the cases. This equals an estimated net gain of ten minutes. Half of the fatalities occur exactly within these first minutes after an accident and only about a third of them occur in the hours following the crash [21]. Transforming this value into the number of saved lives, it may be expected, that up to 2,500 lives could be saved every year with the eCall technology [20]. Knowing that there have been 26,100 road fatalities solely in the year 2015 in the entire European Union [22], eCall technology could lead to a significant reduction in this number. If focusing only on Austria, the number of road casualties in 2015 was 479. Considering it is 1.84% share of the number of fatalities in the entire EU, depending on the actual number approximately 46 people could be saved each year just in Austria [22].

The eCall system positively contributes in terms of saving lives by sharing the vehicle's location with the PSAP. However, even without the use of eCall, the legislation of the European Union anticipates such situations and obligates mobile network operators to provide the location of all 112-callers if asked by the PSAP [16]. Thus one could claim that the eCall system is redundant as this piece of information could be retrieved as well differently. When considering the system more thoroughly it becomes obvious that the IVS supply not only location data but a whole bundle of highly important information (see Sect. 2).

One of the biggest reservations in connection to the implementation of a technology such as eCall is the concern for an individual's privacy. Thus it must be ensured that personal data cannot leak and that the system only communicates with a PSAP when triggered [9]. In addition, any continuous tracking achieved through the IVS should be impossible [11] and, even if it is transmitted to the responsible PSAP, the MSD should be handled according to the data protection rules that are valid within the European Union. However, despite these reassurances, official EU documents indicate that the information gained from IVS could be used for further developments in numerous areas. Besides other things, it could be of use to telecommunication-system providers and the car-industry to develop and propose new services and products [23].

Still, the extent of both positive and negative effects will greatly depend on the system's penetration rate [21]. In the year 2013, the European Union expected the penetration rate to reach 100% by the year 2033 [14]. Given the delayed starting date of implementation however, the maximum penetration rate will be reached only by 2035, at the earliest [15].

3.2 Monetary Value

Apart from the non-monetary value of eCall, the system also brings a lot of aspects that can be expressed in monetary values. These are factors which might not be obvious at first sight but are of great importance. Talking about monetary consequences associated with eCall not only involves the costs of installing eCall devices into vehicles but also the effects of reducing congestion, saving people's lives, and even using the data the system provides. In the following, this subsection will attempt to shed light on the less defined economic consequences of eCall. It will do so from two perspectives. First, there are areas within the private sector which are likely to be influenced through the introduction of the eCall system. Secondly, there is an even greater sphere of the economic background of automated emergency call systems in vehicles within the public sector. Thus, eCall does not only affect individuals but is important for society as a whole. Moreover, these monetary values are foreseen by the European Union itself [7] and are included in the "decision [...] of the European Parliament and of the council [...] on the deployment of the [...] EU-wide eCall service" [7] from the year 2014, stating that the service is also expected to provide monetary benefits to society by reducing congestion and secondary accidents as well as improving incident management [7].

The extent of the monetary values also greatly depend on the penetration rate of the system [21]. Once the desired 100% penetration rate is achieved and given the mandatory regulatory measures of eCall, the benefit-cost ratio is expected to reach 1.74 [14].

3.2.1 Public Sector

It is estimated that the public incurs costs of more than €160 billion per year due to vehicle accidents [16]. Through the introduction of eCall, the EU expects not only a reduction in the number of casualties, as outlined in the previous section, but also a significant reduction of these costs [20]. The cost reduction is estimated to reach up to €20 billion per year [10] and can be explained by a series of individual factors. The most

prominent factors, frequently discussed in related literature include: the reduction of congestion, enhancing traffic services, improving accident management and saving human lives. Each of these factors can be measured in monetary values and will be further discussed in the following.

Traffic congestion is a very cost-intensive factor with regards to this evaluation. Disregarding the European Union, yearly costs of traffic congestion can amount to up to US\$23 billion for a single city. The city of Los Angeles for instance has to face this amount of costs on a yearly basis [24]. With regards to the EU, the costs of traffic congestion per city are not as staggering, however when investigating traffic congestion costs on a national level they can even surpass the frontier set by Los Angeles e.g. the United Kingdom estimated its costs caused by traffic congestion in the year 2003 to have reached up to US\$23.7 billion [25]. As it is hoped that eCall is to be used as a source of information for traffic management hubs, the number of secondary accidents could be lowered, and other drivers could be informed well in advance. Thus, eCall is also expected to help to reduce traffic jams by redirecting traffic to roads that are less congested [9]. Given these outlined scenarios, the portion of public costs due to traffic problems caused by congestion is expected to decrease.

Another aspect that can be expressed as monetary value is the price of a human life itself. Through the goal of reducing the number of fatalities and the severity of injuries [9], further savings could be achieved. Calculating the value of a human life however is a rather complex task. Even though it is a topic that has been targeted by scientists for more than several decades, a generally accepted equation for expressing this value does not exist. However, existing studies agree that the value of a human life depends on the contribution to society that would have been achieved by that individual [26]. Some of the determining values are age, gender, and the length and quality of education obtained by the potential victim [27]. Based on these values, it is possible to ascertain a monetary value expressing the loss to society caused by losing a human life, measured over the achievable lifetime earnings of the victim. Exact numbers will not be calculated at this point and the claim rather serves the goal of expressing the fact, that even the loss of a human life can be expressed as monetary value.

The total amount of costs to society caused by traffic accidents can be expressed as GDP-shares and amounted to 3.3% of the Austrian GDP in the year 2012 [28]. This implies that without any costs related to traffic issues, the GDP of Austria would have not amounted to US\$407.45 billion [29] in 2012 but would have amounted to US \$421.3547 billion. Without doubt, a comparable improvement of Austria's GDP through the introduction of the eCall system, is not realistic. However, considering the forecasted savings' cap of €20 billion on the total costs of €160 billion per year, it could be expected that a similar positive change could have been witnessed in Austria as well. The positive change in the Austrian GDP would have accounted for up to 12.5% and could have led to a net gain of US\$1.74 billion on the GDP back in 2012. Since these amounts depend on the GDP value and the assumption that the 12.5% savings target is equal throughout all EU countries, the sums presented in this paragraph are only an approximation of the possible net-gains achievable through eCall. The exact equations are presented in Appendix.

Still, an introduction of eCall would not only be limited to savings, but would also offer countries new opportunities in terms of collecting revenues, e.g. collecting road charges, which is potentially accomplishable through the IVS [23].

3.2.2 Private Sector

The introduction of eCall presents as well opportunities for changes within the private sector. Some of these are also foreseen by the European Union itself [11].

PSAP can be either publicly or privately owned [21]. With private investors, the costs of updating these answering points would be relocated away from the state budget. Even though the costs of establishing and maintaining such a PSAP are not exactly defined, they are assumed to be substantial, given the fact that alone the needed servers are priced up to €20,000. Further costs include, among others, the expenses for the MSD-decoding software and training costs [14]. Regardless, the decision of letting private organizations take over the role of otherwise public sector-provided services must be taken well in advance and is both a money and time consuming matter, keeping in mind that these eCall-ready PSAP had to be fully operational by 1 October 2017 [7].

Despite the fact that emergency calls, even if made by eCall systems, are free of charge [7], the technology is of interest for various market players. Not only the automotive industry itself, but also telecommunication providers and many more could enhance their business models based on the mandatory introduction of this system [16]. Furthermore, eCall offers an opportunity for Third Party Services (TPS) which are also foreseen by the EU and are expected to work side-by-side with the PSAP [11].

Considering automated manufacturing processes and the desired degree of implementation, the costs for installing one fully operational eCall system into a vehicle will not exceed €100 [23]. Still, manufacturing the software and the hardware could be an interesting business opportunity and is already followed by some well-known companies, e.g. Bosch [30].

Yet, it is an interesting fact that there are no exact conditions under which the system has to be triggered automatically. The sensors have specified limits between which the system surely needs to be activated and when it cannot be activated. These frontiers are defined through UN regulations, but the exact amount of triggering-force expressed in any quantity is not exactly defined. Thus, there remains a small area between the two extreme points and it is the role of the producers of these systems to fine-tune the systems [17, 31]. Based on this, a wide range of differently sensitive products can be expected whereby some might be automatically triggered after a small bump and other models which will not act in the same way. Therefore, it is likely that different segments within the market for eCall IVS will come into existence, influencing the revenues of automobile manufacturers.

An example of this influence could occur if it would become publicly known that the eCall systems of a certain car manufacturer are more likely to be triggered whereas those of another manufacturer might require greater forces to be activated. Considering two cars, each with a different eCall system, it is to be assumed that the mark and thus the type of the IVS installed might be of importance for people when deciding between the two vehicles. These can be substituted by the IVS and thus a decision between these two cars might be based on the preferences of a driver to own a car which either can be

seen as safer since the triggering force set in the IVS is lower or one which has a less sensitive system and thus does not report a little bump when parking the car. Even though this example might be exaggerated and these diverse preferences surely go hand in hand with the different types of drivers, even this could influence customers' choices when purchasing a vehicle. Consequently, car manufacturers could be affected as well through the choice of which IVS they use. Such a connection seems more than plausible. However, the relevance of it would have to be further examined.

Another sector, which could be heavily influenced through the introduction of eCall is the insurance market. Retrieved data from IVS could be used for determining accident causes, it could speed up the process of handling individual cases and it could even be used for revealing insurance scams. Further, it could be used for the automatic processing of administrative procedures and could even help to fight car-theft related crimes [32]. It is a fact that such improvements would influence thousands of people but, since insurance companies are profit-oriented and, not state owned organizations, even this effect is to be classified as influencing the private sector above all.

4 Conclusion and Outlook

The eCall system is a complex solution that provides matters of assisting to make an emergency call in the case of an accident in Europe. Not only does it show its potential in terms of saving lives and reducing the consequences of car accidents in general, it can be further seen in a different light if looked at by an economist. There are many areas of the private sector influenced by this system, and there are also numerous areas within the public sector that are not independent of the eCall system. Furthermore, these influences can be categorized, subject to the sector they affect. Thus, we can diversify between private and public sector related consequences of the mandatory introduction of eCall.

Future developments promise further improvements and enhanced features. This very fact also applies to eCall, which is expected to provide even a two-way data transfer in the future. Not only will the PSAP be able to receive data from vehicles, it will also have the option of sending instructions to the car itself. This means that the PSAP will be given the option of remotely sounding the horn of the crashed vehicle, unlocking its doors, switching its lights on and off, or letting them flash [9]. The two-way data transfer could also be used for establishing a video transfer between the vehicle and the corresponding PSAP [33].

The development and procuring of its operability were not uncomplicated and have been going on for more than a decade now. Nevertheless, as well the eCall technology is expected to further develop and offer more features in the future.

Appendix

For Austria:

US\$407.45 billion GDP in the year 2012 [25]

3.3% GDP loss due to traffic accidents [24]

US\$421.3547 billion = potential GDP without these accidents

$$x = \frac{407.45}{1 - 0.033} = 421.3547$$

€20 billion possible savings [6] on the €160 billion costs throughout the entire EU [12]

12.5% = expected upper bound on the cost savings

$$x = \frac{20}{160} = 0.125 = 12.5\%$$

US\$13.9047 billion = the potential net gain by 100% reduction in the costs of traffic accidents

$$x = 421.3547 - 407.45 = 13.9047$$

US\$1.7380875 billion = the expected upper bound for the net gain in GDP

$$x = 13.9047 * 0.125 = 1.7380875$$

References

1. Brasseur, T.M., Mladenow, A., Strauss, C.: Business model innovation to support smart manufacturing. In: Proceedings of American Conference on Information Systems 2017 Workshop on Smart Manufacturing (AMCIS), p. 9 (2017). <http://aisel.aisnet.org/sigbd2017/9>
2. Bauer, C., Mladenow, A., Strauss, C.: Fostering collaboration by location-based crowdsourcing. In: Luo, Y. (ed.) CDVE 2014. LNCS, vol. 8683, pp. 88–95. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10831-5_13
3. Mladenow, A., Novak, N.M., Strauss, C.: Internet of Things integration in supply chains – an Austrian business case of a collaborative closed-loop implementation. In: Tjoa, A.M., Xu, L.D., Raffai, M., Novak, N.M. (eds.) CONFENIS 2016. LNBIP, vol. 268, pp. 166–176. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49944-4_13
4. Mladenow, A., Novak, N.M., Strauss, C.: Online ad-fraud in search engine advertising campaigns. In: Khalil, I., Neuhold, E., Tjoa, A.M., Da Xu, L., You, I. (eds.) CONFENIS/ICT-EurAsia -2015. LNCS, vol. 9357, pp. 109–118. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24315-3_11
5. Ernst, C., Mladenow, A., Strauss, C.: Collaboration and crowdsourcing in emergency management. *Int. J. Pervasive Comput. Commun.* **13**(2), 176–193 (2017)
6. Mladenow, A., Bauer, C., Strauss, C.: “Crowd logistics”: the contribution of social crowds in logistics activities. *Int. J. Web Inf. Syst.* **12**(3), 379–396 (2016)
7. EU Decision No. 585/2014/EU: Decision No 585/2014/EU of the European parliament and of the council of 15 May 2014 on the deployment of the interoperable EU-wide eCall service. <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32014D0585>. Accessed 30 Aug 2017
8. Maslow, A.H.: A theory of human motivation. *Psychol. Rev.* **50**(4), 370–396 (1943)
9. Iparraguirre, O., Brazalez, A.: Communication technologies for vehicles: eCall. In: Mendizabal, J., et al. (eds.) Nets4Cars/Nets4Trains/Nets4Aircraft 2016. LNCS, vol. 9669, pp. 103–110. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-38921-9_11

10. Cabo, M., Fernandes, F., Pereira, T., Fonseca, B., Paredes, H.: Universal access to eCall system. *Procedia Comput. Sci.* **27**, 104–112 (2014)
11. EC: Commission Delegated Regulation: European Commission Delegated Regulation (EU) of 12.9.2016. <http://ec.europa.eu/transparency/regdoc/rep/3/2016/EN/C-2016-5709-F1-EN-MAIN-PART-1.PDF>. Accessed 30 Aug 2017
12. CotEC: Commission of the European Communities: The 2nd eSafety Communication, Bringing eCall to Citizens/Druhé sdělení o e-bezpečnosti, Zpřístupnění systému eCall Občanům. <http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52005DC0431>. Accessed 30 Aug 2017
13. CotEC: Commission of the European communities: Bringing eCall back on track – Action Plan (3rd eSafety Communication). <http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52006DC0723>. Accessed 30 Aug 2017
14. EC: eCall Deployment proposal: “Proposal for a decision of the European parliament and of the council on the deployment of the interoperable EU-wide eCall”. <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52013PC0315&from=EN>. Accessed 30 Aug 2017
15. EENA: eCall, Everything you wanted to ask, but did not know how... http://www.eena.org/download.asp?item_id=111. Accessed 30 Aug 2017
16. Chochliouros, I.P., Spiliopoulou-Chochliourou, A.S., Lalopoulos, G.K.: Emergency call (eCall) Services based on approved E-112 regulations and infrastructures: a solution to improve security and release of road help. In: FITCE Congress, pp. 76–84 (2005)
17. Harnischmacher, F., Cosyns, C., Grugl, K., Moerbe, M., Portouli, E., Savaresi, S.M.: State of the art assessment powered Two-Wheeler (P2W) eCall. In: 11th ITS European Congress. https://dl.dropboxusercontent.com/content_link/Nv6xtkIjnWqeZ5dyZdKS7EiY2RTK6AfnHkAHkOI4hGE1BCqkB4FyRtxk0gBCYROz/file. Accessed 30 Aug 2017
18. HeERO: About eCall, eCall - saving lives through in-vehicle communication technology. <http://www.heero-pilot.eu/view/en/ecall.html>. Accessed 30 Aug 2017
19. Carutasu, G.: Further challenges of eCall service and infrastructure. In: MIT 2016 Conference Proceedings, pp. 68–72 (2016)
20. Chariete, A., Bakhouya, M., Nait-Sidi-Moh, A., Ait-Cheik-Bihi, W., Gaber, J., Kouta, R., Wack, M., Lorenz, P.: A study of users’ acceptance and satisfaction of emergency call service. *Int. J. Commun. Syst.* **29**, 2279–2291 (2016)
21. Sihvola, N., Luoma, J., Schirkoff, A., Salo, J., Karkola, K.: In-depth evaluation of the effects of an automatic emergency call system on road fatalities. *Eur. Trans. Res. Rev.* **1**(3), 99–105 (2009)
22. EC Road Fatalities: “EU road fatalities”. http://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/observatory/trends_figures.pdf. Accessed 30 Aug 2017
23. EU: Summary for citizens: “Shrnutí určené občanům: Návrh EU zavést systém palubního tísňového volání eCall pro oznamování dopravních nehod”. http://ec.europa.eu/information_society/activities/esafety/doc/comm_20090821/citizens_sum_cs.pdf. Accessed 30 Aug 2017
24. McKinsey&Co: An integrated perspective on the future of mobility. <http://www.mckinsey.com/business-functions/sustainability-and-resource-productivity/our-insights/an-integrated-perspective-on-the-future-of-mobility>. Accessed 30 Aug 2017
25. Lindsey, R., De Palma, A.: Traffic congestion pricing methodologies and technologies. *Transp. Res. Part C Emerg. Technol.* **19**(6), 945–1400 (2011)
26. Card, W.I., Mooney, G.H.: What is the monetary value of human life? *Br. Med. J.* **1977**(2), 1627–1629 (1977)

27. Rice, D.P., Cooper, B.S.: The economic value of human life. *Am. J. Public Health Nations Health* **57**(11), 1954–1966 (1967)
28. WHO: Country profiles. In: *Global Status Report on Road Safety 2015* World Health Organization, pp. 76–256. http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. Accessed 30 Aug 2017
29. Worldbank: GDP (current US\$). <http://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2012&start=1960>. Accessed 30 Aug 2017
30. Bosch: Das Automatische Bosch Notrufsystem eCall für Automobile. <http://www.bosch-press.de/pressportal/de/de/das-automatische-bosch-notrufsystem-ecall-fuer-automobile-42801.html>. Accessed 30 Aug 2017
31. Spreitzer, S., Mladenow, A., Wagner, G.: IT-getriebenes Instandhaltungsmanagement im After Sales Bereich. *HMD Praxis der Wirtschaftsinformatik* **54**(3), 437–451 (2017)
32. Hornung, G.: Verfügungsrechte an fahrzeugbezogenen Daten. In: *Datenschutz und Datensicherheit*, pp. 359–366 (2015)
33. los Santos Aransay, A., Reina Nieves, A., Rueda Morales, C., Ares, F., Saez Gomez, J., Martinez Madrid, N., Sanz Velasco, P., Seepold, R.: Integration of an advanced emergency call subsystem into a car-gateway platform. In: *Design, Automation and Test in Europe Conference and Exhibition*, pp. 1100–1105 (2009)



An IoT-Big Data Based Machine Learning Technique for Forecasting Water Requirement in Irrigation Field

Fizar Ahmed^(✉)

Corvinus University of Budapest, Budapest, Hungary
fizarbd@yahoo.com

Abstract. Efficient water management is a major concern in rice cropping. Controlling the use of excessive water in irrigation field is essential for the protection of underground water that will also be the part of climate change adaptation. The sustainable use of water resources is the prior task in Bangladesh. Imbalances between demand and supply are the main region for degradation of surface and groundwater. The human readability of checking the water level on irrigation field is considerable for these circumstances. In this paper I discussed the procedure for monitoring of surface water level in irrigation field, continuous monitoring of weather condition like temperature, air pressure, sunlight, rainfall etc. by using sensor network. The aim is to create a machine learning mechanism for farmers that can be given a forecast of water demand of irrigation field by the collection of IoT based data. In turn, this will help the farmer to prepare them to give water and on the other hand it will be helpful to use appropriate ground water and also it can be used for predict energy utilization. In this research Multiple linear regression algorithm is used for this prediction. Data from the irrigation field of North-West part in Bangladesh is used here to find the result of prediction.

Keywords: Internet of Things (IoT) · Irrigation · Ground water
Machine learning · Multiple linear regression

1 Introduction

Water management is important for the adaptation of climate change. Shortage of water resources are directly affects the vulnerability of ecosystems, socio-economic activities and human health. On the other hand climate change is likely to lead to major changes in water availability across Bangladesh with increasing water scarcity and droughts mainly in North-West part of this country.

It's assessed that as much as 50% of irrigation water is wasted due to evaporation or runoff. This happens because most irrigation systems still rely upon simple human reading. However, Internet of Things technologies can provide "Smart" irrigation systems. It can be useable for monitoring soil conditions, surface water level in real time with low power, wireless sensor networks. The wireless sensor networks send the data to a central network gateway, and the network gateway sends the data to the cloud platform. The gateways have the ability to connect via both wired and cellular data

connections, so that can be point them from anywhere. In the internet cloud platform machine learning applications can be used for sending the application result to the end users mobile phone or personal computer.

This research mainly focused on the utilization of ground water in weather-based irrigation field. Weather-based irrigation determines the amount of water needed by the landscape based on the current weather conditions, such as precipitation, solar radiation, temperature, relative humidity, and wind speed. Weather data is provided by IoT based land weather station. Data from the weather station matching with measuring the level of water by using distance sensor where mainly measure the level of water loss from the soil due to evaporation and plant transpiration. This water level data in millimeter will be as a class data with other weather perimeter. A machine learning technique multiple linear regression algorithm is used here for prediction of water loses due to this weather condition in near future.

2 Related Works

In Bangladesh, mainly in the north-west part of this country, ground water is the main source of irrigation. Shahid and Hazarika (2010) investigated groundwater scarcity and drought in three northwestern districts of Bangladesh. They proposed a Cumulative Deficit approach from a threshold groundwater level has been used for the computation of severity of groundwater droughts. Their research shows that groundwater scarcity in 42% area is an every year in the region. The daily evapotranspiration from rice field will increase by an average of 31.3 mm and 0.33 mm/day respectively by the year of 2100 (Shahid 2011). The main finding of this research is that climate change will increase the daily use of water for irrigation by an amount of 0.8 mm/day in the end of this century.

In their research finding Qureshi et al. (2014) shows that 35,322 deep tubewells, 1,523,322 shallow tubewells and 170,570 low lift pumps are working in Bangladesh to provide water for irrigation. About 79% of the total cultivated area in Bangladesh is irrigated by groundwater, whereas the remaining is irrigated by surface water. More than 90% of the pumps within Bangladesh are run by diesel engines. The remaining 10% use electricity. Despite subsidies on electricity, diesel pumps are preferred by farmers due to low capital cost and mobility ease within small and fragmented farm lands. Each year, on average, about 980 million kWh of electricity is used by electric tubewells with an estimated subsidized cost of USD 50 million. The annual diesel consumption for groundwater extraction is of the order of 4.6 billion liters, costing USD 4.0 billion in aggregate.

40 million people are at risk of arsenic poisoning-related diseases because the ground water in these wells is contaminated with arsenic. Alam et al. (2002) reviews the arsenic infection of ground water, hydrological systems, groundwater potential and utilization and environmental pollution in Bangladesh. They discussed the main actions required to ensure the sustainable development of water resources in Bangladesh. Safiuddin and Karim (2001) also highlighted the causes and mechanism of arsenic contamination and presented several measures to remedy the arsenic contamination in groundwater. Another survey by Meharg and Rahman (2003) shows that paddy soils

throughout Bangladesh showed that arsenic levels were elevated in zones where arsenic in groundwater used for irrigation was high, and where these tube-wells have been in operation for the longest period of time. The finding of another research of Meharg (2004) is “Arsenic is sequestered in iron plaque on root surfaces in plants, regulated by phosphorus status, and that there is considerable varietal variation in arsenic sequestration and subsequently plant uptake, offers a hope for breeding rice for the new arsenic disaster in South-East Asia – the contamination of paddy soils with arsenic”.

For reducing the wastage of ground water smart irrigation system is now the most prioritize topic in agriculture research. Mathurkar and Chaudhari (2013) focused on optimizing water management for agriculture through the physical and socioeconomic conditions that inspired the success of an “indigenous technology” which has for spanning exploited the potential for excess harvesting. Monda et al. (2011) described a Precision Agriculture (PA) concept was initiated for site specific crop management as a grouping of locating system. By using this way of proper resource utilization and management, to a environmental friendly sustainable agriculture is possible that they focused. Nandurkar and Thool (2012) designed a sensing system is based on a “feedback control mechanism” with a integrated control unit which standardizes the flow of water on to the field in the real time based on the rapid temperature and moisture values. They also prepared a table that discover the amount of water needed by that crop. Roy and Ansari (2014); Awasthi and Reddy (2013) developed the irrigation control system to avoid wastage of water and increase irrigation efficiency by using a PLC based irrigation system with the help of soil moisture sensor, water level sensor, and GSM controller. Their system can be used for sending message to farmer on mobile through GSM network for controlling actions.

Many machine learning techniques have been developed for learning rules and relationships automatically from various agricultural data sets. McQueen et al. (1995); Ozdogan et al. (2010) described a project that is applying a range of machine learning strategies to problems in agriculture and horticulture. They experimented and described some software requirements on real-world data sets. They also explored the value of archived data that enable comparison of images through time. Ozdogan and Gutman (2008) presented a dryland irrigation mapping methodology that relies on remotely sensed inputs from the MODerate Resolution Imaging Spectroradiometer (MODIS) instrument. They proposed different steps for mapping expected patterns where the dividing of majority of irrigated areas is concentrated in the dry lowland valleys. Image processing is an effective tool for analysis of the agriculture data sets (Vibhute and Bodhe 2012). This paper focussed on the survey of application of image processing in agriculture field such as imaging techniques, weed detection and fruit grading.

A machine learning technique Support Vector Machines (SVMs) was used for classified various crop types in a complex cropping system in the Phoenix Active Management Area (Zheng et al. 2015). They used “Landsat time-series Normalized Difference Vegetation Index (NDVI)” data using training datasets selected by two different approaches: stratified random approach and intelligent selection approach using local knowledge. For weather prediction (Radhika and Shashi 2009), long-term prediction of lake water levels (Khan and Coulibaly 2006), SVM is the most promising technique for better expectation. SVM can also be used for time series application in

many application areas from financial market prediction to electric utility load forecasting to medical and other scientific fields (Sapankevych and Sankar 2009).

3 Multiple Linear Regression Algorithm

A multiple linear regression (MLR) model that describes a dependent variable y by independent variables x_1, x_2, \dots, x_p ($p > 1$) is expressed by the equation as follows, where the numbers α and β_k ($k = 1, 2, \dots, p$) are the parameters, and ϵ is the error term.

$$y = \alpha + \sum_k \beta_k x_k + \epsilon$$

For example, in the built-in data set *stackloss* from observations of a chemical plant operation, if we assign *stackloss* as the dependent variable, and assign *Air.Flow* (cooling air flow), *Water.Temp* (inlet water temperature) and *Acid.Conc.* (acid concentration) as independent variables, the multiple linear regression model is:

$$\text{Stack.Loss} = \alpha + \beta_1 * \text{Air.Flow} + \beta_2 * \text{Water.Temp} + \beta_3 * \text{Acid.Conc.} + \epsilon$$

4 Methodology

4.1 Hardware Specifications

To read real-time data is typical of a weather station, using different sensors, and capable of communicating via LoRa. After a review of all known hardware available on the market, all the components strictly necessary to the solution were defined, which in turn fulfilled the requirements of the above: The hardware chosen was:

- Adafruit Feather 32u4 RFM95 LoRa Radio with female pin headers²
- Adafruit RTC DS3231³
- Sparkfun Weather Shield with RJ11 female connectors⁴
- Wind and Rain sensors kit⁵
- Antenna 868 MHz and SMA cable

4.2 Core System Controller

In terms of the core **system** within the Weather Station solution, it is composed by the Feather32u4, which that takes a specialized role in the system where it performs control functions through software, with processing power, enabling the sensory devices to gather data from the environment, using specific libraries. This system also has built-in communication capabilities.

4.3 Data Acquisition

The weather shield is an integrated module with several built-in sensors capable of collecting data, such as temperature, humidity, luminosity, barometric pressure and altitude. Along these sensors, the weather shield also enables the integration of three more different sensors to collect data regarding wind direction, wind speed and amount of rain. Based on the proposed model, it becomes clear the connection between the controller and the Weather Shield. This connection is established with the I2C protocol that allow this digital integrated circuit to communicate with one or more masters. It is used this type of protocol because it's only intended short distance communications within a single device and only requires two signals to exchange information. The software controller uses the library "Wire.h" that is dedicated to the I2C logic protocol. The embedded software requires the "SparkFunHTU21D.h" and "SparkFunMPL3115A2.h" libraries in order to call all the functions responsible for activating and reading the sensors signals coupled to the weather shield.

4.4 Data Communication

Like as expected in the proposed system model, the controller will send data to the outstation, based on the information collected from the weather shield module. For this it makes use of the "SPI.h" library to run the communication with the radio module RFM9x LoRa 868/915. The LoRa radio must communicate with the LoRa gateway, specified by the system, and for that will interact with the "featherLora.h" library. The data will be collected according to the time windows described, already considered in the project. At the end of each time window will be sent the package with the message containing the information collected. In data sharing with outstation it was established to send an acknowledge information packet like a result of the incoming data from the different sensors. The typical message to be sent from one gateway to another is based on the type message as described in the following example:

Example of the message send in the package:

\\!TC/18/HU/85/LU/0.56/ WD/90/WC/5.55

The following table outlines the type and content of the information sent in each package (Table 1):

Table 1. Communication package.

TC	Temperature
HU	Humidity
LU	Luminosity
WD	Wind Direction
WC	Wind Speed

4.5 Water Level Measurement by Ultrasonic Sensor

Ultrasonic distance measurement sensor is used to observe the level of water and radio communication of this value via license free LoRa devices is used under different

package simultaneously with weather station. Temperature compensated distance sensor US-100 (Voltage 2.4–5.5 V), Arduino Pro Mini 3.3 V (8 MHz) version, energy source (2 or 3 battery type AA), Radio module used in this package.

The architecture of proposed model and it’s operating flow with the set specifications (Fig. 1):

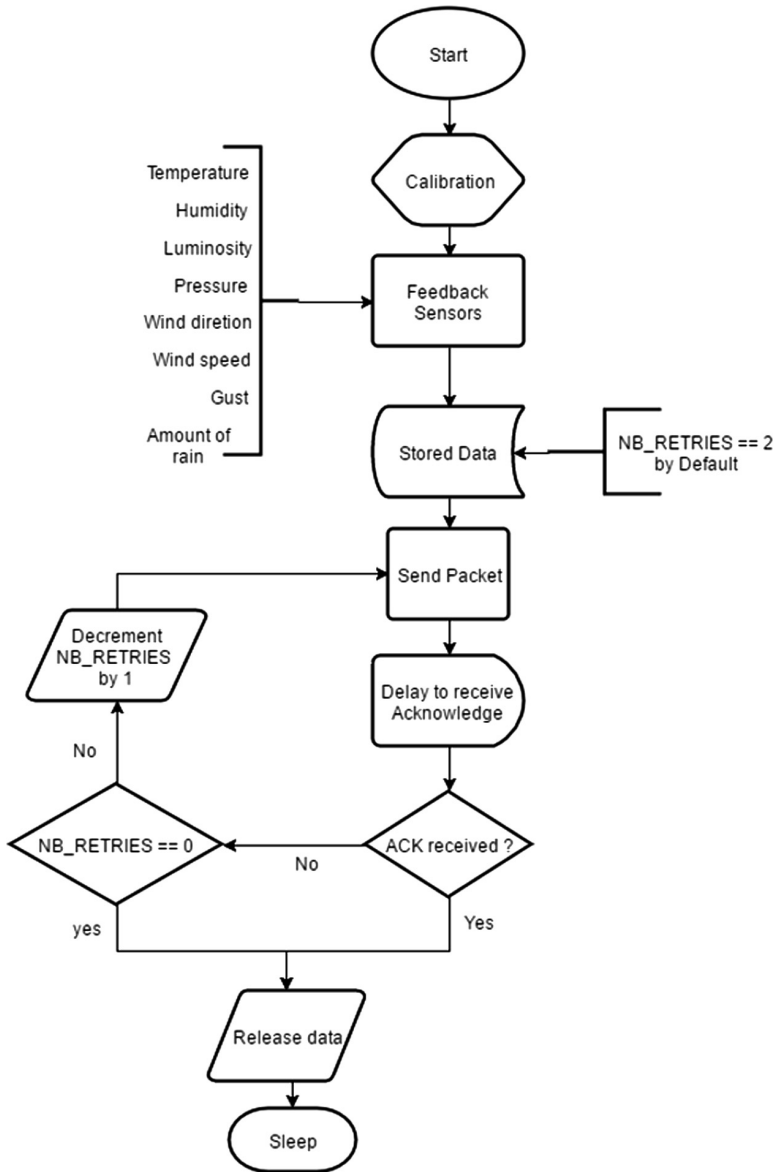


Fig. 1. Architecture model with the set specifications.

5 Result and Analysis

The following result has been found from this testing dataset:

```
Call:
lm(formula = DM ~ WD + WC + HU + LU + TC)

Residuals:
    Min       1Q   Median       3Q      Max
-16.4617  -5.1362   0.6289   6.6590  10.3010

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.011e+02  4.206e+02  -0.478   0.6420
WD           1.965e-02  2.405e-02   0.817   0.4312
WC           5.126e-03  1.900e-02   0.270   0.7923
HU          -1.786e+00  1.720e+00  -1.039   0.3213
LU           1.012e+02  1.098e+02   0.921   0.3769
TC           7.071e+00  2.412e+00   2.932   0.0137 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.378 on 11 degrees of freedom
Multiple R-squared:  0.7915,    Adjusted R-squared:  0.6967
F-statistic:  8.35 on 5 and 11 DF,  p-value: 0.001773
```

Multiple R-squared: Approximately 79% variation in water level (Distance in Millimetre- DM) can be explained by this model. (Wind direction -WD, Wind Speed-WC, Humidity-HU, Luminous-LU and Temperature-TC)

F- statistics: These tests are null hypothesis and all the model coefficients are 0.

Residual standard error gives the idea of how far observed water level -DM (Y-values) are from the predicted or fitted DM(the Y-hats). This gives us an idea of a typical size of residual or error $e = y - y'$.

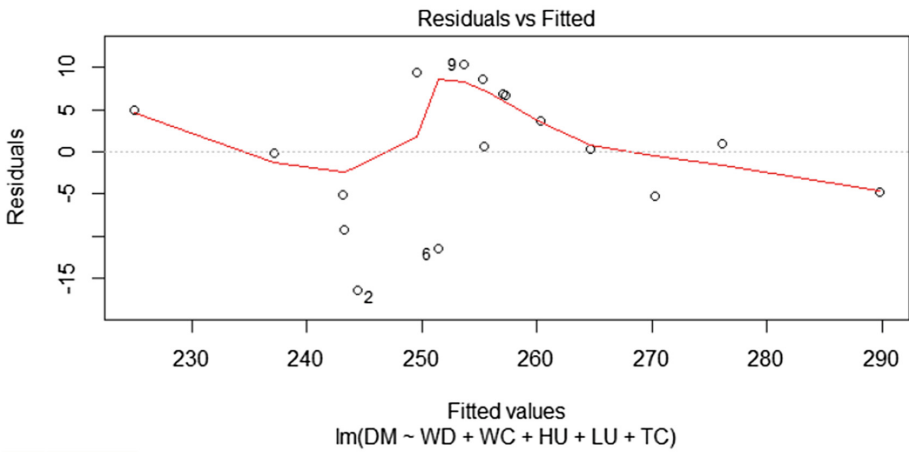
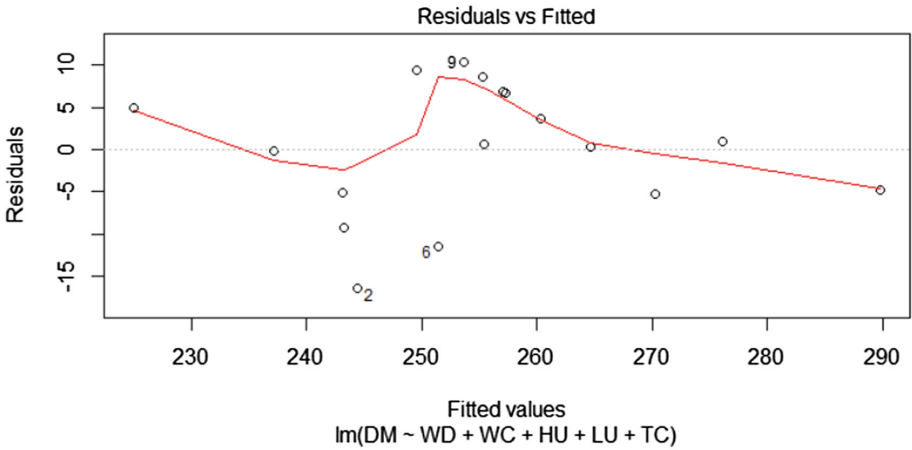
The intercept shows the estimated mean Y value when all Xs are 0. We can associate with increase some values of wind direction, speed or other values water level adjusting or controlling the for Luminous or humidity. The hypothesis test that the slope for WD or others is 0.

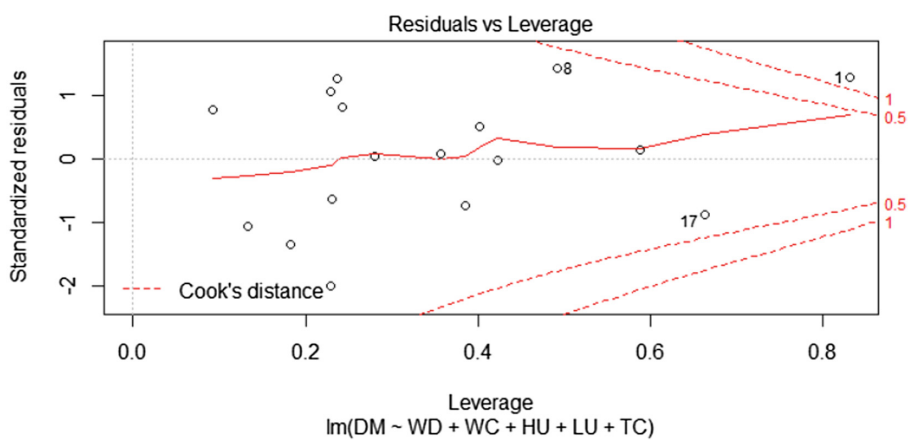
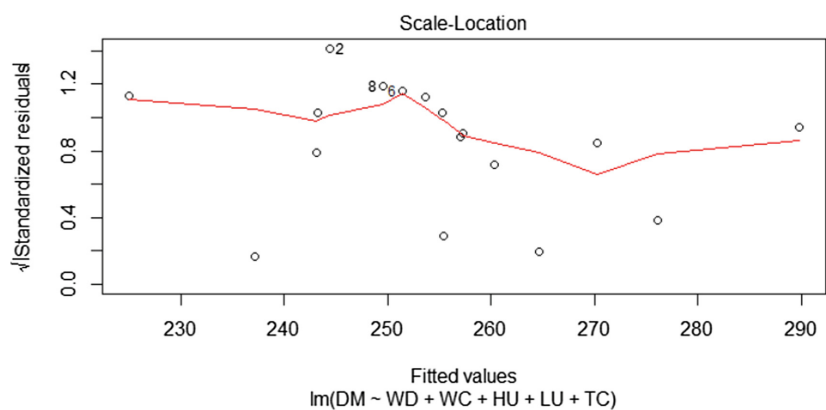
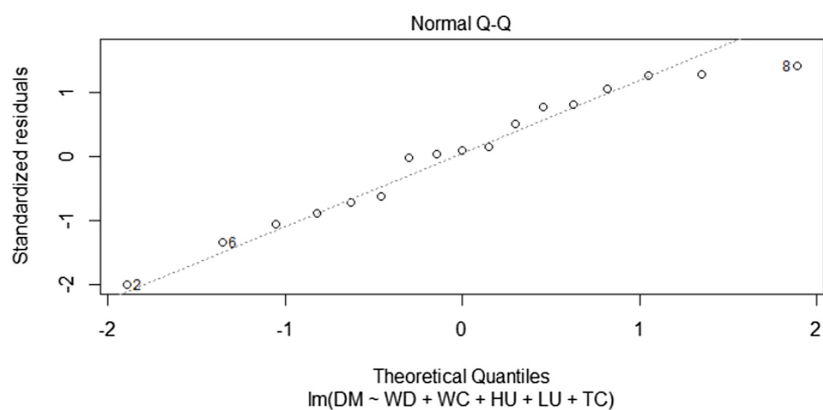
```
> cor(WD, WC, method="pearson")
[1] 0.04384084
>
```

The collinearity between WD and WC means that we should not directly interpret the slope, as the effect of WD on DM adjusting for WC. The high correlation between two values suggests that these two effects are somewhat bounded together.

Here confint values shows the slope for the level of 95%.

```
> confint(model1, conf.level=0.95)
                2.5 %      97.5 %
(Intercept) -1.126757e+03 724.62189877
WD           -3.327983e-02  0.07258805
WC           -3.669872e-02  0.04695124
HU           -5.571183e+00  1.99902046
LU           -1.406134e+02 342.91358614
TC           1.762182e+00 12.38057077
```





6 Conclusion

This IoT based machine learning works was my sample hands-on experience with real time data. Behind this task, it has a lot of preparation for a big part: it involved data understanding, sorting and reframing. That is beyond of this research work. A sample small scale data just used here to show the result for predicting data model. It is definitely challenging to work with this type big data. And finally, as I tried to understand the different correlation relationships between the parameters and the forecasts, I surprisingly also got a better understanding of prediction from the information of weather perimeters and IoT data collection point of view.

References

- Shahid, S., Hazarika, M.K.: Groundwater drought in the northwestern districts of Bangladesh. *Water Resour. Manage* **24**(10), 1989–2006 (2010)
- Shahid, S.: Impact of climate change on irrigation water demand of dry season Boro rice in northwest Bangladesh. *Clim. Change* **105**(3), 433–453 (2011)
- Alam, M.G.M., Allinson, G., Stagnitti, F., Tanaka, A., Westbrooke, M.: Arsenic contamination in Bangladesh groundwater: a major environmental and social disaster. *Int. J. Environ. Health Res.* **12**(3), 235–253 (2002)
- Qureshi, A.S., Ahmed, Z., Krupnik, T.J.: *Groundwater Management in Bangladesh: An Analysis of Problems and Opportunities* (2014)
- Safuiddin, M., Karim, M.M.: Groundwater arsenic contamination in Bangladesh: causes, effects and remediation. In: *Proceedings of the 1st IEB International Conference and 7th Annual Paper Meet. The Institution of Engineers, Chittagong Center, Bangladesh, November 2001*
- Meharg, A.A., Rahman, M.M.: Arsenic contamination of Bangladesh paddy field soils: implications for rice contribution to arsenic consumption. *Environ. Sci. Technol.* **37**(2), 229–234 (2003)
- Meharg, A.A.: Arsenic in rice—understanding a new disaster for South-East Asia. *Trends Plant Sci.* **9**(9), 415–417 (2004)
- Mathurkar, S.S., Chaudhari, D.S.: A review on smart sensors based monitoring system for agriculture. *Int. J. Innovative Technol. Exploring Eng. (IJITEE)* **2**, 76–78 (2013)
- Monda, P., Basu, M., Bhadoria, P.B.S.: Critical review of precision agriculture technologies and its scope of adoption in India. *Am. J. Exp. Agric.* **1**(3), 49 (2011)
- Nandurkar, S.R., Thool, V.R.: Design of a soil moisture sensing unit for smart irrigation application. In: *International Conference on Emerging Technology Trends on Advanced Engineering Research (ICETT 2012)*, Proceedings published by International Journal of Computer Applications (IJCA), pp. 1–4 (2012)
- Roy, K.D., Ansari, H.M.M.: Smart irrigation control system. *Int. J. Environ. Res. Dev.* **4**(4), 371–374 (2014)
- Awasthi, A., Reddy, S.R.N.: Monitoring for precision agriculture using wireless sensor network—a review. *Global J. Comput. Sci. Technol.* **13**(7) (2013)
- McQueen, R.J., Garner, S.R., Nevill-Manning, C.G., Witten, I.H.: Applying machine learning to agricultural data. *Comput. Electron. Agric.* **12**(4), 275–293 (1995)
- Ozdogan, M., Yang, Y., Allez, G., Cervantes, C.: Remote sensing of irrigated agriculture: opportunities and challenges. *Remote Sens.* **2**(9), 2274–2304 (2010)

- Ozdogan, M., Gutman, G.: A new methodology to map irrigated areas using multi-temporal MODIS and ancillary data: an application example in the continental US. *Remote Sens. Environ.* **112**(9), 3520–3537 (2008)
- Vibhute, A., Bodhe, S.K.: Applications of image processing in agriculture: a survey. *Int. J. Comput. Appl.* **52**(2), 34–40 (2012)
- Zheng, B., Myint, S.W., Thenkabail, P.S., Aggarwal, R.M.: A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. *Int. J. Appl. Earth Obs. Geoinf.* **34**, 103–112 (2015)
- Radhika, Y., Shashi, M.: Atmospheric temperature prediction using support vector machines. *Int. J. Comput. Theory Eng.* **1**(1), 55 (2009)
- Khan, M.S., Coulibaly, P.: Application of support vector machine in lake water level prediction. *J. Hydrol. Eng.* **11**(3), 199–205 (2006)
- Sapankevych, N.I., Sankar, R.: Time series prediction using support vector machines: a survey. *IEEE Comput. Intell. Mag.* **4**(2), 24–38 (2009)

EIS for Industry 4.0



Penetration of Industry 4.0 Principles into ERP Vendors' Products and Services – A Central European Study

Josef Basl^(✉)

Prague University of Economics,
W. Churchill Sq. 4, 130 67 Prague, Czech Republic
basl@vse.cz

Abstract. The paper deals with aspects of EIS (Enterprise Information Systems) innovation based on the development of the internet of things. The article presents the main results of a central European study dealing with the penetration of the Industry 4.0 principles into the offers of a representative sample of ERP (Enterprise Resource Planning) vendors. The results show the current strategies of ERP vendors, the integration of the new principles of Industry 4.0 into ERP applications and the position of ERP systems in the roadmap of Industry 4.0 implementation.

Keywords: ERP · ERP market · Internet of things · IoT · Industry 4.0
4th Industrial Revolution · ICT innovation

1 Introduction

Based on the growing number of conferences, seminars, workshop and articles, it seems that we are still in a growing wave of Industry 4.0. Some authors are often underlining a certain “revolutionary” character of these changes and therefore they speak about the 4th Industrial Revolution. On the other hand, it is possible to agree with those who perceive it as a further evolutionary step towards the digitizing of products and business processes based on IT as described by G. Moore in his book “Dealing with Darwin” [16]. Many publications are focused on Industry 4.0. One of them to be mentioned is “The Second Machine Age” [3]. This book describes how technology has influenced human society in recent times and deals with new business models and services that are emerging or are emerging with new technologies. Last but not least, the concerns of human and robot collaboration are mentioned.

One of the main innovation features is related to the Internet of Things (IoT) and many national strategies of leading industrial countries are dedicated to this topic – a good example is “Made in China 2025” [15], “Industrial Internet” in the US [23], Germany [17] and Great Britain. They emphasise the national level, but there are many ways to deal with the readiness for Industry 4 in companies (www.industrie40-readiness, www.przemysl4.pl, www.firma4.cz).

There is no doubt that the new digitalization process is inconceivable without business software applications – ERP systems. These have played a significant role in

the digitalization of companies since the 1990s. ERP systems reacted lately for example to innovations like the first internet wave, social networks, mobile devices, and even now companies believe that ERP systems are a crucial foundation stone of Industry 4.0 enterprise architecture [1].

Many analyses have been published about the readiness of companies to adopt new 4.0 principles. The perspective of the “demand” side of companies is mostly used. But the “offer” side of ERP vendors is not so often taken by researchers. This is one of the reasons why readiness of ERP vendors and their products and services were analysed in May/June 2017 in a very representative sample of a highly industrialized country in Central Europe – the Czech Republic. The following paper describes the main results of this survey.

2 Theoretical Background – ERP Systems and Trends in IoT

2.1 ERP Systems as a Main Long-Term Player in Digitalization of Manufacturing

Industry 4.0 is a very dynamic link between IT and manufacturing companies. It has great prospects, because these two categories are also the basic strategic technologies for the next 15 years [10]. However, many current IT trends in manufacturing companies do not have their origins today but they are being promoted over the longer term, perhaps in only a newer form with new data and in a more integrated and more user-friendly approach.

The digitalization of manufacturing had already started in the 1980s - more than 30 years ago. ERP systems have played a major role in the applied concepts and platforms since the beginning.

Table 1. Key role of ERP in digitization of manufacturing companies

	80’s	90’s	2000	2010	2015+
Main concept	CIM FMS	ERP CAD/CAM	ERP + CRM + SCM PLM	ERP + MES + APS	ERP in Industry 4.0 concept Smart factory Digital Twins
Main technology driver	Relational database	Relational database	Internet Portal role of ERP (.com)	Mobile applications	IoT Apps
Main integration area	Product and production systems	Customer order	Horizontal chain from suppliers to clients Product Lifecycle	Vertical connection from all plans to realization of production and logistics flow	Sensors Machine-machine communication Man-machine communication
Main data role	Product and production system data integration	Customer data integration	BI (Business Intelligence) Analytics of internal data	CI (Competitive Intelligence) Analytics of external data	AI (Artificial Intelligence) Big Data Analytics

Table 1 presents all the main concepts, technology drivers and integration areas of IT penetration in manufacturing over the last 35 years. It is unfortunately not possible to describe all these aspects and relations in proper detail in this paper.

The most important feature is that the ERP systems followed the implementation stages of CIM (Computer Integrated Manufacturing) [21] and FMS (Flexible Manufacturing Systems) [25] concepts. All these concepts were based on relational databases. The digital factory and digital enterprise information systems based on ERP systems represent the next steps, followed by integration with CRM and SCM, MES and APS and BI finally in the following decades. The last step is an ERP system in Industry 4.0 aimed at smart factories [24] and digital twins.

2.2 IoT as One of the Key Trends in ERP Development

Many significant consultancy companies such as Gartner Group [9], BCG [2] or Deloitte [6] present their typologies of what the current trend towards Industry 4.0 should include. Mostly the following trends are mentioned:

- cloud
- big data
- internet of things
- extended reality
- simulation, digitization
- digital twinning, various autonomous solutions, human and robot collaboration
- a wide range of sensors and their evaluation leading to artificial intelligence

The cloud solution and big data are now already relatively widely used and exploited and they are something like a key enabler of current changes. But the real symbol of the new trends is the internet of thing (IoT). The increasing availability of Internet connectivity, declining Internet connection costs, and a growing number of devices that include Wi-Fi technology and other sensors are perfect for creating IoT.

The integration of ERP and IoT is important globally too. For example the world ERP leaders like SAP, Microsoft and Oracle are today also in the top ten of IoT leaders. (Microsoft in 3rd position, SAP in 8th position and Oracle in 10th position (IoT report, 2015).

It is clear that the concept of Industry 4.0 is based on industrial integration mediated by information technology. This integration involves real-time or near-real-time data sharing, information sharing, and continuous communication. This is also a potential for the further development of ERP.

3 Methodology – Formulation of Aim and Research Questions

This paper deals with a survey of the penetration of Industry 4.0 principles into ERP companies, their products and services. The important questions concern the role of selected IT trends and enterprise information system software applications within

Industry 4.0 now and in the near future (the next 2 and 5 years). An important question is also the preparation of ERP companies for this new trend in their strategies.

The motivation for this survey was not only the current technological trends but the published manufacturing study oriented towards Industry 4.0 penetration at the global level (Infosys, 2015) and on the national level in Germany [7] and Perspective [18]. These surveys were the motivation for our own survey in the Czech Republic of certain companies [1]. The results of this survey from last year confirmed the key role of ERP systems (65.2%) in the integration of plans of companies during the preparation for Industry 4.0. The next most important package was MES applications (43.5%).

Another reason for the survey described in this paper is to obtain a more detailed view of current ICT trends that are somehow connected with ICT, such as mobile devices, clouds and big data on the one hand, and ERP, MES, APS and BI applications on the other. Last but not least, trends like robots, smart logistics and flexible production planning are also analysed.

The main research questions in this survey are:

- (1) Have ERP vendors already integrated the new principles of Industry 4.0 in their products?
- (2) Do ERP systems still play a major role in enterprise architecture which should be implemented in Industry 4.0 strategy?
- (3) Are the main Industry 4.0 trends already integrated in current ERP vendor products and services or are they planned in the following 2 or 5 years?
- (4) Do ERP vendors already have their own Industry 4.0 strategy?

4 Sample Description and Data Collection

The subtitle of the article is “a central European study” because the survey was carried out in the Czech Republic. This country provides a good example of ERP trends. It has the highest proportion of industrial production in the economy as a whole from the European Union, namely 47.3%. This is more than the most industrialized EU countries such as Germany (40.2% with EU share 27%), the United Kingdom (41.7% with the EU share 12.7%) or Italy (37.9% with EU share 12.1%).

The penetration of ERP in the Czech Republic is also very high, reaching 21.4% for small companies, 57.8% for medium sized companies and 81.8% for large companies [5].

A complete list of all ERP systems and their vendors is available at the portal systemonline.cz. There are 88 ERP products from 75 ERP vendors. All these vendors were addressed by the survey. A special questionnaire form was created with the formulated research questions and it was made available for ERP companies on a website. Data collection was carried out from the online survey in May/June 2017.

Fifteen ERP companies answered the survey, meaning that the survey had a 22.7% response rate. It is important that the sample of companies reflects the profile of the whole Czech economy. In addition, the main market share belongs to SAP and Microsoft applications and these two companies also responded to the survey. It means that even though the response rate was not so high, the results gained from the survey are representative. Furthermore, some other reactions were gained from the survey.

There were also answers from companies who only sent an email informing us that they were not able to complete it because they were not yet considering Industry 4.0 principles. They also admitted that receiving the survey form was an inspiration for them. There were five such companies.

5 Research Results

Generally, the majority of companies that participated in the survey declared that they have dealt with Industry 4.0 for either more than three years (60%) or more than one year (20%). There were no companies delivering ERP products on the Czech market that said “that they know this new trend but they do not want to implement it” or declared “that they have not met the Industry 4.0 so far”.

5.1 Integration of 4.0 Principles in Enterprise Applications

The results for the first research question “Have ERP vendors already integrated new principles of Industry 4.0 in their products” confirm the key role of ERP systems in integration plans of ERP vendors during their preparation for Industry 4.0 (Table 2).

Table 2. Integration of 4.0 principles in enterprise applications

Integration of 4.0 principles in enterprise applications	Already integrated	Integration in plan	No plan to integrate
ERP (Enterprise Resource Planning)	80%	13%	7%
MES (Manufacturing Execution System)	53%	27%	20%
APS (Advanced Planning and Scheduling)	53%	27%	20%
WMS (Warehouse Management System)	53%	27%	20%
PLM (Product Lifecycle Management)	47%	20%	33%
CRM (Customer Relationship Management)	53%	27%	20%
BI (Business Intelligence)	40%	20%	40%
BPM/BPMS (Business Process Management Suites)	33%	20%	47%
Apps	20%	27%	53%

80% of ERP systems declare that the principles of 4.0 are integrated. The following most integrated are MES and APS applications. CRM with PLM applications are also applications with a high integration rate. In terms of smaller or not planned integration, APPS, but also BMP and BI are positioned on the opposite side of the application spectrum with a very low level even in the future.

5.2 Key Role of ERP in Implementation of Industry 4.0 Strategy

The second research question “Do ERP systems still play a major role in enterprise architecture which should be implemented in “Industry 4.0” strategy” was oriented towards the sequence of implementation of main enterprise software applications.

The results show possible sequences that are very diverse (see the examples below). Anyhow, the answers to proper “implementation order” can be divided into the following three types:

The first type of implementation order uses ERP systems as the first step of gradual Industry 4.0 implementation followed by MES application:

- ERP then MES then APS then WMS
- ERP + MES + WMS + PLM (together) then APS then CRM

In the second type ERP systems are associated with CRM, when CRM was also nominated as the first application:

- ERP + CRM then APS then WMS + MES + BI + PLM
- CRM then ERP then WMS then BI

The third type applied MES in the first position and ERP was one of the last steps:

- MES then APS then PLM then WMS then CRM then ERP then BI

The survey confirms the key role of ERP as the first and the most important foundation stone of Industry 4.0 architecture. ERP was mostly in first place.

5.3 Industry 4.0 Trends Applied in the ERP Offer

The third research question was: Are the main Industry 4.0 trends already integrated in the current ERP vendor products and services or are they planned in the following 2 or 5 years?

The results show that all the selected topics are already being applied today. Cloud computing and mobile devices are in first position, big data is in second and the industrial internet of things is in third position (Table 3).

Table 3. Industry 4.0 trends applied in ERP offer

Industry 4.0 trends applied in enterprise applications	Used now	Planned to be used in following 2 years	Planned to be used in following 5 years	No plan for usage
Mobile devices	73%	0%	20%	7%
Cloud computing	73%	0%	0%	27%
Big data	60%	7%	20%	13%
Industrial Internet of Thing	47%	13%	33%	7%
Digital production	40%	0%	20%	40%

(continued)

Table 3. (continued)

Industry 4.0 trends applied in enterprise applications	Used now	Planned to be used in following 2 years	Planned to be used in following 5 years	No plan for usage
Additive manufacturing	40%	0%	13%	57%
Cyber security of data	33%	0%	33%	33%
Wearables (glasses, watches, ...)	27%	7%	27%	53%
BYOD concept (bring your own device)	20%	13%	20%	47%
Voice control	27%	20%	20%	33%

Trends such as Digital production, Additive manufacturing (known also as 3D printing) and Cyber Security have less positive answers for using today but they have higher potential for the future.

The second group of answers could be called “further industry 4.0 trends”. This second group shows the lower level applied trends today and at the same time the very high level of answers that they will not be in plan even in the future. Trends like augmented reality, digital modelling and new sensors and their relations to the ERP could be an idea for a separate survey (Table 4).

Table 4. Further Industry 4.0 trends applied in ERP offer

Further Industry 4.0 trends	Used now	Planned to be used in following 2 years	Planned to be used in following 5 years	No plan for usage
Machine learning	20%	7%	33%	40%
Artificial intelligence	13%	7%	53%	27%
Augmented reality	13%	7%	33%	47%
Digital modelling	13%	7%	27%	53%
Energy harvesting	13%	13%	13%	60%
Autonomous robots	13%	7%	33%	47%
Virtual assistants	7%	7%	47%	40%
Human-robot interaction	7%	7%	33%	53%
New sensors, incl. Biosensors	7%	13%	33%	47%
Quantum computing	0%	20%	33%	47%
Brain interfaces	0%	7%	33%	60%

5.4 Existence of Industry 4.0 Strategy

The fourth research question was: Do ERP vendors already have their own Industry 4.0 strategy?

A high percentage of ERP vendors (60%) have already a strategy for Industry 4.0 and some vendors still do not have a strategy but they are working on preparing it (13%). And finally still 27% of ERP vendors do not have a strategy for Industry 4.0. The results show that more ERP vendors have own Industry 4.0 strategy than companies using their products. The similar survey from the last year (Basl, 2016) showed that a high percentage of Czech enterprises (39.1%) did not have a strategy for Industry 4.0. Nearly the same percentage of enterprises was preparing such strategy (30.4%). And finally, nearly 25%, it means each fourth company, already had a strategy for Industry 4.0 (Table 5).

Table 5. Industry 4.0 trends applied in enterprise applications

Strategy for Industry 4.0	Ratio
We do not have a strategy for Industry 4.0	27%
We do not have a strategy for Industry 4.0 now but we are preparing it	13%
We have a strategy for Industry 4.0 and it is a part of business strategy	60%
We have a strategy for Industry 4.0 but it is not a part of business strategy	0%

It is very similar to the answers from firms in the global survey (Infosys, 2015). The reason for this could be that companies have to intensively care about the Industry 4.0.

6 Conclusion

Industry 4.0 seems to be a topic with high potential, especially at a time when the digitalization of production is growing. The survey results indicate many similarities between Industry 4.0 penetration in ERP systems in the Czech Republic and leading developed countries.

The survey identified a big potential for further analyses and surveys of the obstacles why Industry 4.0 is not applied more widely. The main reason for ERP companies is high costs connected with Industry 4.0 implementations (53%) and then little awareness of the issues of Industry 4.0 (47%). The low rate of usage of Industry 4.0 in companies (27%) and not so clear business effects (7%) were also mentioned as the obstacles.

The survey results also show that companies perceive the level of penetration to be higher than the level in companies that expressed the need for the existence of the proper methodologies and road maps for Industry 4.0 implementation so far. These aspects are again possible areas for research in the future.

References

1. Basl, J.: Enterprise information systems and technologies in Czech companies from the perspective of trends in Industry 4.0. In: Tjoa, A.M., Xu, L.D., Raffai, M., Novak, N.M. (eds.) CONFENIS 2016. LNBIP, vol. 268, pp. 156–165. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49944-4_12
2. BCG Perspectives: Winning in IoT: It's All About the Business Processes. <https://www.bcgperspectives.com/content/articles/hardware-software-energy-environment-winning-in-iot-all-about-winning-processes/> Accessed 10 June 2017
3. Brynjolfsson, E., McAfee, A: The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies (2016)
4. Computer Sciences Corp: CSC - Studie Industrie 4.0: Ländervergleich Dach (2015). http://assets1.csc.com/de/downloads/Ergebnisse_CSC-Studie_4.0.pdf. Accessed 27 Apr 2016
5. Czech Statistical Office: ICT usage in companies. https://www.czso.cz/csu/czso/podnikatelsky_sektor. Accessed 07 June 2016
6. Deloitte – Industry 4.0. <https://www2.deloitte.com/nl/nl/pages/consumer-industrial-products/topics/industry-4-0.html>. Accessed 01 June 2016
7. Eisert, R.: Sind Mittelständler auf Industrie 4.0 vorbereitet? (2014). <http://www.wiwo.de/unternehmen/mittelstand/innovation-readiness-index-sind-mittelstaendler-auf-industrie-4-0-vorbereitet/10853686.html>. Accessed 27 Apr 2016
8. Tao, F., Zuo, Y., Xu, L.D., Zhang, L.: IoT-Based intelligent perception and access of manufacturing resource toward cloud manufacturing. *IEEE Trans. Ind. Inform.* **10**(2), 1547–1557 (2014)
9. Gartner - Top 10 Strategic Technology Trends for 2016. <http://www.gartner.com/technology/research/top-10-technology-trends/>. Accessed 27 Apr 2016
10. Global Trends 2030: Alternative Worlds, National Intelligence Council (2012). <https://globaltrends2030.files.wordpress.com/2012/11/global-trends-2030-november2012.pdf>. Accessed 27 Apr 2016
11. Industry 4.0 - The State of the Nations, INFOSYS. http://images.experienceinfosys.com/Web/Infosys/%7Bf0e3bb53-176a-4b5a-991b-0708c00fc0a9%7D_Industry_4.0_-_The_State_of_the_Nations_2015_-_Research_Report.pdf. Accessed 17 Apr 2016
12. The Industrial Internet of Things, PwC. <https://www.pwc.com/gx/en/industries/technology/publications/industrial-internet-of-things.html>. Accessed 27 May 2017
13. The Industrial Internet Consortium: A Global Nonprofit Partnership of Industry, Government and Academia (2014). <http://www.iiconsortium.org/about-us.htm>. Accessed 27 Apr 2016
14. IoT Analytics, 2015 - IoT Report: Ranking-IoT-Companies-Q3-Q4-2015. <http://iot-analytics.com/wp/wp-content/uploads/2016/01/Ranking-IoT-companies-Q3-Q4-2015-Dec-2015-v6.pdf>
15. Kenedy, S.: Made in China 2025, Center for Strategic and International Studies (2015). <http://csis.org/publication/made-china-2025>. Accessed 27 Apr 2016
16. Moore, G.A.: Dealing with Darwin: How Great Companies Innovate at Every Phase of Their Evolution. Penguin, New York (2005)
17. National Initiative – Industry 4.0, Ministry for Industry and Trade, September 2015. <http://www.spcr.cz/images/priloha001-2.pdf>. Accessed 27 Apr 2016
18. Perspektive Mittelstand: Industrie 4.0 macht Mittelstand zu schaffen (2015). <http://www.perspektive-mittelstand.de/Industrie-40-macht-Mittelstand-zu-schaffen/managementwissen/6093.html>. Accessed 17 Apr 2016

19. Premier of the State Council of China, Li, K.Q.: Report on the work of the government. In: Proceedings of the 3rd Session of the 12th National People's Congress, March 2015. Accessed 27 Apr 2016
20. Report: Accessed: Research and Markets: Enterprise 2.0: Is It Time for Your Organization to Make the Transition (2008). <http://search.proquest.com/docview/446162456?accountid=149652016-04-27>. Accessed 27 Apr 2016
21. Scheer, A.W.: CIM – Computer Integrated Manufacturing. Springer, Heidelberg (1987). <https://doi.org/10.1007/978-3-642-73458-8>
22. Soliman, F., Youssef, M.A.: Internet-based e-commerce and its impact on manufacturing and business operations. *Ind. Manag. Data Syst.* **103**(8–9), 546–552 (2003)
23. USA: Industry 4.0 the American Way. <http://www.process-worldwide.com/usa-industry-40-the-american-way-a-536602/>. Accessed 31 May 2017
24. Wang, S., Wan, J., Li, D., Zhang, C.: Implementing smart factory of Industrie 4.0: an outlook. *Int. J. Distrib. Sens. Netw.* **2016** (2016). <https://doi.org/10.1155/2016/3159805>. Article ID 3159805, 10 pages
25. Xu, X.: From cloud computing to cloud manufacturing. *Robot. Comput. Integr. Manuf.* **28**(1), 75–86 (2012)



Systematic Analysis of Future Competences Affected by Industry 4.0

András Gábor¹, Ildikó Szabó²(✉), and Fizar Ahmed²

¹ Future Internet Living Lab Association,
Közraktár u. 12/A, Budapest 1093, Hungary
agabor@flab.hu

² Corvinus University of Budapest,
Fővám tér 13-15., Budapest 1093, Hungary
ildiko.szabo2@uni-corvinus.hu, fizarbd@yahoo.com

Abstract. Digital transformations boosted by new technological innovations entail restructured industrial processes and requalified skilled workers. Educational institutions must provide qualifications with learning outcomes fitting to these requirements. Nowadays skill gap analysis between both sides of labor market is a crucial research topic, but researchers mostly draw consequences from experts' visions, trends in past data and not from systematic analysis. Educational institutions must gather information about competences required in the future to start transferring them these relevant knowledge in time. This paper presents an information system dedicated to estimate the importance of actual competences in the future based on different business scenarios.

Keywords: Industry 4.0 · Future competence · Skill gap analysis
I/O model

1 Introduction

Researchers emphasize different aspects of Industry 4.0. Lu [17] summarized it as “an integrated, adapted, optimized, service-oriented, and interoperable manufacturing process which is correlate with algorithms, big data, and high technologies”. Nevertheless, it has social aspects as well, because it takes effect on the demand side of labor market due to the emerging technologies (e.g. Smart Systems, Blockchain, Virtual Reality, Internet of Things etc.) applied by organizations to improve the effectiveness of their business processes. Supply chain processes including human aspects are transforming. The communication between human and machine is getting more interactive, presenting two-way interaction due to the machine learning developments. Technological competences required to execute procurement, production or sales activities are continually changing due to the short technological lifecycles. Adaptive, innovative, responsible decision makers desiring for knowledge are ideal employees in this variant environment. Educational institutions must rethink how the students' attitudes, knowledge, skills, autonomy, responsibility, meaning competences have to be improved during their studies.

New pedagogical approaches, new teaching methods are not so efficient if the labor market needs different training scope than an educational institute can provide. Future needs have to be predicated not just detected, because the approval or accreditation processes are long-term process in the world of education.

From a long-term research point of view, it is more interesting the future demand compliance with supply, because this approach fits better to regional development, managing links between academia and economy. Several directives, surveys, studies try to envisage the future skill demand, just to mention a recent EU wide initiative: S3 strategy [1]. In addition, demography can be mentioned – the growing silver economy, the economic growth, investments create new jobs, and the development of technology, a.o. ICT. According to a popular and scientifically less grounded topos many foresees the AI ‘hostile’ takeover of labor market. (There is life after AI winter.)

One possible solution is to model the economy and try to conclude the future workforce demand. Doing this via well-known and approved input-output analysis model and using standardized statistical data, mainly from open sources (CSO, Eurostat open data), the model mirrors with good accuracy the economy sector-wise. The overall output (GDP), and labor intensity are important results for the next phase of modelling. Labor usage is presented in occupational distribution. In the next step, there are three main tasks to be performed: (1) translate the occupational structure into competence structure, (2) using the model create several future scenario in order to generate the future demand; (3) match the demand with the supply.

The overall purpose of modelling and analysis is to conclude what skills (competences) will be required by the world of labor on the selected time horizon, in region, sector, in selected job roles and/or occupations. After having learnt the requested skills (competences), the supply can be compared with the demand. In our investigation, we narrowed the supply to the fresh graduates. In reality, the supply is bigger if we take into account the number of potential workforce having in mind career exchange, re-integrated unemployed people and the mobility. On one hand, this the bias of the analysis results, on other hand it is very difficult to get realistic and exact figures related to the mentioned subgroups. In order to conduct the analysis several steps need to be performed.

This paper presents a system met these above-mentioned requirements. Section 1 deals with other research approaches to predict future competences and their relation with our solution. Having clarified the related main concepts in Sects. 2–5, a systematic analysis performed with this system is presented in Sect. 6. Conclusions about limitations and future work are drawn in Sect. 7.

1.1 State of the Art

Different studies deal with forecasting future competences. Systematic analysis is our main goal, so Approaches using models or other methods for synthesizing knowledge sources fall into our scope. Hence two research groups and two international institutions (OECD and CEDEFOP) specialized on this topic were selected to present their approaches. Their main characteristics are collected in the following table.

- Authors identify the research approach unequivocally.
- Scope presents that this a general forecast or focuses on a specialized area.

- Input shows which sources were used to build a model or determining predictions.
- Method is the key element of a research, because it reflects the reality in a restricted manner.
- Flexibility means that this research is capable of evaluating the small changes of factors.

Authors	Scope	Inputs	Methods	Flexibility
Hartmann and Bovenschulte [2]	Skill needs prognosis specifically on Industry 4.0	Experts' opinions	Virtual Technology Roadmap, Organizational scenarios, Quantitative and qualitative skill needs analysis	Roadmap has to be renewed
Institute for the Future for the University of Phoenix Research Institute [3]	Six key drivers and ten most relevant skill areas	Experts's foresights collected during a workshop	IFTF's signals methodology	No. There are not quantitative correlations between the factors
OECD report [4]	Mainly ICT skills	National databases	Skills strategy for managing national skill systems Statistical analysis	Possible, but mainly analyzing at national, not skill level
CEDEFOP's European skills and job survey [5, 6]	Skill gap analysis	Questionnaires filled by 48 676 adult employees in the 28 EU Member States	Statistical analysis	Analysis at general skill levels

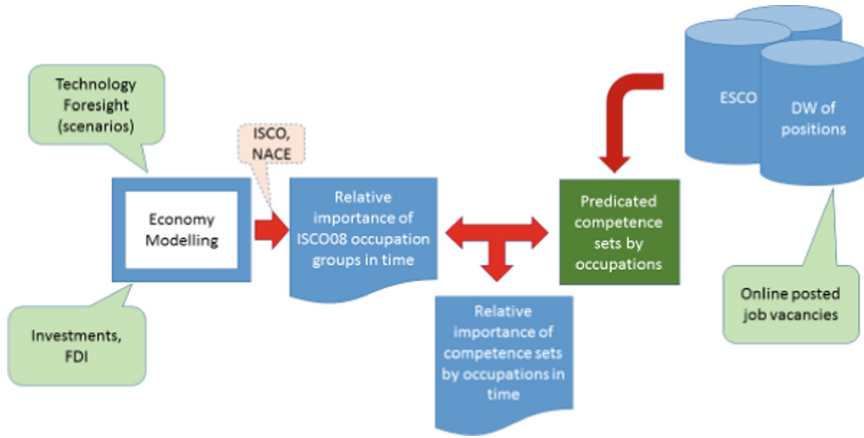
These studies highlighted future competences based on the experts' actual opinions about the future or statistical analyses. The first two studies contained mostly qualitative analyses, diminishing the chance to examine the role of different influencing factors. However, they provide us new insights with defining new skills. The last two studies used quantitative analyses based on past data, so the changes of influencing factors can be examined within these databases. Nevertheless, the current skill categories served as a basis for these analyses.

The goal of this work is to develop a system for revealing the relative importance of different competences in the future. Hence, it exploits the result of the previous studies, but it complements them as well. The qualitative studies must be repeated in order to detect the changes in the environment. The quantitative studies have no interest to modify their categories, because trend analysis requires strict meta-data catalogues to discover patterns or correlations along time dimensions. These approaches can complement each other and information system can synthesize their advantages. It is capable of monitoring current situations, collecting open data as past data, performing

forecasts based on them, presenting the results of different business scenarios due to the changeable parameters, so it is dynamically changeable but the other ones are not.

1.2 Beyond the State of the Art

These studies use models to collect the experts' current thoughts about the future. They do not capitalize the benefits of information systems which are capable of formalizing experts' thinking processes, opinions to create models, continually detecting the actual situations and mixing these outcomes to get a more precise estimation.



Different factors can influence the relative importance of future competences. Economic factors - like GDP, imports and exports can, sectoral outputs etc. – and technological innovations – like smart systems, blockchain etc. - can influence the labor intensity of different sectors. This takes effect on occupations as well. The relative importance of occupations belonging to a more labor-intensive sector will increase. But competences are required to execute different tasks of these occupations. Therefore, the relative importance of these skills will grow. Input-output model developed by Leontief suits this problem, because it is widely used to model industrial outputs, cross-consumptions and labor intensity. Input data of this model can be downloaded due to the Open Government Data initiative. Sectors categorized by NACE represents the industries in this model. Occupations classified by ISCO-08 are connected to these NACE sectors and their distributions are published as open data. The relative importance of different ISCO-08 groups is determined on this way.

Unfortunately, the distributions of ISCO-08 subgroups per sector are not available therefore, another source have to be used to estimate them.

The aim of SMART project (2012-1-ES1-LEO05-49395) [7] was to monitor actual competence needs via online posted job vacancies and compare them with the competences acquired during VET studies in the tourism industry. The SMART+ project (2016-1-ES01-KA202-025304) [8] as its successor is to provide a comprehensive system for detecting skill gap for students and for institutions as well. Besides a comprehensive skill mismatch report, it provides a report concerned on just some

positions. This system is capable of downloading and categorizing job vacancies by position, company, time and regional dimension, and extracting, storing competences required by the position. Another but connected research deals with creating a data warehouse from these datasets in order to monitor and analyze skill mismatches anytime and anywhere. At the end of the development process, monthly data will reside in this data warehouse (DW). ESCO [9] helps to connect the stored positions to occupations and it presents competences per occupation, which can be used to extend the above-mentioned extracted competence sets with new ones. The predicated competence sets per occupation will be calculated within the DW. Their relative importance means that we can state that a competence will be more important than another one, but we cannot state how many times it is. This relative importance will be calculated based on the predicated competence sets per occupation and the aggregated relative importance of the occupations requiring these competences within the ISCO-08. We assumed that the competence can inherit the importance of an occupation, because it is needed to execute a task and if more this kind of position is published then more this kind of task and its competences are required.

Main concepts and areas are clarified in the following sections.

2 Competence

One of the objectives of European Union long term strategy (EU2020) is to create the Single European Market (SEM). In the context of SEM, the macro-regional mobility highlights the questions of free movement of manpower. One aspect of the free movement of manpower is the workforce knowledge, skill and attitude – competences in short - compliance from the perspective of the different Member States. Member States are not only geographically different, but differ significantly in terms of work culture. On corporate level, in the context of employees and employers it is a vital interest to get evidence of existing competences of the new employees, or having a timely monitored competence evolution of the staff.

There is a big terminology bonanza in this area. We are talking about competency and competences, knowledge, skill and attitude, autonomy and responsibility, job role, position and occupation, match and mismatch skills, job seekers, free movements of manpower, migrants and refugees, mobility, career development, employability and unemployment, labor market integration and re-integration ... the list is endless.

Competences has many interpretations in the literature and academia, especially in accreditation processes. In this paper, in our understanding *competence* is a mix of knowledge, hard/soft and transversal skills, attitude, autonomy, responsibility. This interpretation is often mentioned as *employability skills*. On micro level, required competences are linked to the job to be performed. From modelling point of view, jobs emphasized in tasks, and tasks are organized into business processes. Job roles, and their descriptions (in the sense what performers need to know and be able to perform the task) are part of one or more positions, connecting this way process and organizational views together.

While employers, corporates look for competences in connection with positions and connected job roles, on macro level the competences grouped and associated with

occupational structure. An occupation is an element of the statistical nomenclature, positions and occupations are strictly linked together by administrative means. The mutual assignment is not without problems: the competence structure mirrors the *present* and *future* demand of industry, service sectors, and changes dynamically, however occupational structure is a rather rigid, follows a statistical nomenclature and changes are much less frequent. The occupational structure depends on macro structural variables, educational and training systems.

3 Stakeholders/Interested Parties

Competence matching is a wider issue, than many of us would have thought. Many types of stakeholders are interested in it, from different perspectives.

On supply side, first of all employers, owners and managers of individual firms play the most significant customer roles. In addition, different (traditional and online) manpower services either from recruiting or selection purposes are significant stakeholders, but many professional associations offers also guidelines, conduct surveys in this direction. Finally, education sector, as one of the largest suppliers of graduates must know how effective they are.

On demand side, we must mention graduates, employees with mobility or career exchange motivation, unemployed to be re-integrated, any job seeker, in general.

4 Supply/Demand

4.1 Supply

Available workforce may come from several sources. Academia, education sector is one of the main source, the output of the sector is well planned and it is relatively easy to forecast. From available or potential competences, we must take into consideration other forms of education/training, like informal, non-formal education, on the job-training, continuous education. From motivational point of view, the career exchange is an important driver, resulting extra job seekers with new or changed competences. Another source is the internal or macro-regional mobility, especially within EU, although this is not without any problem (cf. Brexit). We may mention also the re-integration of the unemployed people. Finally, the migration plays important role, which is a more complex problem, than just a manpower issue.

4.2 Demand

Demand is also complex. Employers are quite different regarding their expectations, in the first group there are employers who request well-trained workforce whom they can use on the next day with full capacity, while another group of employers demand intelligent, ‘smart’ people, who are able and know how to learn, while training remain the employers’ responsibility. The scale between the two groups is long and colorful.

We need to highlight three specific drivers, which influences the demand: replacement due to the mobility, additional workforce due to economic growth and workforce with different competences due to the technological development.

5 I/O Model

5.1 Model Selection: Why Input-Output Model?

The Leontief model, known as I/O analysis emphasizes the effects of change in the final demands for goods and services on particular industry with respect to its sales and purchases.

“The input-output method is an adaptation of the neoclassical theory of general equilibrium to the empirical study of the quantitative interdependence between inter-related economic activities. It was originally developed to analyze and measure the connections between the various producing and consuming sectors within a national economy... The specific structural characteristics of the system are thus determined by the coefficients of these equations. These coefficients must be determined empirically; in the analysis of the structural characteristics of an entire national economy, they are usually derived from statistical input-output tables” [10].

Leontief gave an extended interpretation to the coefficients the most important among them were the following: the coefficients have a statistical character, therefore they can be estimated; different coefficients based on the estimation statistically are quite stable, hence the model is suitable for different kinds of analyses, like assumption of different economic growth, changes in industry structure, etc.; the analyses may lead to quantitative evaluation of different economic policies, comparison of their indirect effects, accelerator effects or counter effects.

Later in the light of quite different economic theories and also due to the radical changes of global economy, many argued that I/O model has not reflect anymore objectively to the real life of national economy. However, the coefficients – in other word technological matrix – are based on statistical data and still the most reliable although not the only data source for modelling. Adding to the model the capital investments, taking into account the modification effects of export-import activities, the I/O model still gives a good starting point for analysis. On the basis of the model results environmental effects, ecological considerations, the strengthening of third sector will be more understandable and lead to a more complex approach [11].

From the perspective of this paper, we must emphasize the biggest advantages of using I/O model that is it can be built on official statistical data. The validity of data ensured by the national and macro-regional statistical data collection systems, and most of them is available in ‘open data’ format. The later mentioned feature means, the experiments; analyses can be reproduced in an automated fashion.

5.2 Data Sources

Open Data/OGD/GOD

Central Statistical Office (CSO), Hungary. The data on persons employed in the Hungarian economy, according to NACE Rev. 2 (2008)¹ classification, is published by the Hungarian Central Statistical Office on its website together with other information of the national accounts. The information on persons relating to national account is important mainly for the calculation of the ratio, value added per worker in each industry of the national economy.

Eurostat. EUROSTAT provides industry-by-industry symmetric input-output tables. Hungarian data in ESA 1995 format was used from their data source. The output matrix is an object as it is structured by industry. This organization presents supply and use tables and symmetric input-output tables that are a fundamental part of the European System of Accounts (ESA 1995).

OECD. The Organization for Economic Co-operation and Development (OECD) is involved for preparing Inter-Country Input-Output (ICIO) tables which based on different International Standard Industrial Classification of all economic activities (ISIC) revised version. The previous OECD national Input-Output tables present matrices of inter-industrial flows of goods and services (produced domestically and imported in current prices (USD million), for all OECD countries including 28 members of European Union and G20 economies, covering the years 1995 to 2011 based on the ISIC Revision 2.

The latest version of ICIO tables are based on ISIC revision version 3. The better integration with collections of statistics accumulated according to industrial activity such as research and development expenditure, employment, foreign direct investment and energy consumption. The OECD I/OT database is a very useful experiential tool for economic research and structural analysis at the international level as it highlights inter-industrial relationships covering all sectors of the economy.

World Input-Output Database (WIOD). The World Input-Output Database (WIOD) is the first public database that contains new information on the nature of international trade and trends and provides the opportunity to analyze the consequences of division for shifting patterns in demand for skills in labor markets. These tables have been put up in a clear conceptual framework on the basis of officially published input-output tables in concurrence with national accounts and international trade statistics. In addition, the WIOD provides data on labor and capital inputs at industry level.

Concluding to Demand

As it was outlined in the previous sections, the I/O model is used to do predictions on the changes of occupational structure, due to the economic growth, changes in

¹ NACE is the statistical classification of economic activities in the European Community which imposes the job classification uniformly within all the member states of European Union. NACE Rev. 2 reflects the technological developments and structural changes of the economy, enabling the modernization of the community statistics and contributing, through more comparable and relevant data at both community and national level.

productivity and the expected technological developments. In order to get the results, the gross domestic output per sector, used labor force/output unit, distribution of labor force/sector/occupation will be used as variables. With the help of ESCO ontology and database, we conclude to the expected competences/occupation. Difference between the expected and supplied competencies already provides sound basis of portfolio decisions.

6 Comparison (Matching) Competences

6.1 Requested Competences/Occupation

The first problem is how to conclude from the I/O model to the requested skills/competence structure. As we saw in the previous sections, I/O model, the coefficients results output per sectors, labor intensity and quantity per occupations, as they are classified by NACE. The granularity of occupations is very rough. We increase the specificity of occupations combining the occupations with sector. At this point, the ESCO database can be of great help, and with the help of ESCO the requested competences can be better positioned per job roles.

6.2 Available (Provided) Competences/Occupation - EQF/National QF

The other – supply – side of competences seems to be easier issue. The EU accepted the European Qualification Framework (EQF) [12], similar frameworks exist elsewhere, too. The EQF is very general, the national qualification frameworks are more detailed and localized, and on institutional level, we can get very precise and concrete competence (often called as learning outcome) lists – at least what institutions claim to provide. For the sake of simplicity less assume, graduates have those competences.

The comparison of supply and demand will undertake by using SMART and SMART+ system^{2,3}.

6.3 Analysis: ‘What If’ Scenarios

Time Horizon/Preference Selection. The next problem is to select the suitable time horizon. The lower limit will be the minimum time, during which any change is becoming ‘visible’ on the national accounts, become manifest statistically. Theoretically, there is no upper limit, only limitation how far we can see in the future to keep the possible scenarios still realistic. If we consider the ‘lead time’ of a typical higher education institution, the most appropriate time horizon is between 3–5 years.

The Growing Economy (Foreign Direct Investment). In general, economy may grow due to several reasons, increasing productivity, growing export, growing domestic demand, large volume of international aid (e.g. EU Structural and Cohesion

² SMART - Supporting dynamic MAtching for Regional development, 2012-1-ES1-LEO05-49395.

³ SMART PLUS 2016-1-ES01-KA202-025304.

Funds), etc. A special case is the Foreign Direct Investment, the investor is not selling or buying something but creates production sites, jobs, and the economic growth is based both on the direct investment and the additional gross domestic products due to the accelerator effect⁴ [13]. I/O model is suitable to reflect both (direct and indirect) effect. In the first scenario we assume, the recently experienced fast development of electric car manufacturing will effect on the volume and structure of FDI. Retail sale of fossil energy is expected decreasing, energy sector as a whole need to be restructured, more investments, and technological development is needed, while the agro-based renewable energy production will decrease.

As a result, the inter-relations of the sectors will change, the overall domestic output will be increased. As one of the consequence, the labor-part of the GDP will change as well, both in terms of quantity and occupation structure, hence the change requests different skill-set and follows the changes in occupational structure (for the sake of simplicity linearity is assumed, which results some bias).

Changing the Requested Labor Force (Productivity, Unchanged Structure). In the EU and G20 countries, a sound increase of productivity is monitored [14]. The annual average is around 20%, with very big differences among the countries. The largest improvement happened in Ireland and 135% productivity index is expected by 2018 (2010 = 100). From the point of occupational structure and demanded skill set point of view we may model the quantitative changes through the labor coefficients. The question is, in what extent will follow the occupational structure the increasing productivity (the assumption is the less-skilled workers' ratio will decrease).

Changing the Requested Labor Force (Technology, Different Structure). The new technological phenomenon, the fast growing ratio of electric cars among the vehicles. Norway e.g. expects electric or hybrid cars make up half of new vehicle registrations in 2017 [15]. E-car manufacturing needs significantly less skilled jobs in the traditional machinery sectors, although less but better trained and skilled workers in the designer and constructing parts of car manufacturing. These changes will lead to the changes in the occupational structure.

On the other hand, the forecasted technology breakthrough will affect not only the manufacturing sector but significant changes are expected in the energy sector, too. Electric Vehicles (EVs) promise technology for reducing the environmental burden of road transport. Other energy types like renewable energy production provides the largest market swing over time: from 19% of production in 2010, 32% is expected in 2020 and will continuously grows up to 50% by 2050 [16].

Technology plays a vital role for changing labor market trends. Several industry studies in equal extent say that fully autonomous vehicles are to be commercially available before 2020. In 2030, the share of electrified vehicles could range from 10% to 50% of new-vehicle sales.

As a consequence the technological impact on the demand side of labor market implies structural changes of required competencies. Educational institutes need time to

⁴ However in 2016 the FDI fall 7% according to the OECD data, on a longer term increase is expected [13].



Fig. 1. Number and distribution of managerial position by sectors

change their educational portfolio due to the lead-time of formal education. A system, which is capable to predict the future occupational structure and conclude to the required competencies, can facilitate decision making processes both in the educational institutes and in the world of labor.

A business scenario reflecting the influence of growing number of electric car was used to present the working of this system. Figure 1 shows the expected changes in terms growth of output, improving productivity and creating new (skilled) jobs will change the localization of the managerial positions. The result is almost double the demand for skilled workers, and the relative need is bigger in the productive sectors (first economy) than in the second and third.

Nowadays ESCO collected the main managerial aspect like “plan, organize, coordinate, control and direct the work done by others.” Figure 1 presents that managerial jobs and related competences will be important in the future. There are managerial competences specialized in a given sector e.g. monitoring fields and managing agricultural staff by a crop production manager or monitoring technological trends and managing contracts by an ICT production manager and there are general competences related to different managerial job roles [9]. Figure 2 presents how these general competences are distributed among 19 selected occupations. It shows that the “adhere to organisational guidelines” competence is required by seven different managerial occupations.

The system presented in this paper is capable of estimating the importance of occupations affected by changes in industrial structures and the importance of specific and general competences belonging to these occupations.

The selected most important competences can be compared with the competences provided by a training program via Smart+ system. This gives very clear indication to the education sector how and in which direction develop their portfolio.

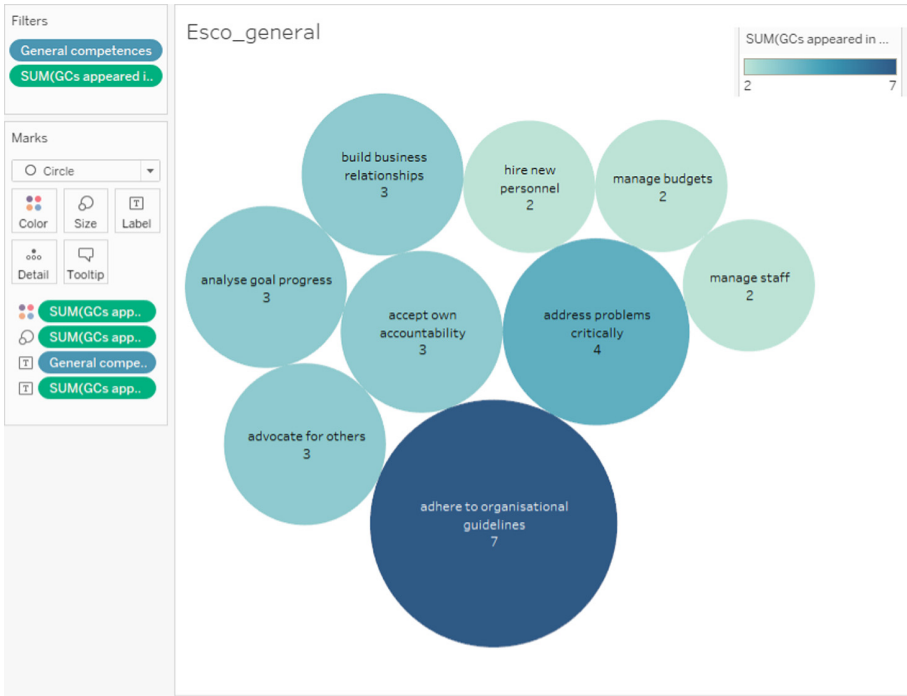


Fig. 2. The distribution of general competences

To execute deeper analysis along the hierarchy of ISCO occupations, the distributions of these occupations per main group are needed. If publishing governmental data on Internet will be a best practice, it solves this problem.

7 Discussion, Conclusion

It is difficult to validate that our prediction is appropriate or not, because different factors are counted in by different research groups. The goal of this paper is not to predict the future. Its goal is to present a conception about developing an information system in order to formalize experts' thoughts via business scenarios by input-output model and check the influence of labor-intensive sectors on related occupations and competences.

The development of this system is ongoing. The concept of Open Government Initiative (OGI) has to be adapted widely in order to get relevant databases from every year in order to make this system capable of monitoring actual open data and modifying results based on them, meaning to make it dynamically adaptive to the variant environment. The distributions of occupations per ISCO-08 main group, per industry, per country are missing. OGI can solve this problem or a data warehouse that can provide statistical analysis based on past data as we plan.

The novelty of this system is to collect input-output and labor market data dynamically from Internet, if these data are available, apply directly input-output model to these data, provide opportunity to test future scenarios about the effects of technological changing on the labor market and determine competence sets for further analysis. It helps to redesign competence sets provided by training programs in the light of business scenarios derived from the business climate of Industry 4.0.

Acknowledgement. cLINK (Centre of excellence for Learning, Innovation, Networking and Knowledge) ERUSMUS MUNDUS Project, Ref: 372242-1-2012-1-UK-ERA MUNDUS-EMA21.

References

1. Home - Smart Specialisation Platform. <http://s3platform.jrc.ec.europa.eu/>. Accessed 28 Aug 2017
2. Hartmann, E.A., Bovenschulte, M.: Skills needs analysis for “Industry 4.0” based on roadmaps for smart systems. In: Using Technology Foresights for Identifying Future Skills Needs. Global Workshop Proceedings, pp. 24–36 (2013)
3. Davies, A., et al.: Future work skills 2020. Institute for the Future for the University of Phoenix Research Center (2011)
4. OECD: Skill for a Digital World, <http://www.oecd.org/els/emp/Skills-for-a-Digital-World.pdf>. Accessed 21 Aug 2017
5. Cedefop’s European skills and jobs survey data released! <http://www.cedefop.europa.eu/en/news-and-press/news/cedefops-european-skills-and-jobs-survey-data-released>. Accessed 21 Aug 2017
6. Briefing note - People, machines, robots and skills. <http://www.cedefop.europa.eu/en/publications-and-resources/publications/9121>. Accessed 21 Aug 2017
7. Smart-Project. <http://www.smart-project.org/>. Accessed 28 Aug 2017
8. Smart Plus Project – Official Website. <http://smartplus-project.org/>. Accessed 28 Aug 2017
9. ESCO - European Commission. <https://ec.europa.eu/esco/portal/home/>. Accessed 28 Aug 2017
10. Encyclopedia.com, Input–Output Analysis - Dictionary definition of Input–Output Analysis | Encyclopedia.com: FREE online dictionary. <http://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/input-output-analysis/>. Accessed 02 Aug 2017
11. Stilwell, L., Minnitt, R.: Input-output analysis: its potential application to the mining industry. *J. S. Afr. Inst. Min. Metall.*, 455–460 (2000)
12. European qualifications framework (EQF). <http://www.cedefop.europa.eu/hu/events-and-projects/projects/european-qualifications-framework-eqf/>. Accessed 02 Aug 2017
13. Foreign Direct Investment Statistics: Data, Analysis and Forecasts – OECD. <http://www.oecd.org/corporate/mne/statistics.htm/>. Accessed 17 Aug 2017
14. Level of GDP per capita and productivity. http://stats.oecd.org/Index.aspx?DataSetCode=PDB_LV. Accessed 18 Aug 2017
15. Half of all new cars in Norway are electric or hybrid | World Economic Forum. <https://www.weforum.org/agenda/2017/03/norway-is-leading-the-charge-towards-electric-vehicles-and-just-hit-another-milestone-along-the-way-d69a8170-cbdc-4d8a-95cd-f9bdf3c8e3ae/>. Accessed 18 Aug 2017
16. van Essen, H., Kaupman, B.: Impacts of electric vehicles: summary report (2011)
17. Lu, Y.: Industry 4.0: a survey on technologies, applications and open research issues. *J. Ind. Inf. Integr.* **6**, 1–10 (2017). <https://doi.org/10.1016/j.jii.2017.04.005>



Process-Based Analysis of Digitally Transforming Skills

Ildikó Szabó^(✉) and Katalin Ternai

Corvinus University of Budapest, Fővám tér 13-15., Budapest 1093, Hungary
{ildiko.szabo2,katalin.ternai}@uni-corvinus.hu

Abstract. In Industry 4.0 a lot of jobs will be replaced by machines due to the technological revolution. Digital transformation entails new skills required to possess by people. This paper presents a solution to create data warehouse to assess future job skills based on the actual industrial business processes. The solution collects time series data from job portals and transforms them into the data warehouse to analyse skill sets. The structure of the data warehouse and the algorithm of extracting data from job vacancies have been introduced.

Keywords: Digital transformation · Business process · Process ontology
Skilled workforce

1 Introduction

The essential goal of Industry 4.0 is to push the manufacturing sector to its next transformation - in order to become more competitive, need to embrace emerging technologies, such as advanced analytics and cyber-physical system-based approaches to improve the efficiency and productivity. The key benefits of Industry 4.0 are to make business processes faster, more efficient and more customer-centric, while at the same time going beyond automation and optimization and detect new business opportunities and models. The human and social dimension are general in Industry 4.0. Tasks and demand for employees may change, however, different skills will be required [1].

New prophecies speak about jobs will be replaced by machines due to this technological revolution. As we learned from the prior industrial revolutions, job roles are transformed into new ones which require new or improved competences to execute their related activities.

1.1 Skill Relevance of Digitally Transformed Business Processes

OECD and CEDEFOP have been investigating skills of future work since many years. OECD report published in 2016 emphasized the importance of ICT foundation skills to get better jobs and high wages by jobseekers. Inequality in possessing these skills is experienced among different age groups. Young adults are more prepared for the digital economy than the older ones. Jobs requiring more intensive ICT entail needs for other skills like “a solid level of information processing skills as well as the ability to collaborate, share information, give presentations, provide advice, work autonomously, manage, influence and solve problems” and social skills [2]. CEDEFOP survey also

states that “at least moderate-level ICT skills also require a strong level of complementary skills, such as foundation skills (literacy, numeracy), soft skills (planning and organization) and behavioral skills (communication and teamwork). Jobs requiring advanced ICT skills depend heavily on people being able to solve problems, learn, adapt, apply new methods and technologies as well as in-depth technical knowledge”. Highly risked groups for having digital skill gap are women, older-aged and lower-educated workers. Professionals in high skill intensive jobs must update their skill to avoid this gap [3]. The research conducted by the Institute for the Future for the University of Phoenix Research Institute revealed ten new skills based on six key drivers of change: sense-making, transdisciplinary, novel and adaptive thinking, social intelligence, new media literacy, computational thinking, cognitive load management, design mindset, cross cultural competency and virtual collaboration [4]. Organizational culture in the environment as Industry 4.0 must be a facilitator of learning and create behaviors in jobs. This requires changes in management practices to provide an appropriate organizational climate. Managerial skills can be derived from these management approaches: making organizational structure more adaptive, knowledge oriented leadership, HR practices or the cut-off of traditional investment style [5].

1.2 Approaches for Analyzing Skill Needs

The studies conducted from macro perspective emphasizes the importance of ICT and related hard/soft, transversal skills in the future. But the organizational perspective can discover new areas like managerial practices within a learning and innovative organizational climate. Nevertheless, the comprehensive studies highlight the role of information systems to continually monitor and predicate skill needs and the commitment of educational institutions to integrate them into their training curricula and programs [2, 3].

Studies, projects, frameworks, information systems built on key indicators or data tables deal with the analysis of skill needs. The New Skill for New Jobs initiative “is intended to promote an improvement in skills forecasting and matching the supply of skills to the needs of the labor market through better cooperation between the worlds of work and education” [6]. The New Skills Agenda for Europe determines action points to equip people with new skills, transferring information to facilitate their job seeking and to improve their life chance. It is planned to provide upskilling pathways for adults, review European Qualification Framework, strength the co-operation among education, employment and industry stakeholders, develop a new framework for strategic cooperation between stakeholder to bridge short and medium-term skill gap to support the sectoral strategy [7].

More and more governmental data and other relevant data sources are available on the Internet due to the Linked Open Data, Open Data and Open Government Initiative. Skills Panorama created by the European Commission and developing by Cedefop “is a central access point for data, information and intelligence on skill needs in occupations and sectors that provides a European perspective on trends in skill supply and demand and possible skill mismatches, while also giving access to national data and sources.” Research studies are provided about employment and future skills per country, per industry, per occupation [8]. Skills related statistics are available on the Eurostat portal.

Employment, labour market supply and demand can be queried based on time series data along different dimensions (industry, occupation etc.) [9]. ESCO (European Skills/Competences, Qualifications and Occupations) connects these three pillars of labour market and classifies their concepts into a comprehensive, ontology structure which will be queried semantically a SPARQL interface, following the Linked Open Data initiative [10]. These data sources or analytical tools [8, 9] applied mainly the macro perspective and not the organizational perspective, because skill or competence sets are not analysed per industry or per occupation. ESCO ensures the bridge to connect analytical results per occupation to the skills of these occupations. Time series data of these competence sets are required to detect patterns or trends in this context. Job vacancies posted on popular job portals can provide these time series data and the contextual information about industry, occupations/position, activities, region etc. as well. During this research, a data warehouse is building from these information to provide further analysis of skill sets to mix macro and organizational perspective as well. Section 2 presents the related theoretical background of this research. The structure of this data warehouse and the algorithm of extracting these data from job vacancies are introduced in Sect. 3. Its implementation and test run are showed in Sect. 4. Conclusions about future work are drawn in Sect. 5.

2 Theoretical Background

2.1 Semantic Business Process Modelling, Process Ontology

Ontologies as well as semantic web have key role in semantic business process management (SBPM) [11]. Truly the semantic web relies heavily on ontologies to structure data for comprehensive and transportable machine understanding. Ontologies are an integral part of semantic web in facilitating knowledge sharing and reuse. In the field of SBPM, ontologies can be utilized for knowledge representation, knowledge engineering, information modelling, database development and integration, information retrieval and extraction, knowledge management and mining, and agent-based system development. They are used to share technical and business information throughout an organization or even in extended or virtual enterprises. An ontology is used as a mechanism for expressing and sharing enterprise knowledge to support intelligent queries.

Among the wide spectrum of approaches top level, domain and task level ontologies can be distinguished.

- Top-level ontologies - are used to represent the building blocks for a particular domain.
- Domain level ontologies - are used to represent vocabulary related to certain domains.
- Task level ontologies - are used to represent vocabulary related to certain tasks.

A top-level ontology consists of very general terms such as “object”, “property”, “relation” that are common across domains. These ontologies are the first step towards knowledge representation for any domain. Terms in the domain ontology are ranked

“under” the terms in the upper ontology and stand in subclass relations. Zhou and Kuntz [12] have developed a top-level ontology for representing the fundamental terms of a company, i.e. product, organization, activity, actor, facility, method and value. The framework consists of generic entities for providing products and services using material flow, information flow, and cost flow. Gialelis et al. [13] presented an ontological model for applications and systems that participate in collaborative processes. They have proposed an architecture that combines web services, ontologies, and workflows for efficiently carrying out the integration process under a collaborative environment. Here the integration is carried out by describing the enterprise processes by means of workflows and by using web services as a channel of communication within an enterprise. They test case was a procurement process carried out by suppliers for ordering materials, spare parts, tools, etc. Task level ontologies are used to model the entities performing their respective operations and support interaction and interoperability among them. Generally, these ontologies are implemented in the field of multi-agent systems to model the tasks of various agent systems. Merdan et al. [14] presented a knowledge-intensive multi-agent architecture that enabled ontology-based communication and cooperation among a set of autonomous and heterogeneous agents.

Several enterprise information models are present in literature that describe the structure and relationships of data and information elements within enterprise information systems, such as CIMOSA, MOSES, MISSION, FDM [14–17]. These models have been developed for intra-enterprise integration, while some generic ontology models have been proposed for enterprise integration such as Toronto Virtual Enterprise (TOVE), GRAI, GERAM, ARIS, Enterprise ontology, PSL, Electronic business using eXtensible Markup Language (ebXML), Business Process Modelling language (BPML), etc. [18, 19].

The Enterprise Ontology was developed under the Enterprise project to provide a collection of terms and definitions relevant to business enterprises [20]. The Enterprise Ontology was developed as a generic model oriented towards business and organizations. TOVE has been developed to support enterprise integration and communication [21]. TOVE provides a generic and reusable ontology for modelling enterprises.

BPML represents a process definition language intended for expressing abstract and executable processes that address all aspects of enterprise business processes [22]. It provides an abstract execution model for collaborative and transactional business processes based on the concept of a transactional finite state machine. ebXML is a family of XML standards sponsored by OASIS and UN/CEFACT [23]. It enables enterprises to conduct business over the internet.

Ontologies, as general but formalized representation can be used for describing the concepts of a business process. According to our research, process ontologies have no precise definition in the academic literature. Some refer to it simply as a conceptual description framework of processes [24]. In this interpretation process ontologies are abstract and general. Contrary, task ontologies determine a smaller subset of the process space, the sequence of activities in a given process [25].

The process ontology is used to reconcile the heterogeneous semantics of process modeling constructs, i.e. meta-model semantics existing in different process modeling languages. It indicates that process ontology should include a set of meta-concepts that are able to describe the semantics of process models.

In this paper the concept of process ontologies is used, where ontology holds the structural information of processes. The solution of establish the links between process model elements and ontology concepts has been prepared in the methodology. The attempt is to provide an extension for the standard ontology definition in the form of an annotation scheme to enable ontologies to cover all the major aspects of business process definition. The approach is identified as a semi-automatic generation of BPM defined ontology.

The objective of ontology learning is “to generate domain ontologies from various kinds of resources by applying natural language processing and machine learning techniques” [26]. Statistical, rule-based or hybrid ontology learning technique are distinguished based on the technique to detect correlations. The pattern-based ontology learning technique is one of them [27]. Our solution is a pattern-based ontology learning in that process ontologies transformed from industrial business processes are used as patterns.

3 Labour Market Data Warehouse

A job vacancy posted on a given job portal usually contain the following information:

Its link, position, permanent/part-time job type, region, salary, start date, company name and job descriptions. The descriptions describe the tasks, responsibilities, requirements for fulfilling the given job. These information can be structured into the following tables:

- Calendar based on the publishing date
- Region: Region_ID, Region_name (from the job vacancy), Region_country (from geographical database)
- Industry: Industry_ID (NACE, SIC code etc. from national databases), Industry_name (name of the given classification level), Company name (from the job vacancy)
- Occupation: Occupation ID (ISCO code from ESCO based on the position name), Occupation name (from ESCO)
- Job role: Job role_ID and name (from the actual process ontologies stored in databases)
- Skill: Skill_ID (unique ID), Skill name (from the actual process ontologies stored in databases and from ESCO)

Additional data sources are used to complement these information. Having created this structure different queries can be executed in this data warehouse. E.g. the distribution of skills related a given occupation per industry (Fig. 1).

The positions name, job roles and skills have to be extracted from job descriptions. But sometimes they are not so well-detailed. We need additional information in order to get these information.

Companies work along business processes whose activities must be executed by people filling the published positions. We assume that the activities mentioned in job vacancies are from these business processes.

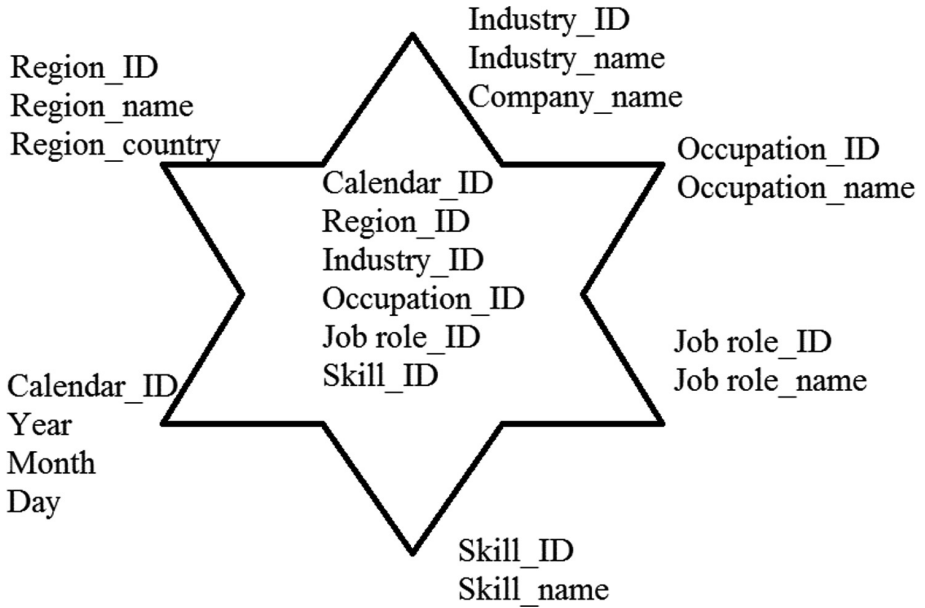


Fig. 1. The structure of this data warehouse

Actual industrial business process models can reflect the organizational perspective and they can be created based on best practices. Skills to execute tasks are also determined by ESCO in these models. A database contains the process ontologies transformed from them (called as reference process ontologies).

We experienced that mainly tasks are presented in job descriptions and not competences. We use these process ontologies to identify the tasks and job roles in the job descriptions, because the relevant skills are identified based on them from the process ontology. The text mining algorithm is detailed in [28].

4 Test

By means of a concrete process from hospitality industry, a job portal called as JobSite from United Kingdom, we show the applicability of the method.

The first step is to determine the structure of a process model (task as process step, role as job role and required skill to execute this task by this role) and transform it into a process ontology.

In the use case the business process models were implemented by using BOC ADONIS modeling platform [29]. Our approach is principally transferable to other semi-formal modeling languages. There are several parameters that can be set or defined when modeling a business process in this tool, and in others as well. The shell of a business process can be easily formed with activities, decision points, parallelism or merging objects, logical gateways and events. The ‘Supplying’ business process model can be seen in (Fig. 2).

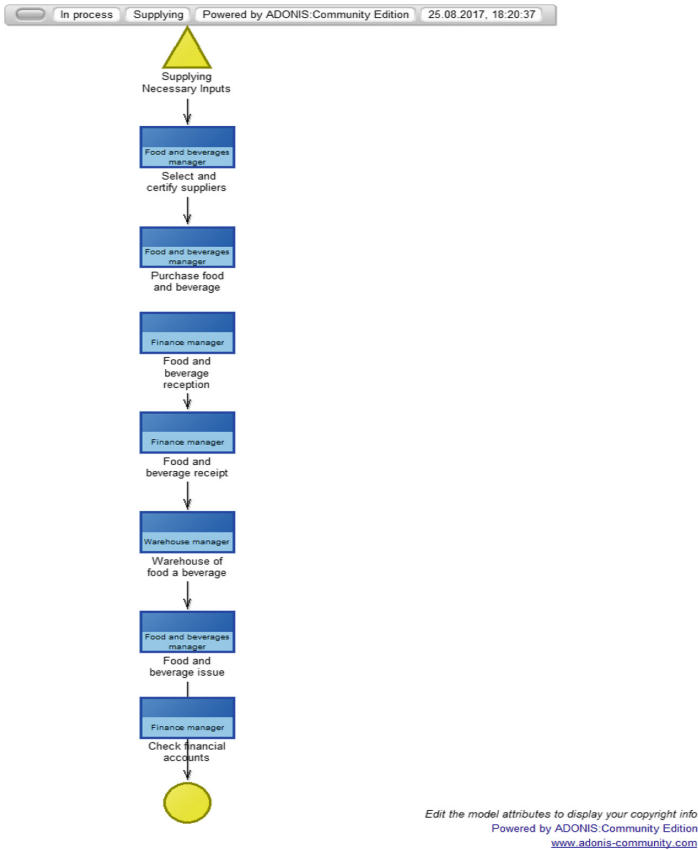


Fig. 2. The ‘Supplying’ process

The activities in the process:

Supplying necessary inputs	
<i>Process start</i>	
Description	Includes providing various kinds of goods and materials for the hotel business as a whole. These elements preferably influence the range and quantity of necessary goods and materials. Consumables character of the material, or the need for its continuous renewal, conditioned by the continuity of the business process. Given that certain types of goods and perishable foods subject, there is no possibility of forming a stock of goods of that kind, or the process requires special conditions for its storage

(continued)

(continued)

<i>Supplying necessary inputs</i>	
<i>Select and certify suppliers</i>	
Description	Selection of the best suppliers; Evaluates and selects suppliers based on their ability to supply product in accordance with the requirements of the hotel
Responsible role	Food and beverages manager
<i>Purchase food and beverage</i>	
Description	Ordering food and drinks; After the ordered food and drinks, with his team performs a set of operations that are related to the acquisition of foods and drinks, such as delivery of merchandise, unloading, the release of the packaging, weighing and storage. Foodstuffs are shipped in appropriate packaging which is an important condition for hygienic and health safety
Responsible role	Food and beverages manager
<i>Food and beverage receipt</i>	
Description	Quantitative and qualitative reception of food and beverage; monitor financial accounts; Receipt of goods from suppliers accompanied by appropriate documentation, conducted by the owner of the process. The acknowledgment of receipt of the goods is evidence that the goods received are reviewed, tested, compared with an order form, and matches all elements in terms of quantity, quality and price
Responsible role	Finance manager
<i>Food and beverage reception</i>	
Description	Quantitative and qualitative reception of food and beverage
Responsible role	Finance manager
<i>Warehouse of food a beverage</i>	
Description	Warehouse of food a beverage; manage stock rotation
Responsible role	Warehouse manager
<i>Food and beverage issue</i>	
Description	Food and beverage issue; Release of the goods production process and serving of food
Responsible role	Food and beverages manager
<i>Check financial accounts</i>	
Description	Monitor financial accounts
Responsible role	Finance manager

To map conceptual models to ontology models we have used meta-modeling approach. Meta-models offer intuitive way of specifying modeling languages and are suitable for discussion with non-technical users. Meta-models are particularly convenient for the definition of conceptual models. In our approach, we establish the links between model elements and ontology concepts. As ontologies provide semantics, they can describe both semantics of the modeling language constructs as well as semantics of model instances [29].

For the mapping the conceptual models to ontology models, the business process models are exported in the structure of ADONIS XML format. All objects from the business process model will be an ‘instance’ in the XML structure, the attributes have the tag ‘attribute’, while the connected objects (such as the performer, or the input/output data, which are stored in another model in the Adonis tool) have the tag ‘interref’.

The “conceptual models - ontology models” converter maps the Adonis Business Process Modeling elements to the appropriate Ontology elements in meta-level. The model transformation aims at preserving the semantics of the business model. The general rule we follow is to express each ADONIS model element as a class in the ontology and its corresponding attributes as attributes of the class. This transformation is done by the means of XSLT script which performs the conversion. The converted OWL ontology in the structure of Protege/OWL XML format is imported into the editor of Protege 4.2 (Fig. 3).

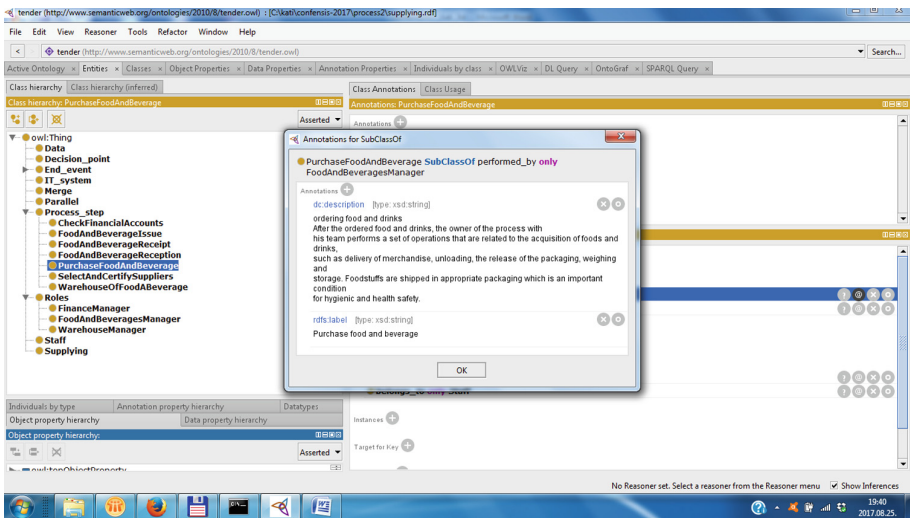


Fig. 3. The process ontology in Protege 5.0

To specify the semantics of ADONIS model elements through relations to ontology concepts, the ADONIS business model first must be represented within the ontology. Regarding the representation of the business model in the ontology, one can differentiate between a representation of ADONIS model language constructs and a representation of ADONIS model elements. ADONIS model language constructs such as “activity”, as well as the control flow are created in the ontology as classes and properties. Subsequently, the ADONIS model elements can be represented through the instantiation of these classes and properties in the ontology. The linkage of the ontology and the ADONIS model element instances is accomplished by the usage of properties. These properties specify the semantics of an ADONIS model element

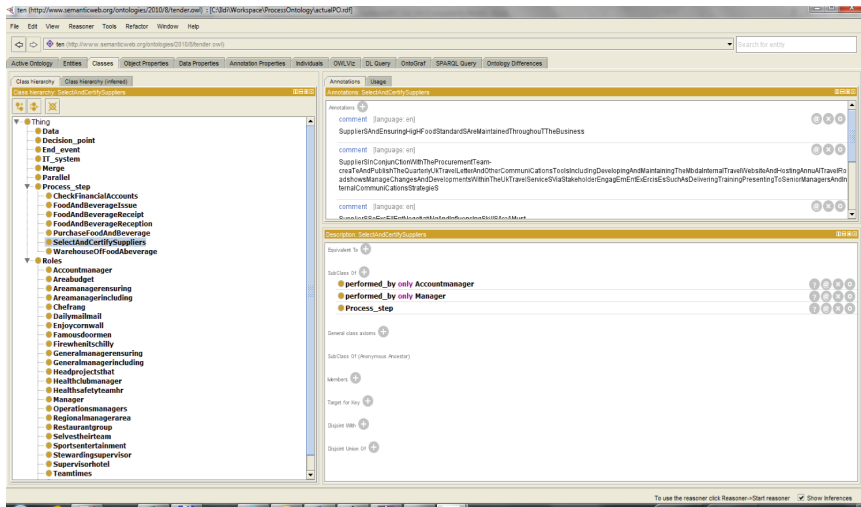


Fig. 4. Results of the extracting algorithm

through a relation to an ontology instance with formal semantics defined by the ontology. The final ontology can be seen in the Protégé editor in Fig. 4.

The second step is to download job vacancies from the job portal and run the above-mentioned algorithm to extract skills. A Java program was written to perform this step. 808 job description were processed by the algorithm. All process steps were identified, but the roles were called by different name in the job description. The accuracy of this algorithm must be tested by a confusion matrix in the future.

5 Conclusions

This paper presented a solution for collecting time series data from job portals and transforming them into a data warehouse to analyse skill sets in a deeply manner than in the case of Skills panorama, or the Eurostat databases. Some difficulties can emerge from data cleansing and identifying skill needs in job descriptions. Specific skills can be appeared in them which have to be extracted with an other algorithm. New scripts of functions are required to overcome these problems.

As future work we have to extend our algorithm with them. We can complement our solution with the results of Cedefop, OECD or the Institute for the Future. The skills mentioned in their report can be stored besides the process ontologies. It is worth to detect them in job descriptions to follow their history in the job vacancies: they are required by which occupation in which industry from when till when.

The occupational trends appeared in the Eurostat data can help to evaluate the future importance of collected skill sets, but this is the topic of an other connecting research.

Acknowledgment. This research was supported by the project nr. EFOP-3.6.2-16-2017-00007, titled *Aspects on the development of intelligent, sustainable and inclusive society: social, technological, innovation networks in employment and digital economy*. The project has been supported by the European Union, co-financed by the European Social Fund and the budget of Hungary.

References


1. Gorecky, D., Schmitt, M., Loskyll, M., Zuhlke, D.: Human-machine-interaction in the Industry 4.0 era. In: 2014 12th IEEE International Conference on Industrial Informatics (INDIN), pp. 289–294 (2014)
2. OECD: Skill for a Digital World. <http://www.oecd.org/els/emp/Skills-for-a-Digital-World.pdf>. Accessed 21 Aug 2017
3. Briefing note - People, machines, robots and skills. <http://www.cedefop.europa.eu/en/publications-and-resources/publications/9121>. Accessed 21 Aug 2017
4. Davies, A., et al.: Future work skills 2020. Institute for the Future for the University of Phoenix Research Center (2011)
5. Shamim, S., et al.: Management approaches for Industry 4.0: a human resource management perspective. In: 2016 IEEE Congress on Evolutionary Computation (CEC), pp. 5309–5316 (2016)
6. New Skills for New Jobs. Policy initiatives in the field of education: short overview of the current situation in Europe. EACEA. http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/125EN.pdf. Accessed 21 Aug 2017
7. New Skills Agenda for Europe - Employment, Social Affairs & Inclusion - European Commission. <http://ec.europa.eu/social/main.jsp?catId=1223>
8. Skills Panorama | Cedefop. <http://www.cedefop.europa.eu/hu/events-and-projects/projects/eu-skills-panorama>. Accessed 28 Aug 2017
9. Eurostat - Data Explorer. <http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do>. Accessed 28 Aug 2017
10. ESCO - ESCOpedia - European Commission. https://ec.europa.eu/esco/portal/escopedia/Main_Page
11. Hepp, M., Roman, D.: An ontology framework for semantic business process management. In: Wirtschaftsinformatik Proceedings 2007, Paper 27 (2007)
12. Zhou, J., Dieng-Kuntz, R.: Manufacturing ontology analysis and design: towards excellent manufacturing. In: Proceedings of the Second IEEE International Conference on Industrial informatics, INDIN 2004, Berlin, Germany, pp. 39–45. IEEE Computer Society (2004)
13. Gialelis, J.V., Kalogeras, A.P., Alexakos, C.E., Papadopoulos, G.: Manufacturing collaborative process integration utilizing state of the art technologies. In: Proceedings of the IEEE International Symposium on Industrial Electronics, ISIE 2005, vol. 4, pp. 1429–1434 (2005)
14. Merdan, M., Kordic, V., Zoitl, A., Lazinica, A.: Knowledge-based multi-agent architecture. In: Proceeding of the International Conference on Computational intelligence for Modelling Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC 2006), Sydney, Australia. IEEE Computer Society, Washington (2006)
15. Canavesio, M.M., Martinez, E.: Enterprise modelling of a project-oriented fractal company for SMEs networking. *Comput. Ind.* **58**, 794–813 (2007)
16. Lin, H.K., Harding, J.A.: A manufacturing system engineering ontology model on the semantic web for inter-enterprise collaboration. *Comput. Ind.* **58**, 428–437 (2007)

17. Fox, M.S., Gruninger, M.: Enterprise modeling. *AI Mag.* **19**, 109–121 (1998)
18. Schlenoff, C., Denno, P., Ivester, R., Libes, D., Szykman, S.: An analysis and approach to using existing ontological systems for applications in manufacturing. *AI EDAM* **14**, 257–270 (2000)
19. Schlenoff, C., Uschold, M.: Knowledge engineering and ontologies for autonomous systems: 2004 AAAI Spring Symposium. *Robot. Auton. Syst.* **49**, 1–5 (2004)
20. Goossenaerts, J.B.M., Pelletier, C.: Ontology and enterprise modelling. In: Eijnatten, F.M. (ed.) *Participative Simulation Environment for Integral Manufacturing Enterprise Renewal*, pp. 41–52. TNO Arbeid, Amsterdam (2002)
21. Fadel, F., Fox, M.S., Gruninger, M.: A generic enterprise resource ontology. In: *Proceedings of the Third Workshop on Enabling Technologies – Infrastructures for Collaborative Enterprises*, pp. 86–92 (1994)
22. Arkin, A.: Business process modeling language (BPML). Working Draft 0.4 (2001). <http://www.bpml.org/>
23. Waldt, D., Drummond, R.: EBXML: the global standard for electronic business. http://www.ebxml.org/presentations/global_standard.htm
24. Herborn, T., Wimmer, M.: Process ontologies facilitating interoperability in eGovernment, a methodological framework. In: *Workshop on Semantics for Business Process Management, The 3rd Semantic Web Conference*. Montenegro (2006)
25. Benjamins, V.R., Fensel, D., Straatman, R.: Assumptions of problem-solving methods and their role in knowledge engineering. In: Wahlster, W. (ed.) *Proceedings of ECAI 1996*, pp. 408–412 (1996)
26. Haase, P., Völker, J.: Ontology learning and reasoning — dealing with uncertainty and inconsistency. In: da Costa, P.C.G., d’Amato, C., et al. (eds.) *URSW 2005–2007. LNCS (LNAI)*, vol. 5327, pp. 366–384. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89765-1_21
27. Zhou, L.: Ontology learning: state of the art and open issues. *Inf. Technol. Manag.* **8**(3), 241–252 (2007)
28. Szabó, I., Ternai, K.: Semantic audit application for analyzing business processes. In: Tjoa, A.M., Xu, L.D., Raffai, M., Novak, N.M. (eds.) *CONFENIS 2016. LNBIP*, vol. 268, pp. 3–15. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49944-4_1
29. BOC Group: Business Process Management with Adonis (2013). <http://www.boc-group.com/products/adonis/>

Big Data Analytics



Big Data Analytics – Geolocation from the Perspective of Mobile Network Operator

Antonin Pavlicek¹ , Petr Doucek¹, Richard Novák^{1,2}, and Vlasta Strizova¹

¹ University of Economics, Prague, Czech Republic
{antonin.pavlicek,doucek, strizova}@vse.cz

² T-Mobile Czech Republic, Prague, Czech Republic
richard.novak@t-mobile.cz

Abstract. The article demonstrates the possibilities of big data analysis of geolocation data of mobile network operator in the Czech Republic. Covers theoretical background of geolocation and then presents case studies conducted in the last four years: National Park visitors' distribution analysis, mountain ski resort usage, use of mobility data for the preparation of city territorial and development plan and use of mobility data for efficient tourism management at Vaclav Havel Airport.

Keywords: Big data · Geolocation · Mobile network operator · Human mobility Tourism

1 Introduction

Modern smart phones have become ubiquitous communications tools—now used not only for phone calls and text messages but also for accessing the Internet, taking pictures and recording videos with integrated camera, navigating with GPS or watching videos and playing games. The proliferation of mobile phones amongst general population is immense. The percentage of active mobile subscriptions within the population reached 96% in 2014 [3]. In developed countries, the number of mobile subscribers has surpassed the total population, with a penetration rate now reaching 121%, whereas, in developing countries, it surpassed 85% and keeps growing.

Analyzing the spatiotemporal distribution of phones geolocated to the base transmitting towers (BTS) may serve as a great tool for population monitoring. With data being collected by mobile network providers, the prospect of being able to map changing human population movement and distributions over relatively short intervals (while preserving the anonymity of individual mobile users) paves the way for new applications and a near real-time understanding of patterns and processes in human geography [2].

1.1 Mobile Phone Location Technology

Mobile network operator (MNO) must be aware of the geographic location of each mobile phone in the network in order to be able to route calls to and from them and to

seamlessly transfer a phone conversation from one base station to a closer one as the user is moving. This originally technical necessity was transformed into a commercial opportunity to increase the Average Revenue Per User (ARPU), through what is now known as ‘Location Based Services’ (LBS). LBS are all services that use the location information of a mobile device to provide a user with location-aware applications and services. Such location information can be provided by the mobile network operator, the mobile phone device, or a combination of both, but this article focuses on the former.

LBS applications initially proposed were very broad, creative and raised quite a lot of expectation. For example, users were offered the possibility to make requests of the type of ‘where is the nearest...?’ (hospital, gas station, bank, restaurant, etc.), identify friends that walk nearby (Foursquare), ask for navigation instructions when lost (Google maps), locate lost phone (device locator), or receive a promotion from a familiar store when walking past it (location based ‘spam’). [4] Nevertheless LBS failed to deliver its promises at the turn of the century, and its huge forecasted market potential did not come to reality [9]. This is partly because early services have been very restricted due to the poor location accuracy available, and the limited capabilities of both the handheld hardware (screen size and quality, processing power and storage capacity) and the network data transfer speeds and bandwidth [1, 4]. However, the second wave of geolocation services comes right now, and this time it seems that it is here to stay.

Mobile networks compose of cells around a base transmitting tower (BTS). Each active mobile phone, therefore, can be located by triangulating the geographic coordinates of its BTS. This network-based positioning method is simple to implement, phone and user independent and its accuracy depends directly upon the network structure; the higher the density of towers, the higher the precision of the mobile communication geolocalization [4].

Records of the time and associated cell of anonymous mobile phone users are valuable indicators of human presence and offer a promising alternative data source for increasing the spatial and temporal detail of large-scale population datasets [2]. Mobile phone geolocation can be therefore used to:

- observe human mobility patterns at the individual level (police and security services only),
- monitor movements and activities of selected population using aggregated data,
- improve responses to disasters and conflicts, [7]
- plan epidemics elimination strategies,
- explore traffic flows and prevent traffic jams,
- study intensity of human activities at different times,
- identify seasonality in both domestic and foreign tourist numbers and destinations.

Legislation in USA and EU also requires mobile network operators to provide an accurate location for calls to emergency services.

2 Geolocation in T-Mobile Czech Republic

T-Mobile is the largest Czech mobile network operator, which is in regular contact with about six million terminals (40% market share) with an aggregate data rate of hundreds of millions of signal records generated daily. T-Mobile had decided to take full advantage of the big data and geolocation potential and over the last three years has developed a series of unique solutions that add value to the customer and provide a competitive edge for the company. In this paper (Chap. 3) we present the sample of the most interesting solutions. But first, let's look into some definitions.

Data Anonymization

Every geolocation project starts with anonymization. The legislation of the Czech Republic and EU stipulates that it is always necessary to make data anonymization before data processing, thus preventing the identification of individual end-users. T-Mobile uses sophisticated encryption algorithms to remove identification and uses aggregated data for processing, so only meta data arise in the calculations, which are the only ones used to interpret the results later.

Technological Background

The source of T-Mobile's geomobile data is residual signaling data from mobile cell identification, which makes possible to know the approximate location of the mobile terminal and thus the distribution of the population in space and time. Further refinement of the position can be calculated if needed. Signaling data arises from typical mobile events such as call, data transmission, SMS message, terminal transfer between individual transmitters, or upon report call to the infrastructure in the so-called periodic specification when the terminal is periodically called for a signaling response. Data from signaling (after anonymization) can be stored in the data warehouse for further processing using classic business intelligence tools or special IT tools supporting large data [5].

2.1 Continuous Online Monitoring System

The current distribution of mobile devices can be mapped through residual signaling data. Random but quite representative pattern of Czech population's mobility can be recalculated in real time into aggregated geodemographic matrix of mobility. Based on both global and local system calibration (according to control check-points) they are recalculated to represent the real number of persons in each area. Specialized software allows displaying the distribution of the population in nearly real time, as well as a historic time lapse sequences (Fig. 1).

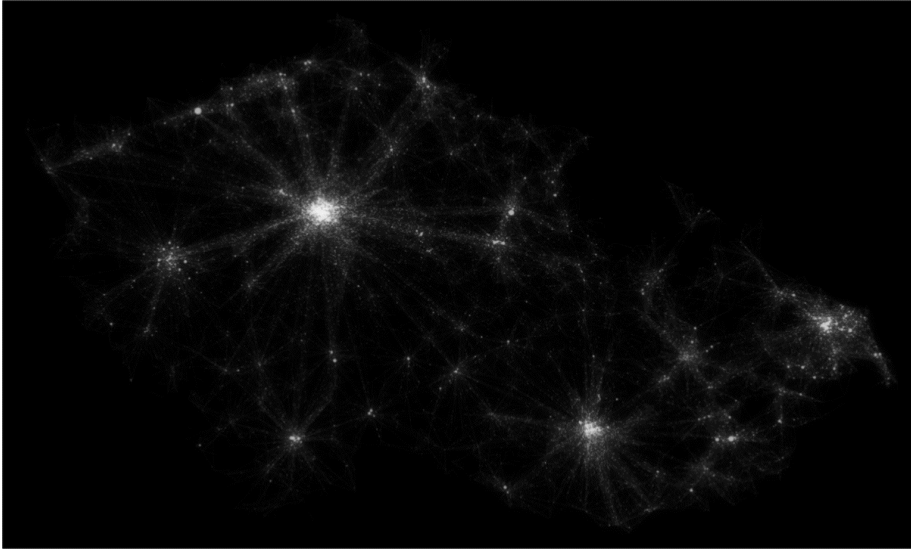


Fig. 1. Online monitoring visualization – movement of population in the Czech Republic

2.2 Business Intelligence and Big Data Tasks

Typical business intelligence and big data processing tasks that need to be handled when working with anonymized data exported from signaling to a data warehouse are as follows:

- Keep, search and archive records of terminals in a given area.
- Position these terminals in required geographic formats such as centroid, square, cadastre, or any given polygon.
- Deal with signal skipping between neighboring BTS.
- Deal with the problems near international border areas (roaming).
- Store the number of people using the mobile phone in the given area in a specific time slot, together with time-lapse data.
- Manage algorithms to count unique terminal approaches versus cumulative access to all terminals.
- Identify the origin and destination matrix, which is important for determining the motion vector.
- Compute the whole population to allow other data layers calibration.
- Solve the non-homogeneity of data in some areas.
- Create enhanced models in locations where network topology does not meet the requirements in terms of precision.
- Modal split, that is to distinguish the movement of the population from the point of view of transport, such as public transport (train, boat or bus) or individual transport [5].

2.3 Fields of Application

Mobile geolocation is successfully used in a number of cases, the most common uses being:

- Crisis management (lost children, information on people in the area of fire, floods or chemical threats) [6].
- Detecting population mobility for state infrastructure and urbanization planning (new roads, P+R areas, public transport, land use plans).
- Commercial statistics (number of visitors to shopping centers, outdoor, festivals, tourism and city and area, replacement or supplementation of Czech statistical office research).
- Optimizing traffic flows.
- Location-based services such as a mobile ad for nearby services.

The list of examples would be unlimited with possibilities enriched by other external data (weather, social networks, CRM systems, etc.) taken into consideration.

3 Case Studies: Big Data Geolocation Analysis

T-Mobile is very active in the big data field and, together with partners from the academic and commercial sectors, is involved in a wide range of research and commercial projects. Our paper will demonstrate above-mentioned possibilities on practical examples of implementation by the biggest mobile network operator in the Czech Republic.

3.1 Pilot Case Study of Šumava National Park

The objective of this pilot tourism-oriented project was to calculate the number of visitors in the Lipno, Kvilda, Modrava and Horská Kvilda regions at the turn of 2013 and 2014, to find out, where visitors came from, how long they stayed and which places they visited. Such data is useful for the national park administration, municipalities and local businesses - hoteliers, restaurateurs, sports facilities, and other entities. They get to know and possibly target the tourists of the Šumava Mountains. National Park, in cooperation with T-Mobile and KPMG, prepared the long-term geolocation analysis.

And what are the results of the case study? [8] Most foreign visitors arrived from the Netherlands (36%), Germany (35%), Austria (6%) and Russia (5%). In total, 260,000 visitors arrived in Šumava during the monitored period, of which 24% were from abroad.

The main destinations for both domestic and foreign tourists are usually near (<10 km) the place where they spent the night. One day trip was made by more than 1,000 (0,5%) domestic tourists and 700 (1,2%) foreign tourists. According to KPMG's analysis, tourists have spent more than 211 million CZK during the two-month survey period of the region, of which about 70% accounted for domestic and 30% for foreign tourists.

3.2 Case Study – Czech Mountain Ski Resorts

Based on a very positive response to the pilot study, a very thorough analysis of the behavior of visitors to Czech mountain resorts was conducted in 2015. Six top Czech and Moravian ski resorts were analyzed: Harrachov (Giant mountains), Pec pod Sněžkou (Giant mountains), Rokytnice nad Jizerou (Giant mountains), Špindlerův Mlýn (Giant mountains), Kohútka (Javorníky), Lipno (Šumava).

The data from the mobile network was only one of the sources - it supplements the information about the profile of the visitors, obtained by the questionnaire survey, and the sample survey of the Czech population.

The results answer some important business related questions: What are the visitors of the mountains doing and what they expect? How many one-day visitors are there and how many tourists sleep in the mountains? Do they differ in behavior? When do the Slovaks begin to travel to the mountains? What services are missing the most numerous visitor groups? Those are valuable information, which would be otherwise hard to get.

The most interesting findings are presented in Figs. 2 and 3.

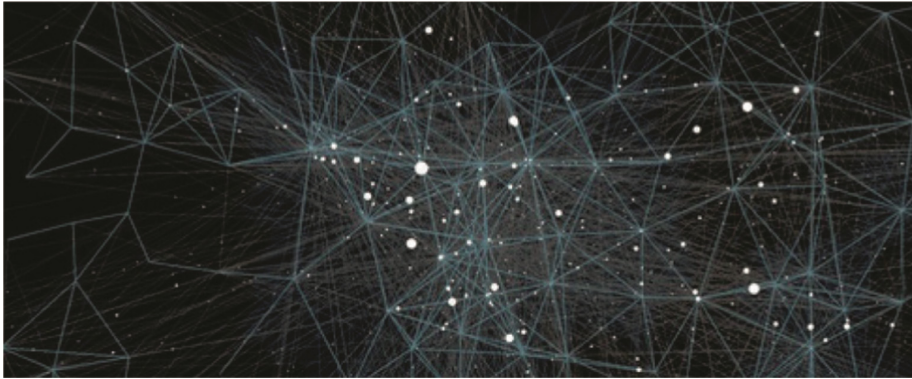


Fig. 2. Example of online monitoring visualization – detail

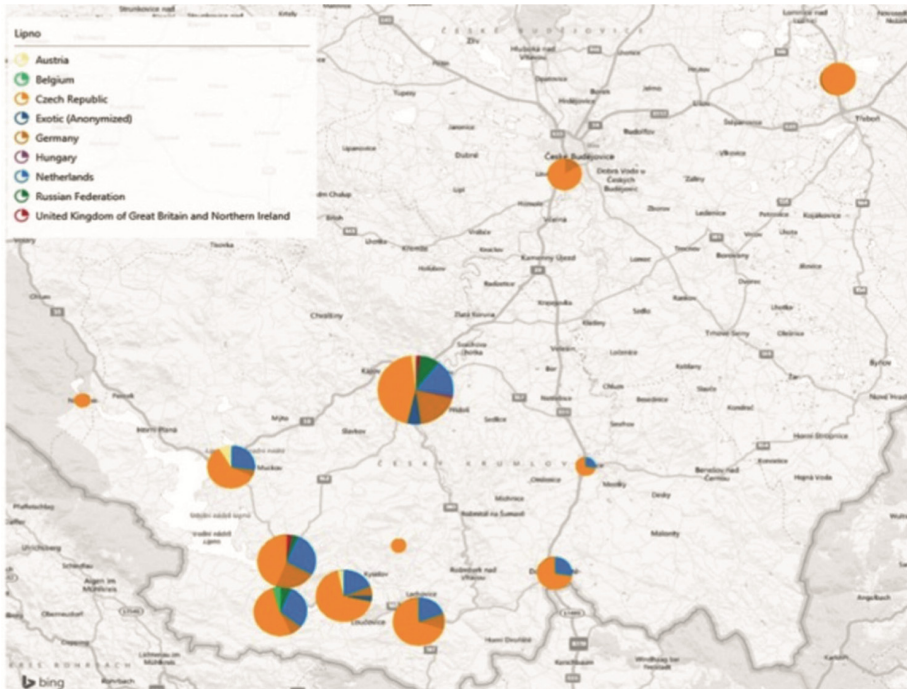


Fig. 3. Distribution of visitors in Sumava National Park (authors, based on data from T-Mobile)

3.3 Use of Mobility Data for the Preparation of City Territorial and Development Plan

The use of geolocation data is not limited to the commercial sector. Public administration also has a number of tasks, where geolocation can be useful. The data make it possible to map the mobility of inhabitants in detail and better understand the mobility dynamics. Detailed monitoring of mobility enables to classify territory in terms of public needs and helps with the sustainable development of communities (Figs. 4, 5 and 6).

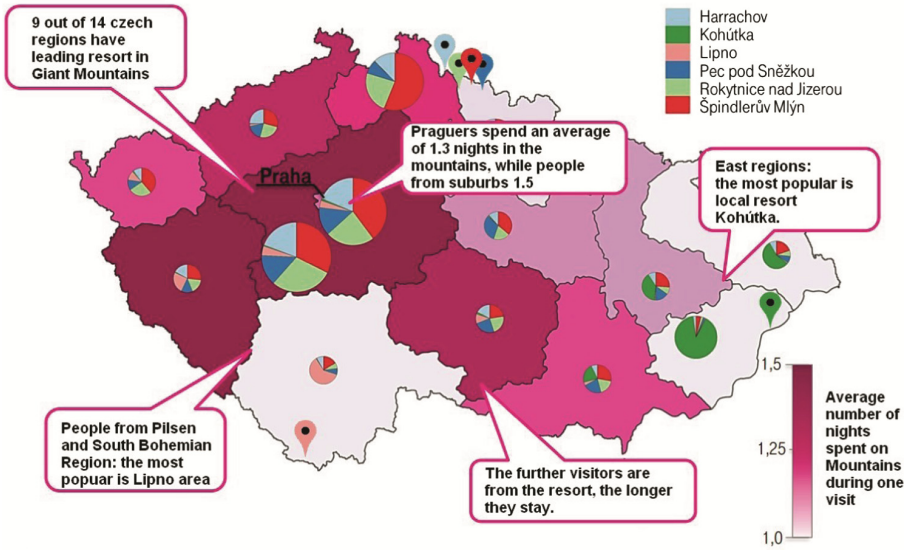


Fig. 4. Czech mountain ski resorts – origin of visitors (authors, based on data from T-Mobile)

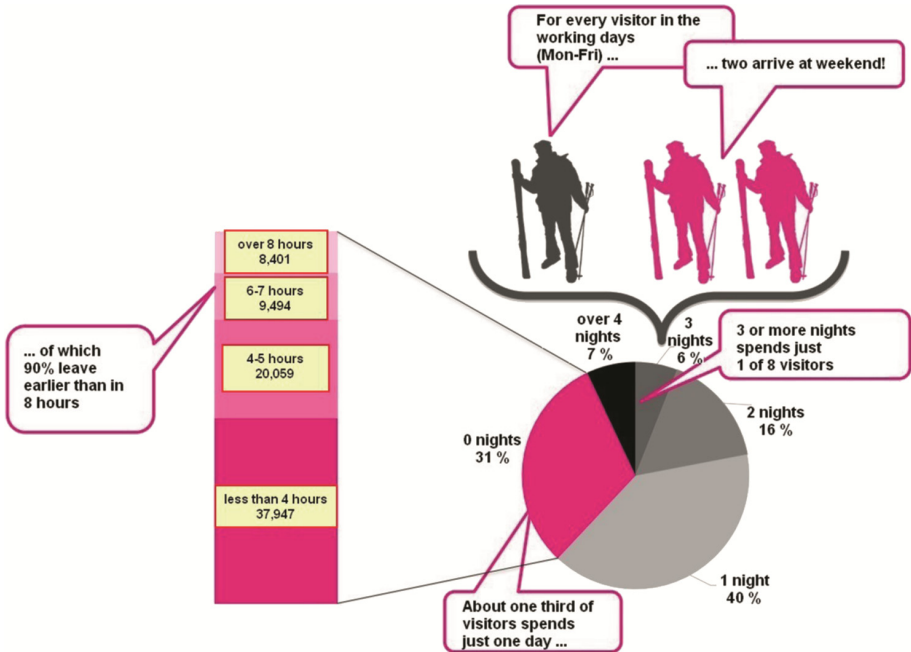


Fig. 5. Czech mountain ski resorts – length of stay (authors, based on data from T-Mobile)

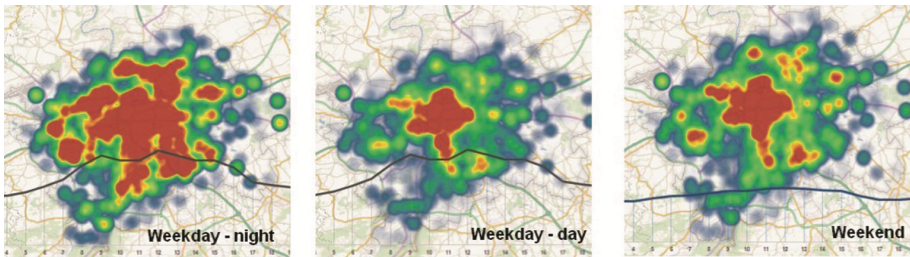


Fig. 6. Use of mobility data for city development coordination – density of population day, night, weekend (authors, based on data from T-Mobile)

Data can help quantify in detail mobility links between the city and its wider surroundings and help, for example, to plan accordingly public transport, housing or new schools and amenities (see Fig. 7).



Fig. 7. Average distribution of the inhabitants of Černý Most area (part of Prague) in the daytime (typical weekday). Purple - inhabitants in the place residence Orange - inhabitants traveling outside of his/her home (Color figure online)

3.4 Václav Havel Airport

Václav Havel Airport Prague, formerly Prague Ruzyně International Airport (IATA: PRG, ICAO: LKPR) is the international airport in the capital of the Czech Republic. It is, with over 13 million passengers in 2016 (over 15 million expected in 2017), the busiest airport in the newer EU member states. It serves as a hub for Czech Airlines, Travel Service, SmartWings, Wizz Air, and Ryanair. Its 2 runways can handle 71,000 t of cargo and 137,000 aircraft movements per year.

The survey was conducted from July 2016 to May 2017 and recorded 12.2 million passengers - 6.218 million arrivals (of which 1.856 million Czech and 4.361 million foreigners) and 5.987 million departures (of which 2.071 million Czech and 3,915 million Foreigners). A significant seasonal component appeared in the airport's operations.

Although these basic statistical data can be gathered by counting arrivals and departures and aircraft occupancy capacities, the use of mobile geolocation brings additional, enhanced capabilities. It is possible to monitor the movement of passengers in the Czech Republic or the destination of Czechs abroad. We can, for instance, easily detect the length of stay of foreigners in the Czech Republic (see Table 1), or even check, which

Table 1. Length of stay in the Czech Republic

Days	Percentage
1	12%
2	20%
3	26%
4	15%
5	6%
6	4%
7	4%
8	2%
9	1%
10	1%
11	1%
12	1%
13	1%
14	1%
More	6%

Table 2. UNESCO sites visited by airport passengers

UNESCO site	%
Prague castle	85,87%
Cesky Krumlov	7,92%
Kutna Hora	3,98%
Telc	0,36%
Olomouc	0,35%
Tugendhat	0,35%
Holasovice	0,32%
Litomyšl	0,30%
Lednice	0,26%
Kromeriz	0,10%

UNESCO monuments they decided to visit (see Table 2). The analysis can go even deeper – we can identify the day of UNESCO visit or even in what order they have been visited. For each monument, we can calculate its popularity by different nationalities (Russians seem to prefer Cesky Krumlov and Kutna Hora – they constitute 44.1% and 44.8% of foreign visitors there, Americans like to go to Lednice – 40.4%)

Last but not least, it is possible to analyze also the movement of the visitors in the defined localities. Figure 8 displays the distribution of tourists from Russia, Germany, and Italy during 24 h in Prague. Data allow measuring quantity and distribution of foreign visitors in the targeted area. Thanks to traffic data, it is possible to identify the potential for further development.

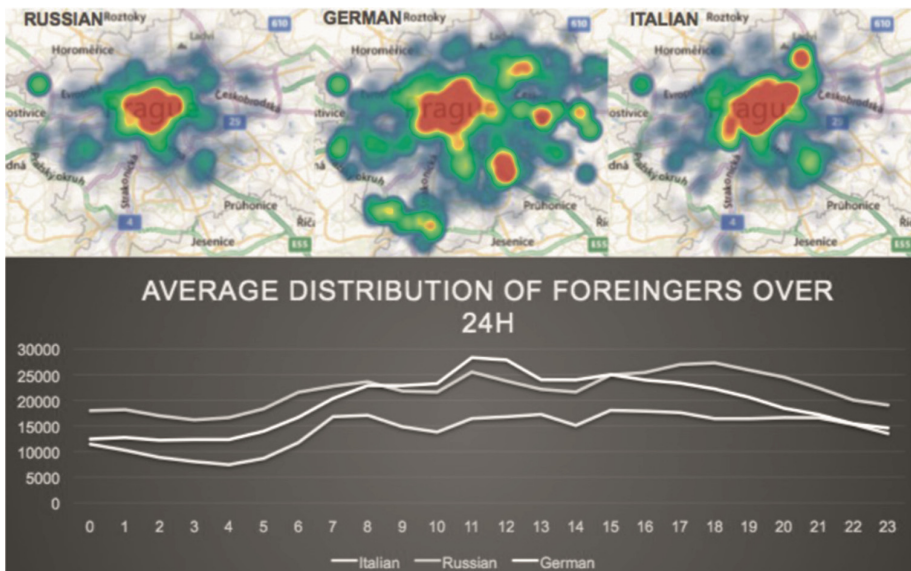


Fig. 8. Distribution of Russian, German and Italian tourists in Prague (authors, based on data from T-Mobile)

Mobile phone geomonitoring allows also determine catchment area of Prague airport – as shown in Fig. 9, it serves mainly central Bohemia region. Mobile monitoring can be (to some extent) used even abroad. Figure 10 pictures destinations of Czech travelers – flying from V. Havel airport – categorized upon the length of their trip.

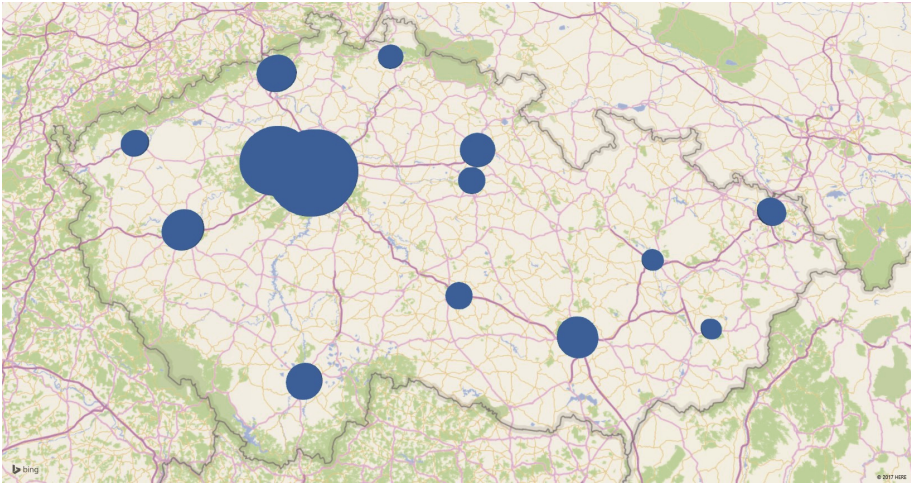


Fig. 9. Origin of Czech travelers – departures from V. Havel airport (authors, based on data from T-Mobile)

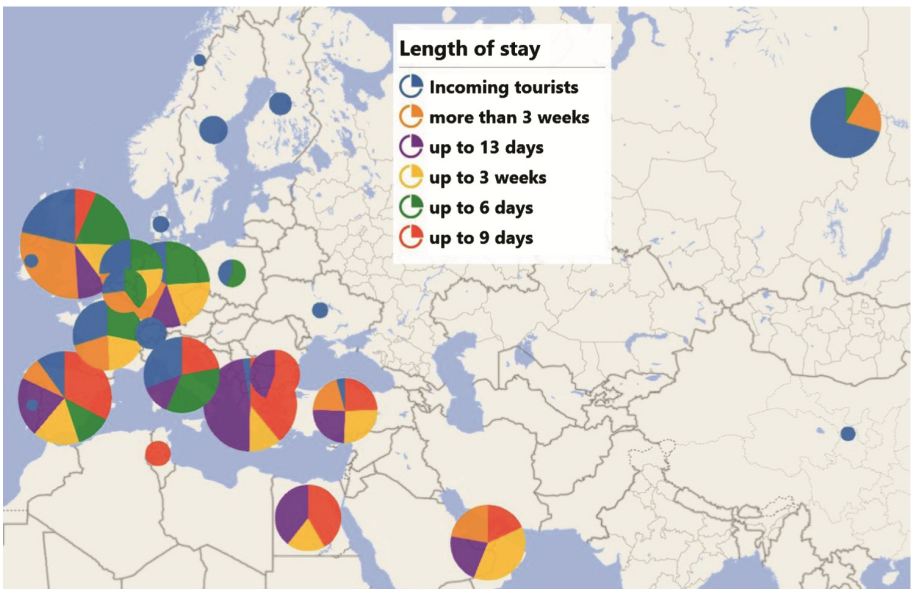


Fig. 10. Destinations of Czech travelers – departures from V. Havel airport and length of the trip (authors, based on data from T-Mobile)

4 Conclusion

This paper presented real-world studies conducted in the Czech Republic that combined mobile device geolocation and big data approach. Despite the relative youth of this field – first studies started less than 5 years ago – it is becoming a standard tool for a near real-time understanding of patterns and processes in human geography. It allows to observe mobility patterns and study intensity of human activities at different times and places. The data thus obtained are very accurate, up-to-date, and can bring previously unpredictable views and facts.

From the perspective of mobile network operators is this analysis also very interesting – the fact that they are the only ones with access to the necessary data puts them into very advantageous position.

In the future, further rapid development and massive expansion of big data geolocation analysis can be expected.



Acknowledgment. Paper was processed with contribution of long term support of scientific work on Faculty of Informatics and Statistics, University of Economics, Prague.

References

1. Mountain, D., Raper, J.: Positioning techniques for location-based services (LBS): characteristics and limitations of proposed solutions. *Aslib Proc.* **53**(10), 404–412 (2001)
2. Deville, P., et al.: Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci. U. S. A.* **111**(45), 15888–15893 (2014)
3. International Telecommunication Union ITU: World Telecommunication Development Conference (WTDC-14): Final Report. <http://handle.itu.int/11.1002/pub/809f5219-en>
4. Mateos, P., Fisher, P.F.: Spatiotemporal accuracy in mobile phone location: assessing the new cellular geography. In: Drummond, J., et al. (eds.) *Dynamic and Mobile GIS: Investigating Changes in Space and Time*, pp. 188–211 Taylor & Francis (2006)
5. Novak, R., Kovarnik, L.: Big Data a T-Mobile. *Computerworld* 2015, 6 (2015)
6. Skrbek, J.: New possibilities of information services for special situations. In: *IDIMT-2009 System and Humans Complex Relationship*, vol. 29, pp. 123–130 (2009)
7. Skrbek, J., Kvíz, J.: Critical areas of early warning system. In: Doucek, P., et al. (eds.) *IDIMT-2010: Information Technology - Human Values, Innovation and Economy*, pp. 193–202. Universitätsverlag Rudolf Trauner, Linz (2010)
8. TTG: Big data: Pomohou cestovnímu ruchu na Šumavě? | TTG - vše o cestovním ruchu (2014). <http://www.ttg.cz/big-data-pomohou-cestovnimu-ruchu-na-sumave/>
9. Zetie, C.: Location Services Find Their Way To The Enterprise. <http://www.informationweek.com/location-services-find-their-way-to-the/26100784>



Pattern Discovery from Big Data of Food Sampling Inspections Based on Extreme Learning Machine

Yi Liu¹ , Xin Li², Jianxin Wang², Feng Chen³, Junyu Wang¹ ,
Yiwei Shi¹, and Lirong Zheng¹

¹ State Key Laboratory of ASIC & System, Fudan University, Shanghai 200433, China
{liuyi13, junynuwang}@fudan.edu.cn

² Beijing Forestry University, Beijing 100083, China
wangjx@bjfu.edu.cn

³ Information Center for China State Food & Drug Administration, Beijing 100053, China
chenfeng@cfd.gov.cn

Abstract. Food sampling programs are implemented from time to time in local areas or throughout the country in order to guarantee food safety and to improve food quality. The hidden patterns in the accumulated huge amount of data and their potential values are worthy to research. In this paper, Extreme learning machine (ELM) is employed on real data sets collected from the food safety inspections of China in recent two years, in order to mine the relationship between food quality and food category, manufacturing site and season, inspection site and season, and many other attributes. Experimental results indicate that the ELM approach has better prediction precision and generalization ability than Logistic regression that was adopted in preceding work. The patterns obtained are helpful for making more effective food sampling plans and for more targeted food safety tracing.

Keywords: Food sampling inspection · Big data · Extreme learning machine
Logistic regression

1 Introduction

Food safety issues have aroused world-wide attention since it is closely related to public and household health and interests [1]. Most countries have implemented systems for food safety supervision and inspection, in order to reduce the quantity, strength, and impact of food safety incidents, and to improve the quality of food finally delivered to the end users [2]. However, food quality testing and food safety inspections are time-consuming, labor-intensive tasks, and they could sometimes be a heavy financial burden. Therefore, much research work has been done in order to improve inspection efficiency and effectiveness without increasing inspection quantity and strength, or even with reduced quantity and strength of food safety inspections [1, 3].

In China, many food safety incidents have occurred in recent years [3]. To deal with these problems, China government has taken a lot of measures to guarantee food safety and quality, and all levels of food testing laboratories in China carry out every day a

great deal of testing work. As a result, a large amount of food testing data is accordingly recorded and collected, and as a matter of fact, after years of accumulation, a huge data warehouse has come into being with rich information about food quality and safety and with many other properties. Initially, these accumulated data were only a matter of recording, and gradually they were utilized for inquiry and statistical purposes. The accumulated data, however, were found to be much more valuable than that [3], since the obtained patterns or rules underlying the data did provide us useful and helpful knowledge about the relationship among the attributes, which are able to help us make more effective and powerful inspection plans to expose more food safety problems, and hence to reduce consumption of time, labor, and financial burden.

Nevertheless, with the size of the data growing steadily in the course of food production, processing distribution and trading, the huge amount of data cannot be handled by conventional computing methods, which are by and by replaced by the technology of big data [4]. After the technologies of cloud computing and internet of things, big data technologies are another profound revolution that have penetrated into a variety of areas and given rise to dramatic changes in these areas. Big data is an abstract concept, with the characteristics of great quantity, rich variety, semi-structured and unstructured data, fast-growing, and that the traditional database management software cannot process it pragmatically with single-node computing resource. Consequently, distributed computing is the core method and key means in the bunch of big data technologies. Reference [5] examined the potential for big data application in the agriculture sector, including the variety and velocity characteristics in the sector and the integration of data and analysis that will be needed for successful implementation.

With the big data of food safety inspections accumulated, managed, preprocessed and analyzed, a variety of applications could be implemented, including dynamic and comprehensive food safety analysis, foodborne disease study, early warning and assessment of food safety, and so on. Fulfillments of these applications are helpful for boosting food quality level and improving food safety tracking. In order to implement these applications, however, a bunch of approaches are needed such as data preprocessing, statistical analysis, machine learning, and data mining [3, 6, 7].

Before applying methods mentioned above, complex processes should be taken for data preparation, including data cleaning, data normalization, and missing data imputation. In most cases, the phenomenon of missing data is inevitable in a real data set, and therefore missing value imputation is an essential preprocessing step in data mining and machine learning. The imputation methods of kNNI [8] in recent years have been widely applied because of its easy operating. The result and hence the accuracy of kNNI, however, are dependent of the parameter k , which means that each k should be tried in order to get an optimal one. Moreover, the result of kNNI is a biased estimation since the neighbors of the targeted point with missing value may lie unevenly around the point. Two variations [9, 10] of kNNI were proposed to overcome the defects of previous versions and they both perform satisfactory. Only after the preprocessing step, are the data sets of food safety inspections ready for further analyzing and mining.

The rest of the paper is organized as follows. Section 2 introduces research work related to this paper, including those on missing data imputation, Logistic regression, neural network, and extreme learning machine. In Sect. 3, the ELM framework is

described in detail that is employed to mine the patterns hidden in the big data of food safety inspections. Section 4 presents the experiments on real data sets and the corresponding results. And Sect. 5 concludes the paper.

2 Related Work

A variety of methods and technologies have been studied, tested and/or implemented to analyze and utilize the data collected from food safety inspections, and many exciting results and conclusions have been obtained.

Khosa and Pasero [6, 7] used an artificial neural network (ANN) as a classifier to predict at an early stage of processing or manufacturing whether important food ingredients, pine and pistachio nuts, are healthy. X-ray images of the nuts were used, and texture features were extracted from the images. In that work, the texture features and the sample labels were used as the training data, and the texture features were independently used as the basis for making predictions and classifications. As a result, the ANN classifier achieved false negative rates of 0% and 6.8% for the pine nuts and pistachio nuts, respectively. The results imply that food quality has good predictability and good describability, at least in certain cases.

Reference [1] focused on a safety risk assessment of dairy products for a single corporation, also in the background of big data. That work used a classic classifier, the support vector machine (SVM). However, instead of using a serial algorithm for the SVM, a parallel cascade SVM was implemented on the platform of Apache Hadoop [11], which is an open-source distributed computing framework that is typically used to process big data by distributing the data in a large-scale cluster platform. The results from [1] demonstrate that when the number of cluster nodes increases steadily, the saved run time decreases steadily compared with the runtime for a single node. The SVM has been a successful classifier in many cases and in many areas due to its good classification accuracy, generalizability and stability. Despite this success, SVM does not perform satisfactorily when the positive and negative samples have more detailed relationships.

Statistical methods are most frequently used to analyze the data obtained from food safety inspections, with [3] being a typical study. Based on the food sampling results of the city of Shenzhen, China, that study first investigated the annual and inter-annual changing tendency of 11 food categories and analyzed the data using the t-test. Then, a logistic regression model was constructed, and the quantitative relationships between food quality and four attributes (namely, food origin, inspection season, sales site, and food packaging) were established. Instead of the result category (qualified/unqualified), the concept of “exceeded percentage” was used to measure the degree of unqualified food. Logistic regression is a powerful classifier that can be applied to both continuous and discrete variables. Although that work is a good application of logistic regression to predict which food products are most likely to be unqualified, the data for both training and testing are simulated data sets, not real data sets, which indicates that the work remains unsatisfactory.

Logistic regression, like many other regression methods, is essentially linear regression; it is aided by some nonlinear transformations, and it can capture the nonlinear relationships between the dependent variable and causative (independent) variables. The ANNs in [1, 7] used a considerable number of nonlinear transformations to capture more detailed relationships. However, as demonstrated earlier, the learning speed of feed-forward ANNs is considerably slower than that of regression learning algorithms, which take the least squares method (LSM) as the core technique. Considering both the speed advantage of the LSM and the nonlinearity advantage of the ANN, Huang et al. proposed the ELM with a single layer of hidden nodes in their two pioneering works [12, 13]. Compared with its predecessor learning techniques, the ELM improves the training speed by hundreds of times by randomly setting the weights between the input nodes and hidden nodes and by computing the weights between the hidden nodes and output nodes using the LSM. Other researchers have supported their work, particularly the random assignment of weights, by mathematical proofs, such as in [14], which provides a geometric perspective.

After these pioneering studies, a variety of variations and improvements in the ELM were presented. Reference [15] proposed an inverse-free ELM that further improved the computational speed of the training process, as computing the inverse of a square matrix is the most time-consuming part of the LSM. Accounting for the architecture of the sub-network nodes, Y. Yang and Q. M. Jonathan Wu designed a variation of the ELM, ML-ELM, that exhibits competitive accuracy and speed compared with other conventional feature learning methods with sub-network nodes [16, 17].

The ELM has a notable defect, namely, that the number of hidden nodes must be manually assigned or assigned by other state-of-the-art methods. In fact, the optimal number of hidden nodes plays a decisive role in the ELM, as an insufficient number of hidden nodes could lead to underfitting, whereas an excessive number of hidden nodes could cause overfitting. Based on this observation, [18] presented an adaptive and automatic selection algorithm that can obtain a suitable or even an optimal number of hidden nodes for each learning case. This method can markedly reduce the degree of artificial participation and hence reduce the burden of human operators.

In addition, there are many applications of the ELM to different types of domains. The ELM was applied to predict soil moisture in an apple orchard [19], taking both the weather factors and the time series of the soil moisture as inputs. Compared to the conventional method of the SVM, the ELM exhibits a higher prediction accuracy over a larger forecast range with a higher speed. Reference [20] proposed a new classification algorithm for food classification based on both spectroscopy and the ELM, and the experimental results indicated that the ELM is typically more precise and robust than its competitors, including k-nearest neighbor, partial least-squares discriminant analysis, back propagation ANNs, and least-squares support SVMs.

3 ELM Approach Specification

In this section, we will present in detail the ELM-based classifier for predicting whether a sample food to be inspected is qualified or not. Firstly, in Sub-Sect. 3.1,

the cause variables are selected according to whether it is likely to affect the food quality. And then data preprocessing techniques are presented in Sub-Sect. 3.2. After that in Sub-Sect. 3.3, the main framework of the ELM method is described based on the discussion of the former two sub-sections.

3.1 Selecting Relevant Factors

According to the food safety inspection data, the result variable that we are most interested in is quite simple: it has binary values for whether the food is qualified or not. However, the causative variables are more complex and involve many factors. We eliminate the factors that are not related to the ability to predict the food quality, such as the sampling number and name of the manufacturer. After the elimination operation, 9 causative variables are retained, as listed in Table 1.

Table 1. Causative (dependent) variables selected for the model^a.

Selected factor/variable	Meaning of the variable
Food category	There are 6 categories ^a in the inspection data
Manufacturing date	The date when the product was manufactured
Manufacturing site	The place where the product was manufactured
Inspection date	The date when the product was sampled and inspected
Inspection site	The place where the product was sampled and inspected

^a The 6 categories are T0, dairy products; T1, aquatic products; T2, infant formula; T3, meat products; T4, liquor; T5: edible oil.

3.2 Preprocessing Technique

When all the factor variables are determined, the data are processed to eliminate the noise data. The missing values are completed with the imputation techniques proposed in [10] while considering the representative point and the densities of the points in each quadrant compared to the targeted point for the missing values.

New causative variables can be generated based on the variables listed in Table 1. For example, from the manufacturing date and inspection date, a new variable, the elapsed days, can be generated; this variable refers to the time span between the manufacturing time and inspection time. Another example is “Whether in the same province”, which is generated from the two variables “Manufacturing site” and “Inspection site”; this variable indicates whether the two sites are in the same province or not.

To date, certain variables have not yet been useful for the models of either logistic regression or ELM, as they are category variables, not numeric variables. For example, the manufacturing date appears to be a numeric variable, but in fact, it is more likely to be a categorical variable because it implies the seasonal information of the manufacturing time. Thus, we transform the variable “Manufacturing date” into four variables, namely, “Spring”, “Summer”, “Autumn” and “Winter”, with each having binary values of true or false. The four new variables are called dummy variables, and they are generated in the same manner as described in [21].

After the preprocessing stage, the data are ready for training, testing, and predicting using both the ELM method and its competitors.

3.3 Framework of ELM Method

In this paper, we also use one-hidden-layered nodes, as shown in [18–20]. The structure of the network is illustrated in Fig. 1.

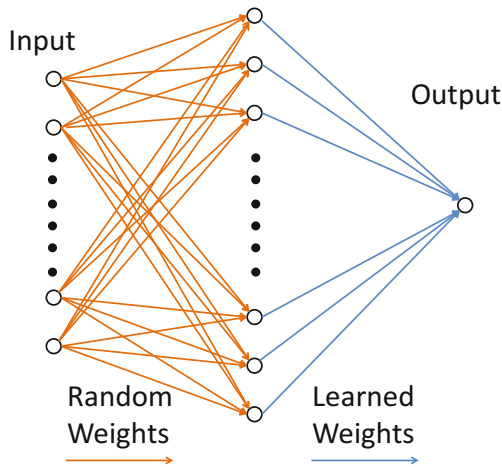


Fig. 1. Structure of the ELM. There is only one hidden layer and only one output node.

Each input node in Fig. 1 represents a causative variable. The causative variables selected and the variables generated by them are each represented by an input node. There are considerably more hidden nodes than input nodes; but it is not a fixed number. Instead, it varies according to the number of inputs and the structure of the training data based on the adaptive strategy given in [18]. As described in [13], the weights between the input nodes and the hidden nodes are set to random values (see Fig. 1), which implies that the output value of one hidden nodes may be proportional (or nearly proportional) to that of another hidden node. Therefore, at least one of them is useless to capture the relationship between the input and the output. The optimization algorithm in [18] first generates a large number of hidden nodes and then selects the nodes one by one, making the newly selected one least linear-correlated to the previously selected node. By taking into account the input data and the output data an optional number of hidden nodes can be obtained. We employ this optimization method to form the structure of ELM. The single output node represents the result of a record, which means whether a food sample is qualified.

The weights between the input nodes and hidden nodes are assigned randomly as described in [12, 13], and they are all set to be in the range $[-1, 1]$. However, all weights between the hidden nodes and output node are obtained by learning and computing based on the training data.

Suppose that the number of input nodes is n and the number of hidden nodes is h , the input of the j th hidden node is calculated as follows:

$$G_j = \beta \sum_{i=1}^n W_{i,j} \quad (1)$$

where β is a parameter that will be discussed later, $W_{i,j}$ is the weight between the i th input node and the j th hidden node.

Each hidden node processes its input by the following equation and then outputs the following:

$$H_i = \frac{2}{1 + e^{-G_i}} - 1 \quad (2)$$

H_i will always lie between -1 and 1 , which makes its value distribution approximately symmetric about the y -axis. Equation (2) is often called the activation function, which is a highly nonlinear function. All activation functions in the hidden nodes together make the system capable of approximating nearly any nonlinear relationship between the input nodes and output node.

Parameter β in Eq. (1) will affect the effectiveness of the system. If β is not sufficiently large, the relationship between the input and output will degenerate to a linear relationship. However, if β is excessively large, all the inputs of the hidden nodes will be transformed by the activation function into either -1 or 1 . Thus, we set parameter β in this paper according to the following empirical formula:

$$\beta = \frac{10}{n} \quad (3)$$

where n is the number of input nodes.

In the step of the LSM for calculating the weights between the hidden nodes and output node, the inverse of a square matrix must be computed, which will not be executable if the matrix is irreversible. If this problem occurs, we will change the square matrix slightly and make it reversible by using the method suggested in [22], which overcomes a significant shortcoming of the ELM.

The overall framework of the ELM approach is shown in Fig. 2.

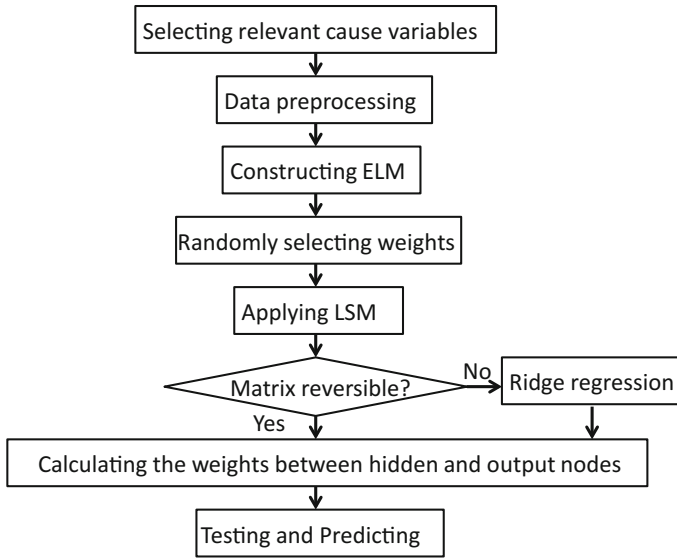


Fig. 2. Overall framework of the ELM approach.

4 Experiments and Results

The data sets used are publicly available from the State Food and Drug Administration of China. For these samples, the manufacturing date ranges from November 26, 2014 to September 1, 2016, whereas the inspection date ranges from October 29, 2015 to September 9, 2016.

The two methods applied to the data sets are logistic regression presented in [3] and the ELM. The variable selection and data preprocessing are same for the two methods. For each category of food, all data are partitioned into training data and testing data. The training set and testing set are identical for the two methods. The testing results are listed in Table 2.

Table 2. Comparison of experimental results.

Category	Number of testing cases	Number of correct cases for LR	Number of correct cases for ELM	Accuracy of LR (%)	Accuracy of ELM (%)
T0	1376	1196	1212	86.9	88.1
T1	873	777	791	89.0	90.6
T2	1403	1260	1310	89.8	93.4
T3	4058	3633	3656	89.5	90.1
T4	2730	2615	2618	95.8	95.9
T5	2063	1785	1798	86.5	87.2

The data listed in Table 2 are shown in Fig. 3.

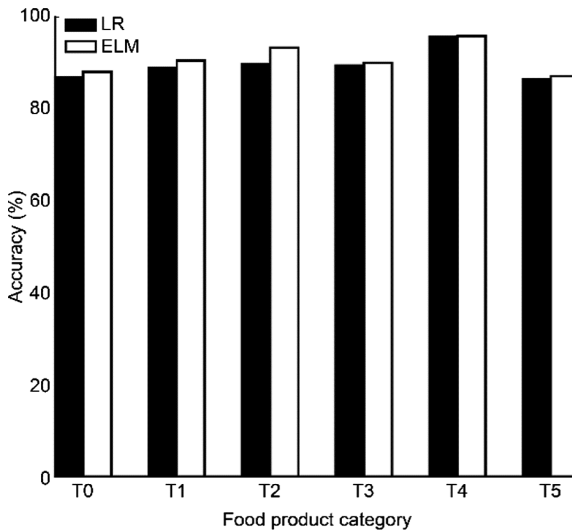


Fig. 3. Comparison of the experimental results from logistic regression and the ELM. The black bar represents the accuracy percentage of the logistic regression, whereas the white bar represents that of the ELM.

Figure 3 shows that the ELM has better accuracy than logistic regression for all food categories, although they perform nearly the same for certain categories, such as T4.

5 Conclusions

ELM is employed in this paper to describe the big data collected from the food safety inspections of China in recent two years. The trained model is used to predict the food quality and it performs better than Logistic regression that was implemented and tested on simulated data sets. Results from a series of experiments show that ELM is better in accuracy than Logistic regression for each of the 6 food categories. And both of the methods run very fast because they all take the advantage of optimized calculating steps. The success of the ELM owes much to the large number of hidden nodes and the nonlinear activation functions in them are able to capture the nonlinear components in the relationship between the inputs and the outputs.

With the ELM model and the according prediction system, food samples can be taken no longer randomly; on the contrary, food products could be filtered by the prediction system and only those with least qualification probabilities will be selected for sampling test. Therefore, aided by the ELM prediction and classification system, more effective inspection plans can be made which mean less labor input and more food safety problems exposed.

Acknowledgments. This work was supported by the National Science & Technology Pillar Program of China (2015BAK36B01), the National Natural Science Foundation of China (61402436), Shanghai Engineering Research Center of Product Traceability.

References

1. Ma, Y., Hou, Y., Liu, Y., Xue, Y.: Research of food safety risk assessment methods based on big data. In: 2nd IEEE International Conference on Big Data Analytics, Beijing, China, pp. 1–5 (2017)
2. Antunovic, B., Mancuso, A., Capak, K., Poljak, V., Florijančić, T.: Background to the preparation of the Croatian food safety strategy. *Food Control* **19**(11), 1017–1022 (2008)
3. He, L., Wang, Z., et al.: The method of food safety sampling inspection based on dynamic weight. *Math. Model. Appl.* **2**(3–4), 4–12 (2013)
4. Li, F., Lv, Y., Zhu, Q., Lin, X.: Research of food safety event detection based on multiple data sources. In: International Conference on Cloud Computing and Big Data, Shanghai, pp. 213–216 (2015)
5. Sonka, S.: Big data and the ag sector: more than lots of numbers. *Int. Food Agribus. Manag. Rev.* **17**(1), 1–20 (2014)
6. Khosa, I., Pasero, E.: Defect detection in food ingredients using multilayer perceptron neural network. In: 2014 World Symposium on Computer Applications & Research, Sousse, pp. 1–5 (2014)
7. Khosa, I., Pasero, E.: Artificial neural network classifier for quality inspection of nuts. In: International Conference on Robotics and Emerging Allied Technologies in Engineering, Islamabad, pp. 103–108 (2014)
8. Kung, Y.-H., Lin, P.-S., Kao, C.-H.: An optimal k -nearest neighbor for density estimation. *Statist. Probab. Lett.* **82**(10), 1786–1791 (2012)
9. Zhang, S.: Shell-neighbor method and its application in missing data imputation. *J. Appl. Intell.* **35**, 123–133 (2011)
10. Wang, J., Zhang, Z., Chen, Z., Yuan, Q.: Imputation missing values with distance- and density-weighted and quadrant-based nearest neighbors. *J. Comput. Inform. Syst.* [1] **11**(18), 6605–6613 (2015)
11. Singh, D., Reddy, C.K.: A survey on platforms for big data analytics. *J Big Data* **2**(1), 8 (2015)
12. Huang, G.B., Chen, L., Siew, C.K.: Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Netw.* **17**(4), 879–892 (2006)
13. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: theory and applications. *Neurocomputing* **70**(1–3), 489–501 (2006)
14. Cervellera, C., Maccio, D.: Low-discrepancy points for deterministic assignment of hidden weights in extreme learning machines. *IEEE Trans. Neural Netw. Learn. Syst.* **27**(4), 891–896 (2016)
15. Li, S., You, Z.H., Guo, H., Luo, X., Zhao, Z.Q.: Inverse-free extreme learning machine with optimal information updating. *IEEE Trans. Cybern.* **46**(5), 1229–1241 (2016)
16. Yang, Y., Wu, Q.M.: Extreme learning machine with subnetwork hidden nodes for regression and classification. *IEEE Trans. Cybern.* **46**(12), 2885–2898 (2016)
17. Yang, Y., Wu, Q.M.J.: Multilayer extreme learning machine with subnetwork nodes for representation learning. *IEEE Trans. Cybern.* **46**(11), 2570–2583 (2016)
18. Mesquita, D.P.P., Gomes, J.P.P., et al.: Pruning extreme learning machines using the successive projections algorithm. *IEEE Lat. Am. Trans.* **13**(12), 3974–3979 (2015)

19. Liu, Y., Mei, L., Ooi, S.K.: Prediction of soil moisture based on extreme learning machine for an apple orchard. In: IEEE 3rd International Conference on Cloud Computing and Intelligence Systems, Shenzhen, pp. 400–404 (2014)
20. Zheng, W., Fu, X., Ying, Y.: Spectroscopy-based food classification with extreme learning machine. *Chemom. Intell. Lab. Syst.* **139**, 42–47 (2014)
21. Changpetch, P., Lin, D.K.J.: Model selection for poisson regression via association rules analysis. *Int. J. Statist. Probab.* **4**(2), 1–9 (2015)
22. Li, G., Niu, P.: An enhanced extreme learning machine based on ridge regression for regression. *Neural Comput. Appl.* **22**(3), 803–810 (2013)



Big Data Analytics Using SQL: Quo Vadis?

K. T. Sridhar^{1,2}(✉)

¹ XtremeData Technologies, Bangalore, India
sridhar@xtremedata.com

² XtremeData, Inc., Schaumburg, USA

Abstract. Big Data processing and analytics are dominated by tools other than SQL based relational databases, which have lost their *numero uno* status. In a world deluged by data, the general perception is that SQL databases play a marginal role even for analyzing structured Big Data despite their inherent strengths in processing such data. Focusing on the most important aspect of Big Data processing, namely analytics for data mining, we examine the validity of this perception through a study of competing technologies, published results on SQL implementations of data mining algorithms, the impact of cloud platforms and the raging debate on SQL vs NoSQL vs NewSQL. Contrary to the general belief, it appears that SQL databases in their parallel, columnar deployments on cloud with UDF support do solve some, if not all, Big Data problems and are not likely to become dinosaurs in Big Data era.

Keywords: Data mining · Big Data · Parallel DBMS · SQL · NoSQL

1 Introduction

There is a churn in the data processing world leading to a metamorphosis of database technology as understood in latter years of 20th century and early 21st century. The deluge of data in a variety of forms, from a connected world driven by internet and mobile technologies, has ushered in Big Data [1,2] and new paradigms of query processing that appear to be beyond the realm of relational model and SQL, the *lingua franca* of database systems. If the data wave from an internet used by humans dislodged DBMSs from *numero uno* status, how would they fare with *data tsunamis* likely to arise from sensor driven Internet of Things and its use in oncoming Industry 4.0 [3], the 4th industrial revolution?

Big Data, with its goal of deriving value through analytics for informed decision making and its envisaged role in Industry 4.0, brings to fore the question of *quo vadis* (*whither going*)? on relational, SQL databases. Are they relevant anymore for processing Big Data using techniques [4,5] of data mining? If yes, on what type of data and at what scale? If no, what are the alternatives?

To understand and answer these questions, this paper presents a state-of-the-art survey of SQL databases and contending technologies for Big Data processing, focusing more closely on Big Data analytics and its solutions realized within a relational database emphasizing algorithms, SQL techniques, platforms and products. Unlike other Big Data analytics studies [6, 7], which adopt only a NoSQL perspective, we examine the role of relational SQL for such analytics.

This paper is organized as follows. Section 2 summarizes evolution and status of contending technologies: SQL, NoSQL and NewSQL. Section 3 addresses issues in Big Data analytics implementation within a SQL relational database, presenting results obtained until now. Section 4 touches upon related topic of platforms and products, in context of cloud. Section 5 summarizes criticisms and limitations of competing Big Data technologies; Sect. 6 concludes the paper.

2 Data Processing Systems

Since the late seventies of 20th century, relational model and products based on it using row stores have dominated data processing applications. The rise of the internet, Big Data, IoT and novel applications have questioned this vice-like grip and given birth to a host of newer data processing technologies. The contending technologies, SQL databases, NoSQL systems and the most recent entrant NewSQL systems are discussed in Subsects. 2.1 to 2.3.

2.1 SQL Databases

SQL DBMSs targeted business data processing for enterprise systems and were very successful in providing OLTP solutions for ERP, SCM, CRM, banking, etc. ERP, the back bone of EIS has been somewhat immune [8] to Big Data, and advances in it, but is likely to be shaken up by Industry 4.0.

Applications used a row store with SQL and relational model supported by indexes, ACID (*atomicity, consistency, isolation and durability*) transactions and cost-based query planners. Based on business intelligence requirements, SQL DBMSs evolved to tackle analytic query processing of data warehouses optimized for dimensional modeling. For performance gains, they adopted parallel programming techniques to run on a shared nothing cluster of nodes as MPP SQL systems partitioning data across nodes. SQL appliances were next, led by Netezza: MPP row store with custom FPGA hardware for query processing.

Despite significant progress in SQL row store DBMSs, over a decade ago, 2014 Turing Award winner, Stonebraker argued [9] that requirements and characteristics of data centric systems vary widely and the then prevalent architecture of databases as all-encompassing, monolithic, “*one size fits all*” systems was no longer relevant or applicable. They categorize extant DBMSs as “*outbound*” systems that must write before processing, and illustrate their unsuitability for

- low latency systems for algorithmic trading, essentially “*inbound*” systems
- data warehousing and OLAP, better served by column stores

- scientific databases that require native support for arrays
- text engines: custom solutions for web (inbound), medical/legal/library data
- semi-structured data such as XML, JSON, etc., common in Web 2.0
- IoT sensor network processing systems, akin to low latency systems

Advocating use of domain specific DB engines, they show performance advantage for such engines over row databases for the first four applications in stream processing, data warehousing, scientific databases and text management. Authors of [9] conclude with the prescient observation “‘*one size fits all*’ theme is unlikely to successfully continue under these circumstances”.

Today, we have a variety of special purpose data processing systems that neither use relational model nor SQL but adopt newer programming paradigms. In data warehousing market, SQL databases changed their underlying store model to columnar and continue to retain their OLAP market share.

Based on an eighties proposal for database page storage by vertical partitioning [10] of columns, research prototypes MonetDB [11,12] and C-Store [13] pioneered column oriented databases that were followed up by successful commercial products. Today, almost all industrial DBMSs support some form of column store, simulated or modern, [14] with performance gains over only-row counterparts due to IO reduction, compression, late materialization, etc.

2.2 NoSQL Systems

The difficulties in scaling up SQL database systems for on-line, web scale processing, and in tune with the thinking against one-size-fits-all, NoSQL [1,15] systems were born. They were built on non-relational models abandoning ACID conformance of databases. The underlying models of NoSQL (*Not only SQL*; not *No to SQL*) systems include key/value, documents, columnar, graphs and streams. For the type of applications targeted, NoSQL proponents believe that

- the rigid relational schema was inflexible, particularly for unstructured data
- guaranteeing ACID properties of transactions reduced performance
- emphasis on high availability is paramount
- horizontal scalability is preferable to expensive vertical scalability of DBMS
- non-procedural SQL was not the most suitable programming language

The most significant impetus for NoSQL systems came from CAP theorem [16] that states an impossibility result for trade-offs in implementing distributed systems: a network shared-data system can have only two of three desirable properties *consistency* (C), high *availability* (A) and *partition* (P) tolerance. The CAP theorem led to an alternative view of transactions favoring “*availability, graceful degradation and performance*” [16] over consistency: BASE standing for *basically available* (BA), *soft-state* (S) and *eventual consistency* (E).

Several NoSQL systems choosing availability and partition tolerance over consistency were built for a diverse set of data processing applications. Table 1 summarizes features of some major NoSQL products. Two other contemporaneous developments contributed to growth and popularization of NoSQL systems:

Table 1. Major NoSQL products

Product	Model	Transn	MapR	Program API	License
Bigtable	Columnar	C+A	Yes	Java, HBase API	Google SaaS
HBase	Columnar	BASE	Yes	Java	Open source
Cassandra	Columnar	BASE	Yes	CQL	Open source
Dynamo	Key/value	BASE	By EMR	Java, Python, Ruby,..	Amazon SaaS
MongoDB	Document	BASE	Yes	C/C++, JavaScript	Open source
Neo4j	Graph	ACID	No	Cypher	Open source

- **MapReduce** [17], Google’s divide-and-conquer software framework for distributed systems: Inspired by LISP like functional programming style, it advocates programming of distributed applications using two functions *Map* and *Reduce* that work on key/value pairs with their execution, including failures, handled by the framework.
- **Hadoop** a distributed file system from Apache: An open source system with the goals of performance, availability and scalability, modeled on the proprietary Google File System underlying MapReduce, it made MapReduce style application development popular in the community.

Relieving the programmer of managing a parallel application running in a distributed environment of commodity clusters, with fault tolerance support, was a major step. Hadoop contributed to the meteoric rise and adoption of MapReduce for solving web scale data processing problems, with several NoSQL products incorporating the MapReduce model as shown in Table 1.

2.3 NewSQL Products

Disputing the importance of BASE over ACID, Stonebraker et al. [18] investigate reasons for under performance of SQL databases in OLTP applications of Big Data era and cite locking, latching, recovery and buffer pool management as the reasons. Eliminating these bottlenecks in their prototype H-Store database they claim 82x performance gain in TPC-C benchmark compared to row DBMSs.

The next few years heralded the term NewSQL [19] coined to refer to a class of products that preserve relational model, ACID transactions and SQL but offer NoSQL like performance and scalability for OLTP read-write workloads. Elaborating on applications of NewSQL systems [19] characterizes them as “*executing read-write transactions that (1) are short-lived, (2) touch a small subset of data using index lookups and (3) are repetitive*”. They also observe that their characterization of NewSQL is in consonance with the more narrow definition of [18]: being lock-free and using a shared nothing distributed architecture.

As of 2016, [19] lists seventeen products as NewSQL systems including SAP HANA, Amazon’s Aurora and both H-Store and VoltDB from Stonebraker et al. It is interesting to note that 15 of the 17 products listed in Table 1 of [19] use

MVCC [20] for concurrency control found in several row store SQL DBs including open source PostgreSQL, and SQL column stores for analytic workloads.

3 Big Data Analytics

The term Big Data is all around us and became part of 21st century English when Oxford dictionary defined it in 2013. Most authors characterize Big Data [1, 2, 6, 7] through 3Vs (*volume*, *velocity* and *variety*) or more (*veracity*, *variability*, *value*). In the user community, as well as lexicon definition, the stress for Big Data has been on data mining, generally understood as *discovery of models for data* using statistics, machine learning and computer science [4, 5].

Though both data mining and analysis techniques predate Big Data, they are current hot topics due to technology challenges and commercial value gained by enterprises through insights gleaned from data. Technology challenges arise from the general perception of SQL databases being inadequate [1, 2, 7, 15] for Big Data processing, due to their difficulty in dealing with the 3Vs:

- **volume:** horizontal scalability, elastic or not, is an issue for DBMSs; the advocated vertical scalability is too expensive
- **velocity:** consequent to being outbound [9] systems, SQL databases do not perform well on streaming data for real-time analytics
- **variety:** heterogeneous data are anathema to SQL databases that deal well with structured data (e.g. number, boolean, varchar), partially well with semi-structured data (e.g. XML, JSON) but are inadequate with unstructured data (e.g. tweets, text, video, audio)

Does this imply the death knell of relational, SQL databases for Big Data analytics? No; there is a contrarian view that we examine in subsections of this section. Our focus is on high volume structured data and analytics on such data; we do not touch upon other aspects of Big Data processing: data collection, cleansing, loading or privacy. Section 3.1 discusses data mining algorithms, 3.2 addresses the question of mining using SQL and Sect. 3.3 surveys published work on implementing data mining algorithms in relational DBMS using SQL.

3.1 Data Mining Algorithms

Data mining algorithms build models to classify data and the models may be used with unlabeled data for prediction or scoring. At a broad level, the learning techniques used by these algorithms may be classified [4, 5] as (a) *supervised* that uses a training set for correct classification and (b) *unsupervised* that discovers a model without any training set or *a priori* knowledge.

A large number of data mining algorithms addressing a variety of topics, clustering, classification, statistical learning, association mining, link mining, bagging/boosting, dimensionality reduction and regression have been published. The often quoted survey [21] discusses about 30 algorithms for the important

unsupervised learning technique of clustering that partitions data into similar groups. Wu et al. [22] conducted a survey to identify the top 10 data mining algorithms ranking them based on votes polled and citations. Table 2 summarizes details of top 10 data mining algorithms of Wu et al. (three with same rank 7).

Table 2. IEEE KDD Top-10 data mining algorithms

Algorithm	Mining Topic	Year	Rank	Notes
(abbrevn: s. \Rightarrow Supervised; u. \Rightarrow Unsupervised; DT \Rightarrow Decision Tree; info \Rightarrow information)				
C4.5	s.Classification	1993	1	DT: info entropy; info gain ratio
k-Means	u.Clustering	1967	2	Partitioning by similarity; iterative
SVM	s.Stats Learning	1995	3	Find best fit hyperplane
Apriori	u.Association	1994	4	Find frequent itemsets
EM	u.Stats Learning	2000	5	Maximize loglikelihood
PageRank	u.Link Mining	1998	6	Network link analysis; graphs mining
AdaBoost	s.Boosting	1997	7	Ensemble learning
kNN	s.Classification	1996	7	k neighbors by <i>nearest</i> criterion
Naive Bayes	s.Classification	2001	7	Non-iterative, Bayesian probability
CART	s.Classification	1984	10	DT: binary recursive; gini index split

Table 2 includes year of publication of algorithm, and a very brief note on nature of algorithm; more details of the algorithms may be found in [22], the original publication references cited therein, and in data mining books [4, 5].

3.2 Why Not SQL for Data Mining?

Some observations on Table 2 algorithms rated highly by mining community:

- formulated on or before 2001, most in 20th century when Big Data was unknown; nothing intrinsic to NoSQL or Big Data in analytics techniques.
- based on mathematics or statistics dealing with numbers or categorical data, both of which are essentially structured data.
- mostly iterative in nature, a programming style that is not supported by a declarative language like SQL.

Though structured data processed by mining algorithms is well handled by SQL databases, iterative nature of algorithms has been a stumbling roadblock. DBMS vendors responded by including imperative style programming with SQL; external to DBMS in C/C++, Java, Python, etc. through ODBC/JDBC interfaces; and as internal database objects: stored procedures in PL/SQL type imperative SQL, or user defined functions (UDF) in C/C++, Java, Python, etc. External programs incur data transfer cost, while code in UDF or stored procedures runs in DBMS environment close to data with performance gains.

Adopting parallel techniques through MPP shared nothing systems, for performance gains with high volume data, originated in DBMS world: first such

commercial MPP system from TeraData was in 1986 [1,23]. Evolution of SQL row DBs into column stores [14], targeting OLAP with better performance, has enhanced their suitability for structured data mining applications.

Ordenez investigates suitability of SQL databases [24] for implementing data mining algorithms and concludes that parallel columnar databases with UDFs can solve important Big Data problems. Row database vendors, Oracle, IBM, Teradata and Microsoft, offer data mining packages tightly coupled to their products modifying internal DBMS code with SQL extensions. As data mining involves development of newer, or modifications to existing, algorithms both needing source code access, such packages are not very popular. Implementations of some of the top 10 algorithms and others exist in user developed SQL.

3.3 Data Mining Algorithms in SQL

k-Means Clustering: The importance of *sufficient statistics*, smaller in size than data, for decoupling mining algorithms from data was highlighted [25], and used in [26] to scale k-Means for large databases. Size in thousands of [26] was scaled to millions [27] on a 4-nodes, parallel row DBMS of TeraData with standard and optimized versions, evaluating performance varying dimensions, clusters and data size. Standard version runs a bunch of SQL statements computing Euclidean distance between points and cluster centroids iteratively until termination. Optimized version improves performance with SQL tricks to reduce joins/groupings, UDFs and uses sufficient statistics of [26], which is defined as a triplet for data of d dimensions and size n to be partitioned into k clusters; sufficient statistics does not eliminate multiple scans due to iterations.

$$N_j = |X_j| \quad \text{cluster } j \text{ size; vector (k x 1) with } n = \sum_{j=1}^k N_j \quad (1)$$

$$L_j = \sum_{i=1}^{N_j} x_i \quad \text{cluster } j \text{ sum; matrix (d x k)} \quad (2)$$

$$Q_j = \sum_{i=1}^{N_j} x_i x_i^T \quad \text{cluster } j \text{ quadratic sum; matrix (d x k)} \quad (3)$$

Using Eqs. (1) to (3), cluster weight W_j , cluster centroid C_j and cluster variance R_j for iteration termination are computed [27] as below:

$$W_j = N_j/n \quad C_j = L_j/N_j \quad R_j = Q_j/N_j - L_j L_j^T / N_j^2 \quad (4)$$

Apriori Association Mining: Association mining algorithm Apriori for market basket problems was programmed in SQL with UDFs [28] on DB2 system. Several alternatives for implementing [29] Apriori in DB2 SQL have also been explored: plain SQL using joins and subqueries, cache-mine, stored procedures and UDFs. Both papers use a non-parallel row DBMS.

Expectation Maximization: EM maximizes loglikelihood and for each point x_i finds its probability for cluster j ; version given in SQL [30] uses Mahalanobis distance on d dimensions and k clusters with C_j being the mean vector of size d and R_j the covariance matrix ($d \times d$) with zeros for off-diagonal elements.

$$\delta_{ij} = (x_i - C_j)^T R_j^{-1} (x_i - C_j) \quad \text{Mahalanobis distance of } x_i \text{ to cluster } j \quad (5)$$

$$P(x_i) = \frac{e^{-\delta_{ij}/2}}{\sqrt{(2\pi)^d |R_j|}} \quad \text{probability of } x_i \text{ for cluster } j \quad (6)$$

Sufficient statistics to improve performance of EM in SQL is suggested in [31].

PageRank: The algorithm that made Google the leader in web search was implemented [32] in SQL with query optimization on 4 parallel nodes of column store Vertica on publicly available real-life data sets (Twitter, Livejournal and YouTube) of large sizes: graphs varying from 81k to 41.6 million nodes, 1.7 million to 1.4 billion edges. They report competitive performance for SQL with NoSQL products GraphLab and Giraph with less system resource utilization for memory and read I/O; extending the comparison of PageRank to mixed graph and relational analysis problems SQL Vertica outperforms Giraph by 17x.

Naive Bayes: Assumes Gaussian classes and independence across dimensions to compute [33,34] sufficient statistics, N_g , L_{gh} and Q_{gh} , for g classes of training set across d dimensions ($h \in 1..d$) like in (1) to (3), and finds class prior π_g , class means C_{gh} and class variance R_{gh} to classify new data by probability $p(x_i)$.

$$\pi_g = N_g/n \quad C_{gh} = L_{gh}/N_g \quad R_{gh} = Q_{gh}/N_g - L_{gh}L_{gh}^T/N_g^2 \quad h \in 1..d \quad (7)$$

$$p(x_{ih}|g)_{h \in 1..d} = \frac{e^{-(x_{ih}-C_{gh})^2/2R_{gh}}}{\sqrt{2\pi R_{gh}}} \quad \text{probability of } x_i, \text{ dimension } h \text{ class } g \quad (8)$$

$$p(x_i|g) = \prod p(x_{ih}|g)_{h \in 1..d} \quad \text{joint probability of } x_i, \text{ all dimensions class } g \quad (9)$$

Final scoring is to class c with maximum probability $p(x_i) = \max(p(x_i|g))$. SQL and MapReduce versions are compared in [34] with better performance for SQL.

kNN: Points in a multidimensional space are mapped [35] to one dimension defining z -value of a point by interleaving binary representation of its coordinates from MSB to LSB. For a point $p_i = (x_i, y_i)$ in 2-d space, its z -value $z_p(x_i, y_i)$ is:

$$z_p(x_i, y_i) = \text{bit}_n(x_i) | \text{bit}_n(y_i) | \text{bit}_{n-1}(x_i) | \text{bit}_{n-1}(y_i) | \dots | \text{bit}_0(x_i) | \text{bit}_0(y_i) \quad (10)$$

where $\text{bit}_k(v)$ is the k th bit of value v . The Z -order of points on z_p is a SQL range query, generally preserving spatial locality. But, for a theoretical guarantee they use random shifts to define a γ -neighborhood and propose algorithms for approximate/exact kNN, distance based θ -join and kNN-joins; analyze complexity, implement in SQL to compare with others (iDistance & Medrank in SQL). z_p is extended to real values, higher dimensions and queries with ad-hoc conditions.

Decision Trees: Two of Top-10, C4.5 and CART, are decision trees, which are greedy, recursive, memory/time intensive algorithms, but intuitive and used widely. Using sufficient statistics (counts: *CC tables*) for splitting, SQL and C++ middleware [36] shows scalable full tree construction with C4.5/CART like entropy measure for selection. Primitives in SQL based on CC tables for C4.5, CART, etc., are given in [37]; C4.5 is implemented [38] as Oracle PL/SQL stored procedure, and decision tree constructed [39] from SQL data cubes.

Graphs Mining: Graph analytics applications, like PageRank, use mining techniques to process graphs. Study in [32] also includes two other graph algorithms in SQL: single source shortest path (SSSP) and HCC to find connected components of a graph. In performance comparison of mixed graph and relational analysis for SSSP on Twitter data, Vertica SQL outperforms Giraph by 4x.

Others: Sufficient statistics is used to build other statistical models in SQL [31], [40]: linear regression/correlation for n variables; dimensionality reduction for preprocessing mining data by principal component analysis (PCA). Regression over 2 variables, multidimensional analysis (cube, rollup, grouping set) and windowing analysis (partitions, order, frames) are part of standard SQL.

4 Platforms and Products

The advent of cloud computing through pay-by-use public services democratizes grid/cluster computing: facilitates scale out, parallel applications and data storage for Big Data processing. Through a browser based GUI any data scientist with web access may harness the power of parallel computing and large stores without recourse to high capital investment or a team of system specialists.

Multiple vendors offer managed IaaS (Infrastructure as a Service) environments with choice of configurations to suit budget and application requirements: Amazon AWS, Microsoft Azure, CenturyLink, Samsung, INAP, Alibaba, etc. Several MPP SQL analytics products are available on public clouds along with competing NoSQL products. Table 3 lists some leading relational SaaS products for Big Data analytics on cloud; all products listed support horizontal scaling.

Table 3 categorizes the listed cloud products on high level criteria for Big Data analytics rather than a detailed evaluation of SQL features support:

1. **DB Store:** Column store has been shown to have better performance for analytics than row store [14]; five of the six products support a native column store with compression; Greenplum is a native row store with restricted, append-only support for column store; dbX is a hybrid store with no serious use-case restrictions: modern column store with compression and row store.
2. **Cloud:** A product available on multiple cloud platforms offers mobility across IaaS platforms and imposes no restriction on the application. Only dbX is cloud agnostic: available on AWS, Azure and other smaller public clouds; Amazon and Microsoft products are tied down to respective vendor clouds; Snowflake is only on AWS; other two on both AWS and Azure.

Table 3. Cloud SaaS: MPP SQL analytics products

Product	DB store	Cloud	Store type	Prem	UDF	Vendor
Redshift	Column	AWS	Attached	No	Yes	Amazon
SQL DW	Column	Azure	Blob+attach	No	Yes	Microsoft
Vertica	Column	AWS, Az	Attached	Yes	Yes	HP/MicroFocus
dbX	Column+row	agnostic	Attach/NW	Yes	Yes	XtremeData
Greenplum	Row+apnd col	AWS, Az	Attach+S3	Yes	Yes	EMC Pivotal
Snowflake	Column	AWS	S3+attach	No	No	Snowflake

3. **Store Type:** Type of cloud storage impacts cost: irrespective of usage, products that run only on attached store must be 24×7 as data is lost on shutdown of compute instances. With network storage (AWS EBS or Azure premium IO), compute and storage are decoupled: shutting down compute instances preserves store for later use. dbX offers services on both store types; Redshift is preconfigured on attached; Vertica/Greenplum use or recommend attached. SQL Datawarehouse and Snowflake, targeting elastic scale-out, use a low cost store with poorer performance (Azure blob or AWS S3) as primary store caching retrieved data on attached store with attendant performance overheads; no shutdown data loss. Greenplum also uses external files on S3.
4. **On Premise:** On premise deployments are sought by users who may not want to store their data on public clouds for security/privacy reasons, or enterprise users who wish to build AaaS (Analytics as a Service) private clouds. Vertica and Greenplum are available as SQL appliances bundled on vendor hardware; dbX may be deployed on commodity clusters and even on other virtualized environments such as VMware; other DBs only on cloud.
5. **UDF:** Table 3 lists UDF support as Sects. 3.2 and 3.3 highlight importance of UDFs for data mining algorithms in SQL. Snowflake is the only product without UDF support; others offer it in different languages: PL/SQL type stored procedures (SQL DW, dbX, Greenplum), C/C++ (Vertica, dbX, Greenplum), Python (Redshift, dbX, Greenplum).

Some products also offer API interfaces to external open source data mining packages such as R and MADLIB, an approach similar to vendor specific mining products. Both Vertica and Greenplum offer customized version of R compatible to their products; additionally, Greenplum SQL may also be used with MADLIB.

5 Discussion

MapReduce, key-enabler of most NoSQL systems, is compared [41] with two SQL parallel databases (Vertica and row DBMS) on clusters of 100 nodes with large, synthetic web crawler type data. The benchmark used 5 tasks; grep task as in [17] and 4 DBMS analysis tasks: selection, aggregation, join and UDF aggregation.

Both DBMSs outperformed MapReduce on all 5 tasks; average values: row store (3.2x); column store (7.4x). Data load was easier and faster on MapReduce.

Based on results of [41], criticisms by Stonebraker et al. [23,41] of MapReduce include (1) repetitive record parsing as data is stored in text form (2) lack of compression advantage: slowing down with block/record level compression (3) pull model of Reduce for data exchange with Map (4) absence of plan optimization (5) lack of high level interfaces and developer eco-system (6) schema less world. They surmise that MapReduce “*is more like an extract-transform-load (ETL) system*” [23] and hence complementary, rather than competitive, to DBMSs. Basing their comparison on a technical evaluation, they perhaps understate the importance of MapReduce framework, Hadoop or GFS, which simplifies distributed application development, by managing everything including failures.

Mohan, inventor of ARIES recovery fundamental to ACID transactions, criticizes [42] NoSQL for oversimplifying complex issues with ad-hoc solutions, being expedient not rigorous, missing interactive query support and ignoring history.

Challenges for MapReduce in Big Data include [43]: (1) data storage issues without schemas (2) coding iterative analytics algorithms in MapReduce (3) performance overheads with correlated data for predictive modeling (4) harder interactive data exploration without high level interfaces like SQL (5) same issues as SQL DBs in low latency applications (6) lack of security and privacy features and its legal impact for proposed privacy regulations.

Examples of on-going work to mitigate the challenges: integration with SQL DBMSs such as Oracle and Greenplum, Spark and HaLoop to deal with iterative algorithms, Storm for low latency applications, SQL like Hive for interactive analytics, Mahout for data mining, etc. It appears that NoSQL systems are evolving like DBMSs into vertically segmented engines to address Big Data.

Revisiting CAP theorem, *raison d'être* of NoSQL systems, its proposer Brewer considers “*‘2 of 3’ formulation was misleading because it tended to oversimplify the tension among properties*” [44], and proposes alternatives to deal with partition tolerance. Stressing consistency-latency trade-off [45] suggests that CAP’s 2-of-3 limitation, applicable only in the context of failures, has been misinterpreted to build limited systems. NewSQL prefers ACID to BASE.

Parallel database theory questions characterization of Big Data through 3Vs and suggests alternative dimensions: communication, iteration and failure [46]. Despite high scale-out of a few NoSQL products, the comprehensive survey [15] discusses issues and limitations in horizontal scaling of other NoSQL products.

Social factors too contribute to perception of SQL not being suitable for Big Data analytics: (1) unlike parallel DBMSs, almost all NoSQL products are open source with no cost (2) limited mathematical exposure of programmers hampers translation of complex and iterative algorithms into declarative SQL.

The 18th KDnuggets poll of 2017 (www.kdnuggets.com) lists percentages of 2900 voters for language use: Python (52.5), R (52.1), SQL (34.9), Java (13.8), C/C++ (6.3); SQL holds one third share! Big Data users [15] of MySQL: Facebook for social graph data, Wikipedia on MariaDB, open source fork of MySQL.

6 Conclusion

It is an axiomatic fact that Big Data and its analysis are decision making drivers in a 21st century world driven by web, mobile and IoT technologies. To understand why 20th century *numero uno* tool for data processing, SQL rdbms, has lost its primacy we have briefly summarized its shortcomings that led to birth of competing technologies, NoSQL and NewSQL, and traced their evolution.

Focusing on the important aspect of analytics in Big Data processing, we examined suitability of SQL relational databases for data mining, and presented published work on data mining algorithms in SQL; majority of the top ten data mining algorithms and a few others in SQL solve Big Data analytics problems on parallel, columnar DBMS with UDFs. Cloud deployments of such products makes them more accessible, at lower cost and easy scale-out, even elastic.

Comparative discussion in Sect. 5 shows that no technology is fully ready for all challenges of Big Data, more so for IoT and Industry 4.0 that could possibly use blockchain based P2P networks for devices [3, 47] along with multiple options on cloud: NewSQL stream DBMSs, NoSQL Storm, or even parallel, columnar MPP DBMSs fed by distributed streaming platform Apache Kafka.

We observe a convergence of technologies, and note that relational model and SQL are unlikely to disappear, a view endorsed by [19]: “*all of the key systems in these groups will support some form of relational model and SQL*”.

References

1. Chen, M., Mao, S., Liu, Y.: Big data: a survey. *Mob. Netw. Appl.* **19**, 171–209 (2014)
2. Khan, N., et al.: Big data: survey, technologies, opportunities, and challenges. *Sci. World J.* **2014**, 18 (2014). Article ID 712826
3. English, M., Auer, S., Domingue, J.: Block chain technologies & the semantic web: a framework for symbiotic development. In: Lehmann, J., et al. (ed.) *Computer Science Conference for University of Bonn Students*, pp. 47–61 (2016)
4. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*, 3rd edn. Morgan Koffman Publishers, Elsevier (2012)
5. Leskovec, J., Rajaraman, A., Ullman, J.: *Mining of Massive Datasets*, 2nd edn. Cambridge University Press, Cambridge (2014)
6. Elgendy, N., Elragal, A.: Big data analytics: a literature review paper. In: Perner, P. (ed.) *ICDM 2014. LNCS (LNAI)*, vol. 8557, pp. 214–227. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08976-8_16
7. Che, D., Safran, M., Peng, Z.: From big data to big data mining: challenges, issues, and opportunities. In: Hong, B., Meng, X., Chen, L., Winiwarer, W., Song, W. (eds.) *DASFAA 2013. LNCS*, vol. 7827, pp. 1–15. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40270-8_1
8. Elragal, A.: ERP and big data: the inept couple. In: *CENTERIS 2014, Procedia Technology*, vol. 16, pp. 242–249. Elsevier (2014)
9. Stonebraker, M., Cetintemel, U.: “One size fits all”: an idea whose time has come and gone. In: *ICDE 2005*, pp. 2–11. IEEE (2005)
10. Copeland, G.P., Khoshafian, S.N.: A decomposition storage model. In: *SIGMOD 1985*, pp. 268–279. ACM (1985)

11. Boncz, P., Kersten, M.L., Manegold, S.: Breaking the memory wall in MonetDB. *Commun. ACM* **51**(12), 77–85 (2008)
12. Idreos, S., et al.: MonetDB: two decades of research in column-oriented database architectures. *IEEE Data Eng. Bull.* **35**(1), 40–45 (2012)
13. Stonebraker, M., et al.: CStore: A Column Oriented DBMS. In: *VLDB 2005*, pp. 553–564 (2005)
14. Abadi, D., Boncz, P., Harizopoulos, S., Idreos, S., Madden, S.: The Design and implementation of modern column oriented database systems. *Found. Trends Database* **5**(3), 197–280 (2012)
15. Strauch, C.: NoSQL databases. In: *Selected Topics on Software-Technology Ultra-Large Scale Sites*, pp. 1–149. Stuttgart Media University (2011). <http://www.christof-strauch.de/nosql dbs.pdf>
16. Brewer, E.: Towards robust distributed systems. In: *19th Symposium on Principles of Distributed Computing, PODC 2000*, pp. 7–10. ACM (2000)
17. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: *OSDI 2004*, pp. 137–149. USENIX (2004)
18. Stonebraker, M., et al.: The end of an architectural era: (its time for a complete rewrite). In: *VLDB 2007*, pp. 1150–1160 (2007)
19. Pavlo, A., Aslett, M.: What’s really new with NewSQL? *SIGMOD Rec.* **45**(2), 45–55 (2016)
20. Bernstein, P.A., Goodman, N.: Multiversion concurrency control: theory and algorithms. *ACM Trans. Database Syst.* **8**, 465–483 (1983)
21. Xu, R., Wunsch, D.: Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **16**(3), 645–678 (2005)
22. Wu, X., et al.: Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**, 1–37 (2008)
23. StoneBraker, M., et al.: MapReduce and parallel DBMSs: friends or foes? *Commun. ACM* **53**(1), 64–71 (2010)
24. Ordonez, C.: Can we analyze big data inside a DBMS? In: *Proceedings of 16th International Workshop on Data Warehousing and OLAP, DOLAP 2013*, pp. 85–92. ACM (2013)
25. Graefe, G., Fayyad, U., Chaudhuri, S.: On the efficient gathering of sufficient statistics from large SQL databases. In: *KDD 1998*, pp. 100–105. AAAI (1998)
26. Bradley, P.S., Fayyad, U., Reina, C.: Scaling clustering algorithms to large databases. In: *KDD 1998*, pp. 9–15. AAAI (1998)
27. Ordonez, C.: Programming the K-means clustering algorithm in SQL. In: *KDD 2004*, pp. 823–828. AAAI (2004)
28. Agrawal, R., Shim, K.: Developing tightly coupled data mining applications on a relational database system. In: *KDD 1996*, pp. 287–290. AAAI (1996)
29. Sarawagi, S., Thomas, S., Agrawal, R.: Integrating association rule mining with relational database systems: alternatives and implications. In: *SIGMOD 1998*, pp. 343–354. ACM (1998)
30. Ordonez, C., Cereghini, P.: Fast clustering in SQL using the EM algorithm. In: *SIGMOD 2000*, pp. 559–570. ACM (2000)
31. Ordonez, C.: Statistical model computation with UDFs. *IEEE Trans. Knowl. Eng.* **22**(12), 1752–1765 (2010)
32. Jindal, A., Madden, S., Castellanos, M., Hsu, M.: Graph analytics using the vertical relational database. In: *IEEE Conference on Big Data*, pp. 1191–1200. IEEE (2015)
33. Ordonez, C., Pitchaimalai, S.K.: Bayesian classifiers programmed in SQL. *IEEE Trans. Knowl. Eng.* **22**(1), 139–144 (2010)
34. Pitchaimalai, S.K., et al.: Comparing SQL & MapReduce to compute naive bayes in a single table scan. In: *CloudDB 2010*, pp. 9–16. ACM (2010)

35. Yao, B., Li, F., Kumar, P.: K nearest neighbor queries and KNN-joins in large relational databases (almost) for free. In: ICDE 2010, pp. 4–15. IEEE (2010)
36. Chaudhari, S., Fayyad, U., Bernhardt, J.: Scalable classification over SQL databases. In: 15th ICDE 1999, pp. 470–489. IEEE (1999)
37. Sattler, K.-U., Dunemann, O.: SQL database primitives for decision tree classifiers. In: CIKM 2001, pp. 379–386. ACM (2001)
38. Taniar, D., D’Cruz, G., Rahayu, J.W.: Implementation of classification rules using Oracle PL/SQL. In: FSKD 2002, pp. 509–513 (2002)
39. Fu, L.: Construction of decision trees using data cubes. In: 7th ICEIS, pp. 119–126 (2005)
40. Ordóñez, C., García-Alvarado, C.: A data mining system based on SQL queries and UDFs for relational databases. In: CIKM 2011, pp. 2521–2524. ACM (2011)
41. Pavlo, A., et al.: A comparison of approaches to large scale data analysis. In: SIGMOD 2009, pp. 165–178. ACM (2009)
42. Mohan, C.: History repeats itself: sensible and NonsenseSQL of the NoSQL hoopla. In: Proceedings of EDBT/ICDT 2013, pp. 11–16 (2013)
43. Grolinger, K., et al.: Challenges for MapReduce in big data. In: SERVICES 2014, pp. 182–189. IEEE (2014)
44. Brewer, E.: CAP twelve years later: how the “rules” have changed. *IEEE Comput.* **45**(2), 23–29 (2012)
45. Abadi, D.: Consistency tradeoffs in modern distributed database system design. *IEEE Comput.* **45**(2), 37–42 (2012)
46. Suciu, D.: Big data begets big database theory. In: Gottlob, G., Grasso, G., Olteanu, D., Schallhart, C. (eds.) BNCOD 2013. LNCS, vol. 7968, pp. 1–5. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39467-6_1
47. Brody, P., Pureswaran, V.: Device democracy: saving the future of the Internet of Things. In: IBM Institute for Business Value Technical report, p. 25. IBM (2014)

Intelligent Electronics and Systems for Industrial IoT



Rethinking ‘Things’ - Fog Layer Interplay in IoT: A Mobile Code Approach

Behailu Negash¹(✉), Tomi Westerlund¹, Pasi Liljeberg¹,
and Hannu Tenhunen^{1,2}

¹ Department of Information Technology, University of Turku, Turku, Finland
{behneg,tovewe,pakrli}@utu.fi

² Department of Industrial and Medical Electronics, KTH Royal Institute
of Technology, Stockholm, Sweden
hannu@kth.se

Abstract. A client-server architecture style is one of the common approaches enabling separation of concerns in distributed systems. In the Internet of Things architecture, this approach exists in different configuration of sensors, actuators, gateways in the Fog layer and servers in the Cloud. This configuration affects the degree of interoperability, scalability and other functional and non-functional system requirements. In this paper, we reflect on best practices in the web and REST style to address IoT challenges; one of the constraints in REST, Code on Demand, is used for IoT to enhance the flexibility and interoperability of resource constrained clients at the perception layer. Scripts written in a domain specific language, DoS-IL, are organized and stored at the Fog layer for sensor and actuators nodes to request and execute the incoming script. A generic application layer protocol and RESTful server are presented along with experimental results.

Keywords: Internet of Things · Architecture · Interoperability
Fog computing · Scalability · DoS-IL · Programmability

1 Introduction

The growing number of cheaper, smaller and embedded devices connected to the Internet is fueling the development of what is known as the Internet of Things (IoT). IoT is expected to bring billions of devices online enabling a wide range of possibilities for everyday life. This has two main perspectives merged as one: the connectivity of the devices and the application running on top of it. In contrast to the development of the current Internet and the World Wide Web, IoT is developing as one it is creating confusion in naming and its vision [1]. This paper reflects on previous approaches and techniques used to address the challenges faced during the development of the Internet and the Web to learn from and adapt it for IoT. The World Wide Web is planned to be a virtual space where

people can interact and information can be stored [2]. This is enabled by the underlying interoperable connectivity provided by the Internet. However, before reaching such level of scale and interoperability, both the Web and the Internet has passed through competing heterogeneous platforms and protocols. The lack of interoperability is even worse in the Internet of Things; diverse communication protocols, architecture, data format and middleware exist. This limits the level of connectivity possible in IoT, and hence the application running over it are usually vertical silos. To bridge the silos and build a global system, a clear IoT vision is required to move forward.

The IoT vision is still evolving and there are contributions from academia, industry and government institutions [1]. Singh *et al.* [3] summarize the IoT vision in three parts; the Internet vision, Things vision and Semantic vision. In general, an IoT vision can be put as smart objects capable of sharing meaningful information over the Internet. This is inline to the vision of the Web, to be shared information space for machines and people [2]. The main contribution of this paper is towards the realization of this vision via the extended architecture and tools discussed here. We start with a general discussion of the nature of the web as the largest distributed system and learn from it. Distributed systems is a general name used to refer to systems that run in multiple separate boundaries, physical or logical, and communicate to accomplish their task. Part of these systems naturally demand for physical separation of its components on to multiple devices - the Web and IoT are typical ones; a server component provides service for a client connected through the Internet. The role of the server in formatting (HTML), and locating (URL) the information requested, and the standard of the communication (HTTP) are the driving reasons for the success of the Web.

An architectural enhancement to the original design of the web came through REST (Representational State Transfer) architecture style [4]. One of the optional constraints put in this style is the code on demand (COD) approach. COD is the process of sending a code for a client to execute, which is a specific case of mobile code pattern [5]. In this paper, we explore mobile code style for the Internet of Things through a semantically organized set of scripts and a novel server in the Fog layer to enable interoperability and enhance the scaling of IoT towards its vision. Devices embedded in physical objects send requests to the server in the Fog layer for instructions on how to present their sensor data or alter built-in actuators. An IoT-lite ontology [6] is used to annotate the devices in our experiment and DoS-IL (Domain Specific IoT Language) [7], an IoT domain specific language is used for the scripts. In summary, the main contributions of this paper are:

- Semantically organized domain specific scripts
- An abstract application layer protocol for communication
- REST like uniform interface and a service bundled together

The rest of the paper is organized in the following manner. Section 2 present the challenges in the Internet of Things, a motivation for this work and state of the art in this area, followed by details of the main work of the paper in

Sect. 3. The results of the implementation and evaluation of the performance are reported in Sect. 4. The final section concludes the paper with the summary of results and discussion of planned continuation of the work.

2 Motivation and Challenges

This research aims to address some of the challenges facing the IoT, such as reconfiguration, interoperability and scalability. To clarify on these difficulties and motivate the work, we consider a simplified IoT use case in a remote patient monitoring in a smart home system and highlight the integration challenges and proposed potential solutions for such systems. There exist multiple challenges to reach an Internet scale and interplay with the state of the art method. Most IoT systems exist as vertical silos forming boundaries of device architecture, programming approach, network protocol and data formats among others. Each application works independently and is connected to the Internet allowing users to interact remotely. The majority of these devices are battery powered, has small memory and limited processing power. Moreover, the network interface associated usually has lower bandwidth and follow different protocols. Looking at the ‘things’ vision of IoT [3], one expects to have a seamless integration across these boundaries with proper authentication. Similarly, the ‘Internet’ vision promises to deliver an Internet scale system of smart systems. Some of the challenges, such as scale and heterogeneity, hindering the realization of this vision are presented by Zorzi *et al.* [8] along with a proposed solution from protocol perspective. Another work presented in [9] shows a horizontal architecture for IoT to help manage the challenge of interoperability and ease of programmability using a software defined networking scheme. At the highest level of abstraction of the systems, the data format and semantic knowledge of the exchanged information has to match for the systems to interoperate. Moreover, the architecture of the systems is a key component for integration. There has been many contributions from industry, academia and public projects to close the gap and hide the heterogeneity. An open survey shows some of these challenges and solutions [10]. One of the proposed solutions is IoT-A [11] - a project aimed to address these challenges with a reference architecture. Another approach is using middleware to bridge the gap in such systems. Razzaque *et al.* [12] present a survey of some of the middleware proposals. In the following sections, we highlight our approach in addressing this challenges and present the experimental results obtained.

3 Enabling Code on Demand

The introduction of Representational State Transfer (REST) as an architectural style in the web simplified the development and consumption of services in distributed systems across the globe. One of the constraints in REST in the area of mobile code architecture style is Code on demand (COD) [4]. It enables a client node to extend its feature through an executable code sent from the server. Code on Demand is one of the ways code mobility is achieved. Fuggetta *et al.* [5]

discuss some of the benefits and uses of mobile code. Our approach resembles the case of remote device control and configuration in [5], where we allow the reconfiguration of devices in the perception layer with DoS-IL [7]. A script stored in the gateways at the Fog layer will be sent to a device on demand to be executed. The result of the execution is also transferred to the gateway via a generic application layer protocol running over heterogeneous network interfaces. To allow this, the overall architecture of the system and its details are discussed in the following subsections.

3.1 System Architecture

Architecture plays a critical role in achieving the desired functional and non-functional requirements of a system. Currently there are two main approaches of connecting devices to a gateway, regardless of the communication protocol: writing the program to read sensors, format, process and send it, or make a service to listen and handle incoming requests. This is similar to push and pull form of communication. In both cases, the node is rich in feature, it has the resources needed as well as the know how to manage it. The role of the gateway in the Fog layer is to provide connectivity and handle incoming or pulled data from the perception layer. In these approaches, maintenance of the application becomes a challenge after deployment. Moreover, the approaches are not scalable to the Internet level as in the case of the Web. For the second case, the nodes are usually resource constrained (processing power, battery or memory) to run a service that is available at anytime. Our approach is different in that the nodes at the perception layer still contain the resources (sensors, actuators or tags), but lack the know how which is the main point of functional or non-functional changes. This know how is written and stored in the Fog layer and sent to the node on a GET request. Figure 1 shows a generic three tier deployment architecture of an IoT system and the different components mapped on it.

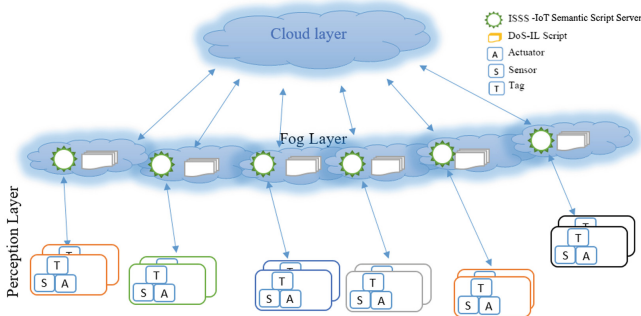


Fig. 1. High level deployment architecture of the proposed system

3.2 Script Organization

The scripts that are stored in the server are written using a domain specific language called DoS-IL. Each one of these scripts has a name and address that is resolved from the domain knowledge represented via an extended annotation using IoT-lite ontology [6] as shown in Fig. 2. Devices have unique names that gets resolved into a URL of the location of the script using a sparql query from the ontology. The ontology is formatted in RDF (Resource Description Format) and stored in the gateway representing part of the Fog layer. This enables easy replication and merging of the ontology in other gateways when necessary. To parse and query the ontology, rdfliib - a python library is used for sparql parsing. In a hierarchical Fog architecture, as proposed by the OpenFog consortium [13], the segments of the ontology can easily be replicated to multiple nodes for easy scaling of the system. This will be part of our future extension of this work towards demonstrating our proposal over distributed hierarchical gateways.

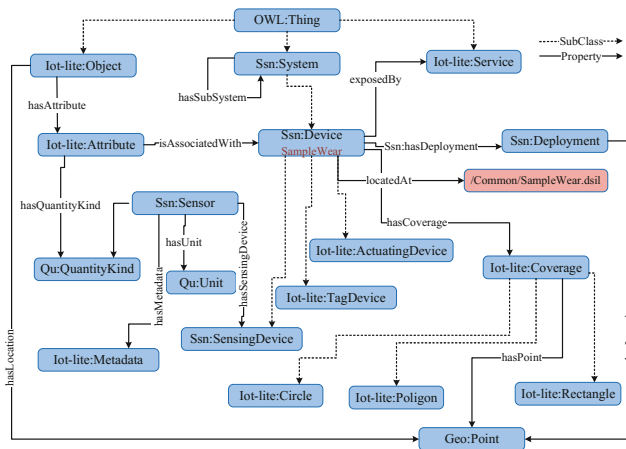


Fig. 2. IoT-lite ontology modified for the server

3.3 Generic Application Layer

One of the components that contributed to the success of the World Wide Web, besides the addressing means and the markup language, is Hypertext Transfer Protocol (HTTP). As an application layer protocol, it runs over the standard TCP/IP protocol. However, in Internet of Things, the majority of the devices, especially those in the perception layer, use various low power communication protocols. Our semantic server is designed to handle incoming requests encoded with this generic application protocol over heterogeneous network standards. An abstract class representing network interfaces is defined, which needs to be implemented for each supported network protocol. Request and response headers are only few bytes to work with networks of low bandwidth. The request header

is two bytes long and it is organized as follow: first four bits indicate the version number, followed by two bits for the verb of the request (REST verbs), one bit to show what type of script format the client accepts and one status bit for notification of script changes. The second byte is for payload size, which is set to 0 for GET requests. In case of the response, the response code takes the first byte followed by the remaining packets in the message and the checksum - a total of three bytes. The response codes and the header formats are shown in detail in the source code repository [14].

3.4 IoT Semantic Script Server

A uniform interface is defined for the communication channel between the clients and the IoT Semantic Script Server (ISSS). Like a RESTful service, a GET request is used for the request of the configuration script while POST is used to send the data from the device for processing at the gateway. The method (GET or POST) is specified in the request header of the generic application layer protocol. Depending on the header, a specific request handler is passed the request to process it. Whenever a GET request is received, the request handler resolves the requested name from the IoT-lite ontology to a proper script location and fragment the script based on the available bandwidth. This server instantiates and listens to all the available network interfaces through the concrete implementations of the abstract class. Regardless of the underlying network protocol, the function of the server and the application layer protocol remains the same. The details of the implementation of our server and its components are shown in more details in the following section.

4 Demonstration and Evaluation

To validate our proposal beyond a conceptual point, we have developed a simplified scenario using our python based server implementation. Figure 3 shows

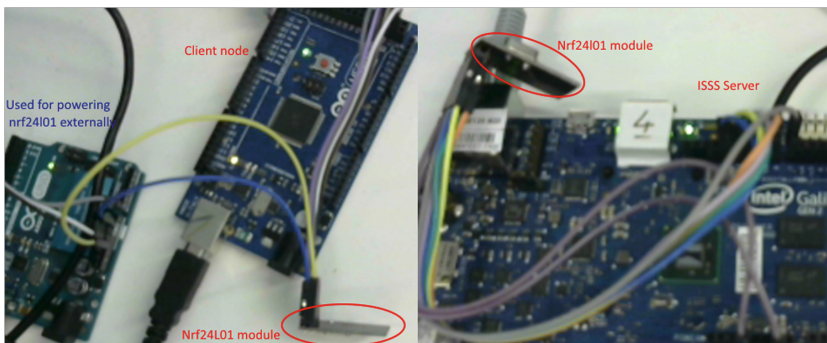


Fig. 3. Simplified demonstrator implementation

a generic application protocol and organized it using a lightweight IoT ontology. An implementation of the generic network interface for nRF24L01 based network is developed and performance evaluation is carried out. The evaluation shows the fragmentation of the script for sending over a low bandwidth network, the round trip time and the performance of the process of resolving a device name to a script path. The initial version of the application layer protocol is also very light and optimized for code on demand approach. In summary, this paper introduced a novel architecture for IoT, with the help of a domain specific language, that enables both semantic and syntactic interoperability and facilitate the growth of IoT to an Internet scale. To further extend this work, we plan to make the service implementation to listen to multiple network interfaces and work with distributed script locations over multiple Fog gateways. Furthermore, we believe that inclusion of standards from some of the big consortium make our efforts comprehensive in pushing forward the integration effort of IoT. One of this standards considered for future works is the data format standard from the open interconnect consortium.

References

1. Minerva, R., Biru, A., Rotondi, D.: Towards a definition of the Internet of Things (IoT). IEEE, Technical report (2015)
2. Berners-Lee, T.: WWW: past, present, and future. *Computer* **29**(10), 69–77 (1996)
3. Singh, D., Tripathi, G., Jara, A.J.: A survey of Internet-of-Things: future vision, architecture, challenges and services. In: 2014 IEEE World Forum on Internet of Things (WF-IoT), pp. 287–292, March 2014
4. Fielding, R.T.: Architectural Styles and the Design of Network-based Software Architectures. Ph.D. dissertation, University of California, Irvine (2000)
5. Fuggetta, A., Picco, G.P., Vigna, G.: Understanding code mobility. *IEEE Trans. Softw. Eng.* **24**(5), 342–361 (1998)
6. Bermudez-Edo, M., Elsaleh, T., Barnaghi, P., Taylor, K.: IoT-lite: a lightweight semantic model for the Internet of Things. In: 2016 International IEEE Conferences on UIC/ATC/ScalCom/CBDCCom/IoP/SmartWorld, pp. 90–97, July 2016
7. Negash, B., Westerlund, T., Rahmani, A.M., Liljeberg, P., Tenhunen, H.: DoS-IL: a domain specific Internet of Things language for resource constrained devices. *Procedia Comput. Sci.* **109**, 416–423 (2017). <http://www.sciencedirect.com/science/article/pii/S1877050917310876>
8. Zorzi, M., Gluhak, A., Lange, S., Bassi, A.: From today's Intranet of Things to a future Internet of Things: a wireless- and mobility-related view. *IEEE Wirel. Commun.* **17**(6), 44–51 (2010)
9. Li, Y., Su, X., Riekkki, J., Kanter, T., Rahmani, R.: A SDN-based architecture for horizontal Internet of Things services. In: 2016 IEEE International Conference on Communications (ICC), pp. 1–7, May 2016
10. Eclipse IoT Working Group: IoT developer survey. IEEE, IoT Eclipse, Agile, Technical report (2016)
11. IoT-A Project, Internet of things - architecture, IoT-A, deliverable d1.5 - final architecture reference model for the IoT v3.0, EU-FP7, Technical report (2013). <http://www.iot-a.eu/public/public-documents/d1.5/view>

12. Razzaque, M.A., Milojevic-Jevric, M., Palade, A., Clarke, S.: Middleware for Internet of Things: a survey. *IEEE Internet Things J.* **3**(1), 70–95 (2016)
13. OpenFog Consortium: OpenFog Reference Architecture for Fog Computing. OpenFog Consortium, Technical report (2017)
14. Negash, B.: ISSS implementation. <https://github.com/behailus/ISSS>



A Security Framework for Fog Networks Based on Role-Based Access Control and Trust Models

Farhoud Hosseinpour^{1(✉)}, Ali Shuja Siddiqui², Juha Plosila¹,
and Hannu Tenhunen¹

¹ Department of Future Technologies, University of Turku, Turku, Finland
{farhos, juplos, hatenhu}@utu.fi

² Department of Electrical and Computer Engineering,
University of North Carolina at Charlotte, Charlotte, USA
asiddiq6@uncc.edu

Abstract. Fog networks have been introduced as a new intermediate computational layer between the cloud layer and the consumer layer in a typical cloud computing model. The fog layer takes advantage of distributed computing through tiny smart devices and access points. To enhance the performance of the fog layer we propose utilization of unused computational resources of surrounding smart devices in the fog layer. However, this will raise security concerns. To tackle this problem, we propose in this paper a novel method using a trust model and Role Based Access Control System to manage dynamically joining mobile fog nodes in a fog computing system. In our approach, the new dynamic nodes are assigned non-critical computing tasks. Their trust level is then evaluated based on the satisfaction rate of assigned tasks which is obtained through different computing parameters. As the result of this evaluation, untrusted nodes are dropped by the fog system and nodes with a higher trust level are given a new role and privileges to access and process categorized data.

Keywords: Fog computing · Cloud · Access control · Trust model

1 Introduction

The benefits achievable by deploying scalable applications serving a large number of users simultaneously are rapidly generating novel innovations and expanding the reach of cloud computing. The cloud computing has replaced the need for owning large private data centers for service providers who want to deploy their projects with minimum infrastructure cost [1]. Cloud computing provides scalability for applications in manifold by enabling addition and removal of processing nodes at runtime as needed. Although cloud computing has deemed itself useful in many scenarios [2], it is not viable for applications that require low latency

and predictable feedback such as Smart Grids, industrial automation systems or intelligent transport systems [3]. This is due to the fact that systems in a cloud service are geographically distributed. For the alleviation of this issue, “fog computing” [4] has been introduced as a complementary concept to cloud computing. Cloud computing can be defined using a layered computation model. Typically there are two layers: a cloud layer and a consumer layer. Recently a new computational layer, called a fog layer, has been introduced to the model. The fog layer resides between the cloud and consumer layers in the network’s edge nodes like sensors and Internet-of-Things devices. Fog computing introduces the concept of location to cloud computing where traditionally non-locational computing has been dominant. Additionally, the fog layer also provides extra computational resources to the cloud layer.

Fog computing is currently an evolving new technology which aims to supplement already established cloud computing platforms to expand their application domain. Fog computing provides a location based expansion of the cloud by using heterogeneous computing devices and access points to which end nodes connect to communicate with the cloud. Bringing the computational intelligence geographically near to the end users provide new or better services for latency sensitive, location-aware and geo-distributed applications that due to their characteristics are not feasible merely through cloud computing. Delegating some simple yet frequent tasks of the cloud to the fog results in better performance for IoT-based applications [5]. In this paradigm, intelligent networking devices with both computation and storage capabilities, i.e., intelligent routers, bridges, and gateways, compose the fog computing platform near to the edge of the network. However, such devices are resource constrained and have computing and storage limitations.

Increasing computing capabilities of fog computing is a major challenge to improve the Quality of Service (QoS). To this end, one possible way is to leverage processing and storage capabilities of surrounding smart devices [6]. Smart devices have become an ubiquitous part of modern life. According to the Global Internet Phenomena Report Spotlight 2016 from Sandvine, the Waterloo-based broadband network equipment company in North America [7,8]: “*The average household was found to have at least seven active, connected devices in use every day, while at the top end of the spectrum, 6% of households tuned in with more than 15 active devices, a marked increase over previous years. Whereas home roaming via mobile devices such as tablets and smartphones accounted for only nine percent of traffic five years ago, it now represents almost 30% of home internet traffic across North America.*” Falaki et al. [9] developed a tool called SystemSens and investigated resource usage such as CPU, memory, and battery in smartphones. According to this study, except for the pick time between 11:00 to 17:00 the average CPU usage in all tested users is below 50%. This amount drops to less than 20% during the night time between 00:00 to 8:00.

Having this motivation, leveraging the available computing power of numerous different smart devices will enhance fog computing. However, utilizing resources of such devices for fog computing will impose some security challenges.

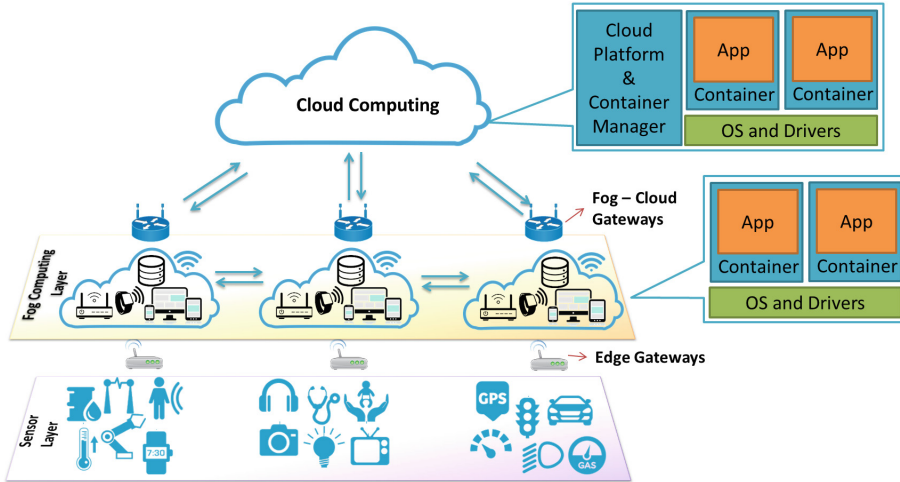


Fig. 1. Fog computing platform.

In this paper, we present a novel approach to tackle this problem by leveraging containerization technology to provide isolation for fog computing tasks in external smart devices. This is further supported by role-based access control and trust models.

The rest of the paper is organized as follows. In Sect. 2, we give an overview of the newly emerging technology of fog computing. In Sect. 3, related works to access control mechanisms for smart devices are presented and discussed. Then, in Sect. 4, we present our proposed framework. The results and discussion on the implemented model are presented in Sect. 5, and, finally, concluding remarks are given in Sect. 6.

2 Application of Fog Computing

Fog computing introduces an intermediate layer between the edge network or the end nodes and the cloud layer (Fig. 1). The fog layer can be implemented using the same components as the cloud layer. The fog layer provides computation in a geographical location. It aims to provide a computing layer physically closer to the end node so that the computing capabilities can be brought near to consumers. The expected benefit is obtaining faster computation times for requests that require low latency. This can play an advantageous role in promotion of the Internet of Things (IoT) [10]. Utilizing fog computing reduces the overhead of communication with cloud through internet and provides a faster response for applications which require lower latency. This is made possible by locally executing such processes in the fog layer and forwarding only those which do not require real-time computation or require higher processing power to the cloud layer. Schulz et al. [3] have investigated different latency critical IoT applications.

According to their study, factory automation applications have the highest critical latency requirements in the range of 0.25 to 10 ms. Process automation, Smart Grids and intelligent transport systems are in the next place in their ranking.

In addition to the requirement of low latency, fog computing as middleware can pre-process raw data coming from the edge nodes before sending them to the cloud. Cloud computing, dealing with Big Data [11], has to process large amounts of data at any time. As a result, the fog layer not only reduces the amount of work needed in the cloud to generate meaningful results, but it can also reduce the monetary cost of computing in the cloud layer.

2.1 Fog Layer Structure

The most important and beneficial aspect of fog computing is the location proximity to the end nodes. The fog layer can be deployed on intelligent access points and gateways that not only connect the edge nodes to the cloud layer but also provide additional computing resources near the edge of the network. In addition to that, independent computing nodes such as smart devices can be added to the fog layer for the sole purpose of computation. Fog nodes can connect to each other to form a mesh. This can also be envisioned as a peer-to-peer (P2P) network with either centralized master controllers or a decentralized implementation without any controllers. Fog nodes cooperate and pool their resources to complete a task. The fog layer can be a dynamic network because some nodes might dynamically join and leave the network due to mobility or power limitations. Or, the other way around, the edge sensors might be mobile and move from one local fog network to another. A robust orchestration system is required to manage the execution of applications in such a dynamic environment without violating QoS and security.

Virtualisation: To support multi-tenancy of different applications and to achieve elasticity in large-scale shared resources, fog computing takes advantages of virtualization technologies. A physical fog node can accommodate several virtual fog nodes. A fog computing platform is composed of several physical and virtual fog nodes that are deployed based on a hierarchical architecture [12] (Fig. 2). Virtualisation technology based on Virtual Machines (VM) is not efficient or even feasible approach for resource constrained fog computing nodes. Containers are a new lightweight alternative for traditional VMs that are ideal for a fog computing platform. Containers provide OS level virtualisation without a need for deployment of a virtual OS. Hence, they are lightweight and significantly smaller in size than VMs. Containers provide a self-contained and isolated computing environment for applications and facilitate lightweight portability and interoperability for IoT applications [13]. Moreover, data and resource isolation in containers offers improved security for the applications running in fog nodes.

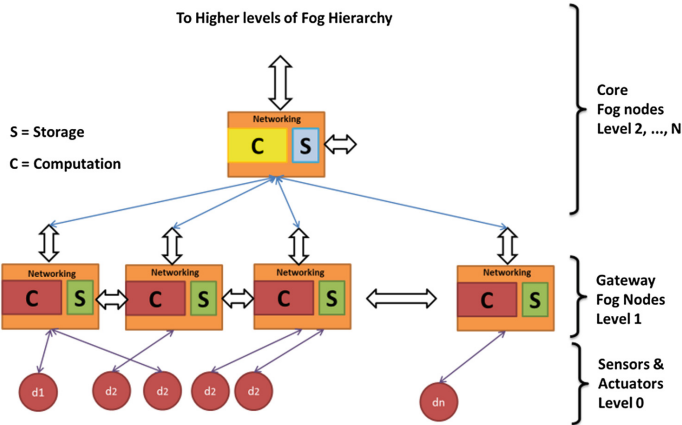


Fig. 2. Architecture of fog computing.

3 Related Works

Fog computing was introduced in 2012 by Cisco [4] as an additional computing layer near to the edge of the network, to complement cloud computing services. We discussed challenges for adoption of this technology in IoT applications as well as its security issues in our review paper [12]. Due to location proximity to the edge of the network, mobility of edge sensors, and also resource limitations, enabling scalable, flexible, and real-time strategies for resource allocation is very challenging. Oueis et al. [14] proposed a cluster-based resource allocation scheme for fog computing in which a cluster of fog computing resources is logically built depending on the profile of computation offloading request from an IoT device or a fog node. Yi et al. [15] investigated security and privacy issues of fog computing and pointed out that unlike the cloud computing that the cloud service provider owns all computing devices, fog computing is more flexible to leverage different computing resources belong to different parties. This flexibility adds more complexity in the terms of trust management and security. Misra and Vaish [16] proposed a cluster-based and multilevel hierarchical architecture for Wireless Sensor Networks (WSN) to establish an authentication mechanism. They deployed a multi-level access control system for each logical cluster using Role-Based Access Control (RBAC) model. They proposed a reputation-based trust model to assign a role for a node and form the logical cluster in WSN. They calculated the reputation value based on the behaviour of a node for successful transmission of data. Salonikias et al. [17] addressed access control issues in fog computing for an intelligent transport system as a case study. They pointed out that fog computing has dispersed nature and sensors can enter and leave the network arbitrarily or the other way around, fog nodes could also be mobile. Hence, traditional identity-based authentication is not a feasible approach in this case. To cope with this problem, they proposed utilizing Attribute-Based Access

Control (ABAC) model in which the authentication is based on the attributes of the subject (in this case, fog node) trying to access a data rather than their identity. In [18] the author discussed the importance of granting access based on the level of trust to individuals. They innovated a mobile device called MS-Ro-BAC for implementation of role-based access control. The MS-Ro-BAC manages access and network authorizations through of role-based access control with no dedicated hubs, servers, special hard-drives or local administrators.

4 Proposed Framework

In this paper, we propose a framework for secure utilization of surrounding smart devices' processing capabilities in a fog computing platform. We use containers as virtualization technology in our fog platform. The reason for this is that they are: (1) lightweight and require less computing and storage, (2) easily portable, (3) platform independent and provide interoperability in a heterogeneous network and, (4) provide isolation of the application that utilize shared resources, which results in better security. We also design and develop an access control system based on the RBAC model to provide authentication for dynamic fog nodes joining the fog computing network. We consider three different kinds of fog nodes according to their capabilities and trust levels. As discussed earlier, a typical fog network is composed of smart communication nodes with the capability of acting as access points. This way they can communicate with edge sensors and also forward preprocessed data to an upper level in the cloud. Also, we propose utilization of dynamic nodes, each of which provides either processing resource only or combined processing and access point resource. Such nodes, after having been identified within the fog network, can join fog computing and share their resources.

Our framework employs trust models in transactions pertaining to data transfer and administration. This adds an extra layer of security and guarantees that untrusted nodes are not able to access sensitive data over the network. Dividing trust into levels will allow segregation of operations and data based on their criticality.

Whenever a node is made part of the fog for the first time, it is assigned the lowest level of trust and the least access privileges to the data to be processed as no knowledge of its previous transactions exists. However, after some transactions, the dynamic nodes can improve their reputation and gain a higher level of trust. They might also be disjoined from the network if any malicious actions are detected, or if they will no longer be in the vicinity of the computing environment. In cases like this, the nodes' access privileges need to be revoked. To make this possible, a manager node is required for managing the task allocation and participating nodes. Figure 3 illustrates the proposed framework in which the fog layer consists of four types of nodes: Fog Manager Node (FMN), Static Node (SN), Dynamic Node (DN) and Processing Node (PN). Any of these nodes have different roles and hence different privileges are assigned to them. A role for a node is defined based on its capability (Processing only or Processing and

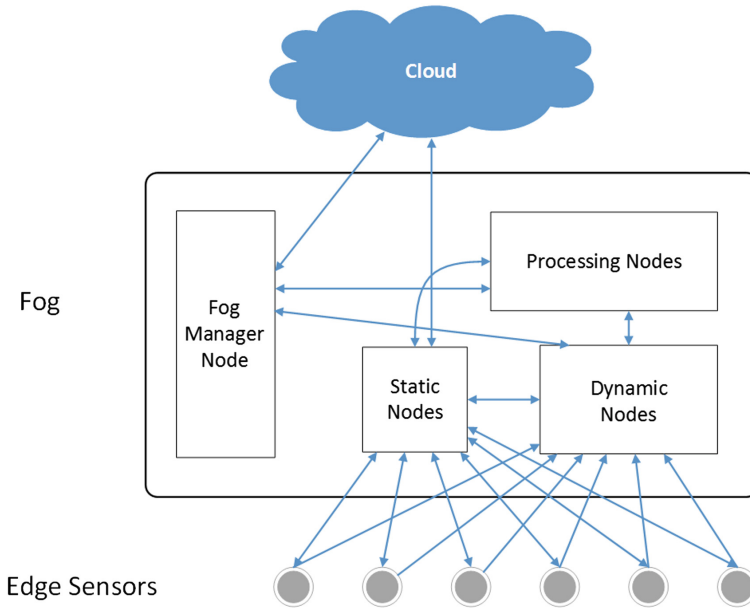


Fig. 3. Proposed framework.

communication) and its current level of trust. The following section describes the definition of roles and trust in more detail.

4.1 Roles

A role for a node defines its privileges for accessing different kinds of data and for participating in processing tasks. In this framework we assume three categories of information based on criticality: *non-critical*, *moderate* and *critical* data. The FMN assigns the roles to nodes based on their reputation and trust levels as well as their capabilities. We define four roles according to the node types in this framework. With each role within fog layer, a set of permissions is assigned. This set limits the nodes' access to certain types of data and defines their privileges and responsibilities in processing certain tasks. Table 1 summarize the security privileges of each node according to its role.

A detailed description of each node is presented in the following.

Fog Network Manager is the central overseer of the fog network. Whenever a node wants to join the network, it must contact the FNM. If the connecting node is an edge node, the FNM will send this node the address of the active fog nodes based on location proximity. The edge node will then connect to the nearest fog node and start sending its data. In case a fog node goes offline, the edge node will be provided with the address of the next suitable nodes to connect to. If a smart device attempts to join the fog network, the FNM will assign the connecting node with the lowest trust level. On the other hand, if a

Table 1. Assignment of privileges according to roles.

Role	Privileges						
	Processing	Edge communication	Cloud communication	Verifying the tasks	Assigning a task	Adding new nodes	Revoking access
FNM		×	×	×	×	×	×
SN	×	×	×	×			
DN	×	×					
PN	×						

node needs to be deleted from network (due to malfunctioning or permanently disconnecting), the FNM will revoke access rights from that node and update the list of active fog nodes to the edge sensors as well as the cloud layer. The FNM is also responsible for promoting or demoting the roles of participating fog nodes according to their trust levels.

Static Nodes: These nodes are used for connecting edge devices to the fog layer. They are static in nature and are expected to be available at all times. By default, they are assigned the trust level *High*. They can either process data themselves or they can request the FNM to initiate a task on some other processing node or an access point. Upon completion of a task, they can forward the data themselves to the cloud layer.

Dynamic Nodes are dynamic and are intended to be used as access points as well as processing nodes. They start with trust level *Low* and gain more levels as trusted more by the FNM. They do not themselves send data to the cloud layer but, instead, they use the static access points for the purpose.

Processing Nodes are dynamic and are used exclusively for processing data. Since they are dynamic, their initial trust level is *Low* and it is increased as the node gains more trust. They cannot themselves connect to the cloud layer nor act as access points but, instead, they use static access points for sending their data. These nodes only share their processing resources. Therefore, their tasks are assigned by the network manager or static nodes. After processing the task the results needed to be sent to a static nodes to be forwarded to the cloud layer.

4.2 Trust Management

A trust level is a measure of the reliability of a participating node. Trust management is applied to dynamic fog nodes of the network. Static fog nodes at any time are considered to have the highest level of trust. Trust in our system is defined in terms of a nodes' privilege to process a certain type of data. It can be divided into multiple levels but in this paper, we will consider the division into three levels:

- *Low*: This is the lowest level. Dynamic fog nodes are initially assigned this level upon joining the network. The FMN assigns tasks of the lowest priority

and criticality to these nodes. Data computed by nodes with this trust level is sent to one of the static nodes to be verified before sending to the cloud layer.

- *Moderate*: This is the second level of trust. On this level, the data is considered to be of moderate criticality. The fog node handling this data is assumed to be reliable, and the result generated by the node will be sent directly to the cloud layer. Dissatisfaction in the service of nodes in this level will demote the node to the low level. However, dissatisfaction up to a pre-determined level can still be tolerated.
- *High*: This is the highest level of trust. The data which is considered to be most critical by the application is handled by the nodes at this level. The requirement of processing is not only that the data be processed correctly, but also that the nodes maintain the highest level of service. The data processed by nodes in this level is sent directly to the cloud layer.

The trust level of each dynamic fog node evolves over time and can change on interaction with other nodes. We utilize already established trust algorithms for our implementation. There can be several ways to calculate the trust level for a node. Manuel [19] investigated different factors to evaluate trust value of a resource in cloud computing. They claim that combination of multiple trust factors such as availability, reliability, data integrity, and turnaround efficiency should contribute to the trust model of a resource. According to this study, in our proposed framework we calculate the trust value based on all attributes mentioned above.

In the following we discuss each of these attributes and present a formula to compute the trust value of a resource based on those attributes:

Availability is a measure to ensure that a resource is operational and accessible to authorized parties whenever needed. A resource is deemed unavailable if (1) it is too busy to process and responds a task request, (2) it denies a task request, or (3) it is just shut down. Availability of a resource Av_R is calculated based on the following equation over a period of time:

$$Av_R = \frac{Ac}{Sb} \quad (1)$$

where Ac denote the number of computing tasks accepted by a resource and Sb denote the total number of tasks submitted to that resource.

Reliability or success rate of a resource is a measure and quality of a resource in consistently performing according to its specifications in specified time. Reliability of a resource Re_R defines its success rate in the completion of the tasks that it has accepted and is calculated based on the following equation over a period of time:

$$Re_R = \frac{Cs}{Ac} \quad (2)$$

where Cs denote the number of accepted tasks completed successfully by a resource, and Ac is the total number of accepted tasks by that resource.

Data Integrity involves maintaining the consistency, accuracy, and trustworthiness of data over its entire lifecycle. Integrity ensures that information is not modified by unauthorized entities. Data Integrity of a resource Di_R is calculated based on the following equation over a period of time:

$$Di_R = \frac{Cm}{Ac} \quad (3)$$

where Cm denote the number of tasks that a resource successfully preserves data integrity, and T is the total number of accepted tasks completed successfully by a resource.

Turnaround Efficiency is a quality that a resource accomplishes a task within the time that it promises. Turnaround is a time frame that starts from when a broker sends a processing request to a resource till the time that the resource completes the task successfully. Turnaround Efficiency of a resource Te_R is calculated based on the following equation over a period of time:

$$Te_R = \frac{Pt}{At} \quad (4)$$

where Pt denote the Promised Turnaround time by a resource for completion of a task and At is the Actual Turnaround time by a resource for the completion of a task.

Trust Value of a Resource: The overall trust value for a resource is calculated based on composition of all attributes of a resource with following equation:

$$TrustValue_R = (a * Av) + (b * Re) + (c * Di) + (d * Te) \quad (5)$$

where $a + b + c + d = 1$ are coefficient positive numbers that define the weight of each attribute and Av , Re , Di , and Te are respectively average value for Availability, Reliability, Data Integrity, and Turnaround Efficiency over determined time T .

Task assignment done by the FNM is also dependent on the trust levels. To ensure that each trust level would have the required number of nodes to perform all tasks defined for that trust level, we will use the weight function to evaluate the need to increase the trust level of a dynamic node. It would be preferred for a node to be promoted to a higher trust level if there is a shortage of nodes at a higher level. For each dynamic fog node, the weight is calculated as:

$$Weight = \begin{cases} 1 - \frac{N_{req}}{N_{avail}} & \text{if } N_{req} < N_{avail} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where N_{req} is the number of nodes required at the next higher trust level, and N_{avail} is the number of nodes available at the next higher trust level.

5 Results and Discussion

The fog computing platform is implemented in SystemC environment. Each processing unit is modeled by a SystemC module which can communicate with all

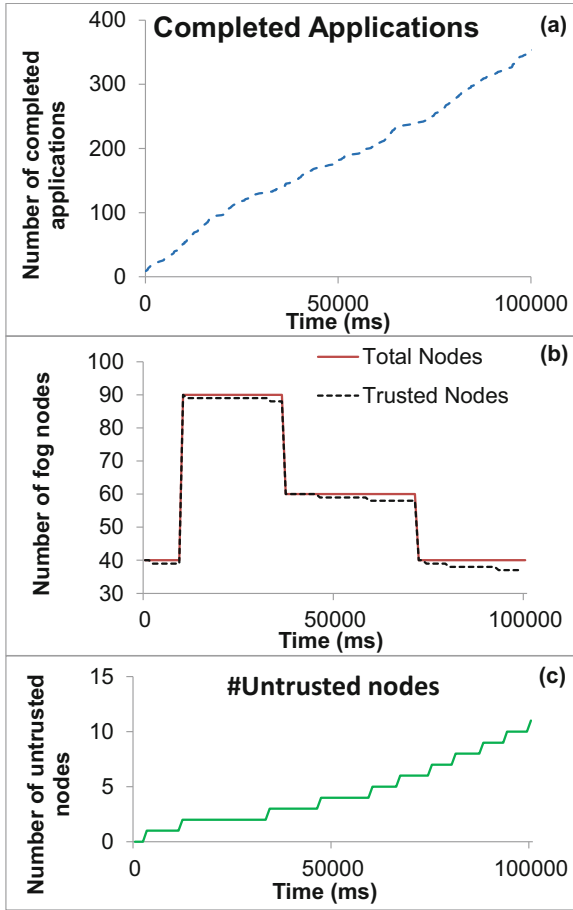


Fig. 4. Experimental results.

the other processing elements in its domain through a SystemC channel. We have considered heterogeneous nodes with different processing capabilities. The execution frequency for each processing elements varies between 500 MHz up to 4 GHz. Applications enter and leave the system based on a randomized amount of workload during the time. Each application is modeled as a task graph where each task should be assigned to a processing element exclusively. Execution of the tasks are independent of each other, and only the data transfer between tasks connects two tasks to each other. Therefore each task can be run at a different frequency. The fog system comprises of a number of fog nodes which include static nodes, processing nodes, and a fog manager node. Along the time, a group of dynamic nodes joins and leave the fog system. The fog manager assigns the tasks to the newly joined dynamic node and calculates their trust level based on

the Trust formula. So, once any of the new dynamic nodes reaches to the desired trust level, then the fog system upgrades their role to become trusted fog node.

Figure 4(a) shows the number of completed applications in the fog system during the time. Figure 4(b) illustrates the total number of the nodes once the dynamic nodes join and leave the fog system. The dashed line shows that the system is able to detect and eliminate the untrusted nodes in each interval. As it can be seen, while the number of nodes in the fog system increases, the rate of the completed applications also increases. And finally Figure 4(c) shows the total number of identified untrusted nodes during the execution time.

6 Conclusion

The fog computing paradigm extends cloud computing and services to the edge of the network to support geographical distribution and mobility of end users. In this paper, we presented a security framework for fog computing infrastructure. After discussing the potential application of fog computing, we argued that to increase the performance of the fog layer we can take advantages of vacant resources of surrounding smart devices such as smart phones and tablets. To tackle the security issues that are imposed by this technique we proposed an implementation of role-based access control in conjunction with trust models. We presented how our proposed framework can contribute to solving security issues of a fog network. In our framework, we defined a method to calculate trust levels based on computing tasks assigned to the nodes. Moreover, we presented algorithms for implementation of our framework. According to our implementation results, the fog system was able to distinguish the trusted and untrusted dynamic nodes. However, in addition to secure access control and authentication methods, secure computation schemes need to be undertaken to guarantee the security and integrity of data in a fog network.

Acknowledgment. This work was supported by University of Turku Foundation, EIT Digital and the Department of Information Technology - University of Turku.

References

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: A view of cloud computing. *Commun. ACM* **53**(4), 50–58 (2010)
2. Rimal, B.P., Choi, E., Lumb, I.: A taxonomy and survey of cloud computing systems. In: 2009 Fifth International Joint Conference on INC, IMS and IDC, pp. 44–51. *IEEE* (2009)
3. Schulz, P., Matthe, M., Klessig, H., Simsek, M., Fettweis, G., Ansari, J., Ashraf, S.A., Almeroth, B., Voigt, J., Riedel, I., Puschmann, A., Mitschele-Thiel, A., Muller, M., Elste, T., Windisch, M.: Latency critical IoT applications in 5G: perspective on the design of radio interface and network architecture. *IEEE Commun. Mag.* **55**(2), 70–78 (2017)

4. Bonomi, F., Milito, R., Zhu, J., Addepalli, S.: Fog computing and its role in the Internet of Things. In: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, ser. MCC 2012, pp. 13–16. ACM, New York (2012)
5. Sehgal, V.K., Patrick, A., Soni, A., Rajput, L.: Smart human security framework using Internet of Things, cloud and fog computing. In: Buyya, R., Thampi, S.M. (eds.) Intelligent Distributed Computing. AISC, vol. 321, pp. 251–263. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-11227-5_22
6. Shi, W., Zhang, Q., Li, Y., Xu, L.: Edge computing: vision and challenges. *IEEE Internet Things J.* **3**(5), 637–646 (2016)
7. Sandvine Intelligent Broadband Networks: 2016 global internet phenomena report, Latin America & North America. Sandvine - Intelligent Broadband Networks, Technical report (2016)
8. MacLean, J.: Households now use an average of seven connected devices every day. Cantech Letter, Technical report (2016)
9. Falaki, H., Mahajan, R., Estrin, D.: SystemSens: a tool for monitoring usage in smartphone research deployments. In: Proceedings of the Sixth International Workshop on MobiArch, ser. MobiArch 2011, pp. 25–30. ACM, New York (2011)
10. Ashton, K.: That ‘Internet of Things’ thing. *RFiD J.* **22**, 97–114 (2009)
11. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.: Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute (2011)
12. Hosseinpour, F., Meng, Y., Westerlund, T., Plosila, J., Liu, R., Tenhunen, H.: A review on fog computing systems. *Int. J. Adv. Comput. Technol. (IJACT)* **8**(5), 48–61 (2016)
13. Bellavista, P., Zanni, A.: Feasibility of fog computing deployment based on Docker containerization over RaspberryPi. In: Proceedings of the 18th International Conference on Distributed Computing and Networking, ser. ICDCN 2017, pp. 16:1–16:10. ACM, New York (2017)
14. Oueis, J., Strinati, E.C., Barbarossa, S.: The fog balancing: load distribution for small cell cloud computing. In: 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), pp. 1–6 (2015)
15. Yi, S., Qin, Z., Li, Q.: Security and privacy issues of fog computing: a survey. In: Xu, K., Zhu, H. (eds.) WASA 2015. LNCS, vol. 9204, pp. 685–695. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21837-3_67
16. Misra, S., Vaish, A.: Reputation-based role assignment for role-based access control in wireless sensor networks. *Comput. Commun.* **34**(3), 281–294 (2011)
17. Salonikias, S., Mavridis, I., Gritzalis, D.: Access control issues in utilizing fog computing for transport infrastructure. In: Rome, E., Theocharidou, M., Wolthusen, S. (eds.) CRITIS 2015. LNCS, vol. 9578, pp. 15–26. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-33331-1_2
18. House, T.: Mobile secure role based access control (MS-Ro-BAC) device. In: Proceedings of the SoutheastCon, pp. 542–546. IEEE, April 2005
19. Manuel, P.: A trust model of cloud computing based on quality of service. *Ann. Oper. Res.* **233**(1), 281–292 (2015)



IoT Platform for Real-Time Multichannel ECG Monitoring and Classification with Neural Networks

Jose Granados^{1(✉)}, Tomi Westerlund^{2(✉)}, Lirong Zheng^{1(✉)}, and Zhuo Zou^{1(✉)}

¹ School of Information Science and Technology,
Fudan University, Shanghai, China

{jose16,lrzheng,zhuo}@fudan.edu.cn

² Department of Information Technology,
University of Turku, Turku, Finland
tomi.westerlund@utu.fi

Abstract. Internet of Things (IoT) platforms applied to health promise to offer solutions to the challenges in healthcare systems by providing tools for lowering costs while increasing efficiency in diagnostics and treatment. Many of the works on this topic focus on explaining the concepts and interfaces between different parts of an IoT platform, including the generation of knowledge based on smart sensors gathering bio-signals from the human body which are processed by data mining and more recently, deep neural networks hosted on cloud computing infrastructure. These techniques are designed to serve as useful intelligent companions to healthcare professionals in their practice. In this work we present details about the implementation of an IoT Platform for real-time analysis and management of a network of bio-sensors and gateways, as well as the use of a cloud deep neural network architecture for the classification of ECG data into multiple cardiovascular conditions.

Keywords: IoT · ECG · Healthcare · AI · Neural networks

1 Introduction

The advancement of Internet of Things (IoT) platforms has promised to solve many of the challenges that healthcare systems worldwide face today. The IoT platforms, which refer to a comprehensive system involving the interconnection of smart sensors to cloud computing services, has spawn a new paradigm in the way healthcare services can be administered. The traditional setup in which a patient visits the physician's office, where he or she has to recall the symptoms from memory is being re-imagined from a reactive stand point to a proactive one. This is possible thanks to smart sensors that gather bio-signals from the human body which are transmitted to cloud services via intelligent web-enabled gateways as shown in Fig. 1. The health signals then pass through an analysis phase consisting

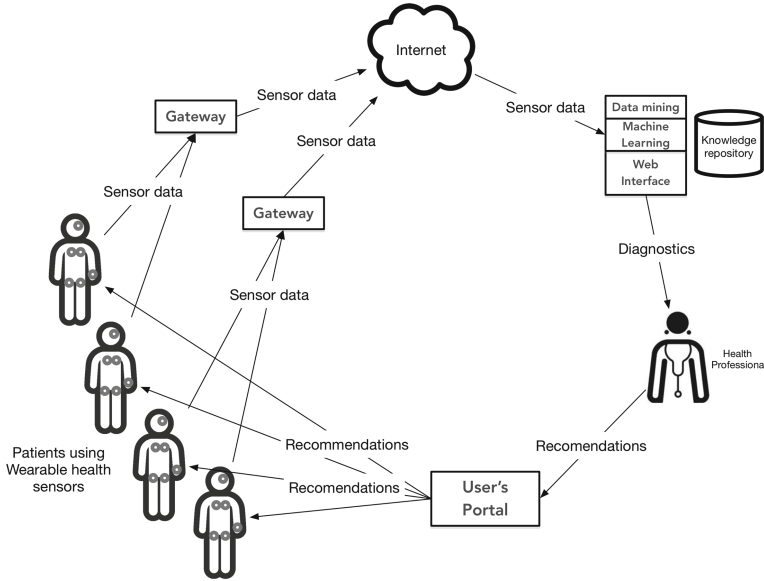


Fig. 1. IoT health workflow

in applying data mining and machine learning algorithms which can diagnose conditions with human comparable accuracy [1]. This way, intelligent systems can provide continuous secure monitoring of physiological parameters in order to detect patterns that might represent an abnormality. This methodology allows to automatically suggest lifestyle changes learned from previous cases or if the condition is critical, alert the corresponding healthcare professionals to act on time before the ailment turns into a dangerous or expensive disease. Considering that most hospital visits are due to over-treatment, redundant, inappropriate, or unnecessary tests and procedures that can be diagnosed and treated at home, these intelligent systems enable the optimization of resource allocation to the more demanding cases.

IoT Platforms for healthcare still face significant obstacles in terms of performance, power consumption and functionality in order to unleash all the possibilities that they can provide. Many of the works in this topic study the implementation of specific cases for monitoring different physiological parameters. These bio signals often require distinctive transmission and processing needs which depend upon a particular solution. However, a comprehensive system that allows the monitoring of a network of sensors through embedded connected devices, as well as the management of a set of gateways from a cloud service in real-time with the hardware and software engineering details has been more or less still a work in progress. Data mining, and more recently Deep Learning has also played the fundamental role of classifying signals for diagnostic purposes with human-like accuracy. The potential of implementing Artificial Intelligence modules to

the real-time processing of bio-sensor data in the context of an IoT platform for Healthcare therefore could have beneficial implications which could ameliorate healthcare services and revolutionize the insurance industry.

2 Design Considerations

2.1 Workload Balance

An important aspect to consider is the workload balance between different system components, namely the edge node, the gateway and the cloud computing modules. In order to ensure readiness and reaction time against health risks, the wearable devices must have up to date detection algorithms available or have permanent access to them on the cloud in real time. Using the cloud approach, the process is similar to constantly ask the cloud service provider for the latest diagnostics and treatments so that they can be presented to users immediately, where the cloud is the knowledge repository that has been learned so far from previous cases and is continuously expanding. Eventually and as edge node microprocessors become more powerful and energy efficient, the devices could train the neural network locally and then share the trained package with other devices in a peer to peer, decentralized fashion. In this scenario, there is less need to regularly upload personal bio-signals which can reduce privacy concerns, and instead only the learning, represented as neural networks weights, is communicated [2]. The goal is then to achieve the optimal balance in terms of power consumption, performance, scalability and security between the different components of the system, while providing quality diagnostics and treatment in a constantly growing knowledge database. With sufficient data, the intelligence repository could achieve human-like detection accuracy of multiple conditions and propose the best evidence-based treatment while monitoring health improvement.

2.2 Mobile Device Gateway

The incorporation of mobile devices such as smartphones and tablets into people's the daily life can be leveraged to manage a network of smart sensors. These mobile devices are usually equipped with Bluetooth 4.1, 4.2 and 5 Bluetooth Low Energy (BLE) radios that can interface with low-power smart sensors in Body Area Networks (BAN). These sensors are generally accessed using the Generic Attribute Profile (GATT) protocol which presents the device properties as a database of services and characteristics. This is specially suitable for defining profiles which are definitions of possible applications and general behaviors that Bluetooth enabled sensors use to communicate with gateways as well as to characterize what kind of data a Bluetooth module is transmitting. Profiles allow the standardization of intercommunication between different kinds of sensors and gateways, even if the type of data is different, as long as they respect the way to organize use case behaviors. A mobile device gateway can also leverage established Web technologies such as Websockets, in order to send sensor data

to cloud servers in real time. Websockets can also serve as a base protocol on which to implement device management frameworks that can be used to control gateways and end devices. In addition, the gateway can perform other tasks such as preprocessing for computation offloading from the node devices as well as serve as a firewall for protection against attackers outside and inside of the sensor network.

2.3 Multichannel ECG

The IoT framework proposed in this work is exemplified in a intelligent multi-channel electrocardiography (ECG) monitoring and analysis system. ECG signals represent the heart activity and are used by cardiologists to diagnose cardiovascular and other diseases. The ECG signal has a particular relative high data rate and serves as a challenge for low-power embedded systems design. ECG also has been subject to different types of data mining and machine learning classification techniques particularly to detect arrhythmia, which once incorporated into a real-time IoT platform, can serve as a valuable companion tool for heart specialists. One of the goals of this work is to use ECG signals to diagnose not only cardiovascular diseases but also other kinds of correlated conditions such as pulmonary disorders. We believe that by using deep correlation and invisible features detection, abnormal signal patterns akin to multiple conditions can be exposed. An additional result from this methodology is the ability to infer context activities from the ECG that can be used to suggest better lifestyle habits for the user.

3 Related Work

The IoT for Healthcare research community has been active in offering solutions to the lack of efficiency and high cost of healthcare systems around the world. [3] describes the goal and the components of a complete IoT Platform for healthcare, including sensing, cloud service, data mining and end user's perspective. The focus of the IoT platform is to move the treatment to home care in order to save hospital costs and increase efficiency of physician resources. The authors use data clustering mechanisms for improving classification. However, they use a software emulator to generate the ECG data, not an embedded device capturing bio-signals, also most of the software engineering details are left out. A health monitoring system using a smartphone case with integrated electrodes and an Android application to display the data is presented in [4]. However, no integration with cloud services or analytics are depicted. [5] shows an Android application connected to an ECG device and to a cloud database. The details of the implementations are explained including the use of the Bluetooth Serial profile to transmit data between the device and the Android application. In addition, the authors describe the use of a SD Card to store the data locally on the smartphone. The ECG data is then uploaded to the cloud using a FTP server that saves the data on a *SQL Server Filetable*. This implementation nonetheless

lacks real-time communication between the smartphone and the cloud server and uses MATLAB for visualization instead of user-friendly Web application. In addition, this solution does not adopt open source components, and seems to require an external Linux or Windows computer to transfer the data from the smartphone to the cloud server using *FileZilla*. An integrated solution in the form of an ASIC for measurement and analysis of ECG including keying for security purposes is presented in [6]. The authors perform Verilog simulations using *Modelsim* and *Synopsys* tools to verify the functionality of the design. The end-to-end system is designed using Verilog-HDL, and a test-bench is created to simulate it by modeling the input data. The resulting ASIC achieves a reduction of 62.2% in power consumption and a 16% reduction in area when compared to similar state-of-the-art processors. The work in [7] describes a smart gateway for e-health which serves as a bridge between numerous sensor protocols and to execute preprocessing of health data in order to generate health indicators before uploading to the cloud. Fog computing is introduced at the gateway network layer in order to reduce the latency of the decision making process which is performed locally instead of in the cloud. A wearable sensor node with energy harvesting and Bluetooth interface with HTML5 based smartphone app is presented in [8]. The battery-less device is able to work 24 h on adequate sunlight conditions and can transmit sensor data to a Web based smartphone application designed to display signals and send emergency notifications.

The current literature makes known the general components and concepts of an IoT platform for Healthcare such as bio-sensors, cloud computing, data analysis and visualization. However, there is still room for empirical studies in building such a system. We consider that it is practically important that the details of the actual software and hardware engineering are explained, as there is a pragmatic need to build a solid infrastructure in a way such that cost, performance and scalability assessments of real world implementations can be achieved. Current research is focused on describing specific use cases like transmitting a physiological parameter variable to the cloud using definite methodologies as well as explaining the main components of an IoT platform such as the edge nodes and cloud services without digging too much into the details of the implementation. We propose to create a general IoT framework that can be mapped to well established technologies such as Bluetooth LE, CoAP, Websockets, NoSQL databases, etc. with a solid reasoning behind the selection and development of technologies used in order to obtain a comprehensive, well structured solution.

4 Real-Time IoT Platform for Healthcare

We propose an IoT Platform for ECG analysis as shown in Fig. 2 consisting of a smart ECG sensor, Web-enabled gateway based on a smartphone and cloud server for monitoring and managing devices integrated with a deep neural network for classification of signals. Emphasis is done on the implementation of real-time capabilities by using Bluetooth LE GATT notifications on sensor nodes and Websockets on the gateway and cloud server link. That way we can achieve

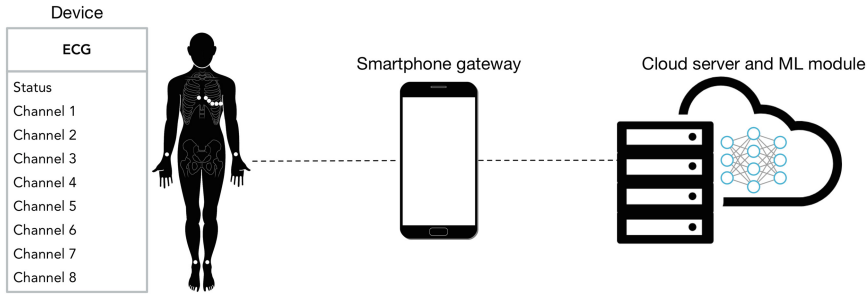


Fig. 2. IoT Platform for ECG analysis

not only near real-time data transmission but also assemble a base communication channel on which a device management and control framework can be built. Also, the use of GATT profiles is taken to define and translate sensor resources and services into other Internet protocols. This allows scalability by standardizing access from the gateway to the sensors and easy mapping of sensor profiles to specialized web transfer protocols similar to the Constrained Application Protocol (CoAP), which include RESTful interaction models between application endpoints, support for service and resources discovery, and other Web concepts such as Uniform Resource Identifiers (URIs) and Web Application Programming Interfaces (APIs). Finally, the ECG classification into different abnormal patterns is implemented on a multi layer Convolutional Neural Network realized using *TensorFlow* and hosted at the cloud. The supervised and unsupervised learning methods are used to classify the signals into normal, abnormal and as part of a cardiovascular disorder such as particular types of arrhythmia. Also, the platform incorporates a Web tool where expert ECG technicians can annotate ECG recordings which can serve as input to the supervised training network. The set of software stacks used in this implementation is shown in Fig. 3.

4.1 Multichannel ECG Monitoring Device

The ECG embedded device consist of ten electrodes that capture electrical signals from the human body. This data gathering node is shown in Fig. 4. The device was designed and built for this work and is branded as *Horizon Medical IoT Holter*. The 12-lead ECG graph can be realized with this configuration which, contrary to single lead configurations, allows to not only recognize a cardiovascular condition but also helps to pinpoint its location on the heart surface. The signals sensed by the electrodes are amplified and converted to 24-bit, 250 samples per second digital representation by a Texas Instruments ADS1298 analog front end, which is then inputted via Serial Peripheral Interface Bus (SPI) into a Nordic Semiconductor nRF52832 microcontroller unit (MCU) with integrated 2.4 GHz Bluetooth Low Energy compatible radio. The embedded software programmed on the device allows to configure the device as a BLE peripheral

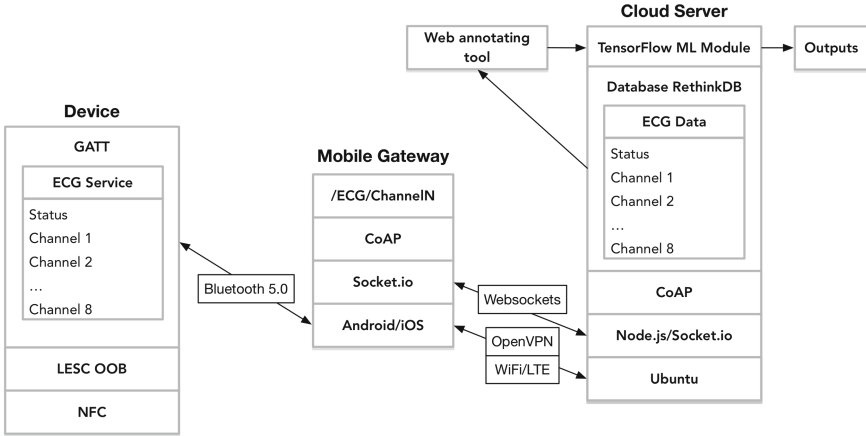


Fig. 3. Software stacks used on the device, smartphone gateway and cloud server.

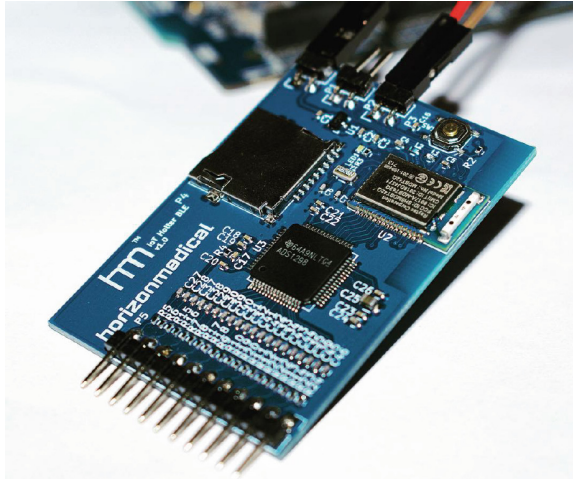


Fig. 4. 8-channel bluetooth 5 embedded monitoring ECG device

with one custom GATT service and nine characteristics, one for the status information and one for each of the eight signal channels. The characteristics are defined with a notification property which allows to push the uncompressed signal data into the central device (smartphone) and makes use of the higher data throughput feature of Bluetooth 5 in order to support the higher bandwidth requirements. The ECG device allows secure connection to the smartphone via Near Field Communications (NFC) using the LE Secure Connections with Out of Band (OOB) Pairing. In this method, a large temporary key (128 bits) is exchanged when the device is in close range of the smartphone which makes it difficult for an attacker to perform Man-In-The-Middle (MITM) attacks. If the

OOB channel is immune to eavesdropping during the pairing process, then the BLE connection will also be immune from passive eavesdropping. Preliminary results obtained for the device energy consumption are shown in Table 1.

Table 1. ECG monitoring device energy consumption

	Avg Current (mA)
Advertising	4
Transmitting	5.4
Sleeping	0.8
Est battery life (800 mAh)	4.5 days

4.2 Mobile Device Gateway

The IoT gateway is a mobile device running Android or iOS operating system. An hybrid mobile app has been developed for visualization of the multichannel ECG signal as well as to forward the sensor data coming from the embedded device to the cloud server. In order to transmit the ECG signal data, the gateway creates a WebSocket connection between itself and a cloud server. The WebSocket protocol is secured with an OpenVPN encrypted connection. In addition, a device management framework based on top of the WebSocket protocol is used to control the gateway and sensors from the remote server using an event-based API. This allows a remote server to scan for devices near each gateway and collect a list of available resources. In addition, the device management framework grants the possibility to ask specific sensors to change its operating modes such as to go from live streaming to local storage for power saving purposes. The mobile application user interface consisting of 8 live charts of sensor data is shown in Fig. 5. The mobile application is built using the Ionic Framework which allows to integrate a HTML5 based application together with native functionality into an hybrid app. The BLE Native plugin is used to establish the communication between the smartphone app and the peripheral device and allows to enable notifications from it. In the background, the WebSocket connection between the app and the cloud server is implemented using the Socket.io client library. The services and characteristics are then translated into messages similar to the ones used by the CoAP protocol in order to leverage the use of URI based access to device resources as well as to perform service discovery. Preliminary results obtained for the gateway performance are shown in Table 2.

4.3 Convolutional Neural Network Classifier

The learning module responsible for classifying incoming ECG signals from the device as well as allowing training of the neural network is implemented using the TensorFlow version 1.3 library running on a GPU accelerated machine.

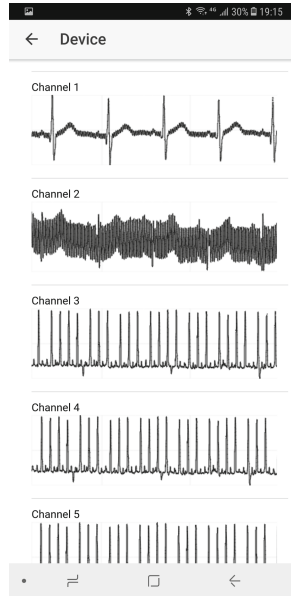


Fig. 5. Smartphone app user interface showing multichannel sensor data.

Table 2. Smartphone gateway performance

Avg data rate in/out	15.35 KB/s
CPU usage %	47%
Est energy use	66 mAh
Connection time	15 m 51 s
Data sent	9.92 MB
Device:	Samsung Galaxy S8+

A Socket.io server running on an Ubuntu 16.04 instance receives the signal data coming from the gateway. The server is implemented on Node.js version 8.4.0 and the Socket.io server used has version 2.0.3. The server also creates a connection to a RethinkDB version 2.3.6 database instance running on the same computer and stores the ECG sensor data as JSON documents containing arrays of 28 samples. On parallel, a TensorFlow engine is implemented which connects to the same RethinkDB database and queries the original raw ECG signal data. When the network is being trained, the ECG signal is first displayed on a Web based annotation tool which allows a heart technician to select waveform segments and label them according to different types of arrhythmias. The selected signals serve as inputs to train the network. The neural network is designed as successive convolutional, pooling, reshape, fully connected and dropout layers which once trained, takes the live incoming signal from the device and automatically

classifies the ECG patterns into possible arrhythmias. The classification outputs are saved into the same RethinkDB instance in order to be used by a Web application designed to show results to the physician. The Web application allows to forward automatic preprogrammed alerts depending on conditions detected or personalized recommendations prepared by the physician via chat messages displayed on the previously mentioned smartphone app used by the patient.

5 Conclusion

In this paper we have presented a practical IoT platform for monitoring and classification of the 12-lead ECG. We have designed a Bluetooth 5 multichannel ECG monitoring device which connects to a cloud service via smartphone gateway using the GATT protocol and enabled notifications. The data is forwarded to a cloud service using a WebSocket protocol which also serves as a base for a device management framework. The cloud service receives the data and passes it through a convolutional neural network which classifies the incoming signal into different types of arrhythmias. The same platform is also used to train the network by taking the signal incoming from the device and displaying it in a Web based annotation tool where heart specialists can select segments of the signal and label them. The system takes advantage of the protocols' real-time capabilities and allows scalability and workload balance between the different components of the system. This IoT platform allows to increase the efficiency of healthcare services by conducting home treatment instead of occupying valuable hospital resources. Future work includes optimizing energy efficiency on the device and gateway levels as well as implementing a FPGA accelerated neural network system for coprocessing on the cloud server.

References

1. Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C., Ng, A.Y.: Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks (2017). CoRR, vol abs/1707.01836
2. Liu, M., Jiang, H., Chen, J., Badokhon, A., Wei, X., Huang, M.C.: A collaborative privacy-preserving deep learning system in distributed mobile environment. In: 2016 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, pp. 192–197 (2016). <https://doi.org/10.1109/CSCI.2016.0043>
3. Abawajy, J.H., Hassan, M.M.: Federated Internet of Things and cloud computing pervasive patient health monitoring system. *IEEE Commun. Mag.* **55**(1), 48–53 (2017). <https://doi.org/10.1109/MCOM.2017.1600374CM>
4. Mahmud, M.S., Wang, H., Esfar-E-Alam, A.M., Fang, H.: A wireless health monitoring system using mobile phone accessories. *IEEE Internet of Things J.* PP(99), 1. <https://doi.org/10.1109/JIOT.2016.2645125>

5. Mohammed, J., Lung, C.H., Ocneanu, A., Thakral, A., Jones, C., Adler, A.: Internet of Things: remote patient monitoring using web services and cloud computing. In: 2014 IEEE International Conference on Internet of Things (iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom), Taipei, pp. 256–263 (2014). <https://doi.org/10.1109/iThings.2014.45>
6. Yasin, M., Tekeste, T., Saleh, H., Mohammad, B., Sinanoglu, O., Ismail, M.: Ultra-low power, secure IoT platform for predicting cardiovascular diseases. *IEEE Trans. Circuits Syst. I Regul. Pap.* PP(99), 1–14. <https://doi.org/10.1109/TCSI.2017.2694968>
7. Rahmani, A.M., Gia, T.N., Negash, B., Anzanpour, A., Azimi, I., Jiang, M., Liljeberg, P.: Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach. *Future Gener. Comput. Syst.* (2017). ISSN 0167–739X. <https://doi.org/10.1016/j.future.2017.02.014>
8. Wu, T., Wu, F., Redouté, J.M., Yuce, M.R.: An autonomous wireless body area network implementation towards IoT connected healthcare applications. *IEEE Access* **5**, 11413–11422 (2017). <https://doi.org/10.1109/ACCESS.2017.2716344>



Deep Ensemble Effectively and Efficiently for Vehicle Instance Retrieval

Zhengyan Ding¹, Xiaoteng Zhang¹, Shaoxi Xu¹, Lei Song^{1,2(✉)}, and Na Duan¹

¹ The Third Research Institute of the Ministry of Public Security, Shanghai 201204, China

² Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China
dzy_wlw@163.com, zxt_wlw@126.com, gbzfx_sunny@163.com,
songlei9312@126.com, naduan323@163.com

Abstract. This paper aims to highlight instance retrieval tasks centered around ‘vehicle’, due to its wide range of applications in surveillance scenario. Recently, image representations based on the convolutional neural network (CNN) have achieved significant success for visual recognition, including instance retrieval. However, many previous retrieval methods have not exploit the ensemble abilities of different models, which achieve limited accuracy since a certain kind of visual representation is not comprehensive. So we propose a Deep Ensemble Efficiently and Effectively (DEEE) framework, to preserve the impressive performance of deep representations and combine various deep architectures in a complementary way. It is demonstrated that a large improvement can be acquired with slight increase on computation. Finally, we evaluate the performance on two public vehicle datasets, VehicleID and VeRi, both outperforming state-of-the-art methods by a large margin.

Keywords: Instance retrieval · Vehicle · Deep ensemble

1 Introduction

Visual instance search is one of the core tasks in the field of computer vision and has been evolving rapidly in recent years. Given a query image example, the basic goal of instance-level retrieval is to search for images that contain the same instance, also viewed as a re-ID task.

With the ground-breaking success of deep learning based methods, image descriptors produced by CNN are significantly improving state-of-the-art performance for various problems including image classification [1–4], object detection [5, 6], etc. It comes as no surprise that pre-trained CNN models for a source task (e.g., classification) are capable of being applied to another target domain (e.g., retrieval), due to the generalization power of transfer learning [7].

Motivated by these advances, instance search approaches based on deep features have attracted sustained attention [7–9] both from academic research and from industrial applications. In this paper, we focus on the problem of instance-level retrieval centered around ‘vehicle’. As shown in Fig. 1, an image is considered to match the query if it contains the same vehicle across different surveillance camera views.

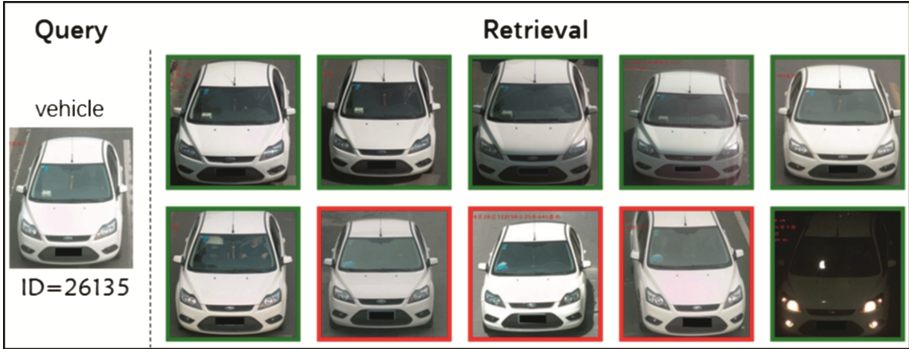


Fig. 1. Example retrieval results on VehicleID [15] dataset. The images in the first column are the query images. The retrieval images are sorted according to the similarity scores from left to right, which are color-coded as (green): correct, (red): incorrect. (Color figure online)

Traditional CNN-based retrieval methods [10–13] have not emphasized the importance of fusing various deep architectures into an ensemble model with slight increase on computation. In contrast, our proposed method has paid more attention to the complementarity of different models and implemented the deep ensemble framework via an effective and efficient way. To this end, this paper addresses two main challenges: (1) Model selection: How to ensemble various deep architectures to obtain an evident performance boost with marginal extra cost; (2) Feature selection: How to exploit multi-level features to generate a more comprehensive and compact fusion feature.

Firstly, we improve the retrieval performance by fusing various deep architectures into a single model. After comparing the advantages of different CNN models, residual-like network [4] is selected, as it can avoid the vanishing gradient problem significantly and some recent works [14] indicate that ResNet behaves like ensembles of relatively shallow networks, with the fusion strengths inherently. In terms of training loss functions, we finetune the network with both the verification and identification losses, inspired by ‘Mixed Difference Network’ [15]. Secondly, we utilize the Feature Pyramid Network (FPN) [6] method to ensemble multi-level features by a top-down architecture with lateral connections. After FPN fusion, we not only improve the retrieval performance significantly, but also obtain a more compact and efficient feature representation.

The rest of this paper is organized as follows: In Sect. 2, we review some related works and in Sect. 3, we introduce the proposed approaches in details. The experimental results are presented and analyzed in Sect. 4. Finally, we draw a conclusion in Sect. 5.

2 Related Work

In this section, we will describe previous works relevant to the approach discussed in this paper. Most of the works utilize CNN models for feature extraction and shed light on ‘vehicle’ retrieval tasks, especially in real-world surveillance scene.

2.1 Deep Representation for Instance Retrieval

Many works in the literature have proposed CNN-based representations for image retrieval. Razavian et al. [7] first investigate the use of CNN features for various computer vision tasks, including image retrieval. A typical CNN consists of several convolutional layers, followed by fully connected layers and ends with a softmax layer producing a distribution over the training classes. However, different from classification task, the pooled convolutional features often perform better than the fully connected layers [10]. Local convolutional features are similar to traditional hand-crafted features, e.g., SIFT, and various aggregation methods are proposed to improve the retrieval performance. To our best knowledge, most of the previous retrieval works focus on aggregating convolutional features from a single layer, but overlook the complementary properties of features from multiple layers. In fact, a series of excellent approaches have improved detection and segmentation performance by fusing different layers in a CNN model. For example, the FCN [16] algorithm sums partial scores for each category over multiple scales to compute semantic segmentations. Hypercolumns [17] uses a similar method for object instance segmentation. FPN [6] develops a top-down architecture with lateral connections for building high-level semantic feature maps at all scales. In this paper, our proposed method firstly introduces an extension of the FPN architecture to instance retrieval task and improves the results significantly.

2.2 Vehicle Instance Retrieval

Vehicle-related research includes detection, tracking, joint detection and 3D parsing. The growing explosion in the use of surveillance cameras highlights the importance of vehicle search from a large-scale image or video database. Liu et al. [15] released a carefully-organized largescale image database ‘VehicleID’, which includes multiple images of the same vehicle captured by different real-world cameras in a city. To facilitate progressive vehicle Re-Id research, Liu et al. [18] collect vehicle instance dataset named VeRi-776 from large-scale urban surveillance videos, which contains not only massive vehicles with diverse attributes and high recurrence rate, but also sufficient license plates and spatiotemporal labels. In this paper, comprehensive evaluations on the above two datasets have shown that ensemble of various deep architectures (e.g., verification and identification loss) and multi-level deep features will make a good contribution to boost the retrieval performance, compared with state-of-the-art results in previous works.

3 Proposed Method

3.1 Baseline Framework

Our baseline framework is illustrated in Fig. 2. It can be divided into offline stage processing and online stage query. In the online stage, a query is provided by a user

based on his intension. Then, the image is transform to the corresponding feature representation through a deep CNN model. Finally, the retrieval results from image dataset are generated, ranking by similarities with the query image feature.

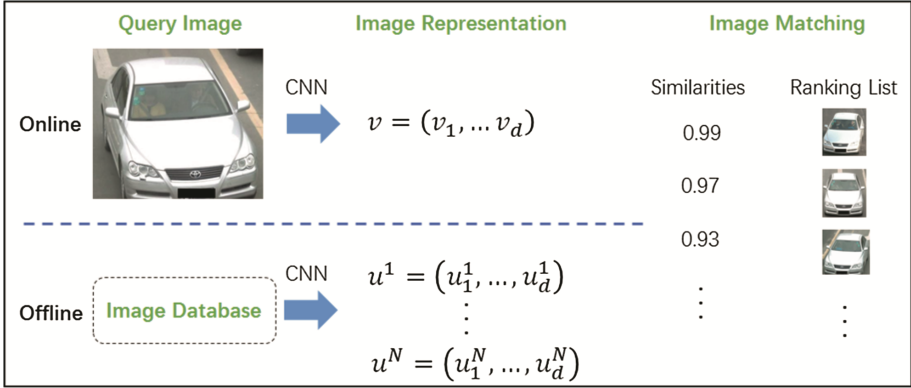


Fig. 2. Our baseline framework with online stage query and offline stage processing

As shown in Fig. 2, image representation is the core part of our framework, in which we adopt a compact representation pooled from activations of convolutional layers. This kind of global feature is very effective for instance-level retrieval, as formulated in Eq. 1: Query Feature:

$$v = (v_1, \dots, v_d) \tag{1}$$

where d denotes the feature dimension. In this paper, all fully connected layers are discarded in the inference part, and the image representation, called Average Activations of Convolutions (AAC), is simply constructed by average pooling over all dimensions per feature map. The dimension of our pooled feature is equal to the number of feature map channels, e.g. 1024 or 2048. To yield a shorter feature vector and improve retrieval efficiency, previous methods [10] use PCA of a post-processing tool for dimensionality reduction, by analyzing the covariance matrix of all descriptors. After acquiring the final image representation of query image (indicated as v) and a certain database image (indicated as u^i), the corresponding similarity can be calculated as inner product of the two feature vectors.

3.2 Effective and Efficient Ensemble Methods

In this part, we will further introduce our effective and efficient ensemble methods for image representation, which can be divided into two main components: model selection and feature selection as shown in Fig. 3.

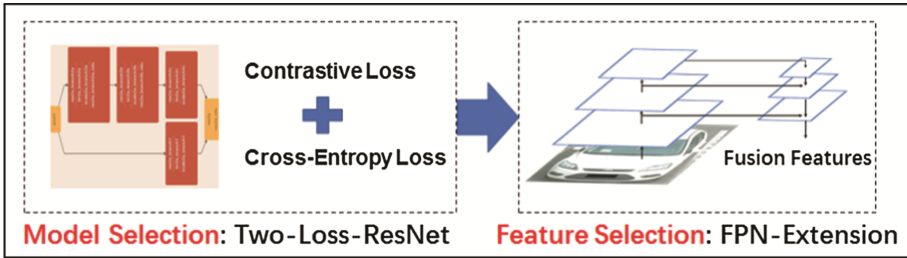


Fig. 3. Ensemble methods for image representation: model selection (Two-Loss-ResNet) and feature selection (FPN Extension)

Two-Loss-ResNet: Ensemble of Shallower Networks and Multiple Tasks

With the rapid development of CNN models on the image classification, more and more excellent CNN models emerged, like VGG-like [2], Inception-like [3] and residual-like [4] models. Compared with other previous architectures, residual networks avoid the vanishing gradient problem by introducing short paths which can carry gradient throughout the extent of very deep networks. Veit et al. [14] even proposes a novel interpretation of residual networks showing that a single ResNet model behaves as an ensemble of shallower networks, which results from ResNet’s additive operations. This property is very important for improving the feature representation since that averaging a set of deep networks is an effective solution to improve the accuracy performance, widely adopted in various computer vision tasks. Considering the above observations, we choose Residual-like architectures as our basic deep pretrained model.

In consideration of the common issues for training deep CNN models, such as infeasible computational cost and model size, we decide to select ResNet-50 as our baseline model, which is a tiny-version residual network and keeps a good balance between effectiveness and efficiency. Compared with traditional VGG-based image retrieval methods [15], ResNet-50 demonstrates a better speed with 3.8 billion FLOPs, which is only 25% of standard VGG-16 model (15.3 billion FLOPs). More significantly, ResNet-50 also yields a much better accuracy for instance retrieval task, as shown in the following experiments (See Tables 1 and 2).

Training a deep model with different loss functions is always a good way to ensemble the representative abilities for multiple tasks [5, 15]. To this end, two commonly-used loss functions are selected, cross-entropy loss and contrastive loss. Similar to conventional multiclass recognition approaches, we use the cross-entropy loss for identity prediction. Meanwhile, Siamese architecture and contrastive loss are adopted, in which we train a two-branch network and each branch is a clone of the other, meaning that they share the same parameters. Then we sum the above two losses together, to measure multiple outputs simultaneously.

Our ensemble method of deep architectures can be denoted as **Two-Loss-Res50**, fusing two loss functions based on the output of ResNet-50 model,

FPN Extension: Ensemble of Multi-level Deep Features

As a basic component in visual recognition systems, feature image pyramids are heavily used to generate scale-invariant representations. In order to fuse multi-level features in the training phase efficiently, we introduce an extension of Feature Pyramid Network (FPN) [6], which has been a popular method used in the object detection model recently. The top-down architecture with lateral connections is developed to exploit the inherent multi-level, pyramidal hierarchy of deep CNN models with marginal extra cost. Our experimental results have demonstrated that such multi-level ensemble representation is very compact, suitable for instance-level retrieval tasks. (See Table 1)

Through the above fusion strategies in training phase, the obtained CNN model is capable of yielding three kinds of features. As shown in Fig. 4, Feature_1 is a linear transformation of output from the 5th stage, which consist of only high-level semantics. On the contrary, Feature_3 is an ensemble of output from three stages (3&4&5), which represents more local and detailed information.

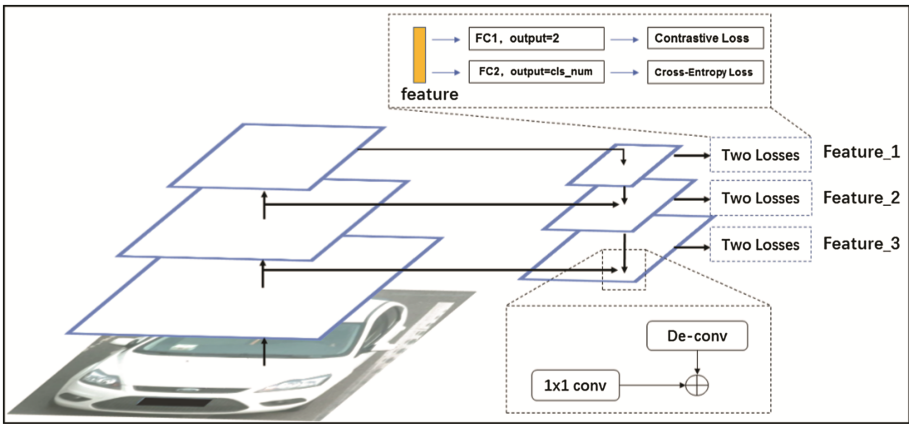


Fig. 4. Extension of FPN [6] to ensemble multi-level deep features

In this paper, we adopt Feature_3 as our final multi-level image representation for instance retrieval. The corresponding ensemble method can be denoted as **Two-Loss-Res50+FPN**.

4 Experiments

The proposed DEEE framework is evaluated by a standard performance metric, i.e., MAP, which is the mean of average precision scores for all query images over all the returned images. In addition, this paper makes most efforts on the retrieval tasks of ‘vehicle’, and two related datasets are used in our experiments. For fair comparison with existing methods, we follow the standard protocol of train/test split. All the results are obtained by **single-query**.

4.1 Datasets

VehicleID [15]: VehicleID dataset is a large-scale vehicle dataset that contains 221,763 images of 26,267 vehicles, where the training set contains 110,178 images of 13,134 vehicles and the testing set contains 111,585 images of 13,133 vehicles. Following the settings in [15], we use 3 test splits of different sizes constructed from the testing set: small, medium and large.

VeRi [18]: VeRi dataset contains over 50,000 images of 776 vehicles captured by 20 cameras covering an 1.0 km² area in 24 h, which makes the dataset scalable enough for vehicle Re-Id and other related research.

4.2 Experiment Setup

According to the above analysis, we choose ResNet-50 as our base model for retrieval tasks. The model pretrained on ImageNet classification dataset is used to initialize the weight parameters. All the experiments are implemented on the Caffe platform. We use the mini-batch SGD algorithm to learn the network parameters, where the batch size is set to 256, and momentum set to 0.9. We only experiment multi-level feature ensemble method with FPN on VehicleID dataset, because that the image samples in VehicleID have higher resolution and the detailed information is abundant, compared with VeRi dataset.

4.3 Comparison with State of the Art

We compare our method with state-of-the-art methods on the two datasets.

For the VehicleID dataset: (1) DRDL [15] method exploits a two-branch deep convolutional network to project raw vehicle images into an Euclidean space. The triplet loss used for deep metric learning is replaced by a novel loss function: coupled clusters loss (CCL). (2) HDC [19] method ensembles a set of models with different complexities in cascaded manner and mine hard examples adaptively. (3) GS-TRS [20] method forms the triplet samples across different categories as well as different groups, through partitioning training images within each category into a few groups. The comparison of statistic results can be found in Table 1.

For the VeRi dataset, we compared with PROVID method proposed in [18], which is a novel deep learning-based approach to PROgressive Vehicle re-ID. Table 2 lists the detailed results.

Table 1. Comparison results for VehicleID dataset (MAP)

Method name	Small	Medium	Large
DRDL+CCL [15]	0.546	0.481	0.455
HDC+Contrastive [19]	0.655	0.631	0.575
GS-TRS [20]	0.746	0.734	0.715
Two-Loss-Res50 (2048d)	0.823	0.775	0.737
Two-Loss-Res50+FPN (256d)	0.843	0.794	0.760

Table 2. Comparison results for VeRi dataset (MAP)

Method name	Standard train-test split [18]
PROVID [18]	0.278
Two-Loss-Res50 (2048d)	0.672

4.4 Ablation Studies

Improving Retrieval Effectiveness by Fusing Multi-scale Features

In this section, we will analyze the typical example that our ensemble framework improves the retrieval result effectively. As shown in Fig. 5, more discriminative features are taken into account after FPN fusion, such as the color of the car roof.

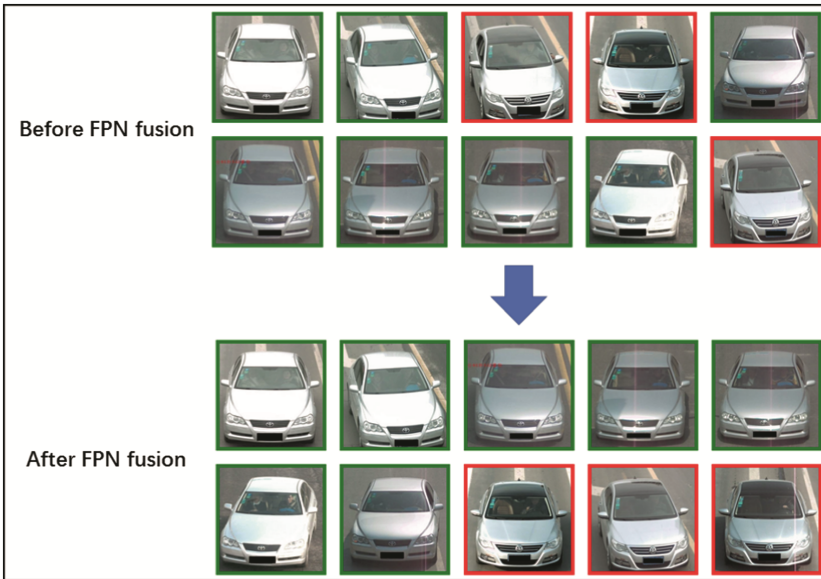


Fig. 5. Results comparison before and after FPN fusion. All the correct images (coded as green) rank before the incorrect images (coded as red) with FPN fusion. (Color figure online)

Improving Retrieval Efficiency by Feature Dimension Reduction

As shown in Table 1, the ensemble image representation using **Two-Loss-Res50+FPN** is 256d, which is only 1/8 of the original image representation using **Two-Loss-Res50**. This property will accelerate the retrieval procedure substantially, especially for large-scale image database.

5 Conclusion

In this paper, we propose a deep ensemble framework that simultaneously considers the effectiveness and efficiency, focusing on the instance-level retrieval task centered around ‘vehicle’. A set of problems are investigated comprehensively, including the selection of base CNN models, loss functions and multi-level features for training a discriminative ensemble model. The final experimental results indicate that the proposed DEEE framework is very effective and achieve the state-of-the-art results with a considerably short image representation. Our study also suggests that an efficient fusion method is capable of generating strong representation for instance retrieval tasks, providing a practical solution for balancing the speed and accuracy issues in the future research.

Acknowledgements. The authors of this paper are members of Shanghai Engineering Research Center of Intelligent Video Surveillance. Our research was sponsored by following projects: the National Natural Science Foundation of China (61403084, 61402116); Program of Science and Technology Commission of Shanghai Municipality (No. 15530701300, 15XD1520200); 2012 IoT Program of Ministry of Industry and Information Technology of China; Key Project of the Ministry of Public Security (No. 2014JSYJA007); the Project of the Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University (ESSCKF 2015-03); Shanghai Rising-Star Program (17QB1401000); The Special Fund for Basic R&D Expenses of Central Level Public Welfare Scientific Research Institutions (C17384).

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *ICLR* (2015)
3. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (2015)
4. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
5. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in Neural Information Processing Systems*, pp. 91–99 (2015)
6. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. arXiv preprint [arXiv:1612.03144](https://arxiv.org/abs/1612.03144) (2016)
7. Razavian, A.S., Azizpour, H., Sullivan, J., et al.: CNN features off-the-shelf: an astounding baseline for recognition. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 512–519. IEEE (2014)
8. Razavian, A.S., Sullivan, J., Carlsson, S., et al.: Visual instance retrieval with deep convolutional networks. *ITE Trans. Media Technol. Appl.* **4**(3), 251–258 (2016)
9. Tolias, G., Sivic, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: *ICLR* (2016)

10. Yue-Hei Ng, J., Yang, F., Davis, L.S.: Exploiting local features from deep networks for image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 53–61 (2015)
11. Kalantidis, Y., Mellina, C., Osindero, S.: Cross-Dimensional weighting for aggregated deep convolutional features. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9913, pp. 685–701. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46604-0_48
12. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1269–1277 (2015)
13. Hoang, T., Do, T.T., Tan, D.K.L., et al.: Selective Deep Convolutional Features for Image Retrieval. arXiv preprint [arXiv:1707.00809](https://arxiv.org/abs/1707.00809) (2017)
14. Veit, A., Wilber, M.J., Belongie, S.: Residual networks behave like ensembles of relatively shallow networks. In: Advances in Neural Information Processing Systems, pp. 550–558 (2016)
15. Liu, H., Tian, Y., Yang, Y., et al.: Deep relative distance learning: tell the difference between similar vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2167–2175 (2016)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
17. Hariharan, B., Arbeláez, P., Girshick, R., et al.: Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 447–456 (2015)
18. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 869–884. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_53
19. Yuan, Y., Yang, K., Zhang, C.: Hard-aware deeply cascaded embedding. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
20. Bai, Y., Gao, F., Lou, Y., et al.: Incorporating Intra-Class Variance to Fine-Grained Visual Recognition. arXiv preprint [arXiv:1703.00196](https://arxiv.org/abs/1703.00196) (2017)

Author Index

- Ahmed, Fizar 67, 91
- Basl, Josef 81
- Bi, Jianshui 21
- Bulyshhev, Alexander 34
- Bulysheva, Larisa 3, 34
- Cai, Hongming 12, 21
- Chen, Feng 132
- Ding, Zhengyan 192
- Doucek, Petr 45, 119
- Duan, Na 192
- Gábor, András 91
- Granados, Jose 181
- Hosseinpour, Farhoud 168
- Jiang, Lihong 12, 21
- Kataev, Michael 3, 34
- Lego, Tomas 56
- Li, Chun 12
- Li, Guoqiang 12
- Li, Xin 132
- Liljeberg, Pasi 159
- Liu, Yi 132
- Loseva, Natalia 3
- Luc, Ladislav 45
- Mladenow, Andreas 56
- Negash, Behailu 159
- Novak, Niina Maarit 56
- Novák, Richard 119
- Pavlicek, Antonin 45, 119
- Plosila, Juha 168
- Shi, Yiwei 132
- Siddiqui, Ali Shuja 168
- Song, Lei 192
- Sridhar, K. T. 143
- Strauss, Christine 56
- Strizova, Vlasta 119
- Szabó, Ildikó 91, 104
- Tenhunen, Hannu 159, 168
- Ternai, Katalin 104
- Wang, Jianxin 132
- Wang, Junyu 132
- Westerlund, Tomi 159, 181
- Xu, Shaoxi 192
- Ye, Congcong 12
- Zhang, Xiaoteng 192
- Zheng, Lirong 132, 181
- Zhou, Bo 21
- Zou, Zhuo 181