Advances in Intelligent Systems and Computing 575

Radek Silhavy
Petr Silhavy
Zdenka Prokopova
Roman Senkerik
Zuzana Kominkova Oplatkova *Editors*

# Software Engineering Trends and Techniques in Intelligent Systems

## Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017), Vol 3

Springer

# Advances in Intelligent Systems and Computing

Volume 575

**Series editor**

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

## *About this Series*

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

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

## *Advisory Board*

More information about this series at http://www.springer.com/series/11156

Radek Silhavy · Petr Silhavy
Zdenka Prokopova · Roman Senkerik
Zuzana Kominkova Oplatkova
Editors

# Software Engineering Trends and Techniques in Intelligent Systems

## Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017), Vol 3

Springer

*Editors*

Radek Silhavy
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

Roman Senkerik
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

Petr Silhavy
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

Zuzana Kominkova Oplatkova
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

Zdenka Prokopova
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-57140-9 ISBN 978-3-319-57141-6 (eBook)
DOI 10.1007/978-3-319-57141-6

Library of Congress Control Number: 2017937147

Printed on acid-free paper

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

# Preface

This book constitutes the refereed proceedings of the Software Engineering Trends and Techniques in Intelligent Systems Section of the 6th Computer Science On-line Conference 2017 (CSOC 2017), held in April 2017.

Particular emphasis is laid on modern trends in selected fields of interest in software engineering. New algorithms, methods, and applications of intelligent systems in a software engineering are also presented.

The volume Software Engineering Trends and Techniques in Intelligent Systems brings and presents new approaches and methods to real-world problems and exploratory research that describes novel approaches in the field of software engineering and intelligent systems.

CSOC 2017 has received (all sections) 296 submissions, in which 148 of them were accepted for publication. More than 61% of accepted submissions were received from Europe, 34% from Asia, 3% from Africa, and 2% from America. Researches from 27 countries participated in CSOC 2017 conference.

CSOC 2017 conference intends to provide an international forum for the discussion of the latest high-quality research results in all areas related to computer science. The addressed topics are the theoretical aspects and applications of computer science, artificial intelligences, cybernetics, automation control theory, and software engineering.

Computer Science On-line Conference is held online, and modern communication technology which is broadly used improves the traditional concept of scientific conferences. It brings equal opportunity to participate to all researchers around the world.

The editors believe that readers will find the following proceedings interesting and useful for their own research work.

March 2017

Radek Silhavy
Petr Silhavy
Zdenka Prokopova
Roman Senkerik
Zuzana Kominkova Oplatkova

# Organization

## Program Committee

## Program Committee Chairs

Zdenka Prokopova, Ph.D., Associate Professor, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: prokopova@fai.utb.cz

Zuzana Kominkova Oplatkova, Ph.D., Associate Professor, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: kominkovaoplatkova@fai.utb.cz

Roman Senkerik, Ph.D., Associate Professor, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: senkerik@fai.utb.cz

Petr Silhavy, Ph.D., Senior Lecturer, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: psilhavy@fai.utb.cz

Radek Silhavy, Ph.D., Senior Lecturer, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: rsilhavy@fai.utb.cz

Roman Prokop, Ph.D., Professor, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: prokop@fai.utb.cz

Prof. Viacheslav Zelentsov, Doctor of Engineering Sciences, Chief Researcher of St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS).

## Program Committee Members

Boguslaw Cyganek, Ph.D., DSc, Department of Computer Science, University of Science and Technology, Krakow, Poland.

Krzysztof Okarma, Ph.D., DSc, Faculty of Electrical Engineering, West Pomeranian University of Technology, Szczecin, Poland.

Monika Bakosova, Ph.D., Associate Professor, Institute of Information Engineering, Automation and Mathematics, Slovak University of Technology, Bratislava, Slovak Republic.

Pavel Vaclavek, Ph.D., Associate Professor, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czech Republic.

Miroslaw Ochodek, Ph.D., Faculty of Computing, Poznan University of Technology, Poznan, Poland.

Olga Brovkina, Ph.D., Global Change Research Centre Academy of Science of the Czech Republic, Brno, Czech Republic & Mendel University of Brno, Czech Republic.

Elarbi Badidi, Ph.D., College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates.

Luis Alberto Morales Rosales, Head of the Master Program in Computer Science, Superior Technological Institute of Misantla, Mexico.

Mariana Lobato Baes, M.Sc., Research-Professor, Superior Technological of Libres, Mexico.

Abdessattar Chaâri, Professor, Laboratory of Sciences and Techniques of Automatic control & Computer engineering, University of Sfax, Tunisian Republic.

Gopal Sakarkar, Shri. Ramdeobaba College of Engineering and Management, Republic of India.

V.V. Krishna Maddinala, Assistant Professor, GD Rungta College of Engineering & Technology, Republic of India.

Anand N. Khobragade, Scientist, Maharashtra Remote Sensing Applications Centre, Republic of India.

Abdallah Handoura, Assistant Prof, Computer and Communication Laboratory, Telecom Bretagne, France

## Technical Program Committee Members

Ivo Bukovsky
Miroslaw Ochodek
Bronislav Chramcov
Eric Afful Dazie
Michal Bliznak
Donald Davendra
Radim Farana
Zuzana Kominkova Oplatkova
Martin Kotyrba
Erik Kral
David Malanik
Michal Pluhacek
Zdenka Prokopova
Martin Sysel
Roman Senkerik
Petr Silhavy
Radek Silhavy
Jiri Vojtesek
Eva Volna
Janez Brest
Ales Zamuda
Roman Prokop
Boguslaw Cyganek
Krzysztof Okarma
Monika Bakosova
Pavel Vaclavek
Olga Brovkina
Elarbi Badidi

## Organizing Committee Chair

Radek Silhavy, Ph.D., Tomas Bata University in Zlin, Faculty of Applied Informatics, email: rsilhavy@fai.utb.cz

## Conference Organizer (Production)

OpenPublish.eu s.r.o.
Web: http://www.openpublish.eu
Email: csoc@openpublish.eu

## Conference Website, Call for Papers

http://www.openpublish.eu

# Contents

# Improving Algorithmic Optimisation Method by Spectral Clustering

Radek Silhavy[✉], Petr Silhavy, and Zdenka Prokopova

Faculty of Applied Informatics, Tomas Bata University in Zlin,
nam T.G. Masaryka 5555, Zlin, Czech Republic
{rsilhavy, psilhavy, prokopova}@fai.utb.cz

**Abstract.** In this paper, a spectral algorithm for effort estimation is evaluated. As effort prediction method the Algorithmic Optimisation Method is employed. Spectral clustering is used in version of normalized Laplacian matrix and k-means algorithm is used for clustering eigenvectors. Results shows that clustering lowers a Mean Absolute Percentage Error by 6% and Sum of Squared Errors/Residuals is decreased by 43,5%. Difference in mean value of residuals is statically significant (p = 0.0041, at 0.05 level).

**Keywords:** Effort estimation · Clustering · Use case points · Algorithmic optimisation method

## 1 Introduction

The Algorithmic Optimisation Methods (AOM) as introduced by Silhavy R. et al. [1] represent an improvement of Use Case Points (UCP) [2]. Typically, in a software development, a project scope description is limited. Therefore, Use Case Model (UCM) based methods should be a winning strategy to accurate prediction of software size or later to derivate a software development effort. UCP is based on UCM analysis. In previously published studies [1, 3–8] authors present methods, which can be adopted for UCP tuning. AOM method is based on multiple-linear regression model (MLR); MLR is faced to lower performance, when data points have no close range. Therefore, method of selecting a proper subset are under investigation. One of those method is clustering analysis. Clustering is adopted for reducing a number of data-points. Azzeh and Nassif [9] recommends a method called bisecting k-medoids.

Azzeh et al. in their paper [10] presents hybrid model that consists of support vector machine and radial basis neural networks.

Bardisiri et al. [11] declares that clustering have significant effect on accuracy of development effort prediction. In [12] Bardisiri et al. improved their method by combination with Particle Swarm Optimization (PSO) [13] algorithm. Hihn et al. [14] investigates that nearest neighbour method has significantly more outliers than spectral clustering. Therefore, the spectral clustering is investigated in conjunction with AOM method. In this paper, we are aiming on question, if applying a spectral clustering improves ability of compared algorithm to predict a software size and secondly minimizing a prediction error. The rest of the article is structured as follows: Sect. 2 defines

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_1

the problem statement. Section 3 describes experiment evaluation. Section 4 presents the results we obtained. Finally, Sect. 5 summarises our conclusions and future work.

## 2 Problem Statement

AOM [1] method is based on linear regression, which is represent in the following formula (1):

$$UCP_{AOM} = a_1(UAW \times TCF \times ECF) + a_2(UUCW \times TCF \times ECF) \tag{1}$$

where attributes are adopted from UCP as follows:

- Unadjusted Actors Weights (UAW);
- Unadjusted Use Case Weights (UUCW);
- Technical Complexity Factors (TCF);
- Environmental Complexity Factors (ECF);
- $a_1, a_2$ correction coefficients, obtained by MLR.

Regression model used in AOM is a MLR where two predictors are employed. In (2) regression model in matrix form can be seen.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \times \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \Rightarrow \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \left(X^T \times X\right)^{-1} \times \left(X^T \times Y\right) \tag{2}$$

Where $x_{i1}$ and $x_{i2}$, $i = 1 \ldots$n, are obtained as follows:

$$x_{i1} = (TUAW_i \times TCF_i \times ECF_i) \tag{3}$$

$$x_{i2} = (UUCW_i \times TCF_i \times ECF_i) \tag{4}$$

AOM predicts new project size by using (1) identically without respect to project types or any other differences in project. Project similarity is important for obtaining more accurate prediction. Employing MLR on non-consistent project data points creates models, which are accurate in average. It is expected that clustering will improve a model accuracy and prediction error will be decreased.

### 2.1 Methods and Experiment Design

As declared in Problem Statement clustering is expected to be improvement to AOM. Therefore, the Algorithmic Optimisation Method with Clusters (AOM-C) algorithm was designed. The experiment of AOM-C is designed as follows:

(1) Create training and testing fold by hold-out validation method (70% testing).
(2) Run spectral clustering on training fold of dataset.
(3) Create AOM formula for each of cluster from step 2.

(4) Classify data points in testing fold.
(5) Predict a value of size by AOM formulas for each project. Formula is selecting according a cluster to which a newly predicted project belongs to.

Spectral clustering algorithm in step 2 is a method of unsupervised clustering, which is based on graph representation, where each data point is a node and edges between data points represents a similarity. Later the adjacency matrix W is created, where each cell in matrix correspond to edge' weight between to data points. Second matrix which is need is a G matrix, which represents a degree diagonal matrix, in which a cell presents a sum of weights corresponding to each node from graph, respectively a cell of matrix W. Finally, a Laplacian matrix $L = G - W$ is calculated and then normalized.

The L matrix is used for spectrum calculation, which is a key point in spectral clustering algorithm. Spectrum is a sorted list of eigenvectors of L matrix. In fact, the eigenvectors represent a data point of dataset and an eigenvalues of L matrix. Spectral clustering uses those eigenvectors as a feature. Clustering of features ca be performed by any of known algorithm. In this paper k-means algorithm is used.

In step 3 the AOM models for clustered training fold are obtained. In fact, a partial regression model for each cluster. The AOM [1] model process workflow can be compared to AOM-C process workflow in Fig. 1 (AOM) and in Fig. 2 (AOM-C).

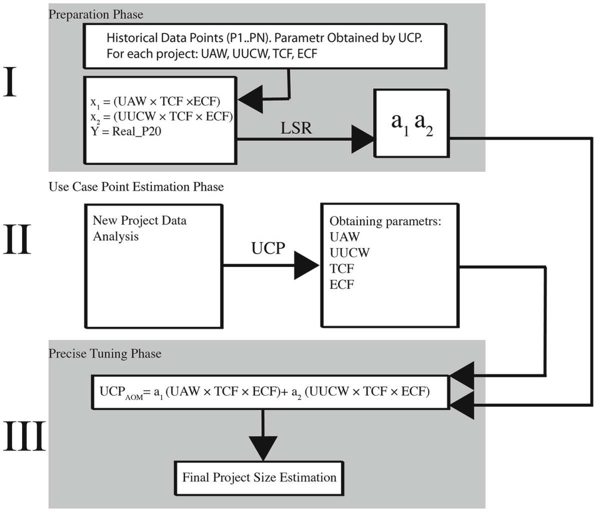

Preparation Phase
Historical Data Points (P1..PN). Parametr Obtained by UCP.
For each project: UAW, UUCW, TCF, ECF
I
$x_1$ = (UAW × TCF ×ECF)
$x_2$ = (UUCW × TCF × ECF)
Y = Real_P20
LSR
$a_1$ $a_2$
Use Case Point Estimation Phase
II
New Project Data Analysis
UCP
Obtaining parametrs:
UAW
UUCW
TCF
ECF
Precise Tuning Phase
$UCP_{AOM}$= $a_1$ (UAW × TCF × ECF)+ $a_2$ (UUCW × TCF × ECF)
III
Final Project Size Estimation

**Fig. 1.** AOM workflow

The AOM – see Fig. 1 method is composed of three phases. Phase I is used for obtain based correction coefficients $a_1, a_2$, which are later used in Phase III for new prediction. Phase II is only used for obtaining variables (UAW, UCW, ECF, ECF), which are used in Phase III for new prediction according (1).

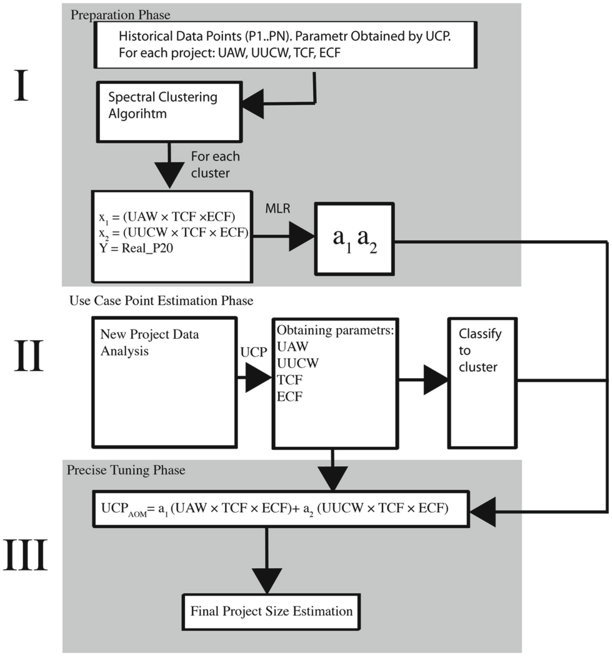

Preparation Phase
Historical Data Points (P1..PN). Parametr Obtained by UCP. For each project: UAW, UUCW, TCF, ECF
I
Spectral Clustering Algorihtm
For each cluster
x1 = (UAW × TCF ×ECF)
x2 = (UUCW × TCF × ECF)
Y = Real_P20
MLR
a1 a2
Use Case Point Estimation Phase
II
New Project Data Analysis
UCP
Obtaining parametrs: UAW UUCW TCF ECF
Classify to cluster
Precise Tuning Phase
UCP_AOM= a1 (UAW × TCF × ECF)+ a2 (UUCW × TCF × ECF)
III
Final Project Size Estimation

**Fig. 2.** AOM-C workflow

AOM-C – see Fig. 2 method differs in Phase I, in which a historical data points are clustered by means of spectral clustering and correction values $a_1, a_2$ are calculated individually for each of defined clusters. In Phase III is therefore mandatory to classify newly predicted project to proper cluster. This a mandatory step, because of values $a_1, a_2$ have to be select from identified.

## 2.2 Hypothesis Formulation and Research Questions

This paper compares the accuracy of the AOM model with that of the AOM-C model using Wilcoxon rank sum test. The Wilcoxon test is used as a test of the null hypothesis that the means $(\mu)$ of two normally distributed populations are equal. The Wilcoxon test will be used for evaluation of residuals.

H$_0$: $\mu_{AOM} = \mu_{AOM-C}$, there is no difference in the capability of prediction between AOM and AOM-C models. No difference between mean Sum of Squared Errors (SSE).

Alternative hypothesis:

H$_1$: $\mu_{AOM} \neq \mu_{AOM-C}$, there is difference in prediction capability between AOM and AOM-C models. There is a difference between mean residuals.

H$_2$: $\mu_{AOM} > \mu_{AOM-C}$, there is difference in prediction capability between AOM and AOM-C models. There is a statistically significant evidence, that clustering brings better accuracy and lower prediction error.

These hypotheses allow us to conclude a research question:

RQ1: Applying spectral clustering we will be able to improve prediction accuracy of AOM method.

### 2.3 Evaluation Criteria

AOM and AOM-C models were evaluated according to (5) Mean Absolute Percentage Error (MAPE), (6) the adjusted coefficient of determination (R$^2$), (7) the Sum of Squared Errors/Residuals (SSE).

$$\mathrm{MAPE} = \frac{1}{n}\sum \frac{|y_i - \widehat{y}_i|}{y_i} \times 100, \tag{5}$$

$$R^2 = 1 - \left[\frac{\sum_i^n (y_i - \widehat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}\right], \tag{6}$$

$$SSE = \sum\nolimits_{i=1}^{n} \varepsilon_i^2, \tag{7}$$

where, $n$ is number of observations, $k$ is the number of independent predictors, $y_i$ is an observed value,$\widehat{y}_i$ is predicted value, $\bar{y}$ is a mean value of predicted values and $\varepsilon$ is a residual error value. For AOM-C model $R^2$ will be presented as mean value, because there is more than one MLR model to cover all data points in the dataset.

## 3 Experiment Evaluation

### 3.1 Project Dataset

The experiment described above was evaluated using real-life dataset, which was collected and firstly used in [15]. For purpose of this study all data points were used. Therefore, this a public historical data experiment, no a company only data were applied. Figure 3 shows a histogram of dataset. As shown, the dataset represents mid and large project size. Dataset characteristics are summaries in Table 1.

Dataset is freely available for replication study, or further experiments in [15]. Value of Real_P20, which is used for histogram (Fig. 3) and for Table 1 is a representation of project size in UCP points. Real_P20 was obtained by dividing a project effort in person-hours by a Productivity Factor (PF), where PF = 20.

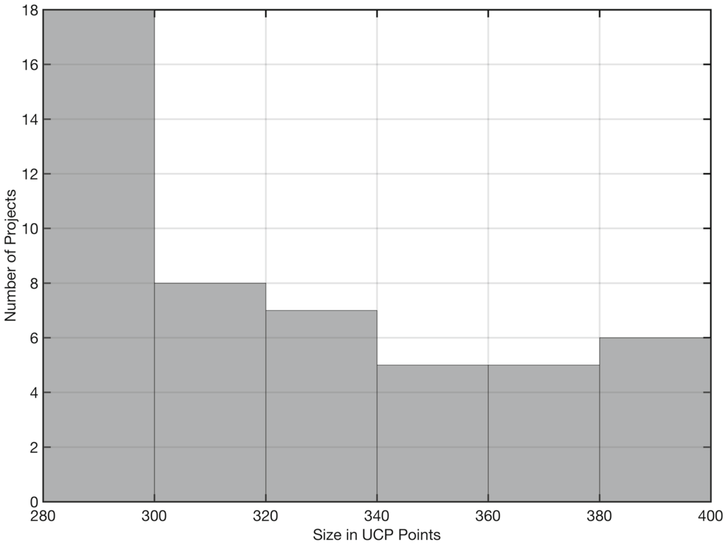


**Fig. 3.** Dataset histogram, Real_P20 size is used

**Table 1.** Dataset Characteristics

| | Median Person-Hours | Median Real_P20 | Range Real_P20 | Standard Deviation | Minimum Real_P20 | Maximum Real_P20 | n |
|---|---|---|---|---|---|---|---|
| Dataset | 7,012.000 | 320.600 | 109.750 | 33.394 | 288.750 | 398.500 | 70 |

# 4 Results

## 4.1 AOM-C, MLR Models

AOM-C method rely on properly selected number of cluster. In this paper for clustering in spectral clustering algorithm is used a k-means algorithm, when cosine distance measurement is applied. Therefore, number of cluster have to be predefined. In Fig. 4 you can see an evaluation of number of cluster. On Fig. 4 it shown a trend of silhouette value for 2, 3, 4, 5, 6 and 7 clusters, where $x_{i1}$, $x_{i2}$ and $Real_{P20}$ were used as a feature for similarity graph construction. As can be seen a number of 2 cluster brings highest silhouette level (0.7592), therefore 2 cluster are the best option for training data (n = 21).

As can be seen in Fig. 5 more than 4 clusters brings surprisingly better solution and prediction capability of models are significantly better. In Table 2 correction coefficients $a_1, a_2$ for each of three cluster can be found. In Fig. 6 classification of testing data points is presented, as can be seen new project are correctly identified within one of available clusters.

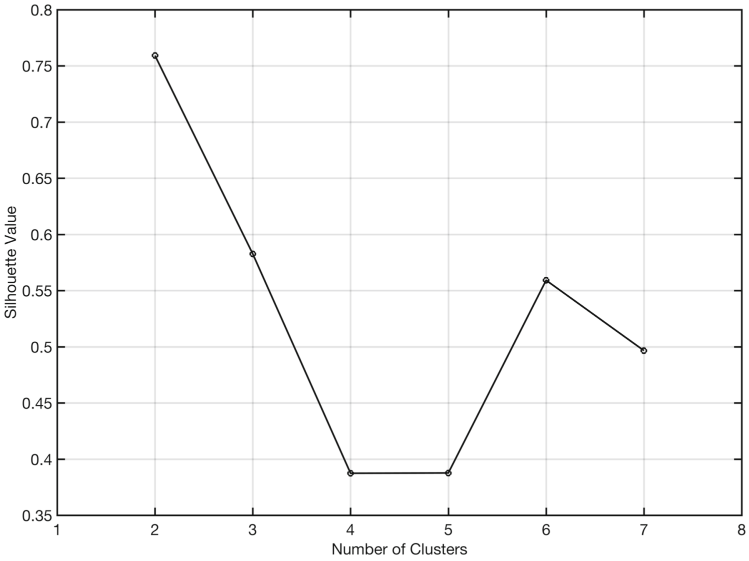


**Fig. 4.** Silhouette values for 2–7 clusters (n = 21, training set)

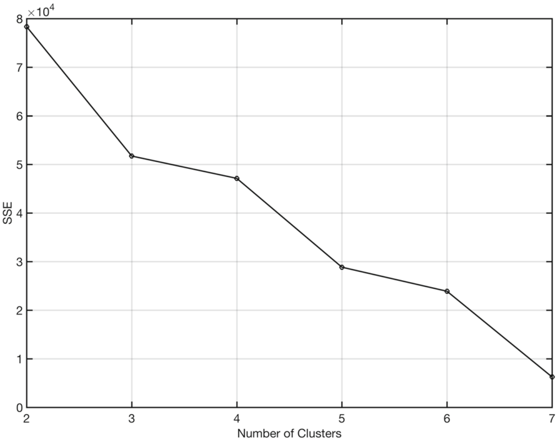


**Fig. 5.** SSE values for 2–7 clusters, (n = 21, training set)

**Table 2.** Values of $a_1, a_2$ for 4-clusters' AOM-C solution (n = 21, training set)

| AOM-C Cluster No. | $a_1$ | $a_2$ | $R^2$ | SSE |
|---|---|---|---|---|
| Cluster 1 | −0,81 | 1,06 | 0.25 | 159.04 |
| Cluster 2 | −5.17 | 0.85 | 0.36 | 427.21 |
| Cluster 3 | −35.81 | 1.48 | 0.02 | 394.64 |
| Cluster 4 | 43.59 | 0.55 | 0.51 | 40,872 |

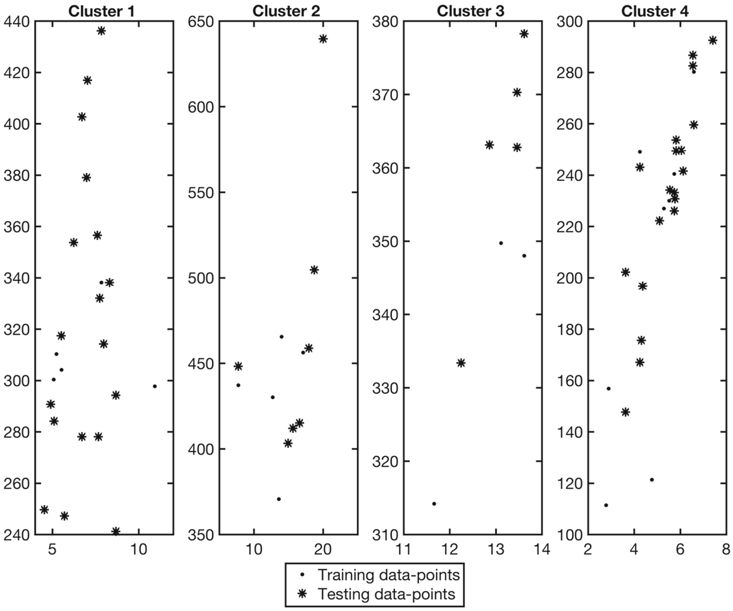


**Fig. 6.** Classification of testing data-points

## 4.2 AOM, MLR Model

AOM model is need for comparison, when for AOM-C four models were created, for AOM only one is needed. In Table 3 same attributes as were presented in Table 2 are used.

**Table 3.** Values of $a_1, a_2$ for AOM solution, training data-points

| | $a_1$ | $a_2$ | $R^2$ | SSE |
|---|---|---|---|---|
| AOM | −14.82 | 1.37 | 0.45 | 222,330 |

## 4.3 Model Prediction Performance

When MLR models for AOM-C and for AOM are created, we can employ a testing data-points (n = 49), which were hold-out from original dataset. In the Table 4 there can be seen a performance of both models. For AOM-C a mean value for $R^2$ is presented. MAPE and SSE values were derivate in same manner as for AOM. This is possible because each data-point is use only once, no duplicity in clusters.

**Table 4.** Comparison of AOM-C and AOM models

| Method | *MAPE* | $R^2$ | SSE |
|---|---|---|---|
| AOM-C | 17 | 0.29 | 243,950 |
| AOM | 23 | 0.45 | 433,510 |

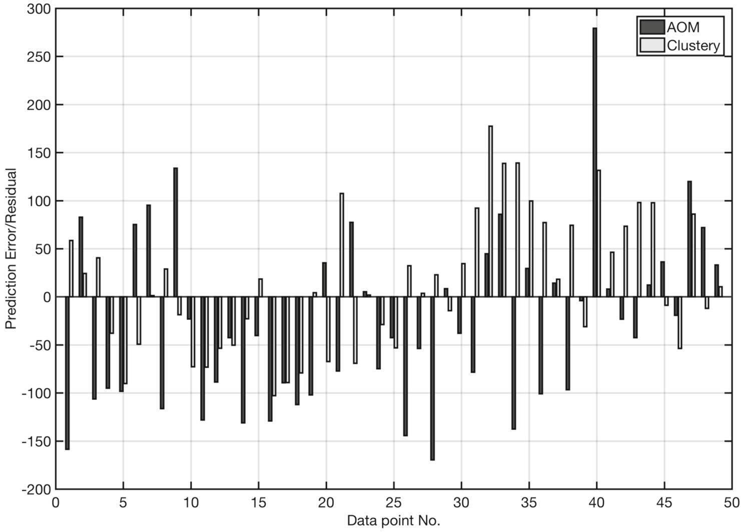


**Fig. 7.** Prediction Error for 49 testing data-points

# 5 Conclusion

The main purpose of this paper is study an effect of spectral clustering on AOM method. As can be seen in Table 4. AOM-C has SSE value lower than AOM (43,7%). It also perform better from point of view of MAPE (−6%), AOM-C is statically better, when Wilcoxon 1-side left tailed test is applied (p = 0.0041, at 0.05 level). Therefore, we $H_1$ $H_2$ can be rejected. There is an evidence that spectral clustering improves AOM.

Regarding a RQ1, we can confirm that using spectral clustering will increase a performance of AOM, but surprisingly better when data points seems not correctly clustered according Silhouette value (see Fig. 4). When Silhouette values should is considered, then 2-clusters solution should be used. Contrastingly, if SSE is employed than 4-clusters should be used. As can be seen in Fig. 7 AOM-C estimates better for majority of data points.

When we compare regression model it can be seen that ability to predict is better when partial models (each cluster using its own model). The reason can be seen in simplicity of regression method of AOM, which useful for practitioners, but brings a lower prediction capability, when clusters are no form linearly. K-means algorithm is used distance measurement from centroid, which is not best option for linear model application.

In our future research we will focused on evaluating other clustering methods and comparing it with method of sub-set selection. Windowing, weighted windowing will be evaluated too. AOM will be compared to Model D approach [15], which is based polynomial form of regression (obtained by stepwise approach), and finally machine learning methods will be considered. Spectral clustering is interesting option, therefore other optimal setting of similarity graph construction and for eigenvectors clustering will investigated.

## References

1. Silhavy, R., Silhavy, P., Prokopova, Z.: Algorithmic optimisation method for improving use case points estimation. PLoS ONE **10**, e0141887 (2015)
2. Karner, G.: Metrics for objectory. Diploma, University of Linkoping, Sweden, No. LiTH-IDA-Ex-9344, vol. 21, December 1993
3. Ochodek, M., Alchimowicz, B., Jurkiewicz, J., Nawrocki, J.: Improving the reliability of transaction identification in use cases. Inf. Softw. Technol. **53**, 885–897 (2011)
4. Ochodek, M., Nawrocki, J., Kwarciak, K.: Simplifying effort estimation based on Use Case Points. Inf. Softw. Technol. **53**, 200–213 (2011)
5. Anandhi, V., Chezian, R.M.: Regression techniques in software effort estimation using cocomo dataset. In: 2014 International Conference on Intelligent Computing Applications (ICICA 2014), pp. 353–357 (2014)
6. Jorgensen, M.: Regression models of software development effort estimation accuracy and bias. Empirical Softw. Eng. **9**, 297–314 (2004)
7. Nassif, A.B., Ho, D., Capretz, L.F.: Towards an early software estimation using log-linear regression and a multilayer perceptron model. J. Syst. Softw. **86**, 144–160 (2013)
8. Urbanek, T., Prokopova, Z., Silhavy, R., Vesela, V.: Prediction accuracy measurements as a fitness function for software effort estimation. Springerplus **4**, 17 (2015)
9. Azzeh, M., Nassif, A.B.: Analogy-based effort estimation: a new method to discover set of analogies from dataset characteristics. IET Softw. **9**, 39–50 (2015)
10. Azzeh, M., Nassif, A.B.: A hybrid model for estimating software project effort from Use Case Points. Appl. Soft Comput. **49**, 981–989 (2016)
11. Bardsiri, V.K., Jawawi, D.N.A., Hashim, S.Z.M., Khatibi, E.: Increasing the accuracy of software development effort estimation using projects clustering. IET Softw. **6**, 461–473 (2012)
12. Bardsiri, V.K., Jawawi, D.N.A., Hashim, S.Z.M., Khatibi, E.: A flexible method to estimate the software development effort based on the classification of projects and localization of comparisons. Empirical Softw. Eng. **19**, 857–884 (2014)
13. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: 1995 IEEE International Conference on Neural Networks Proceedings, vol. 1–6, pp. 1942–1948 (1995)
14. Hihn, J., Juster, L., Johnson, J., Menzies, T., Michael, G.: Improving and expanding NASA software cost estimation methods. In: IEEE Aerospace Conference 2016, pp. 1–12 (2016)
15. Silhavy, R., Silhavy, P., Prokopova, Z.: Analysis and selection of a regression model for the Use Case Points method using a stepwise approach. J. Syst. Softw. **125**, 1–14 (2017)

# FRDF: Framework for Reliable Data Fusion to Leverage Communication Performance in Sensor Network

B.S. Jayasri(✉) and G. Raghavendra Rao

Department of Computer Science,
National Institute of Engineering, Mysore, India
jayasriphdl4ni@gmail.com

**Abstract.** Data fusion technique in wireless sensor network assists in enhancing the data quality in wireless sensor network. Unfortunately, it does it at the cost of energy and uncertainty of data forwarding. We have reviewed various existing data fusion schemes and found that the studies claim to have increase in throughput by using supportability of multipath, but very fewer studies have emphasized on reliable data fusion, considering assurity of data forwarding to destination. Therefore, we introduce a model called as Framework for Reliable Data Fusion in wireless sensor network in order to address these problems. The paper discusses about three core algorithm i.e. algorithm for (i) distributed tree construction, (ii) reliable data fusion, and (iii) data forwarding reliability. The outcome of the proposed system is found to excel superior with respect to energy saving and enhance communication performance in comparison to existing techniques.

**Keywords:** Data fusion · Fuser node · Reliability · Energy efficiency · Wireless sensor network

## 1 Introduction

Wireless sensor network is one of the most pivotal areas of research under wireless network due to its beneficial factors as well as continued unsolved problems associated with it. Such forms of the network consist of three types of nodes member node, cluster head, and sink node [1]. Member node extracts the raw environmental data, uses TDMA scheduling, and forwards the data to cluster head, which in turn accumulates multiple data from the member nodes.

This process is just called as data collection. However, during data collection, it is quite possible that a cluster head collects redundant data and forward it to sink, in which case it may result in additional control overhead. In order to solve this problem, a cluster head perform filtering of the redundant data using statistical computation, to forward the unique data to the base station [2]. This principle is called as data fusion and plays a significant role in enhancing the data quality in wireless sensor network [3]. However, the biggest challenging problem is uncertainty associated with the exact time of event generation, which leads to higher degree of energy variance that further

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_2

significant reduces the lifetime of the wireless sensor network. As the wireless sensor network works on the principle of radio-energy model [4], it can be said that enhanced communication performance demands more amount of energy. Hence, data fusion is also closely linked with energy problems. However, to some extent the available techniques of energy efficient routing can enhance the performance of data fusion, but still the problem lies in few facts e.g. (i) inability to identify and repair an intermittent links over increasing rounds of iteration, (ii) there is assurity that fused data will be reaching its destination (it could be any other node in multihop or base station in single hop). Although, there are many existing systems that has focused on energy efficient data fusion, but still they do not ensure reliability. Here reliability is the term, which relates to forwarding of the fused data from distributed data fusion to the destination node. Reliability factor becomes more important, when compared to energy saving, especially with respect to time-critical applications in wireless sensor network. There are various applications, which need sensitive messages to reach the destination at any cost, and with high degree of reliability. Hence, this paper presents one such novel technique that assists in enhancing the reliability of the data fusion in wireless sensor network.

The rest of the paper is organized as follows. The background of the proposed study is discussed in Sect. 1.1 along with the contribution of some of the existing research work. Section 1.2 briefly describes the problems statement after reviewing the existing system, the proposed solution contribution is described in Sect. 1.3. Section 2 presents elaborate discussions of the algorithms implemented to ensure energy efficiency and reliability during data fusion over large-scale wireless sensor network and are followed by the result discussion in Sect. 3. Finally, the Sect. 4 is used to conclude our paper.

### 1.1 Background

This section discusses about the existing techniques that have been discussed most recently in the area of data fusion in wireless sensor network. Our prior study [5, 6] has already reviewed some of the standard techniques of data fusion. We have also presented a model called as EEDF (Energy Efficient Data Fusion) most recently that focuses on incorporating virtual multipath data fusion in wireless sensor network [7]. Zhu et al. [8] recently carried out the problem of reliable data transmission. Considering up and downlink communication path, the author has presented a model that can perform an energy-efficient data fusion in wireless sensor network. Liu et al. [9] have presented a study to emphasize the need of energy efficiency during data fusion modelling. Li et al. [10] who have used back propagation algorithm in order to enhance the convergence performance of fuser node in wireless sensor network carried out studies towards optimization of data fusion. Jorio et al. [11] have presented a fusion technique using geographic routing approach along with hierarchical techniques. Ma et al. [12] who have introduced a clustering technique along with a unique fusion model using reputation factor also addressed the problem of reliability in data fusion. Dingcheng et al. [13] have presented a technique that uses Bayesian approach to perform data fusion. The study outcome was testified using standard dataset using squared error. Zhou et al. [14] have introduced a technique that uses Dempster-Shafer evidence theory in order to enhance the performance of multi-sensor data fusion. Reliability problem in data fusion was also addressed in the work carried out by Peng et al. [15] where the outcome was found to posses reduced delay. Wang and Dong [16] have adopted unscented Kalman filter to leverage the performance of Multiscale data fusion in wireless sensor network using both simulation and experimental approach. Bangash et al. [17] have introduced a technique that ensures reliable routing over body sensor network considering constraint data packets. Testified with and without relay nodes, the study outcome was found to posse's better delivery ratio. Luo and Li [18] have presented a technique that ensures the better fault tolerance performance for distributed fusion in wireless sensor network. The technique performs quantization of its sensed data considering multiple network constraint. The study outcome was tested with ROC curves. Tan et al. [19] have presented a study to show an impact of data fusion over coverage and connectivity in wireless sensor network. Using explicit scaling laws, the technique performs data fusion. Yuan et al. [20] have developed a technique of data fusion where threshold plays an important role. The author uses local and global thresholding scheme for performing threshold-based data fusion computation.

Hence, it can be seen that there are various techniques for addressing data fusion problems where such study has contributed to new guidelines in enhancing performance, but at the same time, the existing studies are also shrouded with many pitfalls. The next section discusses about the problems being identified in the existing system.

## 1.2 The Problem

The problems explored after reviewing the existing system are as follows:

- Majority of the study was focused on energy efficiency and very less number of studies has actually addressed the reliability problems in data fusion.
- Less focus on distributed data fusion techniques leads to non-exploration of the problems associated with large-scale data fusion.
- Few studies are dedicated towards identification of intermittent links caused due to ineffective design of data fusion in wireless sensor network and to autonomously address it.
- Studies pertaining to energy efficient data aggregation, routing, and spontaneous routing updates are quite less.

The problem statement of the proposed study is –"*It is a computationally challenging task to develop a framework that can formulate the condition of reliability in data fusion along with assurity to forward the fused data to sink in wireless sensor network*." The next part of the study presents the contribution.

## 1.3 The Proposed Solution

The proposed system introduces a model called as Framework for Reliable Data Fusion (FRDF) in wireless sensor network. The prime target of this modelling is to incorporate some conditional feature during data fusion in order to ensure that fused packets once relayed must reach the base station irrespective of the presence of intermittent or broken links over a multiple hop communication system. The main emphasis is laid on to, developing a technique that can estimate the minimum amount of energy, which is required to achieve 100% reliable data fusion operation in wireless sensor network. The research approach will be purely analytical. The architecture of proposed system is shown in Fig. 1

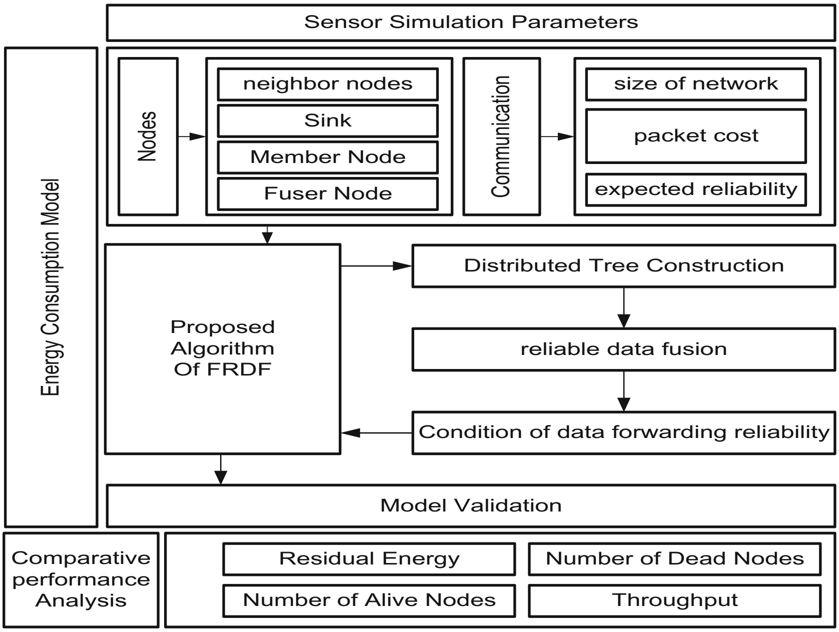


**Fig. 1.** Proposed architecture of FRDF

The contributions of proposed system are as follows:

- A simple graph theory is used for implementing a novel design of data fusion in order to model the routing strategy from nodes to sink.
- A new parameter called as cost of data fusion is designed where cost means the accurate amount of information contained in fused data after performing data fusion. It also relates to amount of information being received by the node and is affected by the tree structure of data fusion.

- An algorithm is developed that can compute the extent of data reliability after the fusion is performed considering amount of fused data received at base station to total number of data being disseminated from the parent node (or cluster head).
- A novel condition for data fusion reliability is introduced that considers that a sensor with maximized cost will also require maximized data forwarding reliability in order to achieve a targeted reliability score as well as minimization of the cumulative energy consumption during transmission.

The next section elaborates the above-mentioned contribution of FRDF with respect to algorithm design implementation.

## 2 Algorithm Implementation

The proposed system implements three different algorithms in order to ensure energy efficient and incorporate reliability feature in data fusion over large-scale wireless sensor network. Current section is used to discuss the algorithms involved in the design of proposed system.

### 2.1 Algorithm for Distributed Tree Construction

This algorithm is mainly an extension to our prior EEDF [7] by strengthening the contraction of the distributed tree in order to assort good number of routes. All the computation is carried out with respect to hops right from each node to base station. We consider that a base station broadcast a particular message called as $DTC_{msg}$ to all the nodes (Line-1). The design of the $DTC_{msg}$ is kept simple by keeping only 2 fields in its data frame i.e. (i) identity $I$ of the node and (ii) hop distance $d$, which will be used by the $DTC_{msg}$ to pass. For simpler computation, we initiate the hop distance d with value 1 at the position of the base station. Only the nodes $x$ which receives the beacon $DTC_{msg}$ (Line-2) performs validation if the hop distance value is less than hop distance value that it has stored (Line-4). The base station S also forwards a query for network size, cost of packet, and anticipated reliability. Cost of packet will mean size of packet in our fusion model and reliability factor can be computed by packet cost towards sink divided by total packet cost. For true condition, the sensor updates the value of next hop i.e. ($x + 1$) with the value of the field $ID$ of $DTC_{msg}$ (Line-5). At the same, it also updates the value of hop distance $d$ of $\mathrm{DTC}_{\mathrm{msg}}$. The sensor also performs relaying of the DTCmsg (Line-1–8). However, identification of false condition will mean that the sensor has already acknowledged the $\mathrm{DTC}_{\mathrm{msg}}$ for which reason it has to reject the $\mathrm{DTC}_{\mathrm{msg}}$.

**Algorithm for Distributed Tree Construction**

```
Input: S (Sink), $DTC_{msg}$ (Distributed Tree Construction message), $\eta$ (size of network), $pkt_{wt}$ (packet cost), $\alpha_{pkt}$ (expected reliability)
Output: Distributed Tree Construction
Start
1. $S \rightarrow broadcast(DTC_{msg, d})$
2. $x \leftarrow received(DTC_{msg})$
3. $S \rightarrow query(\eta, pkt_{wt}, \alpha_{pkt})$
4. if $d(x)>d(DTC_{msg})$ and flag(x)
5. then, $(x+1) \leftarrow ID(DTC_{msg})$
6.    $d_x \leftarrow d(DTC_{msg})+1$
7.    $ID(DTC_{msg}) \leftarrow IDx$
8.    $d(DTC_{msg}) \leftarrow d_x$
9.    $flag_x \leftarrow false$
10. End
11. else
12. x reject($DTC_{msg}$)
13. End
End
```

## 2.2 Algorithm for Reliable Data Fusion

This algorithm is initiated only after the sensors started detected any significant event and results in selection of reliable data fuser node. All the active sensors participate in the selection of best fuser node in this algorithm. The proposed concept assumes that cluster head usually resides near to the base station. However, as we implement random distribution which may result in farthest distance too (from sink to node). In such case, we select the cluster head on the existing established link (Line-6–7). In this algorithm, initially the sensors are shortlisted based on the event sensing capability of a node i.e. $\eta_{evn.}$ All the monitored node $x$ can be considered to be subset of event detected sensors and hence can be referred to be eligible for playing the role of fuser node. In such case, fuser node $x$ will announce the identified event in the form of broadcasting. All the neighbor nodes that have received this broadcasted message will perform a computation shown in Line-5–16. The neighboring nodes checks if the distance value of $x$ is more than distance value of neighbors, than the broadcasting node $x$ can play the role of member node. A member node will perform retransmission of broadcasted message, which are received from neighbor node $\omega$. However, if the distance of $x$ is equivalent to distance of neighbor node $\omega$ than the neighbor node checks identity of $x$ is more than that of neighbor node $\omega$. In such, the broadcasting node will play the role of member node itself. The computation continues until a single node is found to be announced as reliable fuser node. The algorithm allows all the nodes that have sensed the similar event to be acting as member nodes, whose responsibility is to collect the information and then forward it to the fuser node that in turn finally forwards it to the base station. The prime beneficial factor of this algorithm is that the sensors capturing the equivalent events at one node only i.e. reliable fuser node, which consumes less energy, filters more redundancies, and ensures smaller travel period, fuse entire data.

**Algorithm for Reliable Data Fusion**

```
Input: $\eta_{evn}$ (sensors detected an event), $\omega$ (neighbor nodes)
Output: x (reliable data fuser node)
Start
$\eta_{evn} \leftarrow$ event
for each x$\varepsilon$ $\eta_{evn}$
    x$\leftarrow$fuserNode
    x$\rightarrow$announce(identifiedEvent)
    foreach ($\omega\varepsilon$neighborNode)
        if d(x)>d($\omega$) then
            x$\leftarrow$memberNode
        end
        elseif d(x)=d($\omega$)
        ID(x)>ID($\omega$) then
            x$\leftarrow$memberNode
        end
        else
            x rejects broadcast from $\omega$
        end
    end
End
```

## 2.3 Algorithm for Condition of Data Forwarding Reliability

The main purpose of this algorithm is to ensure 100% reliability such that the data will be forwarded among the sensors network. In this case, the new fuser node x that has been selected from previous algorithm (Algorithm for reliable data fusion). Then the selected fuser node x transmits a message for establishing route to its next hop represented as (x + 1). After the acknowledgement is received by the next hop (x + 1), it forwards it to all its descendants which significantly assists in updating the distributed tree on other hand. The iteration continues until and unless the search terminates after obtaining base station. In case it does not reach the base station, then it alternatively searches for a sensor, which is already present in an established route. Hence, it can be seen that proposed system has good supportability of multihop communication system during fusion operation in sensor network. We construct a condition of reliability to establish the routes by selecting the superior neighbouring nodes present at each hops. The contraction of condition is carried out in dual stages i.e. (i) after the preliminary event, the sensor with shortest path to base station and (ii) after an event has already occurred, the sensors that are already in an established path. One of the key advantage of this algorithm is that it enhances the fusion points, in order to ensure that data fusion will occur more to generate more reliable data and also will be transmitted with increased reliability. Finally, the outcome of this algorithm is basically a tree that links the reliable data fuser node with the base station. This algorithm also supports higher frequency of updates during routing, and mostly carried out by relay nodes. The relay node forwards messages for updating the hops. The algorithm also checks, if the sensor does have data packets to be forward to more than one branches $br_y$. In positive case, the sensor waits for a specific period and fuses all the data and then forwards the data to its consecutive hop (x + 1) or else it directly forward it to next hop (x + 1). For every iteration, we identify the need to increase the data reliability by flagging the situation of data delivery. The reliability is increased by ensuring that a transmitting node must wait for pre allocated duration in order to get the acknowledgement of data deliver. In case of failure to receive the acknowledgement, a new receiver node is chosen and the acknowledgement message is reforwarded through that node.

**Algorithm for Condition of Data Forwarding Reliability**

```
Input: x(fuser node of current event), y (node receiving msg from current fuser node x), $br_y$ (branches of y node)
Output: Ensuring each node get data to transmit
Start
x→msg(x+1)
Iterate
    if y=(x+1) then
        d←0
        y←relay
    end
Until sink node is found
Iterate
    If $br_y$>1 then
        fuseddata→(x+1)
        if y←relay
            Flag→Reliability need to be increased
        end
    end
    else
        transmit pkt→(x+1)
        if y←Relay
            Flag→Reliability need to be increased
        end
end
Until sensor get data to transmit.
End
```

The prime advantage of proposed system is to construct a reliable path in the form of the tree that can forward unique data to the base station and thus enhance the data fusion process in wireless sensor network. Another advantage is usage of flagging mechanism to identify which are all the consecutive routes that have been failed to be detected. In such case, the failed route is compensated by exploring new consecutive route. The proposed system uses a simple cost computation, which is the amount of

**Table 1.** Notation used in algorithm design

| Notation | Meaning |
|---|---|
| S | Sink node |
| $DTC_{msg}$ | Distributed Tree Construction message |
| X | Node receiving $DTC_{msg}$ |
| D | Hop distance |
| H | Size of network |
| $pkt_{wt}$ | Packet cost |
| $\alpha_{pkt}$ | Expected reliability |
| $\Omega$ | Neighbor nodes |
| $br_y$ | Branches of y node |
| $\eta_{evn}$ | Sensors detected an event |

non-redundant fused data in order to ensure lesser retransmission attempt. It also displays an efficient data fusion reliability model (Table 1).

## 3 Result Discussion

The result analysis is carried out by simulating 100 sensor nodes across 1200 × 1500 $m^2$ simulation area. The nodes are initialized with 0.5 J of energy with packet length of 20000 bytes. The outcome of proposed Framework for Reliable Data Fusion (FRDF) is compared with our prior model EEDF [7] and LEACH [21] with respect to residual energy, alive nodes, throughput, and dead nodes.

Figure 2 highlights the comparative analysis of the residual energy which shows that sustenance of LEACH [21] is restricted to extremely less simulation rounds, which is exceeded by prior EEDF [7] model. EEDF formulates better link selection criteria based on stabilized energy, which cannot be found in LEACH. Moreover, EEDF also supports multihop resulting with less occurrences of error thereby saving significant amount of energy as compared to LEACH. However, proposed FRDF further optimizes EEDF by ensuring better route and data reliability conditions. Our present model has a mechanism of flagging the unestablished links (i.e. link created between two nodes with less energy) along with identification of failed link discovery, which adds quite value to prior structure of data fusion. Another enhancement in our proposed approach is the inclusion of cost factor, which significantly reduces the redundancies thereby further reducing transmittance energy consumption. The inclusion of new reliability factor is dependent on the data cost of sink to total, which can be obtained during the routing operation and does not require any extra computational steps. Hence, FRDF does not require much computational steps and hence enough residual energy is found to revive the sensors for more rounds.

Figures 3 and 4 shows the graphical outcome of alive and dead nodes respectively. The plotted values in both the figures can be founded to be symmetric to each other with respect to number of alive and dead nodes. Proposed FRDF uses scope-based flooding process, which is not only used for route discovery but also for updating routing information. This process simultaneously enables all the neighbouring nodes to

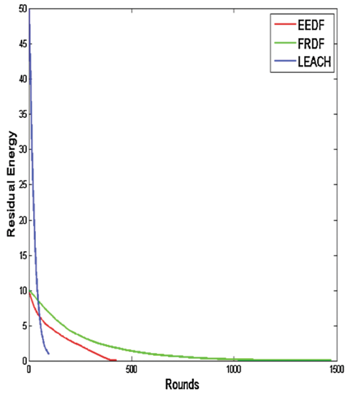


**Fig. 2.** Comparative analysis of residual energy

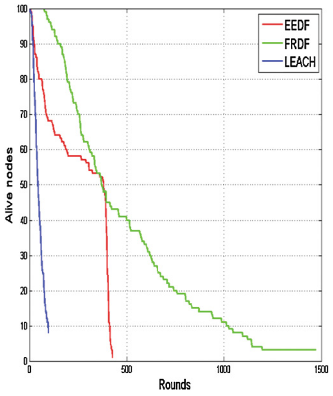


**Fig. 3.** Comparative analysis of alive nodes

establish and update the root along the path at the same time causing significant reduction in energy due to the effort exerted by other nodes to get the update. Moreover, reduction of retransmission positively assists in further increasing the longevity of network lifetime.

As per the observation made in Fig. 4 with respect to dead nodes, FRDF tends to be quite predictable and does not exhibit linearity as LEACH or uncertainty associated similarly to that of EEDF. The reason behind LEACH behaving linearly is due to excessive energy drainage during routing and clustering owing to inefficient selection process or cluster head.

Moreover, LEACH does not support multihop. The limitation of EEDF is basically due to the uncertainty associated with the construction of routes based on topology and current channel condition without including any technique to identify/discover/rectify the broken link, if any and Moreover, the best link selection is based on energy level only. However, whereas in proposed FRDF, there are multiple factors and condition that leverage the fuser node selection process. The route formation is based on three different forms of routing message, which gives more comprehensive information about the nodes and links using distributed tree construction process. Therefore, the death of the nodes occurs very slowly and uniformly in FRDF, when compared to LEACH and EEDF.

Figure 5 highlights the throughput accomplished by the proposed system FRDF with comparison to existing techniques LEACH and EEDF. EEDF provides a better throughput than LEACH, as it uses a mechanism of virtual multipath propagation to increases the data delivery in multifold, however, EEDF does not have the capability to identify the reliable routes and hence leads to sudden drop in throughput. Thus, it indirectly affects the network longevity by using more transmittance energy during path failure. This problem is completely solved by our proposed system FRDF that not only identify and formulate the condition for data reliability, but also rectifies the problems of intermittent links. Hence, throughput of our proposed system FRDF significantly outperforms our prior models LEACH and EEDF.

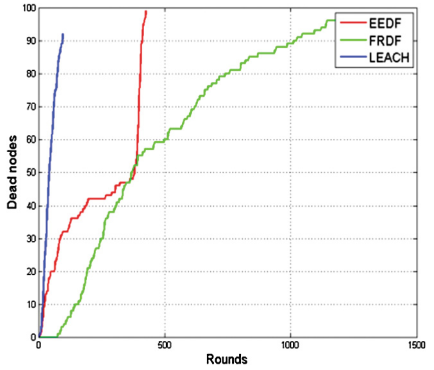


**Fig. 4.** Comparative analysis of dead nodes

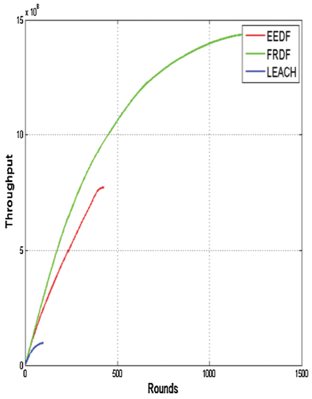


**Fig. 5.** Comparative analysis of throughput

# 4 Conclusion

Data fusion plays a very important role in increasing the data quality by filtering significant amount of redundant data during the data dissemination process in wireless sensor network. We have reviewed some of the recently implemented and relevant techniques that is known to enhance the performance of the data fusion process in wireless sensor network. We find that all the existing techniques are more or less focusing on accomplishing the energy efficiency, but at the same time unable to ensure, if the fused data could reach the destination node over uncertainty of the intermittent nodes in wireless sensor network. Hence, reliability is the less focused technique in this perspective. This paper presents a framework that can enhance the data fusion technique using distributed tree configuration over multihop communication system. The technique has the capability to identify the intermittent links and can perform data transmission through alternative routes to ensure that fused data is perfectly delivered. The outcome of the study was implemented in Matlab and compared with our prior model and existing LEACH algorithm to find that proposed system outperforms both in terms of energy efficiency and escalated communication performance at the same time in wireless sensor network. Our future study will be focusing on further optimizing the process of data fusion in large-scale wireless sensor network.

# References

1. Khan, S., Pathan, A.K., Alrajeh, N.A.: Wireless Sensor Networks: Current Status and Future Trends, 546 pages. CRC Press, Boca Raton (2016). Computers
2. Aggarwal, C.C.: Managing and Mining Sensor Data. Springer, New York (2013)
3. Mahmoud, M.S., Xia, Y.: Networked Filtering and Fusion in Wireless Sensor Networks. CRC Press, Boca Raton (2014). Computers

4. Kuorilehto, M., Kohvakka, M., Suhonen, J., Hämäläinen, P., Hännikäinen, M., Hamalainen, T.D.: Ultra-Low Energy Wireless Sensor Networks in Practice: Theory, Realization and Deployment. Wiley, New York (2008). Technology & Engineering
5. Jayasri, B.S., Rao, G.R.: Need for energy efficient data fusion in wireless sensor networks. Int. J. Eng. Res. Technol. **3**(1), 1663–1668 (2014)
6. Jayasri, B.S., Rao, G.R.: Reviewing the research paradigm of techniques used in data fusion in WSN. In: IEEE International Conference in Computing and Communications Technologies, pp. 83–88, 26–27 February 2015
7. Jayasri, B.S., Rao, G.R.: EEDF: energy efficient data fusion supportive of virtual multipath propagation in WSN. Int. J. Appl. Eng. Res. **10**(86), 54–59 (2015)
8. Zhu, X., Lu, Y., Han, J., Shi, L.: Transmission reliability evaluation for wireless sensor networks. Hindawi Publishing Corporation (2016)
9. Liu, Y., Zeng, Q.A., Wang, Y.H.: Energy-efficient data fusion technique and applications in wireless sensor networks. J. Sens. **2015**, 2 (2015). Hindawi Publishing Corporation
10. Li, Q., Ma, X., Peng, H., Huang, S.: Data fusion optimization model of elastic wave in wireless sensor networks. J. Comput. Inf. Syst. **11**(3), 815 (2015)
11. Jorio, A., Fkihi, S.E., Elbhiri, B., Aboutajdine, D.: An energy-efficient clustering routing algorithm based on geographic position and residual energy for wireless sensor network. J. Comput. Netw. Commun. **2015**, 11 (2015). Hindawi Publishing Corporation
12. Ma, T., Liu, Y., Fu, J., Jing, Y.: A reliable information fusion algorithm for reputation based wireless sensor networks. Int. J. Future Gener. Commun. Networking **8**(1), 281–298 (2015)
13. Dingcheng, Y., Zhenghai, W., Lin, X., Tianku, Z.: Online bayesian data fusion in environment monitoring sensor networks. Int. J. Distrib. Sens. Netw. **2014**, 10 (2014). Hindawi Publishing Corporation
14. Zhou, J., Liu, L., Guo, J., Sun, L.: Multisensor data fusion for water quality evaluation using Dempster-Shafer evidence theory. Int. J. Distrib. Sens. Netw. **2013**, 6 (2013). Hindawi Publishing Corporation
15. Peng, H., Zhao, H., Li, D., Si, S., Cai, W.: Research on reliability-oriented data fusaggregation algorithm in large-scale probabilistic wireless sensor networks. Int. J. Distrib. Sens. Netw. **2014**, 11 (2014). Hindawi Publishing Corporation
16. Wang, H., Dong, S.: Adaptive fusion design using multiscale unscented Kalman filter approach for multisensor data fusion. Math. Probl. Eng. **2015**, 10 (2015). Hindawi Publishing Corporation
17. Bangash, J.I., Abdullah, A.H., Razzaque, M.A., Khan, A.W.: Reliability aware routing for intra-wireless body sensor networks. Int. J. Distrib. Sens. Netw. **2014**, 10 (2014). Hindawi Publishing Corporation
18. Luo, J., Li, T.: Bathtub-shaped failure rate of sensors for distributed detection and fusion. Math. Probl. Eng. **2014**, 8 (2014). Hindawi Publishing Corporation
19. Tan, R., Xing, G., Liu, B.: Exploiting data fusion to improve the coverage of wireless sensor networks. IEEE/ACM Trans. Networking **20**(2), 450–462 (2012)
20. Yuan, Z., Xue, H., Cao, Y., Chang, X.: Exploiting optimal threshold for decision fusion in wireless sensor networks. Int. J. Distrib. Sens. Netw. **2014**, 7 (2014). Hindawi Publishing Corporation
21. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocols for wireless micro sensor networks. In: IEEE Proceedings of the 33rd Hawaaian International Conference on Systems Science (HICSS 2000), January 2000

# RASK: Request Authentication Using Shared Keys for Secured Data Aggregation in Sensor Network

Jyoti Metan[1](✉) and K.N. Narasimha Murthy[2]

[1] Visvesvaraya Technological University, Belagavi, India
jyotimetan2013@gmail.com
[2] Faculty of Engineering, Christ University, Bangalore, India

**Abstract.** Accomplishing a robust security features to resists lethal attacks is still an open research area in wireless sensor network. The present paper review existing security techniques to find that there is still a trade-off between cryptographic-based security incorporations and communication performance. Moreover, we have identified that majority of the existing system has not emphasized on first line of defense i.e. security the route discovery process that can act as a firewall for all forms of illegitimate nodes existing in the network. The proposed study introduced RASK i.e. Request Authentication using Shared Key, which is a novel concept developed using simple quadratic formulation of generating keys for encrypting the message during data aggregation. The study outcome has been significantly benchmarked with recent studies and existing cryptographic standards to find RASK outperform existing techniques.

**Keywords:** Wireless sensor network · Cryptography · Key management · Energy · Data aggregation · Security

## 1 Introduction

Wireless Sensor Network is not the new arena of research. The topic has been consistently been the point of attraction for the research community till last two decade. A sensor node is basically a tiny electronic device that can sense physical information from its surrounding. A sensor node consists of sensing unit, communication unit, and power supply unit [1, 2]. It is capable to sense data like that of smoke, heat, pressure, motion, proximity, etc. and can forward the data to perform some specific task. Generally, sensors are always deployed in a group either using uniform distribution or using random distribution. It is because a single sensor can sense only small amount of data sustained over an inbuilt battery system [3]. Hence, it cannot sustain for longer. Hence, when a sensor is deployed in group, it shares it task using the routing mechanism and than it can sustain longer period. Normally, sensors are deployed in such adverse area where it is infeasible for human to intervene. Application of wireless sensor network ranges from habitat monitoring, healthcare monitoring system, industrial monitoring system etc. [4]. In last two decades, there has been substantial amount of research towards wireless sensor network that are mainly directed towards routing

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_3

issue [5], energy-issues [6], resource allocation issue [7], security issue [8] etc. Such research contribution has resulted in various standard techniques, algorithm, and protocols, which is still helping the research communities to evolve up with more new ideas. However, all these issues are open end and have not yet met the full proof solution. However, one common observation is that energy is the primary point in 96% of research works done till date, be it any problems in wireless sensor network. Out of all the problems, we will highlight the ongoing security problems and discuss some possible measure to solve it. Cryptography has made a considerable amount of progress in last decade and today breaking any novel cryptographic code is quite expensive which demotivate hackers in the world of wired and wireless networking. But implementing such cryptographic code will be quite challenging for resource-constraint node. The prime reason behind this is normally cryptographic implementation included various forms of iterative steps and usually techniques e.g. Elliptical Curve Cryptography, Diffie Hellman, Galois Field, etc. have complex mathematical structure which cost maximum resource of a small sensor just to read them [9]. Hence, encryption mechanism is quite time for a node that consistently depletes energy in every seconds. Hence, more the communication progress, more the sensor depletes its energy. Another problem is invariable energy allocation. It is said that transmit energy of cluster head is quite high as compared to transmit energy for any other node (member node or candidate node) in a cluster. Hence, there is a mis-balance between energy efficiency and potential security algorithm, because energy efficiency can be achieved using non-cryptographic approach, while security features can be strengthened using complex cryptographic approach. Owing to this tradeoff, the problem of security is still an unsolved problem in wireless sensor network.

Therefore, this paper will present such a technique which is essentially meant for keeping a well balance between security features by using a mathematical quadratic approach and communication performance by carefully developing the cryptographic algorithm. The significant novelty of this work is (i) it uses simple mathematical approach using sensor tags, which has not been focused by prior technique, (ii) usage of novel cryptographic hash function called as keccak, which has never been experimented in sensor network in past, and (iii) the outcome shows better security and communication perform with respect to (i) recent research work in same area and (ii) standard cryptographic technique. Section 2 discusses about related work followed by problem identification in Sect. 3. Proposed system is discussed in Sect. 4, research methodology in Sect. 5, and algorithms in Sect. 6. Result discussion is done in Sect. 7 and conclusion in Sect. 8.

## 2 Related Work

This section discusses about the recent studies being carried out in addressing security problems in wireless sensor network. We choose to review the standard papers being published in the duration of 2010–2016. Most recently, wireless sensor network is mostly spoken with respect to Internet-of-Things (IoT), which blends sensor network with cloud applications. Security problems in such issue has been discussed by Wu et al. [10], who have introduced a framework for identifying unknown threats using software defined network. The outcome of the study was investigated with respect to time consumption, memory consumption, and attack detection rate.

Fayed et al. [11] have utilized complex cryptographic approach e.g. elliptical curve cryptography, Diffie Hellman, and Kerberos protocol to investigate its impact on the security in wireless sensor network. The study implements Diffie-Hellman authentication protocol for the sensor nodes while it implements Kerberos protocol to authenticate the cluster head and sink. The study outcome is found to have better energy efficiency with elliptical curve cryptosystem. Similar usage of elliptical curve cryptosystem has been witnessed in the study of Khan et al. [12], where the authors have used it for the purpose of authenticating the nodes as well as to establish the secret key. The study was implemented using Cooja simulator on smaller scale of sensor network. The study was claimed to be resistive of sniffing attack, Denial-of-Service attack. Cheikhrouhoua et al. [13] have introduced a study towards group management using re-keying mechanism. The presented approach performs multiple clusters that perform re-keying using tree-based technique. The technique was also claimed to be resistive against replay attack in wireless sensor network. The study also uses message authentication code for validating two different nodes along with usage of group keys. Guermazi and Abid et al. [14] have implemented a similar technique for securing established routes over wireless sensor network. The authors have used pair wise key management as well as group key management. The simulation is carried out using Tiny OS over random distribution, where the outcome of the study was evaluated with respect to group key distribution and pair wise key distribution. The time was the prominent performance factor that was essentially observed in result analysis.

Hence, it can be seen that there are various scales of research being carried out in securing the communication system in wireless sensor network. Each of the technique has its own advantages towards addressing specific adversaries and their detrimental effect. However, there are also certain pitfalls of existing security techniques, which is briefed in the next section of problem identification.

## 3 Problem Identification

This section discusses about the significant problems that has be briefed after reviewing the existing system on the previous paragraph.

First of all, it has been seen that majority of the existing study have use security techniques which is mainly encryption mechanism. In this regards, there are various studies which has adopted the potential use of cryptographic techniques e.g. Elliptical curve cryptography, Diffie Hellman, and Kerberos. 40% of the existing approaches are based on these forms of cryptography which has quite a complex mathematical structure. These technique, although claims of better security standards, but could not prove the success factor in leveraging the non-repudiation in security. There is still a trade-off between the potential cryptographic techniques and communicational demands. The second problem that has been identified is less adoption of benchmarking or ineffective benchmarking mechanism in this regards. For an example, the work carried out on security problems in sensor network by Fayed et al. [11] have not adopted their outcomes to be benchmarked, in which case it is quite difficult to rely on the accomplished outcomes. The study conducted by Khan et al. [12] doesn't have a substantial evidence to prove that the outcome is better outcomes till date. The third problem identified is usage of confined networking area. Maximum cases, it was found that smaller scale of network is considered for simulation study. The issue is malicious behavior of intruder reacts in different way in bigger to smaller and from dense to sparse network. The existing technique has always focused on using encryption-based mechanism with more iterative steps. However, quite a less number of the studies have discussed about how such algorithm influences the computational complexity. It was also seen that there are few studies which has analyzed the impact on increasing and dynamic traffic on the performance of security protocols. Although, there are good number of mathematical modelling being discussed in prior techniques, but there was no much discussion about if the model has adhered to any standard radio-energy model or if the outcomes of such mathematical modelling has been testified with standard security techniques to measure its effectiveness.

Hence, the problem statement of the proposed study is - '*It is challenging task to develop a simple and lightweight security that maintains well balance between security and communication performance*'. The next section discusses about the contribution of proposed study.

## 4 Proposed System

The prior section summarized the research problem which in a nutshell says that there is still an unsolved issue towards ensuring secure routing or data aggregation in wireless sensor network. As it can be understood that majority of the techniques in past has not emphasized on initial line of defense i.e. during route discovery. As the sensor node uses public key cryptography hence they are bound to broadcast their key information which is highly susceptible for intrusion. Hence, our proposed technique entails its primary target of identifying the form of request from any sensor without using complex cryptographic use. The prime purpose of the proposed study is to introduce a novel technique of secure data aggregation in wireless sensor network called as Request Authentication using Shared Key or RASK. The architecture of the proposed system is as follows:

Figure 1 highlights the architecture of RASK, where the formulation of sensor tags is done to be incorporated in using quadratic approach. The system checks for shared key. It aborts the communication in case of illegal request and compute the new shared key in case of legitimate request. The shared key is computed using novel *keccak* algorithm [15] and is used for encrypting both routing message and data packet during data aggregation in wireless sensor network. The shared key can be given a specific expiry in real-time application in order to avoid misuse of it. Finally shared keys are instantly refreshed and updated to all the communicating nodes. The next section discusses about research methodology adopted in RASK.

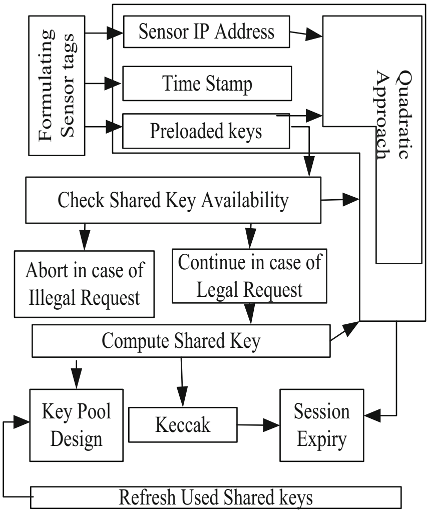


**Fig. 1.** Architecture of proposed RASK

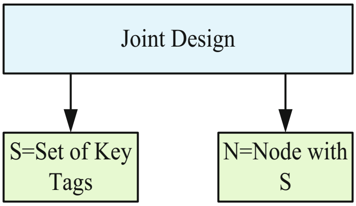


**Fig. 2.** Pictorial description of joint design

## 5 Research Methodology

The proposed system considered analytical research methodology which considers different mobility aspects of sensors i.e. both nodes with uniform and mobile position. The proposed system uses a light weight cryptographic approach using mathematical modelling in order to secure the process of data aggregation. The methodologies adopted by the proposed system are as follows:

### 5.1 Intruder Model

It was studied from the existing literatures that majority of the prior studies considers a specific form of attacks and model the solution accordingly. In this work, we want that proposed technique should be equally resistive to majority of the lethal threats in wireless sensor network. Therefore, we consider that a very safe environment before the simulation. However, after the simulation starts, we assume that an intruder has made its way inside the simulation using malicious means. We also assume that an intruder is potentially stronger in accessing private information of the nodes and hence compromise the nodes. We are least bothered about the specifics of the attacks as in case anyone node is victimized by an intruder, it can be any form of attacks. Our idea of proposed system is - how to stop a sensor be getting compromised in this first level of authentication itself. We have observed that when the first level of authentication among the nodes in poorly design, the possibility of intrusion increases exponentially.

### 5.2 Design Principle

The study uses a mathematical modelling that presents a design and development of finite sets of properties whose agreement must meet the concept of well balance secure authentication using novel key management techniques. We consider that all the sensors are equipped with pre-loaded keys which may be slightly different in number in different sensors. The study consider joint design of (S, N), where *S* is a set of key tags while *N* are those sensors that has possession of key tags from *S*. The joint design is shown in Fig. 2.

The proposed study considers a mechanism where the key are already loaded into the node. The design of the proposed model is completely based on the novel mathematical approach that initiates by considering two-dimensional matrix of single-tier of architecture designs $(\mathrm{K_P}, \mathrm{S_N})$, where $\mathrm{K_P}$ is the set of key tags, while $\mathrm{S_N}$ is the mote that consist of specific key tags from $\mathrm{K_P}$. The tags will consists of mainly three types of data i.e. node IP Address, time stamp, and preloaded key. The process of key-allocation and giving a mathematical shape is not easy as it is quite possible that every sensor nodes have different number of keys. Hence, to ease off the computation, the proposed system considers equal number of key distribution in every sensor. It applies intersection logic in mathematics in order to extract the shared key information as shown in Fig. 2. Figure 3 highlights the presence of 3 sample clusters considering 5 sensors (1 cluster head and 4 member nodes). The figure doesn't show the nodes but it highlights the presence of keys among the sensors. Hence, there are three set of keys i.e. $\mathrm{K_P}$, $\mathrm{K_Q}$, and $\mathrm{K_R}$ such that $\mathrm{K_P} = \{\mathrm{K_1}, \mathrm{K_2}, \mathrm{K_3}, \mathrm{K_4}\}$, $\mathrm{K_Q} = \{\mathrm{K_5}, \mathrm{K_6}, \mathrm{K_7}, \mathrm{K_8}\}$, $\mathrm{K_P} = \{\mathrm{K_9}, \mathrm{K_{10}}, \mathrm{K_{11}}, \mathrm{K_{12}}\}$. It is studied from existing theories that a sensor node is required to broadcast its key information to its destination node in order to assist in authentication. In large scale multihop network, usually, various intermediate clusters assist in establishing the route from source to destination. Therefore, it is essential that such routing information should be protected. Hence, we formulate a condition that:

*Condition:* If any two clusters want to communicate with each other, they must have shared key. The proposed system could use simple mathematical intersection method to find the shared key. Referring to Fig. 2, if there is a shared key $\mathrm{K_4}$ and $\mathrm{K_8}$ than only 1$^{\mathrm{st}}$

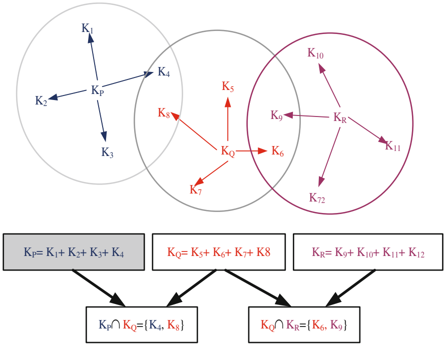


**Fig. 3.** Shared key processing in proposed system

cluster can communicate with 2nd one. Similarly, if there is a shared key $K_6$ and $K_9$ than only 2nd cluster can communicate with third cluster. In absence of shared key, the two clusters will not be able to communicate with each other.

- **Pros:** The positive aspect of this formulation is that pre-loaded keys can be used for embedding certain symmetric information, which can ensure faster authentication. As all the sensors will have common manufacturing units, hence, response for authentication will be quite faster.
- **Cons:** In case of node capture attack, such pre-loaded keys are the first one to be compromised by an intruder.

However, we just deploy shared key exploration process just for preliminary authentication, and the later steps perform implementation of an algorithm to generate shared key using simple cryptography approach. The next section discusses about an algorithm design implementation.

## 6 Algorithm Design

An algorithm for the proposed system is developed and implemented over Matlab. The implementation is carried out considering 1000 sensor nodes on 1500 × 1200 m2 simulation area. The preliminary phase of the simulation is carried out by considering assigning set of tags randomly to all the sensor nodes present in the network. The system also provides a key along with these tags. The communication only initiates of two nodes from different groups are found with shared keys. The proposed system offers security on every level of proposed architecture. The term level means routing stages of data aggregation i.e. 1st level corresponds to securing routes between member nodes and cluster head, 2nd level corresponds to that of cluster head to another cluster head, while third level corresponds to securing link between cluster head to sink. As the cluster head bears enough physical information of its member node along with other information e.g. battery life, buffer etc., hence chances of member node being rogue is not considered. Hence, the study is more dominant for 2nd and 3rd level of data aggregation. For this purpose the system initially considers assuming intruder module that has recently intruded the network either using inside or outside attacking strategy. Existing techniques usually is experimented using 128 or 512 bit of encryption keys. However, the proposed system offers highest level of flexibility by offering differential the key size. The algorithm description is as follows:

As majority of the sensors are bound to broadcast the symmetric key information over wireless channel, it is quite evident that preloaded keys are directly prone to get compromised. Hence, the algorithm performs extraction of new shared key and not the old shared key. The entire algorithm works on the concept of Sensor tag, which is used to represents some unique identical character of a sensor. The proposed system increases the robustness by broadcasting three basic tags i.e. node IP address, time slots, and preloaded keys. The prime task is to ensure how to protect the routing information. The proposed RASK considers a source node *i* communicating with destination node *j*. It is also assumed that both source node *i* and destination node *j* posses three significant tag information i.e. $\alpha$ (Node IP Address), $\beta$ (Time stamp), and $\delta$ (Key). The proposed system considers node capture attack and hence it assumes that its preloaded keys are compromised. Therefore, if at any round of interaction, two nodes are found to posses similar preloaded key (as shown in line-3) than it is a case of spoofing owing to certain malicious intervention from an intruder. Hence, no shared key will be ever computed in such case (line-4). However, even in such case also, it is only possible for an intruder to only posses either $\delta_1$ or $\delta_2$ and never the both. Hence, it truly maintains backward secrecy. The proposed system uses a discriminant of quadratic equation for reshaping the usage of sensor tags in the form of quadratic roots (line-6). Both the roots are basically part of new shared key computation, which can never be null (as highlighted by the condition exhibited in line-6) (Fig. 4).

```
Algorithm: Request Authentication using Shared Key (RASK)
Input: $k$ (individual key), $tag$ (Sensor tags), $\lambda$ (discriminant),
$S$ (Set of key tags), $d$ (data), $\alpha$ / $\beta$ / $\delta$ (Node IP Address / Time stamp / Key)
Output: $K_s$ (shared key)
START
1. Init $k$
2. Init $tags$ of $i^{th}$ and $j^{th}$ node
        $tag_i$=[$\alpha_1$, $\beta_1$, $\delta_1$]          $tag_j$=[ $\alpha_2$, $\beta_2$, $\delta_2$];
3. IF ($\delta_1$= = $\delta_2$)
4.    Return $k_{share}$=0;
5. ELSE
6.    IF (sqrt (abs($\lambda$)) ~ = 0)
      &
7.    IF ($r_1$ && $r_2$ < $k$)
8.        $S_{ij}$= $\arg_{max}$ [concat ($r_1$, $r_2$)]
9.        $d$ = concat ($S_{ij}$, $tag_i$, $tag_j$);
10.       $K_s$=Encrypt (Data, '$keccak$');
11.   End
End
```

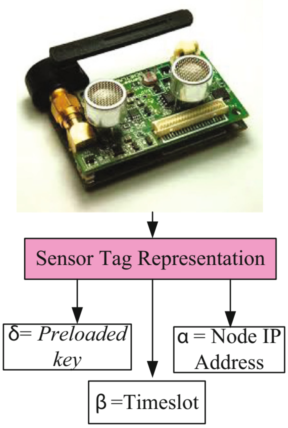


**Fig. 4.** Representation of sensor tag

The next part of the equation (line-7) is meant to ensure the lowered size of the newly generated shared key from the quadratic roots. Hence, it gives better scalability to the stakeholder to maintain any standard value of *k* and can perform better optimization by lowering the value of *k* in order to get shared key of much reduced size. The next step is to perform concatenation of newly generated quadratic roots which will generate a unique concatenated numbers. The set of key tags selects maximum value of it (line-8), which is further used for performing concatenation operation to generated concatenated data *d* (line-9). Finally, the proposed RASK uses a new cryptographic algorithm called as keccak, which is recently launched by NIST. The newly generated shared key $K_s$ is then used for performing data aggregation by cluster head. Hence, encryption is performed only in one step, but usage of simple mathematical operation ensures that proposed algorithm of RASK maintains both forward and backward secrecy in wireless sensor network. The next section discusses about the result being accomplished from the implementation of proposed algorithm.

## 7 Comparative Performance

As the proposed system targets to offer resiliency against lethal threats during data aggregation in wireless sensor network, hence, the proposed RASK introduces an lightweight algorithm with very less usage of complex cryptography in it. The simulation parameters considered for the result analysis are exhibited in Table 1. Hence, it is always expected that proposed system should offer better stability and performance during data aggregation. Hence, the outcome of the proposed system is analyzed with respect to end-to-end delay, residual energy, and processing time. The outcome of the proposed study was compared with the recent work carried out by Roy [16] and Tang [17].

**Table 1.** Simulation parameters

| Parameters | Nodes | MAC protocol | Initialized energy | Traffic load | Simulation round | Message size | Data packet size |
|---|---|---|---|---|---|---|---|
| Value | 500–100 | IEEE 802.15 | 0.5 J | 500–1000 bytes | 1000 | 250 bytes | 1000 bytes |

The reason behind selection of these two recent work are as follows [16] have presented an energy-efficient routing considering the case study of flooding attack, wormhole attack, Sybil attack, and sinkhole attack. The technique used by author is one-way hash chain for the purpose of authenticating base station and message authentication codes. The benchmarked study outcome shows better energy efficiency with less number of packet drop incident. Similarly, work done by [17] have introduced a secure routing mechanism using cost factor and outcome was found to posses better energy efficiency, lower delay, and higher delivery ratio. Unfortunately the outcome was not compared with standard studies to prove its efficiency. This section briefs about the outcome accomplished from the study and some essential findings related to it.

### 7.1 Analysis of Residual Energy

The proposed system adopts the first order radio-energy model [18], which means that communication performance is directly proportional to the energy usage. Hence, we choose to select residual energy as the first performance factor as it can directly represent how much resource (i.e. energy) is being utilized by proposed RASK in order to provide ultimate security. The outcome shown in Fig. 5 shows that RASK posse's better energy conservation factor compared to [17] and [16]. The prime reasons behind this are as follows: [17] have presented security over multihop communication with more focus on energy efficiency. The authors proved his model using energy information as distribution mechanism and performed statistical analysis on it. Although, the mechanism is highly fruitful in energy conservation but usage of random walking principle along with deterministic approach leads to maximum loss of residual energy if the similar model is implemented over large scale sensor network with MAC IEEE 802.15. Similar model when tested with increasing number of queries from new traffic, the trend of [17] was found to be heavily declining. The work carried out by [16] has mainly used one-way hash chain that doesn't require much resource to be utilized to perform validation as compared to inclusion of many statistical calculations in [17]. However, author has applied concatenation operation on multiple values combining both static memory (IP address) and dynamic memory (trust value, MAC). This consumes maximum resources to pull the data from different clusters. Moreover the study uses elliptical curve cryptography which reduces the key size but increases the requirement of maximum resource to compute private keys resulting in energy drainage. Whereas the proposed RASK follows a simple *keccak* algorithm, which works on principle of sponge constructing that results in very low memory consumption and faster response time. Moreover, usage of quadratic approaches and concepts of sensor tags ensure validating the data using lesser dependency on energy consumption. Hence, proposed RASK does dissipate energy very less than existing approaches (Fig. 8).

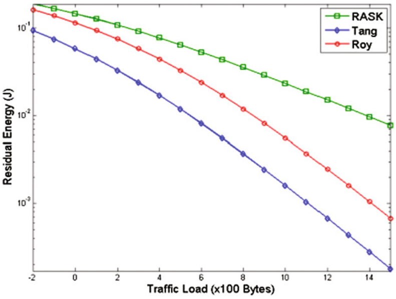


**Fig. 5.** Analysis of residual energy

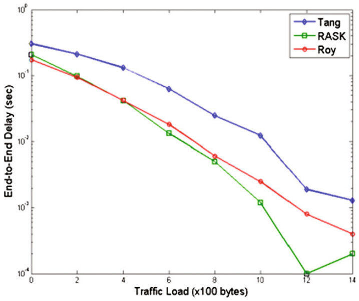


**Fig. 6.** Analysis of end-to-end delay

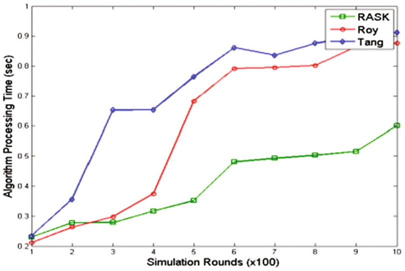


**Fig. 7.** Analysis of algorithm processing time

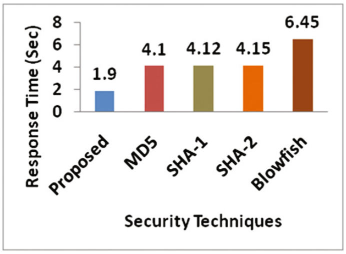


**Fig. 8.** Response time of different techniques

### 7.2 Analysis of End-to-End Delay

Although the delay factor essentially represents time attributed in data aggregation process, but it is one of the suitable parameter to judge if the computation of proposed RASK algorithm results in any delay in communication. For this purpose, we exponentially increase the incoming packet with increasing traffic load to observe its impact on end-to-end delay. The approach of [17] has maximum usage of statistical calculations in order to have energy efficiency over grid topology. However, when the topology is changed to random than trend for [17] is found with increased delay as compared to other two techniques. The performance trend for [16] is really good as it shows less variance and lesser delay as compared to [17]. However, [16] uses MAC as medium of validation which cannot be resistive against purposely modification of message contents by intruder. It is worst in case of node capture attack. Hence, although, the outcome trend of [16] shows better delay but it cannot be considered to have maintained equilibrium of security factor. Moreover usage of multiple steps of concatenation operation may give better forward secrecy but it cannot ensure non-repudiation if the network topology and assize is change. But proposed system improvises this using quadratic approach to generate shared key, which is volatile in nature. This will mean that once the key is used it will be also dispose, which mean no memory consumption resulting in faster response in query processing in terms of acceptance or rejection to queried node.

### 7.3 Analysis of Algorithm Processing Time

Majority of the standard motes can process across 48 kbytes of physical memory. As the proposed algorithm has to run over the physical memory of a sensor node, it is important to understand its response time. Therefore, algorithm processing time will significantly exhibit how much time is consumed by proposed RASK to complete up its mission objective. The outcome shown in Fig. 6 shows that proposed RASK has lower algorithm processing time as compared to [17] and [16]. Usage of maximum computational steps exists in the work of [17] for following purpose i.e. getting shortest route, energy efficient route, route with maximum security factor. Although, it is good in energy efficient viewpoint, but this also consumes maximum time, which is absolutely detrimental for any type of routing-based intrusions. On the other hand, [16] have used blowfish algorithm along with extended mathematical steps of concatenation which is another cause of time consumption. The proposed system only uses two steps of concatenation over the quadratic roots to generate a volatile shared key that is used in encrypting aggregated data by the cluster head. Hence, trend for algorithm processing time is extremely lower as compared to [17] and [16] approach over increasing simulation rounds.

### 7.4 Security Analysis

It has to be understood that RASK is designed to provide front line of defense which happens during the first step of data aggregation i.e. validating the query or beacon.

Hence, in this direction, it is essential that proposed RASK should have better resiliency against any form of threats, which can be only ensured if the algorithm is computational intensive for a longer run. The storage complexity of proposed system can be signified as O ($\sqrt{N}$) and communication complexity as O (log $N$)), where $N$ can be represented as highest amount of sensors. With this lowered computational complexity, the mechanism of proposed RASK can be also used in adhoc networks as well as any mobile networks too. From Fig. 7, it can be seen that proposed system offers faster response time in comparison to standard cryptographic algorithm. The interesting point is proposed system has only one step of cryptographic usage with highly reduced response time. Hence, it can be said that proposed system offers a better balance between security, computational complexity, and all forms of communications performance in large scale wireless sensor network.

## 8 Conclusion

A closer look into the problems of wireless sensor network will show that almost all the problems are interlinked with each other. Similarly, this paper shows the similar trend with energy and security. It is discussed in the paper that achieving potential security standards with best communication performance is still a gap to be bridged. Hence, we designed RASK using lightweight cryptographic approach. We develop a concept of sensor tags, which will be used for the purpose of authentication by checking if both the nodes have similar preloaded keys, which can be only the case of spoofing or any attacks pertaining to identity. Hence, we use quadratic approach to develop solution which can generate shared keys for further securing the transmission. We use the shared key to encrypt the data using most recently launched keccak that further secures the data aggregation in wireless sensor network. The outcome of the study is compared with the most recently done work as well as some frequently used cryptographic standard to find RASK is better security protocol in every respect. Another novelty in RASK is probably it is one of the first security techniques that ensure lower size of the key. However, our future work will be to investigate further optimization of key size and further betterment in the encryption techniques.

## References

1. Hesselbach, J., Herrmann, C.: Glocalized Solutions for Sustainability in Manufacturing. Springer, Heidelberg (2011)
2. Athawale, R.P., Rana, J.G.: Wireless sensor network based environmental temperature monitoring system. In: International Conference on Information Engineering, Management, and Security (2015)
3. Mostafa, M., Azim, A., Jiang, X.: Wireless Sensor Multimedia Networks: Architectures, Protocols, and Applications, 279 p. CRC Press, Boca Raton (2015). Computers
4. Cagn Gungor, V., Hancke, G.P.: Industrial Wireless Sensor Networks: Applications, Protocols, and Standards. CRC Press, Boca Raton (2013)

5. Kumari, J., Prachi: A comprehensive survey of routing protocols in wireless sensor networks. In: 2nd IEEE-International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 325–330 (2015)
6. El-Aaasser, M., Ashour, M.: Energy aware classification for wireless sensor networks routing. In: 15th IEEE-International Conference on Advanced Communication Technology (ICACT), PyeongChang, pp. 66–71 (2013)
7. Chitnis, M., Pagano, P., Lipari, G., Liang, Y.: A survey on bandwidth resource allocation and scheduling in wireless sensor networks. In: IEEE-International Conference on Network-Based Information Systems, Indianapolis, IN, pp. 121–128 (2009)
8. Chelli, K.: Security issues in wireless sensor networks: attacks and countermeasures. In: Proceedings of the World Congress on Engineering (2015)
9. Hankerson, D., Menezes, A.J., Vanstone, S.: Guide to Elliptic Curve Cryptography. Springer, New York (2006)
10. Wu, J., Ota, K., Dong, M., Li, C.: A hierarchical security framework for defending against sophisticated attacks on wireless sensor networks in smart cities. IEEE Access, **4**, 2169–3536 (2016)
11. Fayed, N.S., Daydamoni, E.M., Atwan, A.: Efficient combined security system for wireless sensor network. Egypt. Inf. J. **13**(3), 185–190 (2012)
12. Khan, S.U., Pastrone, C., Lavagno, L., Spirito, M.A.: An authentication and Key establishment scheme for the IP-based wireless sensor networks. Procedia Comput. Sci. **10**, 1039–1045 (2012)
13. Cheikhrouhou, O., Koubâa, A., Dini, G., Alzaid, H., Abid, M.: LNT: a logical neighbor tree for secure group management in wireless sensor networks. Procedia Comput. Sci. **31**(5), 198–207 (2011)
14. Guermazi, A., Abid, M.: An efficient key distribution scheme to secure data-centric routing protocols in hierarchical wireless sensor networks. Procedia Comput. Sci. **5**, 208–215 (2011)
15. The Keccak Sponge Function Family. http://keccak.noekeon.org/. Retrieved 8 Feb 2017
16. Roy, S., Das, A.K.: Secure hierarchical routing protocol (SHRP) for wireless sensor network. In: Mauri, J.L., Thampi, S.M., Rawat, D.B., Jin, D. (eds.) SSCC 2014. CCIS, vol. 467, pp. 20–29. Springer, Heidelberg (2014). doi:10.1007/978-3-662-44966-0_3
17. Tang, D., Li, T., Ren, J., Wu, J.: Cost-aware secure routing (CASER) protocol design for wireless sensor networks. IEEE Trans. Parallel Distrib. Syst. **26**, 960–973 (2013)
18. Zheng, L.-Z., Gao, L., Yu, T.-G.: An energy-balanced clustering algorithm for wireless sensor networks based on distance and distribution. In: Qi, E. (ed.) Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation, pp. 229–240. Atlantis Press, Paris (2016). doi:10.2991/978-94-6239-145-1_23

# Warehouse Stock Prediction Based on Fuzzy-Expert System

Radim Farana[1](✉), Ivo Formánek[2], Cyril Klimeš[1], and Bogdan Walek[3]

[1] Department of Informatics, Mendel University in Brno,
Zemědělská 1, 613 00 Brno, Czech Republic
{radim.farana, cyril.klimes}@mendelu.cz

[2] Department of Entrepreneurship and Management,
University of Entrepreneurship and Law,
Michálkovická 1810/181, 710 00 Ostrava, Czech Republic
ivo.formanek@vspp.cz

[3] Institute for Research and Applications of Fuzzy Modeling,
University of Ostrava, 30. Dubna 22, 701 03 Ostrava, Czech Republic
bogdan.walek@osu.cz

**Abstract.** Actually, a lot of companies have tried to optimize their systems of warehouse stock management to minimize the production costs. The main goal is evident – not to spend too much money for stock. To predict the behaviour of the system, there are usually used the methods of time series analysis. They are able to determine the main trend very well, seasonal influences, etc. But they do not take into account the internal and external influences acting on the system. For their destination there is appropriate to use the expert knowledge from companies that are often vague. As an appropriate tool, therefore appears to be a fuzzy expert system. Its use, however, causes problems if the system exhibits a significant trend. It turns out that the system for determining the main trend is appropriate to use methods known from the time series analysis and then followed by taking advantage of the expert system. This paper presents a fuzzy expert system that combines expert knowledge with the analysis of the trend of the system. The presented expert system was also verified in a practical application.

**Keywords:** Warehouse stock · Time series · Expert system · Fuzzy logic · Analysis · Optimization · Prediction

## 1 Introduction

Actually, a lot of companies have tried to optimize their system of warehouse stock management to minimize their production costs. The optimization means mainly optimization of processes like resources adjustment, resources planning, purchasing, deliveries, sales etc. The main goal is clear – not to spend too much money for stock. There are various information systems more or less successfully anticipating and predicting the quantity of resources that should be ordered.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_4

There are generally used different approaches to the sales prediction and thereby the production planning [1, 2]. These can be equalized on statistical methods, especially the analysis of time series, but in practice we often come across with very simple approaches that are very robust at the same time, such as method of moving average. Our approach is based on the use of fuzzy logic expert systems [3, 4]. Experts systems, in particular using fuzzy logic are in this area used by a number of authors for different applications [5–7]. The applications of artificial neural networks [8–11] or tools of soft computing are also very interesting. As advantage of Rule-Based Expert Systems is a particular opportunity to use the knowledge of experts and their simple expressions by rules. Fuzzy logic then helps us especially with easy expression of dependences among the values which is poorly expressed using crisp values. The problem of practical deployment of the expert system is a situation where the course of a monitored variable is affected by a global trend [12]. It shows that in such cases it is advisable to use the methods known from the time series analysis [13] to identify this trend, normalize the data by subtracting and expert system used to determine the additional values and the description of the dynamic of system behaviour) [14]. An example of such an approach is demonstrated in this paper.

## 2 The Major Trends Determination

As an example of the use of the expert system for prediction of sales we use the specific example of the sale of the two specific products of a mechanical engineering company. We have sales data for each of the weeks in 2014 and 2015 (Fig. 1), which are split into two parts. We have used the data 2014 and the first 42 weeks in 2015 to determine the knowledge. The remaining data, then we have used to verify the behaviour of the expert system compared to the prediction based on moving average method.

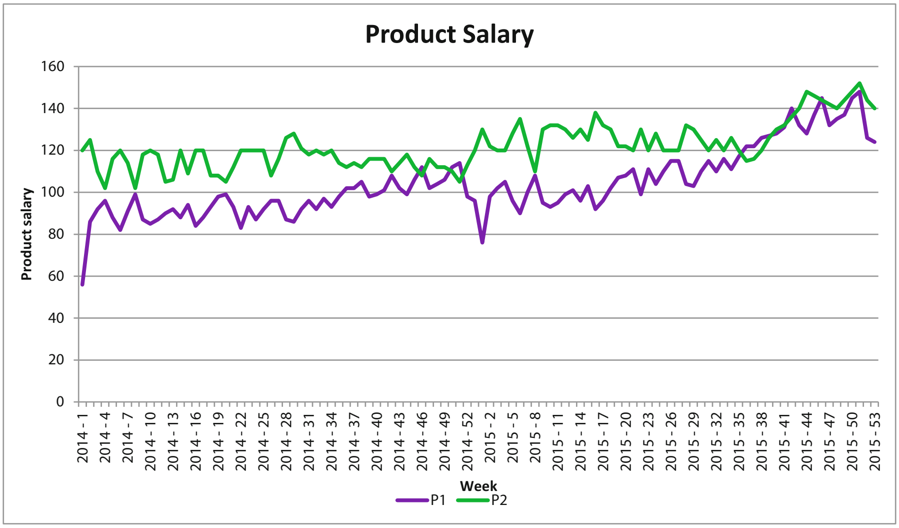


**Fig. 1.** The progress of the sales of the two products of the engineering company

In the first step, we determine the main trends of the development of the sales of the two products P1 and P2. We will use the standard optimization method of least squares. For the calculation we use the discrete time *kT* for the axis of the independent variable where *T* – is a period of one week, *k* = 0, 1, … So we get the major trends:

$$\begin{array}{c} \mathrm{P1}: y_1(t) = 0.435(kT) + 81.1 \\ \mathrm{P2}: y_2(t) = 0.246(kT) + 109 \end{array} \tag{1}$$

After deduction of major trends we collect the data for the expert system, see Fig. 2.

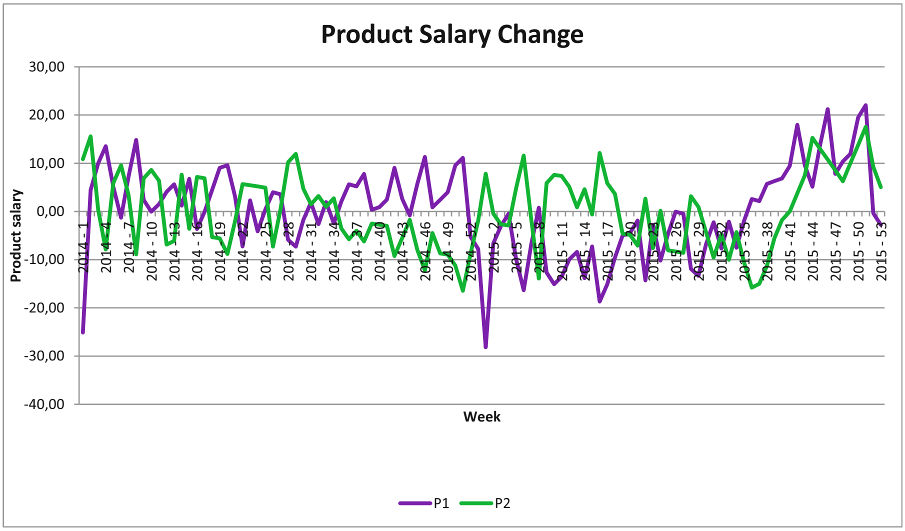


**Fig. 2.** The standard data source for expert system

## 3 Creation of an Expert Model

The first task is to determine the parameters of expert system. Because no information available about the behaviour of the market or competition, we are going to focus only on the information available within the company. From previous works, we have already known that the more parameters affecting the sales we can describe, the more accurate the prediction is, see for example [14]. Thus, as the parameters we set the sales of individual products in the previous two weeks and the status of the negotiation of contracts in the previous two weeks, which the most affect the current sale. For the realization of the expert system, we will use the Linguistic Fuzzy Logic Controller (LFLC) [3], which is very convenient for practical applications.

LFLC is the result of application of the formal theory of the fuzzy logic in broader sense (FLb). The fundamental concepts of FLb are evaluative linguistic expressions and linguistic description. Evaluative (linguistic) expressions are natural language expressions such as high, medium, deep, about thirty-one, roughly one thousand, very long, more or less deep, not very tall, roughly cold or medium warm, roughly strong, roughly medium important, and many others. They form a small, but very important, constituent of natural language since we use them in common sense speech to be able to evaluate phenomena around. Evaluative expressions have an important role in our life because they help us determine our decisions; help us in learning and understanding, and in many other activities. Simple evaluative linguistic expressions (possibly with signs) have a general form:

$$<\text{linguistic modifier}> <\text{TE} - \text{adjective}> \tag{2}$$

where <TE-adjective> is one of the adjectives (also called gradable) “small – sm, medium – me, big – bi” or “zero – ze”, the <linguistic modifier> is an intensifying adverb such as “extremely – ex, significantly – si, very – ve, rather – ra, more or less – ml, roughly – ro, quite roughly – qr, very roughly – vr”).

A very important feature is the ability of setting the context of respective variables for applying the compiled knowledge base for a different range of values, see Fig. 3.

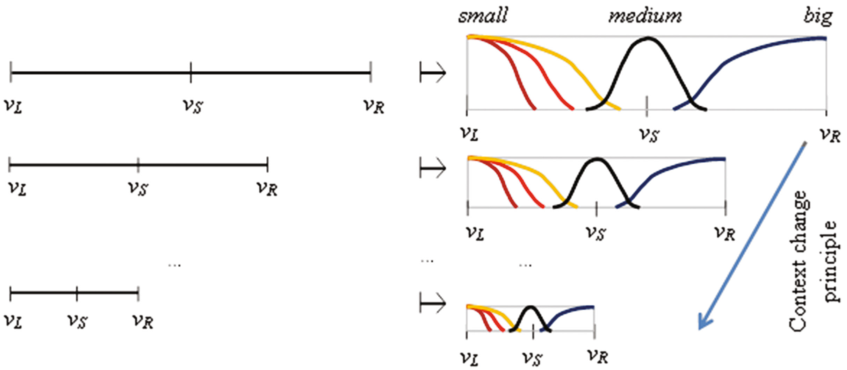


**Fig. 3.** A general scheme of intension of evaluative expressions (extremely small, very small, small, medium, big) as a function assigning to each context a specific fuzzy set [5] and the automatic context change principle

This set of linguistic expressions has been drawn up on the basis of the experience of the experts, but it does not always suit the particular situation. Figure 4 shows the frequency of each value of sales for the product P1. We can see that most of the values are concentrated in the middle of the interval, which covers little linguistic expressions, so when compiling a system of rules for the expert system, there have often appeared the same value (ze). LFLC tool offers the possibility of user-set assembly of evaluative linguistic expressions that will better respond to the current situation.

Now we can build a base of knowledge about the behaviour of the system (standardized system) on the basis of its prior development (marked P1-2, and P1-1, P2-2 and P2-1). For technical reasons, the range of values used is shifted, so that we work only with positive numbers. Activity in the preparation of future contracts will be assessed with a 5 degree scale with the values 1–5 and evaluated in the two previous weeks (marked A-2 and A-1).

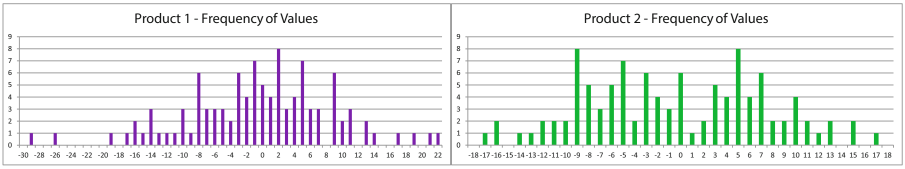
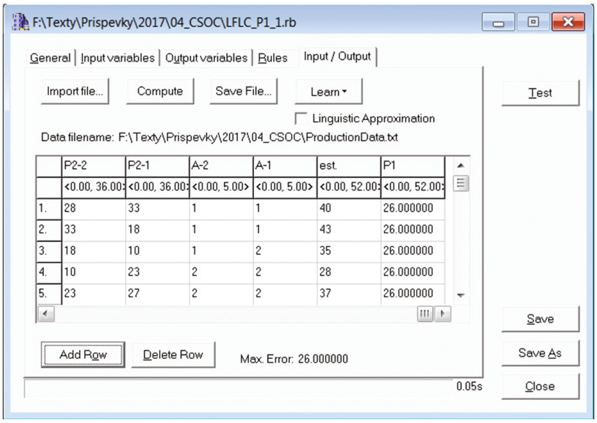


**Fig. 4.** The frequency of (normalized) sales of the products P1 and P2

At the same time the system is ready for a future change of contexts based on the found trend (1). If we do not have the expert knowledge available, we can use the tool for automation of learning that is a part of LFLC, see Fig. 5.



□ Linguistic Approximation

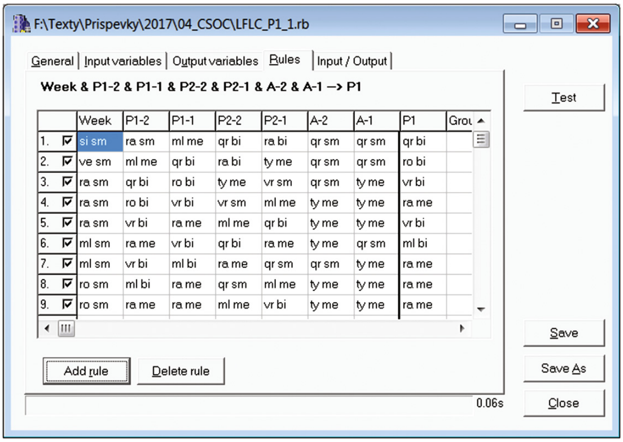
Data filename: F:\Texty\Prispevky\2017\04_CSOC\ProductionData.txt

| | P2-2 | P2-1 | A-2 | A-1 | est. | P1 |
|---|---|---|---|---|---|---|
| | <0.00, 36.00: | <0.00, 36.00: | <0.00, 5.00> | <0.00, 5.00> | <0.00, 52.00: | <0.00, 52.00: |
| 1. | 28 | 33 | 1 | 1 | 40 | 26.000000 |
| 2. | 33 | 18 | 1 | 1 | 43 | 26.000000 |
| 3. | 18 | 10 | 1 | 2 | 35 | 26.000000 |
| 4. | 10 | 23 | 2 | 2 | 28 | 26.000000 |
| 5. | 23 | 27 | 2 | 2 | 37 | 26.000000 |

Max. Error: 26.000000

0.05s

**Fig. 5.** Data uploaded into the automated learning system in LFLC



Week & P1-2 & P1-1 & P2-2 & P2-1 & A-2 & A-1 –> P1

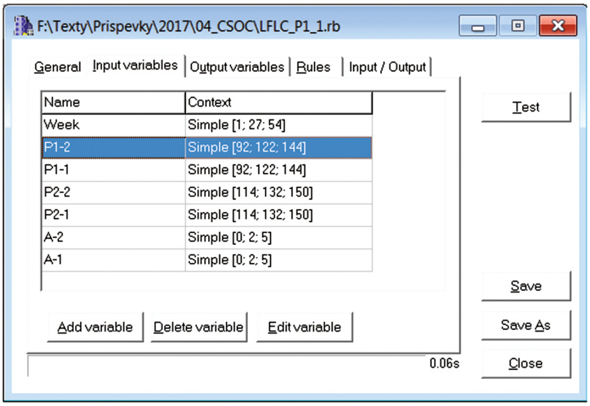
| | Week | P1-2 | P1-1 | P2-2 | P2-1 | A-2 | A-1 | P1 | Grou |
|---|---|---|---|---|---|---|---|---|---|
| 1. ☑ | si sm | ra sm | ml me | qr bi | ra bi | qr sm | qr sm | qr bi | |
| 2. ☑ | ve sm | ml me | qr bi | ra bi | ty me | qr sm | qr sm | ro bi | |
| 3. ☑ | ra sm | qr bi | ro bi | ty me | vr sm | qr sm | ty me | vr bi | |
| 4. ☑ | ra sm | ro bi | vr bi | vr sm | ml me | ty me | ty me | ra me | |
| 5. ☑ | ra sm | vr bi | ra me | ml me | qr bi | ty me | ty me | vr bi | |
| 6. ☑ | ml sm | ra me | vr bi | qr bi | ra me | ty me | qr sm | ml bi | |
| 7. ☑ | ml sm | vr bi | ml bi | ra me | qr sm | qr sm | ty me | ra me | |
| 8. ☑ | ro sm | ml bi | ra me | qr sm | ml me | ty me | ty me | ra me | |
| 9. ☑ | ro sm | ra me | ra me | ml me | vr bi | ty me | ty me | ra me | |

0.06s

**Fig. 6.** Example of rules for product P1 prediction

Figure 6 presents an example of established rules for a basic range of input and output variables. Thanks to the use of application of linguistic expressions, you can easily change the input and output context, and use this expert system throughout the time interval.

Figure 7 shows the input contexts setting for the week of 43 in 2015 (the validation data set) designed according to the trend function (1). Figure 8 shows the follow-up assessment for the output value of product P1. Note that after setting the contexts, the

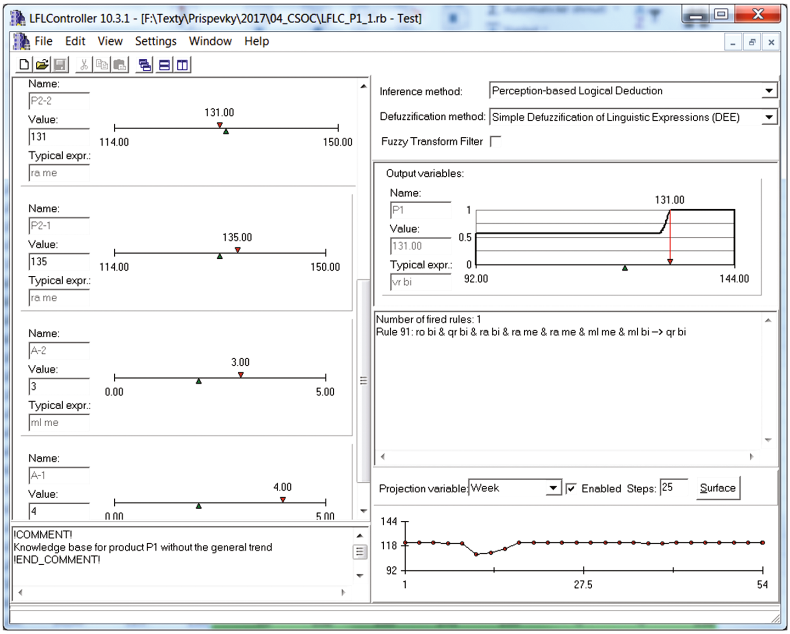


F:\Texty\Prispevky\2017\04_CSOC\LFLC_P1_1.rb

| Name | Context |
|---|---|
| Week | Simple [1; 27; 54] |
| P1-2 | Simple [92; 122; 144] |
| P1-1 | Simple [92; 122; 144] |
| P2-2 | Simple [114; 132; 150] |
| P2-1 | Simple [114; 132; 150] |
| A-2 | Simple [0; 2; 5] |
| A-1 | Simple [0; 2; 5] |

0.06s

**Fig. 7.** Input context for the week 43, year 2015



LFLController 10.3.1 - [F:\Texty\Prispevky\2017\04_CSOC\LFLC_P1_1.rb - Test]

Name: P2-2
Value: 131
Typical expr.: ra me

Name: P2-1
Value: 135
Typical expr.: ra me

Name: A-2
Value: 3
Typical expr.: ml me

Name: A-1
Value: 4

!COMMENT!
Knowledge base for product P1 without the general trend
!END_COMMENT!

Inference method: Perception-based Logical Deduction
Defuzzification method: Simple Defuzzification of Linguistic Expressions (DEE)
Fuzzy Transform Filter □

Output variables:
Name: P1
Value: 131.00
Typical expr.: vr bi

Number of fired rules: 1
Rule 91: ro bi & qr bi & ra bi & ra me & ra me & ml me & ml bi –> qr bi

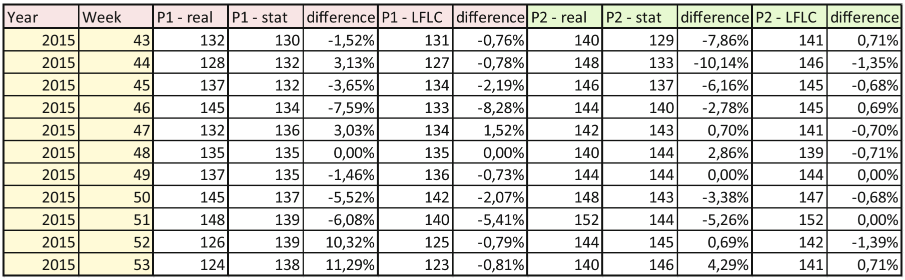
Projection variable: Week ☑ Enabled Steps: 25

**Fig. 8.** Testing the fuzzy set of rules in LFLC environment for week 43, year 2015

expert system works directly with the actual values. This is important for practical use. The user is not burden with the recalculating of values; this will ensure the system automatically.

## 4 Verification of an Expert System Prediction

As it has been said before, the part of the available data was used to create the expert system. The other part of the data was used to verify its activities. Figure 9 presents the results provided by the expert system compared to the moving average. We can see that the expert system achieves better results overall. That is particular about the knowledge about the impact of the various parameters on the behaviour of the entire system, which has been stored in the knowledge base.



| Year | Week | P1 - real | P1 - stat | difference | P1 - LFLC | difference | P2 - real | P2 - stat | difference | P2 - LFLC | difference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2015 | 43 | 132 | 130 | -1,52% | 131 | -0,76% | 140 | 129 | -7,86% | 141 | 0,71% |
| 2015 | 44 | 128 | 132 | 3,13% | 127 | -0,78% | 148 | 133 | -10,14% | 146 | -1,35% |
| 2015 | 45 | 137 | 132 | -3,65% | 134 | -2,19% | 146 | 137 | -6,16% | 145 | -0,68% |
| 2015 | 46 | 145 | 134 | -7,59% | 133 | -8,28% | 144 | 140 | -2,78% | 145 | 0,69% |
| 2015 | 47 | 132 | 136 | 3,03% | 134 | 1,52% | 142 | 143 | 0,70% | 141 | -0,70% |
| 2015 | 48 | 135 | 135 | 0,00% | 135 | 0,00% | 140 | 144 | 2,86% | 139 | -0,71% |
| 2015 | 49 | 137 | 135 | -1,46% | 136 | -0,73% | 144 | 144 | 0,00% | 144 | 0,00% |
| 2015 | 50 | 145 | 137 | -5,52% | 142 | -2,07% | 148 | 143 | -3,38% | 147 | -0,68% |
| 2015 | 51 | 148 | 139 | -6,08% | 140 | -5,41% | 152 | 144 | -5,26% | 152 | 0,00% |
| 2015 | 52 | 126 | 139 | 10,32% | 125 | -0,79% | 144 | 145 | 0,69% | 142 | -1,39% |
| 2015 | 53 | 124 | 138 | 11,29% | 123 | -0,81% | 140 | 146 | 4,29% | 141 | 0,71% |

**Fig. 9.** Comparison of prediction results – moving average, LFLC expert system

## 5 Conclusion

The paper has introduced a very convenient combination of classic methods known from the time series analysis, which allowed identifying the main trend of development of the system. The expert system, which builds time series analysis, allows describing the effect of different parameters on the final value of the output from the system and it was therefore achieved very good estimates of the further development of the system. This procedure has been validated on real data that was added by a cooperating company. The obtained results have showed the correctness of the chosen strategy.

**Acknowledgment.** This work was supported by the project "LQ1602 IT4Innovations excellence in science" and during the completion of a Student Grant with student participation, supported by the Czech Ministry of Education, Youth and Sports.

## References

1. Brown, S.A.: Customer Relationship Management: A Strategic Imperative in the World of E-Business. Wiley Canada, New York (2000). ISBN 0-4716-4409-9
2. Swift, R.S.: Accelerating Customer Relationships: Using CRM and Relationship Technologies. Prentice Hall PTR, Upper Saddle River (2001). ISBN 0-1308-8984-9
3. Novak, V.: Linguistically oriented fuzzy logic control and its design. J. Approximate Reasoning **12**, 263–277 (1995). ISSN 0888-613X
4. Pokorny, M.: Artificial Intelligence in Modelling and Control. BEN - technická literatura, Praha (1996). ISBN 80-901984-4-9
5. Bin, X., Zhi-Tao, L., Feng-Qiang, N., Xin, L.: Research on energy characteristic prediction expert system for gun propellant. In: IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), vol. 2, pp. 732–736 (2010). ISBN 978-1-4244-6582-8
6. Bofeng, Z., Na, W., Gengfeng, W., Sheng, L.: Research on a personalized expert system explanation method based on fuzzy user model. In: Fifth World Congress on Intelligent Control and Automation, WCICA 2004, vol. 5, pp. 3996–4000 (2004). ISBN 0-7803-8273-0
7. Bofeng, Z., Yue, L.: Customized explanation in expert system for earthquake prediction. In: 17th IEEE International Conference on Tools with Artificial Intelligence ICTAI 2005, vol. 5, p. 371 (2005). ISBN 0-7695-2488-5
8. Wang, J.: Data warehousing and mining: concepts, methodologies, tools, and applications. Information Science Reference: Hershey, PA, c2008, vol. 6, (lxxi, 3699, p. 20) (2008). ISBN 978-1-59904-951-9
9. Khosrow-Pour, M.: Encyclopedia of information science and technology, p. 10384, 3rd edn. IGI Global (2014). ISBN 978-1-46665-889-9
10. Vaisla, K.S., Bhatt, A.K., Kumar, S.: Stock market forecasting using artificial neural network and statistical technique: a comparison report. (IJCNS). Int. J. Comput. Netw. Secur. **2**(8), 50–55 (2010). ISSN 2076-2739
11. Vaisla, K.S., Bhatt, A.K.: An analysis of the performance of artificial neural network technique for stock market forecasting. (IJCSE). Int. J. Comput. Sci. Eng. **2**(06), 2104–2109 (2010). ISSN 0975-3397
12. Baker, P., Canessa, M.: Warehouse design: a structured approach. Eur. J. Oper. Res. **193**(2), 425–436 (2015). doi:10.1016/j.ejor.2007.11.045. ISSN 0377-2217, Online ISSN 1872-6860
13. Jemelka, M., Chramcov, B., Kriz, P.: Design of the storage based on the ABC analyses. In: Proceedings of the International Conference on Numerical Analysis and Applied Mathematics (IVNAAM-2015), Greece, 23–29 September 2015. doi:10.1063/1.4951909. ISBN 978-0-7354-1392-4, ISSN 0094-243X
14. Walek, B., Farana, R.: Proposal of an expert system for predicting warehouse stock. In: 4$^{th}$ Computer Science On-line Conference 2015, CSOC 2015, UTB ve Zlíně, Zlín, pp. 85–91, 27–30 April 2015. ISSN 2194-5357

# New Framework Model to Secure Cloud Data Storage

Beldjezzar Leila[1(✉)], Zitouni Abdelhafid[1], and Djoudi Mahieddine[2]

[1] Lire Labs, Abdelhamid Mehri Constantine 2 University,
Ali Mendjli, 25000 Constanine, Algeria
{leila.megouache,abdelhafid.zitouni}
@univ-constantine2.dz

[2] Techne Labs, University of Poitiers,
1 Rue Raymond Cantel, 86073 Poitiers Cedex 9, France
mahieddine.djoudi@univ-poitiers.fr

**Abstract.** Nowadays companies are increasingly adopting the technology of cloud computing. This technology is subject to a lot of research and continuous advances are made. The use of cloud computing in the companies advantages such as: reducing costs, sharing and exchange of information between institutions, but the data in the Cloud computing are susceptible to be compromised and the companies are exposing to see their data loss. In this study, we address the subject of security in cloud computing; we expose and discuss some researches that had been proposed to secure the data stored in the cloud. And then we will present our new frameworks that ensure confidentiality of data storage in the cloud environment.

**Keywords:** Cloud computing · Security · Authentication

## 1 Introduction

Cloud computing is one of the new technologies that appeared in these last years. Its main objectives are to deliver different services for users, such as infrastructure, platform or software with a reasonable and more and more decreasing cost for the users. However, cloud computing is still in its initial stage [1]. The lack of standards, the security and the interoperability issues hamper the growth of cloud computing [2, 3]. Thus, the choice of clouds made by companies is usually based on the quality of services, but measuring the quality of cloud providers' approach to security is difficult because many cloud providers will not expose their infrastructure to customers.

Therefore, the security is an important factor that should be taken into account by cloud service providers.

Before making the transfer of the data towards Cloud, the company owes classify them and choose Cloud adapted according to the following categories.

The deployment models in cloud computing are:

(a) Public cloud: the service is provided by a third party via the internet. The physical infrastructure is owned and managed by the service provider. This cloud is less

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_5

secure compared to other models [4], since all applications and data are available to the public and accessible via the Internet.

(b) Private cloud: it is a dedicated cloud that is managed internally or by a third-party and can be hosted generally on premises or even externally [5]. The physical infrastructure is exclusively used by one organization. This Cloud offer a higher degree of security since only users in the organization have access to the private cloud.

(c) Community cloud: the physical infrastructure is controlled and shared by several organizations and is based on a community of interest [6].

(d) Hybrid cloud: combine two or more distinct cloud infrastructures. This cloud must be linked by a standard technology for data and applications portability.

The main characteristics of Cloud computing are [6]:

(a) On-demand self-service: users can access and control their services automatically without the intervention of the service provider.

(b) Broad network access: Services are available over the internet, they accessible from anywhere regardless of the device used.

(c) Resource pooling: the computing resources are pooled to serve multiple consumers, with different physical and virtual resources [6].

(d) Rapid elasticity: The computing resources are rapidly and dynamically provisioned. Users can increase and decrease the functionality available according to their needs. The capacities appear to be unlimited.

(e) Measured service: The resource use is controlled by leveraged metering capability.

This paper is organized as follows: Sect. 2 introduces cloud computing and cloud security. In Sect. 3, the related works is presented. In Sect. 4, we propose our framework. Finally, In Sect. 5, the conclusions of this work are presented.

## 2 Cloud Computing Security

In this section we will present the different security problems in cloud computing such as if the data loss or leakage.

Moving sensitive data to the cloud involves moving the control of the data to the service provider [7]. Therefore, the security and confidentiality of information becomes a major concern data security and authentication in cloud computing is similar to data security and privacy in traditional environments [8].

However, because of the characteristic of opening and multi-location of cloud computing, data security and privacy face to more risks.

The user data need to be protected [6]:

(1) ***Personally identifiable information:*** Includes any information that can be used to identify an individual, such as name, address, Tel…

(2) ***Sensitive information:*** requires additional protection. Such as personal financial information on job performance information and information considered being

sensitive personally identifiable information such as biometric information or collections of surveillance camera images in public places

(3) ***Usage data:*** consists of information collected from computer devices such as printers or habits of researchers.

(4) ***Unique device identities:*** information that could be attached to a user device, such as IP address, unique hardware identities, for example the health data is personal and sensitive information.

(5) ***Implementation of a program of data protection*** [9]:

Optimization of security interne by setting up devices internal for protect their systems.

Identification of Accommodated according to the exigency concerns the statutory particular requirements or the jobs which it is necessary to identify [10].

Means of piloting the customer has to have ways of piloting and operational follow-up supplied by the service provider Cloud.

Committee of safety between both parts (customer, supplier of service) where the person receiving benefits creates a "mutualzed and secure environment".

(6) ***The selection of the offer Cloud:*** the most adapted to the need must be made with operational consideration of the precautions and the contract employees.

In terms of protection, Cloud Security Alliance has proposed two action methods [11, 12]:

- The first method (and the simplest) is to ensure fine control of data access, fully through identity and access management

(e.g. ensuring that the raw data couldn't be accessed by human users, monitoring the access to requesters, authenticating its users).

- If necessary (regulatory perspective), encrypting the most sensitive data, but to be efficient, an encryption solution must appoint means of access control to fine-grained data and encryption key management, while maintaining a high level of performance.

## 3 Related Works

There are a number of work concerning the security, the privacy and the authentication of companies data in the cloud computing.

[8]: Indeed the cloud is a virtual space, which contains data that is fragmented; the data fragments are always duplicated and distributed on physical storage media in addition the cloud contains a restitution function to restore the data. But in this solution granularity of the selected fragmentation is important (the fragments may be too big or too small)

[11]: In this article, the authors present the models to maintain the confidentiality of data handled at the data integration system. The draft PAIRSE addresses the challenge of flexible and preserving privacy in data integration system.

To ensure protection of the data, the authors proposed execution model preserve privacy for data services (data services) that enable service providers to respect their privacy and security policies. The advantage of this model is added only a small increase of the execution time of the service and that model to protect the confidentiality of the data handled in the data integration system.

[12]: Proposes an access control mechanism to ensure confidentiality of data in the cloud. The mechanism is based on two protocols: ABE (Attribute Based Encryption) for data privacy and ABS (Attribute Based Signature) for user authentication. ABE is combined with ABS to ensure anonymity of users that store their data in the cloud. Key attributes and distribution is done in a Decentralized Manner.

[13]: Provides an overview of Common approaches to preserve confidentiality in e-Health Cloud. These approaches are classified into two categories:

Cryptographic approaches (based on encryption techniques) and non-cryptographic approaches (mainly use on access control). They also point out the advantages and disadvantages of each approach.

[14]: In this, the authors present a contribution to protecting the privacy of Web users. The objective of this work is to allow a client to query the search engine in a way to preserve privacy. This means that the search engine, which receives the request, or any opponent who listens to the network, cannot deduce the identity of the applicant (the user). The authors aim to generate false application (Fake query) that cannot be identified by the opponents (or engine research).

A major disadvantage is the introduction of irrelevant answers to protect the applications in this solution because they added noise to the search request to perform obfuscation (interference). This solution decreases the precision of the results and causes overload on the network.

## 4 Proposed Security Model

Major concerns and issues in security have been discussed in the previous sections. It has been observed that, despite quality research on security data outsourcing and data services for almost a decade, existing approaches on database encryption, certification, digital signatures [12], contractual agreements etc. have not gained much success in operations.

To date, there is minimal work done in the field of security of data as compared to traditional data storage. Different approaches are discussed with assorted categories of confidentiality, privacy, integrity and availability.

Our proposal framework be to create one networks virtual deprived between the customer and the Cloud of such goes out that the customer little to reach his space Cloud in a secure way, to add has it a double cryptography authentication. The Fig. (2) demonstrates our framework.

A software given to the supplier to the customer called ***VPN*** (virtual privacy network) *customer,* who allows establishing the connection of the customer his virtual network with a users and a password. This information shall be encrypted automatically.

Our framework consists of five steps, that each has a function explained as follows:

- ***Establishment of the connection***: The customer opens the application which is installed on it computer called "***vpncustomer***". Made a connect, a window opens asking him of entered the user name and the password this password is encrypted.
- ***Open the http address***: Where the application will once be connected, the customer goes on the internet page and entered the URL of the cloud; this URL is delivered by the provider.
- ***Identification and password:*** Once the user name and the word of pass seized for second time, at the same time this information is encrypted by the application according to the algorithm of encryption.
- ***Check of the certificate***: The supplier of the cloud verify has every entered the validity of the right of the customer as well as its certificate which includes everything piece of information of the customer (the last update, the expiry date of its contract…).
- ***Secure Access:*** The Access of the customer in their space is totally secured.

A software given to the supplier to the customer called ***VPN*** (virtual privacy network) *customer,* who allows establishing the connection of the customer his virtual network with a users and a password. This information shall be encrypted automatically.

Our framework consists of five steps, that each has a function explained as follows:

Wherever, we will use the public key cryptography just to exchange the symmetric key between the restitution and encryption program. Advanced Encryption Standard or ***AES*** (**S**ymmetric **E**ncryption **A**lgorithm) is a symmetric encryption algorithm. He won in October 2000 the AES competition, launched in 1997 by the NIST and became the new encryption standard for US Government organizations.

The following procedure describes the overall operation of AES. It takes as input a data table T (clear text) that is modified by the procedure and returned output (ciphertext).

The algorithm used to encrypt data:

```
  Input : table T and key K Out-
put : table T modified Function
AES (T, K)
  Begin
  KeyExpansion (K, TK);
  AddRoundKey (T, TK [0];
  for (i = 1; i<nr; i + +)
  Round (T, TK [i]);
  FInalRound (T, TK [nr]);
  end
```

The algorithm used to decrypt data:

Decryption with AES:

The encryption routine can be reversed and rearranged to produce a decryption algorithm.

```
AES_Decrypt(T, K) {
KeyExpansion(K, RoundKeys); /*
Initial addition */
AddRoundKey(State, RoundKeys[Nr]);
for (r=Nr-1; i>0; r--) {
InvShiftRows(T);
InvSubBytes(T); AddRoundKey(T,
RoundKeys[r]); InvMixColumns(T);

}
/* FinalRound */ InvShiftRows(Out);
 InvSubBytes(Out);
 AddRoundKey(Out,RoundKeys[0]);
}
```

When the connection will be establishes, the customer goes on the internet page and opens him on the address URL, Which is delivered by provider, when the page opens another user and password are required Fig. 2. The customer little to reach his workspace Cloud, to put this data in the daytime, and to consult them, or used the available applications. The work of the customer will be protected by two authentications in entries and even the little protected customer a copy of its work at his home, and nothing will be lost or destroys even in the case of cut or maintenance.

– *Authentication Protocol*

This diagram of sequence explains better our architecture which is in the Fig. 1. Our process follows the following steps:

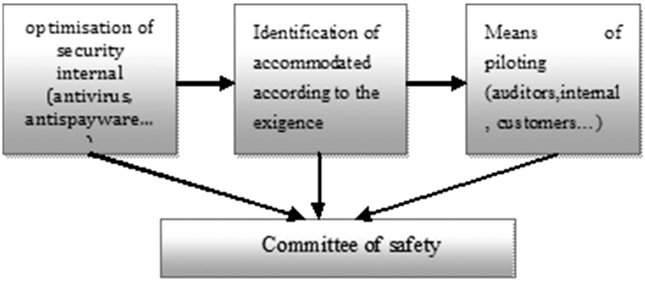

optimisation of security internal (antivirus, antispayware...)
Identification of accommodated according to the exigence
Means of piloting (auditors,intemal , customers...)
Committee of safety

**Fig. 1.** How to secure data [10]

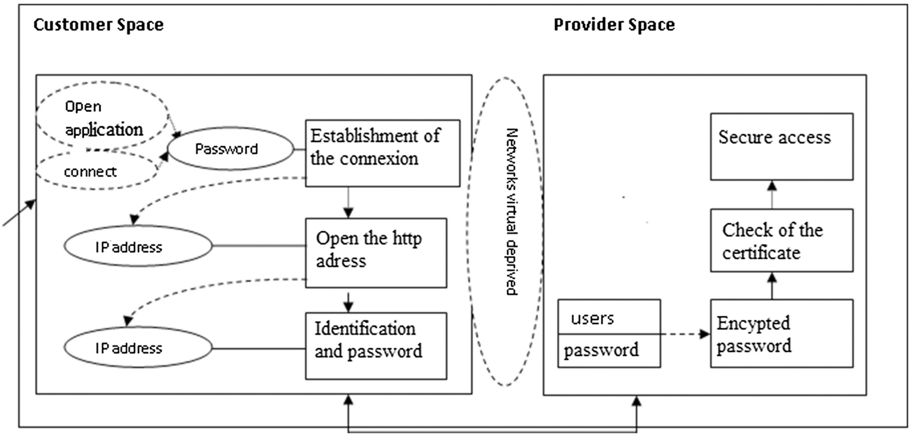


**Fig. 2.** Prototype of safety

*Connect;*

**Step 1:** it the phase of registration, the connection is established between the customer and the private network

$$Customer\ g\ \boldsymbol{of}\ private\ network$$

**Step 2:** the user's name and key is verified (by encrypted and decrypted algorithm)

$$Encrypted/decrypted\ \text{https : // address}$$

**Step 4:** generate Contract, in this step the second user's name and code shall be seized by the customer, and Right of customers shall be verify
**Step 5:** the key is encrypted. Just after this step, if Valid key: Accept Starting Service cloud else Reject

$$Customer\ \mathbf{g}\ \boldsymbol{Of}\ service\ provider$$

**Step 6:** decrypt key by the service provider, the customer can access in his private space

$$Access\ to\ the\ space$$

*Disconnect;*
*Disconnect;*

In this protocol, the user will have only to present his user name and the key in Step1 and step 5 to obtain a service. Unlike other solutions this identification is encrypted, the encrypted and the decrypted will not be seen by the customer.

By this solution only user can be access to services and, it's better to encrypted all the data which will be transfer in the cloud, it sets a lot of time. (Fig. 3)

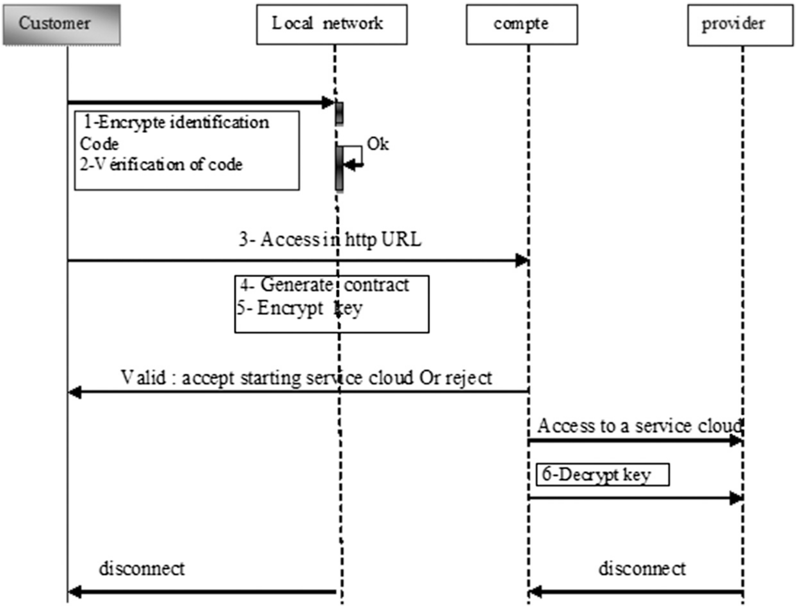


**Fig. 3.** Authentication protocol

## 5 Conclusion and Future Works

The cloud computing allows companies not only to protect their data and transform their spending investment into operational spending, but also to manage better their budget, because they pay only what they use. Numerous companies turn at present to Cloud computing for the saving, the archiving and the off-site restoration. However, Cloud computing can be also used to have solutions of real time replication and high availability, To reduce at the most the interruptions of service and the losses of data. The companies which envisage the appeal to the cloud computing for the resumption after breakdown have to verify how their data, application and suppliers will handle diverse questions. Work is currently going on the frame work implantation where it will be applied to a specific case study. Further research could be realized to improve and to extend the present work by including and resolving the interoperability issues in the cloud.

## References

1. Nithiavathy, R.: Data integrity and data dynamics with secure storage service in cloud. In: Proceedings of the International Conference on Pattern Recognition, Informatics and Mobile Engineering, pp. 125–131. IEEE (2015)
2. Clement, T.: The security in the cloud computing, presented in CQSI, October 2012

3. Nithiavathy, R.: Data integrity and data dynamics with secure storage service in cloud. In: 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME). IEEE (2013)
4. Alzain, M.A, Pardede, E.: Protection of data and cloud computing Europe (2011)
5. Kumbhare, A., Simmhan, P., Prasanna, V.: Cryptonite: a secure and performant data repository on public clouds. In: IEEE 5th International Conference on Cloud Computing (CLOUD), pp. 510–519. IEEE (2016)
6. Huang, H., Liu, K.: Efficient key management for preserving HIPAA regulations. J. Syst. Softw. **84**, 113–119 (2015)
7. Gupta, A., Verma, A., Kalra, P., Kumar, L.: Big data: a security compliance model. In: IT in Business, Industry and Government (CSIBIG), pp. 1–5, Indore (2014)
8. Cigref, cloud computing and protection of data, network companies enterprises (2015)
9. Idrissi, H.K., Kartit, A., El Marraki, M.: A taxonomy and survey of cloud computing, presented at the Security Days (JNS3), pp. 1–5(2013)
10. Merkle, Ralph C.: A Certified Digital Signature. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 218–238. Springer, New York (1990). doi:10.1007/0-387-34805-0_21
11. Benslimane, D., Barhamgi, M., Cuppens, F.: PAIRSE: a privacy- preserving service-oriented data integration system. SIGMOD Record **42**(3) (2013)
12. Heffetz, O., Ligett, K.: Privacy and data-based research. No. w19433. National Bureau of Economic Research (2013)
13. Deyan, D.C., Zhao, H.: Data security and privacy protection issues in CC. In: IEEE International Conference on Computer Science and Electronics Engineering (ICCSEE) (2015)
14. Petit, A., Ben Mokhtar, L.B., Kosch, H.: Towards Efficient and Accurate Privacy Pre-serving Web Search. In: MW4NG 2014, December 8–12, Bordeaux, France (2014)

# Proposal of Complex Software Applications

Cyril Klimeš[1], Vladimír Krajčík[2](✉), and Radim Farana[1]

[1] Department of Informatics, Mendel University in Brno,
Zemědělská 1, 613 00 Brno, Czech Republic
{cyril.klimes, radim.farana}@mendelu.cz
[2] Faculty of Informatics, Pan-European University,
Tomášikova 20, 821 02 Bratislava, Slovak Republic
vladimir.krajcik@paneurouni.com

**Abstract.** Designing complex software tools can be advantageously supported by using modelling tools based on fuzzy algorithms and principles used in modelling of decision-making processes. Nevertheless, it is necessary to dispose of a system enabling to acquire knowledge from reactions of users of these complex systems. It primarily concerns simulation of decision-making processes as well as ways of knowledge acquisition. The solution is a proposal of an automated tool which enables to substitute an expert when selecting a suitable action based on symptoms of the current way of work. This tool will suggest which practices are suitable to use in SW development. Therefore, the tool supports decision making on practices relevancy, whereas implementation itself is up to the team.

**Keywords:** Information search · Decision under indeterminacy · Modelling of information systems under indeterminacy · Adaptive model

## 1 Introduction

When designing complex software applications, there are frequently issues that can be described by the following characteristics:

- the problem is not algorithmizable,
- the problem is new, not repeatable, unique, usually very important,
- there are more factors influencing the solution (cannot be expressed by numbers),
- some of the factors are not known at all or there are complicated relations,
- changes of some elements in the software application environment where the solution is carried out are random,
- there is no routine solution,
- there are no analytical methods to find an optimal solution,
- there are more criteria to assess the solution, some of which are of a quantitative nature,
- interpretation of information necessary for the decision is difficult,
- a human is usually an active element of the system (they create and change the system by deliberate activity).

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_6

Thus, we have to decrease the risk in a proposal of a solution of a **badly structured problem**, when a wrong decision of the project team can result in creating an unsuitable application or even in an application which does not meet customer's idea. It is then necessary to decrease the risk of a wrong decision to the lowest possible level. It primarily concerns SW applications are crucial from the point of view of maintaining company strategy in the future, i.e. so-called **strategic applications**.

The solution is a proposal of an automated tool which enables to substitute an expert when selecting a suitable action based on symptoms of the current way of work. This tool will suggest which practices are suitable to use in SW development. Therefore, the tool supports decision making on practices relevancy, whereas implementation itself is up to the team. This step only reduces the risk of selecting an unsuitable practice, but not the risk of unsuitable implementation and misunderstanding the principle in the practice background. The tool cannot eliminate misunderstanding so support of an experienced member of the team or a mentor is necessary.

In order to propose a suitable decision-making tool, we have to define the decision-making process, input and output data, data base and rules. Simulation of decision-making processes are specific in the following [1]:

- The decision making does not rely only on analytical information, but mostly on knowledge represented by a recognition process and a process of abstraction (which is a privilege of brain activity).
- The decision can be made by several approaches considering how many individuals will assess it.
- It is very difficult to accurately define the algorithm of making a decision.
- Great part of information used in decision making is of external origin with respect to already established and known data base of the decision-making problem.

A decision-making process, in a wider sense, can be defined as an organic unit of three phases:

- Information (knowledge acquisition).
- Planning (considering the alternatives).
- Selecting (selection of a variant).

Making a decision relates only to the last phase, the first two phases (knowledge acquisition and considering the alternatives) are only preparative ones.

In order to define the structure of a decision-making process and thus to create prerequisites for finding effective processes for its algorithmization, we have to deal with decision-making processes in a wider, primarily methodological, perspective. One of characteristic features of decision-making processes is the fact that they often work with indeterminate (but quantitative) information, which often results from the fact that the input quantities of these processes are defined by a human based on their estimates, experience, opinion, etc.

Designing complex software tools can be advantageously supported by using modelling tools based on fuzzy algorithms and principles used in modelling of decision-making processes. Nevertheless, it is necessary to dispose of a system enabling to acquire knowledge from reactions of users of these complex systems.

It primarily concerns simulation of decision-making processes as well as ways of knowledge acquisition.

Implementation of such a decision-making system was published in [1]. Following this published, general decision-making model, the paper discusses its use in the process of proposing complex software tools [2]. An example is provided in a form of a proposal of a safe operating system architecture.

## 2 General Decision-Making Model

This decision-making model stems from the structure described in [1]. Elements of the decision-making process were divided in the following groups:

S – set of situations,
D – set of all possible solutions,
G – set of all objectives (admissible) of subsequent functioning of the given system,
F – of all existence levels (probabilities) of the given object,
K – set of all assessments of the given solution,
T – time interval.

The decision-making process itself is represented in various mappings between those sets. It primarily concerns the following mappings:

1. process of information completion about the given situation and its assessment, i.e. selection of only such information that is relevant for the final solution:

$$M_1 : S \times T \times F \rightarrow S \times T \times F \tag{1}$$

2. process of creation of all admissible solutions which consists of two partial processes

$$M_2 = M_{22} \circ M_{21} \tag{2}$$

where

*$M_{21}$* – formulating objectives of the control based on the description of the given situation $M_{21} - S \times T \times F \rightarrow G \times S \times T \times F$

*$M_{22}$* – formulating admissible solutions $M_{22} - G \times S \times T \times F \rightarrow D \times S \times T \times F$

3. process of modelling of all effects of admissible solutions:

$$M_3 : D \times S \times T \times F \rightarrow D \times S \times T \times (S \times T)^{*} \times F \tag{3}$$

where $(S \times T)^{*}$ determines the set of all strings over $(S \times T)$, (each admissible solution is assigned with a set of situations and their time courses which are created based on the given decision).

4. process of approving the solution itself which consists of two partial processes

$$M_4 = M_{42} \circ M_{41} \tag{4}$$

where

*$M_{41}$* – assessing the behavior of the effects of admissible solutions

$$M_{41} - D \times S \times T \times (S \times T)^{*} \times F \rightarrow D \times K \times T \times F,$$

*$M_{42}$*, selection of the best variants $M_{42} - D \times K \times T \rightarrow D \times T$.

The whole decision-making process consists of a gradual composition of the following processes,

$$M = M_4 \circ M_3 \circ M_2 \circ M_1 \tag{5}$$

as can be depicted in the following diagram (see Fig. 1).

Note that implementation of processes *$M_1$* – *$M_4$* in [1] was carried out by using so-called fuzzy algorithms with the results of the fuzzy sets theory [4–6].

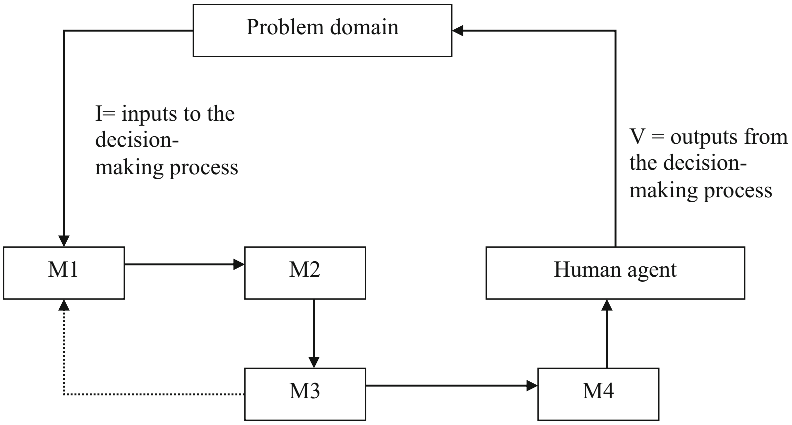

Problem domain
I= inputs to the decision-making process
V = outputs from the decision-making process
M1
M2
Human agent
M3
M4

**Fig. 1.** Decision-making process

## 3 Model of Creating a Complex Software Application

When creating a software application, it is very important to encompass and analyze the key requirements and needs of the future software. Those requirements can be stored in various forms (document with a text description of the requirements, visions containing identification of the key system functionalities depicted by use-case diagrams, etc.). Our objective is to process vague (indeterminate) information created during the analysis of the key requirements and functionalities of the future software.

The above-presented decision-making model is fully usable for creation of complex software applications. Assume that the proposed software application is an information system providing information from a local database in a form described by a user using rules. Those can be described vaguely, most often by linguistic terms characterizing the size of specific quantities or relations between specific quantities. In order to implement those inputs and algorithms on a computer, it is necessary to use a suitable mathematical apparatus. One of the possibilities is the use of the fuzzy sets theory [4–6].

Let's consider an example of input information in a decision-making process about selecting the most important processed information by a software application, which can be divided into the following groups:

- amount of data;
- content of information;
- importance;
- availability;

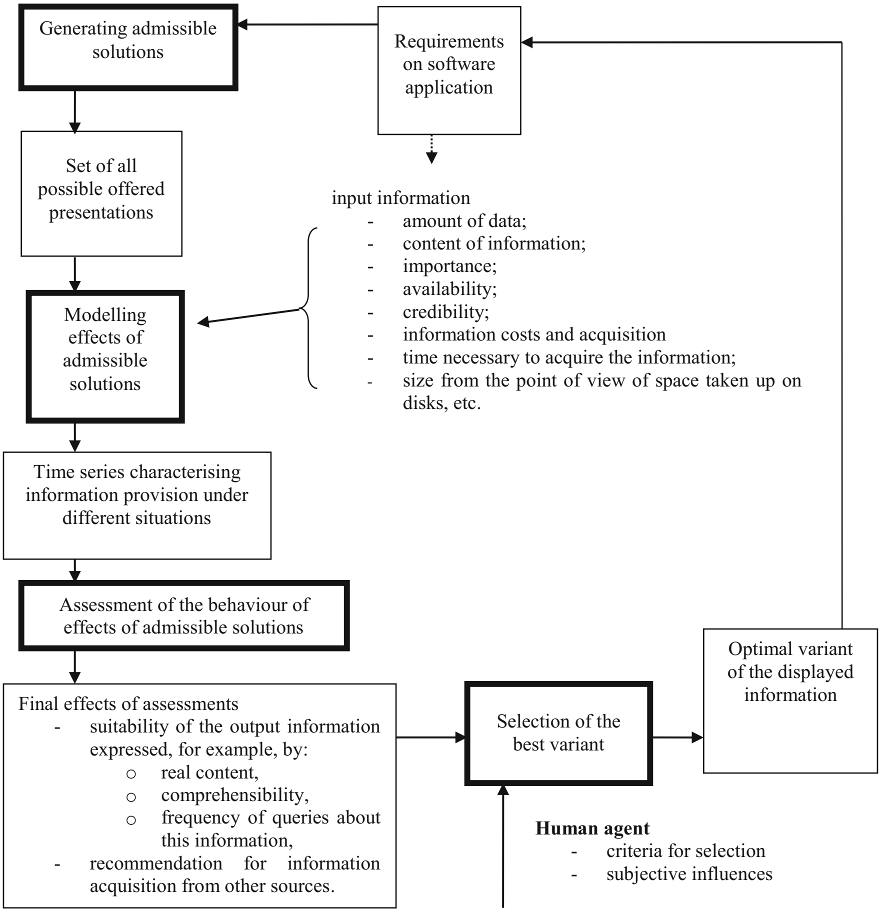

Generating admissible solutions
Requirements on software application
Set of all possible offered presentations
input information
- amount of data;
- content of information;
- importance;
- availability;
- credibility;
- information costs and acquisition
- time necessary to acquire the information;
- size from the point of view of space taken up on disks, etc.
Modelling effects of admissible solutions
Time series characterising information provision under different situations
Assessment of the behaviour of effects of admissible solutions
Final effects of assessments
- suitability of the output information expressed, for example, by:
○ real content,
○ comprehensibility,
○ frequency of queries about this information,
- recommendation for information acquisition from other sources.
Selection of the best variant
Optimal variant of the displayed information
Human agent
- criteria for selection
- subjective influences

**Fig. 2.** Structure of the decision-making system

- credibility;
- information costs and acquisition
- time necessary to acquire the information;
- size from the point of view of space taken up on disks, etc.

The output of the proposal system should be a set of all possible solutions represented by the following characteristics:

- suitability of the output information expressed, for example, by:
  - real content,
  - comprehensibility,
  - frequency of queries about this information,
- recommendation for information acquisition from other sources.

The whole decision-making process is shown in the Fig. 2.

## 4 Decision-Making Model for a Proposal of a Safe Operating System Architecture

A model for a proposal of architecture of a safe operating system stems from the general model for decision-making support under indeterminacy presented in chap. 2 [3].

**Process $\boldsymbol{M_1}$** represents completion, i.e. input data completion which relate to architecture of an operating system, and selection of relevant components of an operating system which are important for the final solution.

Input data completion is realized on the basis of a questionnaire, which is answered by the user (operator) of the expert system. A general architecture of an operating system (a set of all components) is then specified and only components relevant for the final solution are selected.

Process $M_1$ is expressed as:

$$M_1 : K \times A \to K^* \tag{6}$$

where:

$K$ – a set of all components of an operating system which enter the decision-making process. It represents a general architecture of an operating system.
$A$ – a set of answers which describe actual architecture of an operating system. It enters the decision-making process through a questionnaire.
$K^*$ – the final set of components after information completion. It represents the real architecture of an operating system.

**Process $\boldsymbol{M_2}$** generates all admissible solutions. It has 3 sub-processes which create the following (based on security requirements):

1. a set of operating system components that are subject to protection,
2. a set of threats that the components are expose to,
3. a set of vulnerabilities relating to the components and threats (Fig. 3).

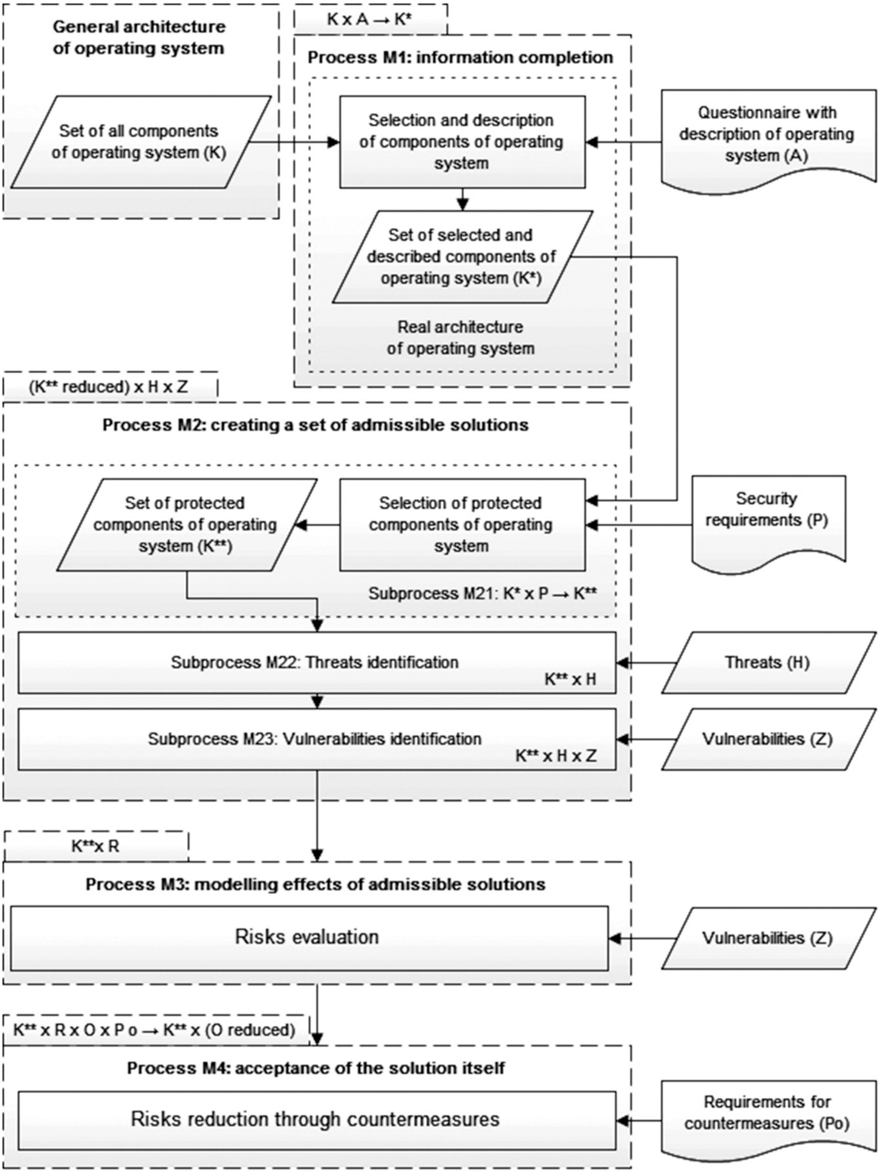

General architecture of operating system
Set of all components of operating system (K)
K x A → K*
Process M1: information completion
Selection and description of components of operating system
Questionnaire with description of operating system (A)
Set of selected and described components of operating system (K*)
Real architecture of operating system
(K** reduced) x H x Z
Process M2: creating a set of admissible solutions
Set of protected components of operating system (K**)
Selection of protected components of operating system
Security requirements (P)
Subprocess M21: K* x P → K**
Subprocess M22: Threats identification
K** x H
Threats (H)
Subprocess M23: Vulnerabilities identification
K** x H x Z
Vulnerabilities (Z)
K**x R
Process M3: modelling effects of admissible solutions
Risks evaluation
Vulnerabilities (Z)
K** x R x O x P o → K** x (O reduced)
Process M4: acceptance of the solution itself
Risks reduction through countermeasures
Requirements for countermeasures (Po)

**Fig. 3.** Model of a proposal of a safe operating system

### Process $M_{21}$

Based on security requirements, this process consists in creating a set of operating system components that are subject to protection. Identification of these components is realized using rules from a knowledge base. Process $M_{21}$ is expressed as:

$$M_{21} : K^{*} \times P \rightarrow K^{**} \tag{7}$$

where:

$P$ – a set of security requirements laid on the operating system by the user. They enter the decision-making process using a questionnaire.
$K^{**}$ – the final set of operating system components which are subject to protection (protected operating system components respectively).

**Process M$_{22}$** represents creating a set of threats affecting the operating system components which are subject to protection. Their identification is realized on the basis of a knowledge base depending on security requirements. Process M$_{22}$ is expressed as:

$$M_{22} : K^{**} \times H \tag{8}$$

where: $H$ – the final set of threats which the components are exposed to, affected by respectively.

**Process *M*$_{23}$** consists in creating a set of vulnerabilities relating to the protected operating system components and their threats. Vulnerability identification is realized on the basis of a knowledge base depending on threats, absence or insufficiencies of appropriate security measures (they are part of an operating system too).

It holds that a threat takes advantage of vulnerability and if there is no threat to take advantage of such vulnerability, there is no risk of breaking security of an operating system. Thus the set of operating system components subject to protection shrinks. Process $M_{23}$ is expressed as:

$$M_{23} : (K^{**} narrowed) \times H \times Z \tag{9}$$

where: $Z$ – a set of vulnerabilities relating to the components and their threats.

**Process *M*$_3$** models effects of admissible solutions. Based on the rules from the knowledge base, it qualitatively assesses risks for the components subject to protection. A security risk is assessed for individual threats to each protected component. The level of such a risk determined on the basis of probable occurrence of the threat and the extent of its impact on the protected component, where:

- the probability of threat occurrence can be determined in relation to the number of vulnerabilities or attributes of the component. It rises if user's IT knowledge is lower, if there are several users, or if the operating system works in unsecured computer network without appropriate security measures.
- the threat impact on the component is assessed by expert.

Process $M_3$ is expressed as:

$$M_3 : K^{**} \times R$$

where: $R$ – a set of risks affecting these components.

**Process *M*$_4$** selects the most suitable measures for operating system components subject to protection. The purpose of such protective measures is to reduce or eliminate security risks existing for the protected components. Identification of protective measures is realized on the basis of rules from a knowledge base depending on threats which the components are exposed to.

Selection of all existing protective measures for particular components narrows on the basis of user's requirements (e.g. level of configuration and use of an antivirus system: easy – advanced – expert, etc.)

Process $M_4$ is expressed as:

$$M_4 : K^{**} \times R \times O \times P_O \rightarrow K^{**} \times (O\,narrowed)$$

where:

$R$ – a set of risks affecting these components.

$O$ – a set of all protective measures relating to the components and their threats.

$P_o$ – a set of user's requirements on the protective measures. They enter the decision-making process through a questionnaire. They are a base for further narrowing the set of protective measures for individual components.

## 5 Conclusion

The presented paper solves implementability of a model of decision making under indeterminacy for implementation of software tools, such as a proposal of a safe operating system architecture [2]. The results showed that the model proposed in [1] is suitable for these tasks.

**Acknowledgment.** This work was supported a Student Grant with student participation, supported by the Czech Ministry of Education, Youth and Sports.

## References

1. Klimeš, C.: Model of adaptation under indeterminacy. Kybernetika **47**(3), 355–368 (2011). Prague, ISSN 0023-5954
2. Klimeš, C., Bartoš, J.: IT/IS security management with uncertain information. Kybernetika **51** (3), 408–419 (2015). Prague, ISSN 0023-5954
3. Masár, J., Bartoš, J., Klimeš, C.: Deployment of mandatory access control policies of operating system under uncertainty. In: WorldCIS-2013 Proceedings, pp. 127–131. Infonomics Society, UK (2013). [2013-12-09]. ISBN 978-1-908320-17-9
4. Novák, V.: Fuzzy sets and their application (in Czech). SNTL, Prague (1986)
5. Novák, V.: Fuzzy Relation Equations with Words, 1st edn, pp. 167–185. Springer, Heidelberg (2004). ISBN 3-540-20322-2
6. Novák, V., Perfilieva, I., Močkoř, J.: Mathematical Principles of Fuzzy Logic, 1st edn., 320 p. Kluwer Academic Publishers, Boston (1999). ISBN 0-7923-8595-0

# Migrating from Conventional E-Learning to Cloud-Based E-Learning: A Case Study of Armangarayan Co.

Mohammad Reza Rasol Roveicy[1](✉) and Amir Masoud Bidgoli[2]

[1] Computer Department, Islamic Azad University of Tehran,
North Tehran Branch, Tehran, Iran
Rasoli@live.co.uk

[2] Department of Computer, Azad University, North Tehran Branch, Tehran, Iran
am_bidgoli@iau-tub.ac.ir

**Abstract.** Cloud-Computing may be viewed as a resource available as a service for virtual data centers, but cloud computing and virtual data centers are not the same, cloud computing is logical evaluation of Information Technology (IT) in a world that is becoming more and more based on the division of the work. In this paper we aim at examining the recent successful trend and efficiency brought about by cloud computing in the Armangarayan company. This company is delivering e-learning materials through cloud computing to the bank of 'Ayandeh' in Iran by considering SaaS model. In the present study a part from cloud E-learning, we came across with interesting findings when shifting from traditional e-learning to e-learning cloud such as total cost reduction to 65%, increase in number of Bank's employee who are really interested in e-learning for reaching a better career, IT labor and development team cost decreased to 66%, and testing time to 66.28%.

**Keywords:** Cloud-computing · Conventional e-learning · E-learning cloud

## 1 Introduction

In the cloud computing system the data is stored on remote servers accessed through the internet. Traditional e-learning can be viewed as web-enabled version computer-based training, focusing on providing a multimedia experience including elaborate animations to simulations. Due to its high Production cost, one hour of tradition e-learning typically takes two to five professionals between four and six months to complete and resulting course is typically as graphically sophisticated as a movie and video game. Traditional e-learning teams, which might consist of writers, SME's, instructional designers, programmers and graphic artists, need to develop storyboard and scripts prior to implementation. The National Institute of Standards and technologies (NIST) is releasing its first guidelines for agencies that want to use could computing in second half of 2009, and groups such as the Jeticho Forum are bringing security executive together to collaborate and deliver solutions. The purpose e-learning cloud is to provide E-learning as service (EaaS) [1]. This service model e-learning provide educational institutions with e-learning systems and e-learning resources as on

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_7

demand services. The e-learning cloud proposes five components based on cloud computing ontology to be dynamic data center, testing platform, security control, operational management and software Platform. Dynamic data center is the physical heart of management cloud to dynamically manage, deploy and secure services. Testing platform provides capability to incorporate new technology and contexts into e-learning platforms and allows educational institutions to test new e-learning applications prior to full launch, security helps to control cases of server crashes, lost of data and applications located in remote site of a vendor. For instance, virtualization, technology allows for rapid and cost effective replacement for a server [2].

## 2 Concept of E-Learning

Some people hold that e-learning is limited to what takes place entirely within a web-browser without the need for other software or learning resources. Such a pure definition, though, leaves out many of truly effective uses of related technologies for learning. There are a lot of complex definitions of e-learning. The simple one is given as below:

***E-Learning is the use of electronic technologies to create learning experiences***
This definition is deliberately open-ended, allowing complete freedom as to how these experiences are formulated, organized, and created. Recollecting that this definition does not mention "courses", for courses are just one way to packages e-learning experiences. It also does not mention any particular authority tool for management system. This attitude is the biggest problem in e-learning: a creator centric attitude rather than a learner-centric one [3].

## 3 What's Cloud Computing?

Cloud computing is a paradigm that focuses on sharing of resources and computations over a scalable interconnected nodes. Cloud computing can transform education and has following desirable properties that can be explored to solve inherent e-learning challenges [4]. Dynamic scalability, self service, measured service, resource pooling, resource sharing, rapid elasticity, mobility support, service availability, fast connection, virtualization, multi-tenacity and pay as you consume. In addition, cloud computing provides a great opportunity for faster processing power, cost effective maintenance, less computing downtime, large storage, maximum resource utilization, maximum return on investment, increase competitiveness, access to latest infrastructures and improves agility by allowing customers to provide products as utility services. Cloud computing is increasingly used for transacting business activities with lot of patronage from ICT based organizations. The continuous refinement of cloud computing by its providers increases possibility of making it an alternative technology for future investment.

Cloud computing is an extension of traditional internet, service oriented architecture, web services and grid computing using virtual shared computing servers to deploy products, resources, software, infrastructures, devices, platforms and databases as utility services. The basic service ontology models of cloud computing are the following. Software as a Service (SaaS) provides opportunity for customers to run cloud applications through web browser, thin computing terminals and hosted desktops and eliminates the necessity to install and run these applications on consumer devices. Platform as a Service (PaaS) is the capability that allows customers to build and deploy specific applications using cloud software development environment (languages, libraries, functions, classes, components, services, packages and tools) supported by a cloud provider. Infrastructure as a Service (IaaS) provides customers with computational resources such as network (Network as a Service, NaaS) and data storage (Data as a Service, DaaS) to perform specific tasks. Cloud computing uses an architecture described in Service Oriented Architecture (SOA). Service consumers have flexibility to choose services in cloud according to their needs. The standard related to services consumption is maintained through Service Level Agreements (SLAs) [5]. The comparison of traditional e-learning and could-based e-learning is presented in Fig. 1.

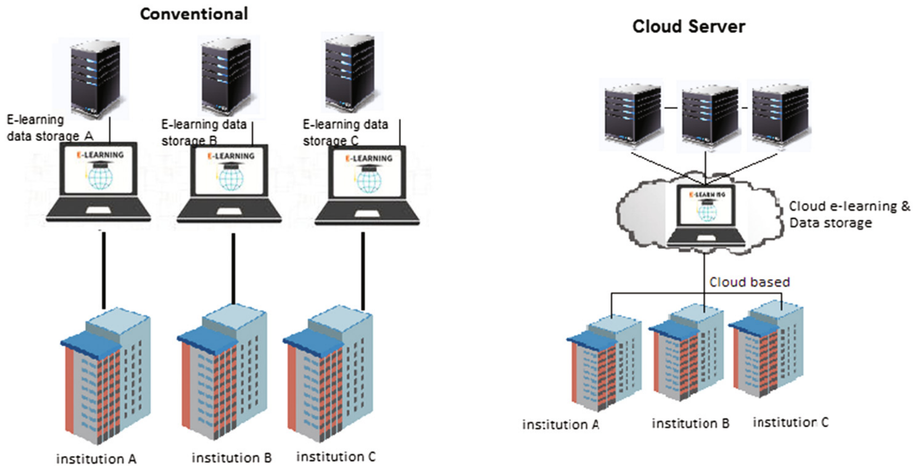


**Fig. 1.** Comparison of architecture of e-learning cloud and traditional e-learning

## 4 Literature Review

The research of e-learning in the cloud environment have been carried out by previous researchers, such as those conducted by Chuang, Chang, and Sung (2011), Dong et al. (2009), Vishwakarma & Narayanan (2011), Pocatilu (2010) and Ghazizadeh (2012). Research on the application of e-learning in a cloud environment is one form of cloud services education services. There are several architectural cloud-based e-learning have been proposed by previous researcher. Phankokkruad (2012) proposed e-earning architecture based on cloud computing consists of three layers: (1) infrastructure layer, (2) platform (middle) layer, and (3) application layer. Infrastructure layer is a hardware layer that supplies the computing and storage capacity for the higher level and this layer, which is used as e-learning and software virtualization technologies, ensures the stability and reliability of the infrastructure. The second layer is Platform layer, this layer is a middle layer consisting middleware that is Web service. It is used for providing the learning resources as a service. This layer consists of two modules: item classification module (ICM) and course selection module (CSM). They are used for accessing the items from the item bank and selecting suitable learning content from the content database. The third layer is Application layer which is responsible for interface provision for the students. Not much difference can be inferred from the comparison of the architecture delivered by Phankokkruad (2012) and Wang, Pai, & Yen (2011). They proposed an architecture of e-learning-based cloud computing consists of three layers, namely: (1) infrastructure layer, (2) middleware layer, and (3) application layer [6].

In Sect. 4 we choose Armangarayan company app. for implementing our e-learning services based on cloud as a case study in Iran. This will be described as the following:

## 5 The Case Study and Discussion

This section describes results from a case study implementation of e-learning on a cloud. SLA's and related parameters are first captured, and then tracked through metrics for our evaluations, a course management service was developed to display available course to the user. The interacting agent collect status of this service and passes information to the parameter collector for storage. According to the inquiry about the performance of Armangarayan in delivering educational training to the bank of Ayandeh in Iran, the important results which have been achieved by shifting to e-learning cloud during different time interval in 2016 will be as the following. In February 2016 at time of beginning traditional e-learning the number of users using Ayandeh Bank E-learning App. was on 100 users, that is because the company at the beginning was not able to anticipate this large increase in number and at the same time it was impossible to estimate the required resources, this caused our self-managed server to be down. The above number increased to 700 users in different time periods. Due to the ignorance in anticipating this amount of increase in number, this led the server resources to be down in different time periods as a result of which many hours of time consumed every week were lost for testing and running the server services (webserver & data base) unnesseraily. This required to have a team of 6 experts to manage and develop the server. Auto Scaling also helped us automatically increase number of instances (Storage, Ram, CPU) during demand spikes to maintain performance and decrease capacity during lulls to reduce costs. Moreover, because of capability of auto scaling cloud computing during demand peak, this could avoid the server from being down as a result of which cost of company decreased. In other words, Our cost in time of traditional e-learning which was $200 USD per month decreased to $75 USD per month while moving to e-learning cloud. While using e-learning cloud the hosting capacity increased simultaneously by 1000 user. When migrating to cloud-based e-learning, the uptime of the server increased to 99.95%. This means that the maximum time that a server could be out of range has been only 1–2 h in a month and the service outage in cloud computing e-learning ranges from 0–0.9.

**Table 1.** Comparison of E-learning systems before and after mobbing on to Cloud

| E-learning characteristics | Before moving to cloud | After moving to cloud |
|---|---|---|
| Need for Deployment | Y | N |
| More Loss of control of any application or resources | N | Y |
| Conflicts between opposing goals of different Clients, either play it together if not need to separate them | N | Y |
| Higher risks of Resource availability and failure | N | Y |
| Lack of trust in data alteration before storing | N | Y |
| Denial of Service attacks in critical server health | N | Y |
| Higher risks of Stress, Load and congestion | N | Y |
| Difficult to audit | N | Y |
| Monitoring of client logs and information by third party | N | Y |
| Need for Technical IT support for Fail over | Y | N |
| Need for e learning system development team | Y | N |
| Need for extra hardware and software resources | Y | N |
| Need to configure latest technology updates | Y | N |
| Need to arrange own extra power and cooling | Y | N |
| Lack of computation and accuracy trust | N | Y |
| Lack of confidentiality | N | Y |
| Lack of trust on security policies and access control | Y | Y |
| Daily storage and backup burden | Y | N |
| Massive cost | N | N |
| High speed internet connection | N | Y |
| Subscription and registration charges | Y | Y |
| Need for requirement gathering and elicitation | Y | N |
| Need for project management | Y | N |
| Need for coding | Y | N |
| Need for testing | Y | N |

The information processing speed also increased tremendously [7]. The summary of above discussion and advantage of e-learning on cloud has been given in [8] Table 1.

## 6 Conclusions and Future Work

In this paper we have conducted a comparative analysis for e-learning before and after moving on to cloud computing environment. We have investigated the issue of cloud computing technology and its deployment in e-learning systems using Armayan-garayan company as vendor to provide cloud-based e-learning by applying SaaS module to Ayandeh bank of Iran located in Tehran city. The investigation results confidentially support moving e-learning to cloud computing environment. Cloud-based e-learning shown in Table 1 can reduce deployment team cost, technical support team cost, testing effort, requirement elicitation, burden of daily backup management, and cost overall project expenditure [8]. As cloud implementation based on SOA architectures functional and non functional SLAs have to be thoroughly reviewed in order to uplift traditional e-learning system to cloud.

## References

1. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. NIST, Gaithersburg (2011)
2. Odunaike, S.A., et al.: Mitigating rural e-learning sustainability challenges using cloud computing technology. In: World Congress on Engineering and Computer Science (2012)
3. Horton, W.: E-learning by Design, 2nd edn. (2012)
4. Riahi, G.: E-learning system, based on cloud computing. In: International Conference on Soft Computing and Software Engineering (SCSE 2012) (2012)
5. Sampat, K.: Disruptive Cloud Computing and IT (2015)
6. Selvandro, N., et al.: Cloud–based e-learning: a proposed model and benefits by using e-learning based on cloud computing for educational institution. In: International Conference on Information and Communication Technology, CT-(ur Asia 2013) (2013)
7. Bank-Ayandeh-E-learning-app. https://itunes.apple.com/us/app/bankayandeh/id1071645417?mt=8
8. Ahmed, F.F.: Comparative analysis for cloud–based e-learning. In: International Conference on Communication, Management and Information, ICCMIT (2015)
9. Masud, M.A.H., Huang, X.: An E-learning System Architecture based on Cloud Computing. IEEE (2012)

# Enterprise Architecture: An Alternative to ArchiMate Conceptualization

Sabah Al-Fedaghi(✉)

Computer Engineering Department, Kuwait University,
P.O. Box 5969, 13060 Safat, Kuwait
sabah.alfedaghi@ku.edu.kw

**Abstract.** ArchiMate is a modeling language developed to provide a uniform representation of enterprise architecture descriptions. It visualizes the different architecture domains and their underlying relationships and dependencies. Many organizations are already using it as the standard for describing enterprise architecture. Because of its inherent holistic nature and as a direct consequence of its coarse-grained language, ArchiMate lacks specificity for in-depth modeling of different perspectives. This paper proposes a flow-based modeling technique as a foundation enterprise architecture. This approach is illustrated by use of a case study from the ArchiMate literature.

**Keywords:** Enterprise architecture · Conceptual model · Diagram

## 1 Introduction

Experience with IT utilization over the last several decades has led to the evolution of concepts related to IT Governance. IT Governance refers to the framework of IT process and service requirements that are important for achieving an organization's goals. Enterprise Architecture (EA) is a major approach that presents principles, methods, and models for design and realization of organizational structure, business processes, information systems, and infrastructure of an organization. *Architecture* here refers to logical constructs used in representing and interpreting things and their behavior, e.g., in programming, logical constructs include sequence, selection, and iteration. A general definition of *architecture* is given in IEEE 1471-2000/ISO/IEC 42010:2007 [1] as "the fundamental organization of a system embodied in its components, their relationships to each other, and to the environment, and the principle guiding its design and evolution."

Architecture … is concerned with understanding and defining the relationship between the users of the system and the system being designed itself. Based on a thorough understanding of this relationship, the architect defines and refines the essence of the system, i.e., its structure, behavior, and other properties [2].

EA has recently become dominant among practitioners [3]. It provides a gestalt view of an organization's concepts and artifacts. It is an important instrument for addressing company-wide integration through use of a coherent total representation that is used throughout the process of designing an organizational structure, business processes, information systems, and IT infrastructure.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_8

A good architecture practice enables an organization to align business and IT operations with its strategy, quickly respond to changes in the environment, and make optimal use of technological opportunities. The development and maintenance of architectures will lead to efficiency, cost reduction and flexibility [4].

An organization is modeled as connected layers, where each of these layers produces artifacts, e.g., assets, people, projects [5]. Accordingly, an architecture description is specified for such uses as expression of the system, analysis of alternative architectures, business planning, communications, production, documentation, operation, and maintenance of a system and negotiations [1].

In current practice, architecture descriptions are heterogeneous in nature: each domain has its own description techniques, either textual or graphical, either informal or with a precise meaning. Different fields speak their own languages, draw their own models, and use their own techniques and tools. Communication and decision making across these domains is seriously impaired [2].

According to Lankhorst et al. [2],

> Important for an architecture description language is that the properties of the system can be represented in their bare essence without forcing the architect to include irrelevant detail. This means that the description language must be defined at the appropriate abstraction level… The language and methods are the basis for unambiguous mutual understanding and successful collaboration between the stakeholders of the architecture.

This has required more "coarse-grained modelling concepts than the finer grained concepts that can typically be found in modelling languages used at the level of specific development projects, such as e.g. UML (OMG, 2009) and BPMN (OMG, 2008b)… Therefore a new language was needed, leading to the development of ArchiMate" [6].

### 1.1 ArchiMate

ArchiMate is a modeling language developed to provide a uniform representation for EA. It is supported by an array of vendors and service providers. "Many organizations are using it already as their company standard for describing enterprise architecture, and its value has been proven in practice" [2]. It describes and visualizes the different architecture domains and their underlying relationships and dependencies [7].

The ArchiMate Enterprise Architecture modeling language provides a uniform representation for diagrams that describe Enterprise Architectures. It includes concepts for specifying inter-related architectures, specific viewpoints for selected stakeholders, and language customization mechanisms. It offers an integrated architectural approach that describes and visualizes different architecture domains and their underlying relations and dependencies. Its language framework provides a structuring mechanism for architecture domains, layers, and aspects [7].

ArchiMate defines three layers: business, application, and technology, based on specializations of its core concepts in terms of relationships:

- "used by" relationships show how the higher layers make use of the services of lower layers.
- realization relationships: elements in lower layers may realize comparable elements in higher layers.

### 1.2 Research Problem

According to Kinderen et al. [8],

> Because of the inherent holistic nature, ArchiMate lacks specificity on how to model the different perspectives in-depth. For example, ArchiMate lacks guide-lines for process modelling, and lacks expressivity for modelling an enterprise from a value exchange perspective [9]. Moreover, as claimed by [10] ArchiMate lacks conceptual clarity and precision. This "lack" is, however, a direct consequence of the coarse-grained, and holistic, nature of ArchiMate. In that sense this freedom of interpretation has been designed into the language on purpose [11].

Kinderen et al. [8] proposed using the DEMO modelling technique and toolset as a front-end for ArchiMate.

In this paper, we propose an alternative solution: to use a Flowthing Machines (FM) modelling technique as a foundation EA. This approach is illustrated by a case study drawn from the ArchiMate literature. FM has been utilized in several software engineering applications [12–17], and the next section briefly reviews some of its features. The example given in Sect. 2 is a new contribution. Section 3 applies FM to model EA and contrasts it with ArchiMate.

## 2 Flowthing Model

A flowthing machine is a diagrammatic abstraction of a system from the perspective that *things* flow through five stages: creation, processing, receiving, release, and transfer (Fig. 1) that are used to support modeling. The basic machine is a generalization of the classical stages of input, process, and output. A machine triggers other machines, denoted by dashed arrows. A sphere is the environment of a flow.

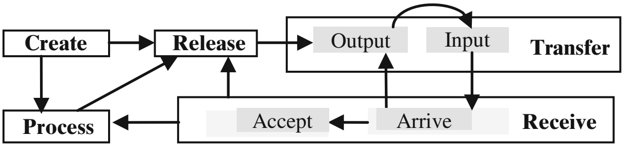


**Fig. 1.** Flowthing machine.

**Example:** According to Lankhorst et al. [2], representation in Business Process Model and Notation (BPMN) is restricted to process modelling. It provides a uniform notation for modelling business processes in terms of activities and their relationships, as shown in the example of insurance claim processing given in Fig. 2. Figure 3 shows the corresponding FM representation.

In Fig. 3, in the sphere of *client* (circle 1 in the figure), the *claim* is created (2) and flows (3) to the *insurer* (4). There, it is processed (5) and, if accepted, is evaluated (6) to trigger (7) the issuing *of payment* (8) that flows to the client (9).

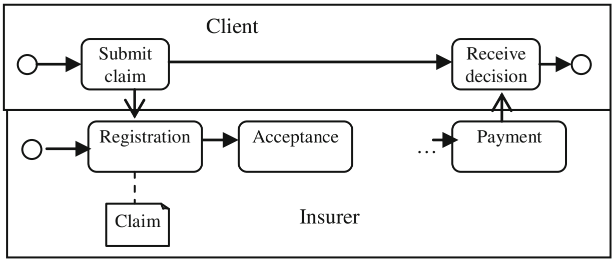


**Fig. 2.** BPMN example (redrawn, partial from [2])

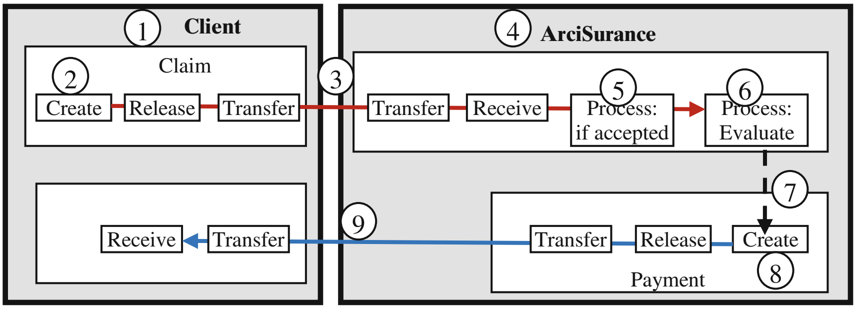


**Fig. 3.** FM representation of the example

# 3 FM as an EA Description

According to Lankhorst et al. [2], different architectures (e.g., information, product, process, application, technical architectures) might be represented in different models, e.g., applications in UML and business processes in BPMN. In these cases, it is unclear how concepts in one view are related to concepts in another view and whether views are compatible with each other [2].

When we talk about the integration of architectural domains, we need a language in which we can describe these domains. For example, some sources refer to entities and relations, as in entity-relationship diagrams. Others refer to classes and objects, like in object-oriented modelling and software engineering. And yet others refer to concepts and instances; for example, in the area of conceptual modelling [2].

In ArchiMate, as a unifying framework, a general structure is adopted within three different layers: business layer, application layer, and technology layer.

## 3.1 Business Layer

The Open Group [18] gives an example of “integration from the technology layer…, via the application layer…, all the way up to the business layer.” The example involves a “client who wants to register an insurance claim and the business process that provides the necessary services.” Figure 4 shows the business layer of this example.

Notice the multiplicity of notations at different levels, a reminder of the UML diagramming method. Figure 4 gives a sense of heterogeneous ontology like a portion of English text that embraces structural terms–words, sentences, and paragraphs–in addition to the text itself. The only way to illustrate this “uneasiness” of the representation is to translate it into FM and contrast the two descriptions.

Figure 5 shows the FM model of the partial view shown in Fig. 4.

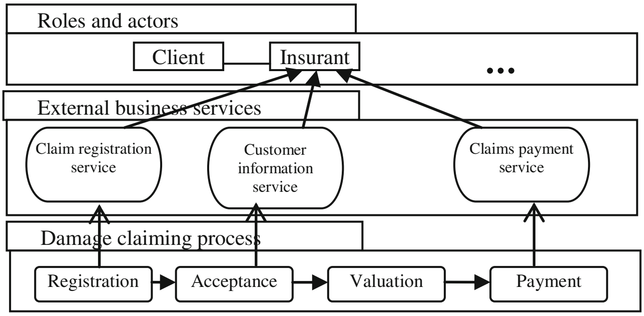


**Fig. 4.** Business layer in ArchiMate for the given example (partial, redrawn from [18])

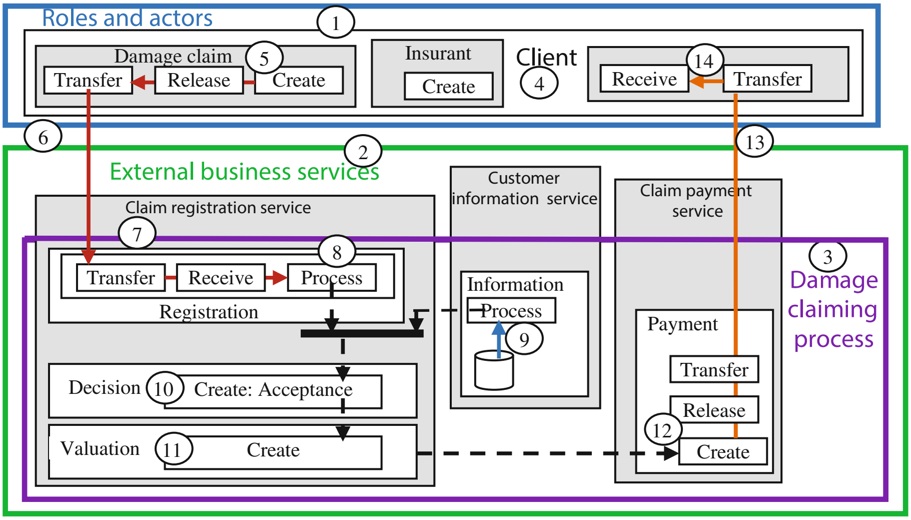


**Fig. 5.** FM representation of the example shown in Fig. 4.

The FM representation keeps all terms mentioned in the original example. It includes three spheres: Roles and actors (circle 1 in the figure), External business services, and Damage claiming process. In the sphere of Roles and actors, the client (4) creates a damage claim (5) that flows (6) to the Claim registration service (7) where it is processed (8). This process along with data coming from the Customer information service (9) triggers the decision to accept the claim (10). This decision triggers the creation of valuation (11) which in turn triggers the creation of payment (12) that flows (13) to the client (14).

It might be observed that the FM representation is too "complex" to be utilized as an architecture description, but FM can be abstracted at various levels of granularity and complexity. The interesting aspect of FM description is the systematic application of the five stages of flow. This application creates more complete specifications by showing the interconnectedness of the flow of things; however, it is possible to simplify and customize the depiction by reducing the level of description. Such simplification is shown in Fig. 6, where the stages of the flow are removed and the remaining diagram shows the Roles and agent, External business services, and Damage claiming process layers of ArchiMate. With few realignments of the construct in Fig. 6, the basic structure of ArchiMate's Fig. 4 can be reached. Hence, a model based on ArchiMate *is contained* in its FM model. FM is mappable to ArchiMate.

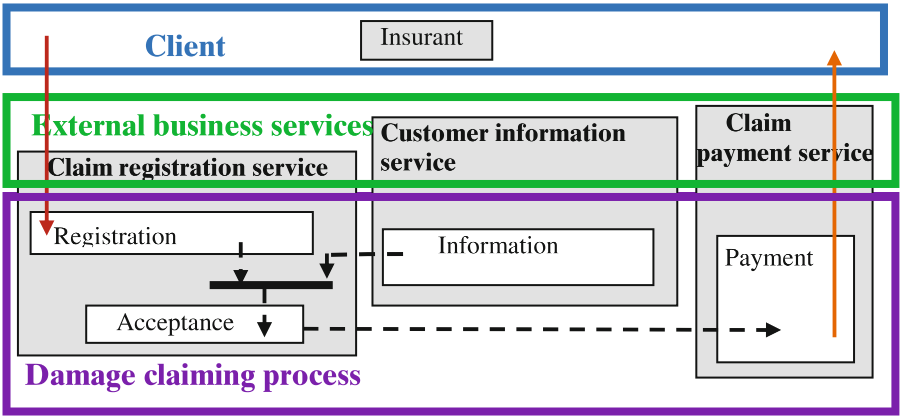


**Fig. 6.** Coarse representation of Fig. 4.

Is the FM representation shown in Fig. 5 complex in comparison with such diagrams as construction blueprints or engineering drawings? In FM, the complexity issue is reduced through use of simple notions (five stages of flow, triggering mechanism, and Venn diagram-like structure of spheres) that are applied, repeatedly, across macro- and micro-levels of detail. The number of *distinct constructs* has been closely linked to measures of simplicity [19] (in the source terminology, "distinct variables"); accordingly, as a language, FM is simpler than ArchiMate.

Complexity is related to *completeness* of the description. Consider three views of a car engine:

- The first view is of the engine as a single inexplicable unit. No explanation is required.
- The second is a view involving a partial knowledge of the engine. The parts are roughly identified as well as some of the interactions.

- The third view is that of mechanics. They have an understanding of the decomposition of the engine into functionally near-independent parts.

The engine as manageable complex, due to the appropriateness and utility of the mechanic's model of it [20].

We claim that production of an FM representation is a prerequisite to any simplification of the EA model to guarantee a complete description.

## 3.2 Application Layer

Continuing with the example of insurance claims given in [18], Fig. 7 shows the application layer of the example. Note that *application* is typically refers to a logical grouping of software components, e.g., Desktop Application: MS Office [21]. Because of space limitations, the focus is on the *application policy* in this layer, shown as rendered in ArchiMate in Fig. 8 and described in the Open Group document "An Introduction to the ArchiMate 2 Modeling Language" [22].

Figure 9 shows the FM model that includes the business layer (upper part of diagram, from Fig. 5), with the application layer limited to the application policy (lower part).

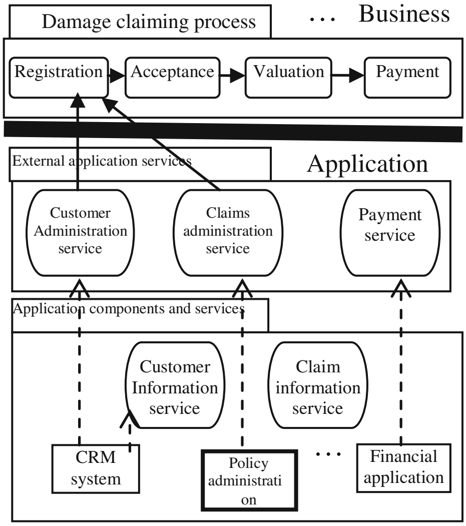

Damage claiming process
...
Business
Registration
Acceptance
Valuation
Payment
External application services
Application
Customer Administration service
Claims administration service
Payment service
Application components and services
Customer Information service
Claim information service
CRM system
Policy administrati on
...
Financial application

**Fig. 7.** Application layer in ArchiMate for the given example (partial, redrawn from [18]).

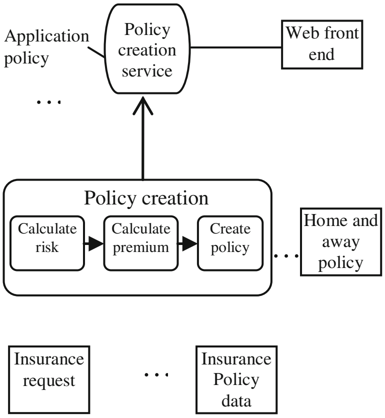


**Fig. 8.** Application policy in ArchiMate for the given example (partial, redrawn from [18]).

In External Application services (circle 15), in the Web front end (16), the application for a policy (17) is downloaded to flow to the client (18). When it is received (19 – upper corner of the figure), it triggers the creation of a completed application (20) that flows to the Web front end (21–22). Policy administration (23) receives the filled

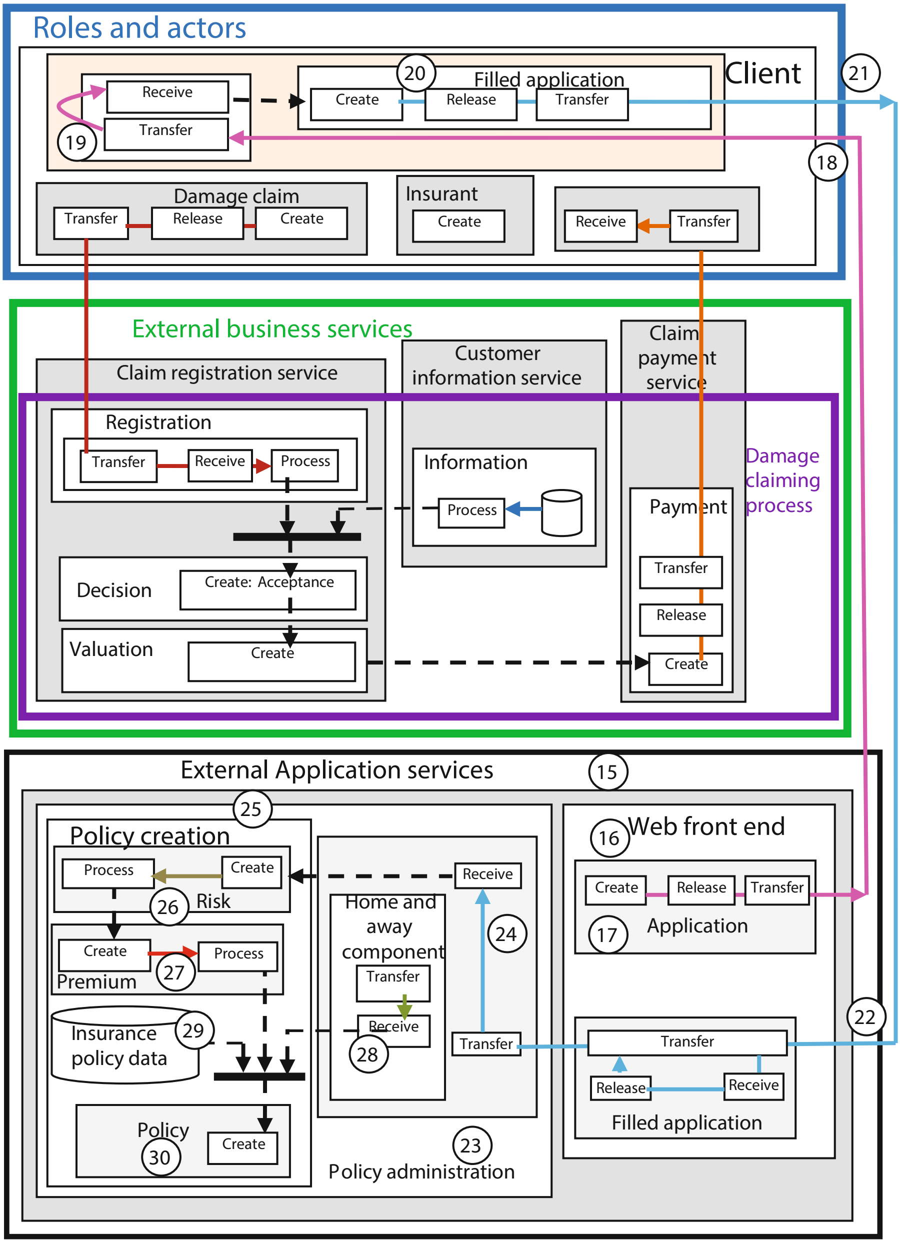


**Fig. 9.** FM representation of part of the example that includes the application policy in the application layer

application (24), triggering creation of a policy (25). Creating a policy starts by calculation of risk (26) and then the premium (27), and depending on these, data about the Home and away component (28 – part of the filled application) and Insurance policy data (29) are also factored in, leading to creation of the policy (30).

Note how using FM to show the two layers results in an integrated description that can be marked, if desired, as two spheres: the business sphere and the application sphere. The description can also be simplified to the level of the ArchiMate depiction.

### 3.3 Application Layer

We will not elaborate on this layer but present a sample segment of a general view, shown in Fig. 10. In the figure, in External infrastructure services (circle 31), File access services (32) requests access to files (33) and returns the results (34).

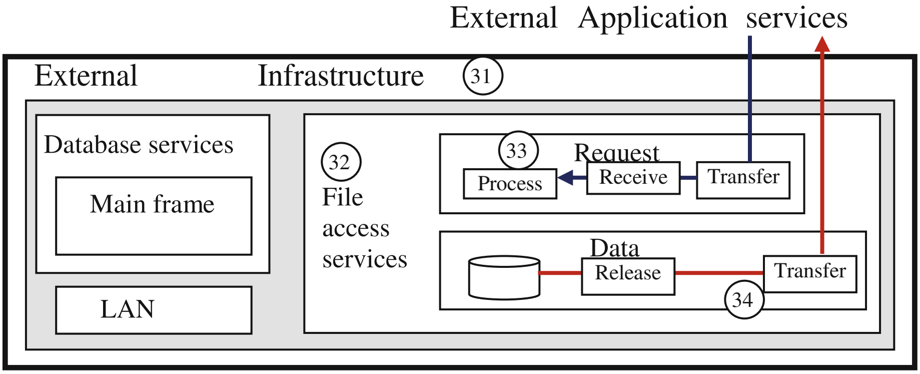


**Fig. 10.** FM representation of part of the example that includes components in the technology layer.

## 4 Conclusions

This paper proposes using the Flowthing Machine modelling technique as a foundation enterprise architecture. This approach is illustrated by a case study from the ArchiMate literature. The examples point to the feasibility of FM in the context of enterprise architecture descriptions as a foundation for a more simplified depiction of the architecture. More study is required to substantiate the appropriateness of FM for this domain. Future research will examine more examples from ArchiMate and similar modeling languages.

## References

1. IEEE Computer Society. IEEE Std 1471-2000: IEEE Recommended Practice for Architecture Description of Software-Intensive Systems. IEEE, New York (2000)
2. Lankhorst, M., et al.: Enterprise Architecture at Work: Modelling, Communication and Analysis, 2nd edn. Springer, Berlin (2009). ISBN: 978-3-642-01309-6

3. Vicente, M.: Information Systems and Computer Engineering, MS thesis, TECNICO Lisboa
4. The Open Group. 2016. ArchiMate Example (2013). http://www.archimate.nl/en/about_archimate/example.html
5. Gama, N., Sousa, P., da Silva, M.M.: Integrating enterprise architecture and IT service management. In: 21st International Conference on Information Systems Development (ISD 2012), Prado, Italy (2012)
6. Lankhorst, M.M., Proper, H.A., Jonkers, H.: The anatomy of the ArchiMate® language. Int. J. Inform. Syst. Model. Des. (IJISMD) **1**, 1–32 (2010)
7. The Open Group. ArchiMate 3.0 Specification (2016). http://pubs.opengroup.org/architecture/archimate3-doc/
8. Kinderen, S.D., Gaaloul, K., Alex, H., Proper, H.A.: Transforming transaction models into ArchiMate. In: CAiSE Forum, pp. 114–121 (2012)
9. Pijpers, V., Gordijn, J., Akkermans, H.: $e^3$ alignment: exploring inter-organizational alignment in networked value constellations. Int. J. Comput. Sci. Appl. **6**(5), 59–88 (2009)
10. Ettema, R., Dietz, J.L.G.: ArchiMate and DEMO – mates to date? Adv. Enterp. Eng. **3**, 172–186 (2009)
11. Lankhorst, M.M., et al.: Enterprise Architecture at Work: Modelling, Communication and Analysis. Springer, Berlin (2005)
12. Al-Fedaghi, S.: Conceptual modeling in simulation: a representation that assimilates events. Int. J. Adv. Comput. Sci. Appl. **7**(10), 281–289 (2016)
13. Al-Fedaghi, S.: Function-behavior-structure model of design: an alternative approach. Int. J. Adv. Comput. Sci. Appl. **7**(7), 133–139 (2016)
14. Al-Fedaghi, S.: Context-aware software systems: toward a diagrammatic modeling foundation. J. Theor. Appl. Inf. Technol. **95**(4), 936–947 (2017)
15. Al-Fedaghi, S.: Flow-base provenance. Inf. Sci. **20**, 19–36 (2017)
16. Al-Fedaghi, S.: Conceptualization of various and conflicting notions of information. Inf. Sci. J. **17**, 295–308 (2014)
17. Al-Fedaghi, S.: System for a passenger-friendly airport: an alternative approach to high-level requirements specification. Int. J. Control Autom. **7**(2), 427–438 (2014)
18. The Open Group. ArchiMate Example, ArchiMate.com. http://www.archimate.nl/en/home/. Accessed Nov 2012
19. Kemeny, J.G.: Two measures of complexity. J. Philos. **52**, 722–733 (1953)
20. Edmonds, B.: Syntactic Measures of Complexity. Ph.D. thesis, Department of Philosophy, University of Manchester (1999)
21. Betz, C.T.: Architecture and Patterns for IT Service Management, Resource Planning, and Governance: Making Shoes for the Cobbler's Children. Elsevier (2011). ISBN: 0123850185, 9780123850188
22. The Open Group. An Introduction to the ArchiMate® 2 Modeling Language (2013). http://www.opengroup.org/archimate

# Fissure Extraction Using Dual Tree Complex Wavelet Transform and Lung Lobe Segmentation from CT Lung Images

M. Jannathl Firdouse[1(✉)] and M. Balasubramanian[2]

[1] R&D Department, Bharathiyar University, Coimbatore, India
jfirdouse@gmail.com

[2] Department of CSE, Annamalai University, Chidambaram, India
balu.junel@gmail.com

**Abstract.** The lungs play a very vital role in human respiratory system. It has five separate lobes which are detached by fissures of three types such as left and right oblique fissure and a horizontal fissure. The way of identifying the fissure lobes in computed tomography scanned lung images are difficult for the medical practitioners because of the incorrect shapes alongside with less contrast and the extraordinary noise associated with it [1]. The last phase of the lung cancer treatment is the elimination of the unhealthy lung by the major surgery. So, it is required to identify the location of the cancer affected part of the lungs by extracting the fissure lobes before making the proposal for the surgery. This paper presents a mechanized process of extracting the left oblique fissures and right oblique fissures by applying the Dual Tree Complex Wavelet Transform from the Computed Tomography lung images. This will help the medical practitioner to identify the lobar fissures from the computed tomography lung images.

**Keywords:** Oblique fissure · Horizontal fissure · DTCWT · Fissure lobes · Discrete wavelet transform · Filter bank and fissure sweep

## 1 Introduction and Literature Survey

Human lungs are having five lobes which are parted by visceral pleura which are known as pulmonary fissure. The right lung comprises of three lobes such as upper, middle and lower. The right minor fissure divides right upper and middle lobes, whereas the right major fissure bounds the lower lobe from the rest of the lung. Because of the incomplete fissures and anatomical variations, the segmentation of pulmonary lobes is tedious. The framework of a human lung is shown in Fig. 1.

For the experienced medical practitioner, the recognition of the fissures from the CT image is hard because of image's different shape beside with low dissimilarity and more noise along with it. In order to do the surgery of lung removal, it is essential to recognize the site by eliminating the lobar fissures before the start of surgical procedure. The lung cancer has beaten the breast cancer as it was taken as the important reason of demises in females due to cancer [2]. The cells form a tumor which is different from the surrounding tissues. The analysis process of lump is created on

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_9

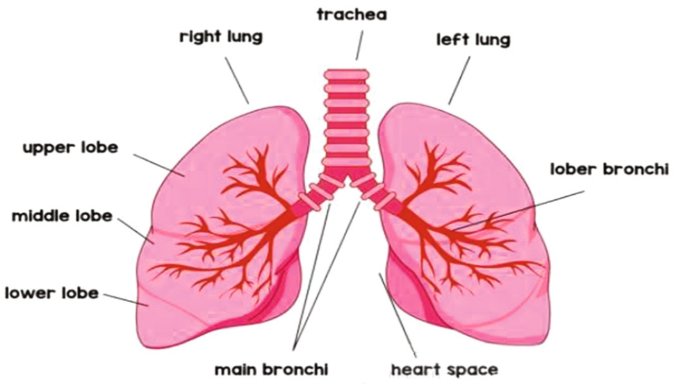


**Fig. 1.** Anatomy of human lung

whether the respiratory nodule is usual or tumorous. It can be identified by inspecting the growing rate of nodule. The tumorous nodule become binary in size on a middling of every quarter year and the usual nodule do not grow much at all. Another way of differentiating the cancerous nodule from the normal nodule is by examining the size and the surrounding surface. Irregular shapes, lumpy surface and color variations are the identification marks of cancerous nodule. Whereas, the normal nodules are regular in shape, smoother and the color is evenly distributed. Mostly, CT scanned images are used for the effective diagnosis.

The CT slice of lung image has three fissures such as right horizontal fissure, right and left oblique fissure. The medical practitioners check the stack of two dimensional CT lung images to identify the diseased lung for the surgical planning. This will take long time to decide and start the surgical procedures. We propose the concept of extracting the boundaries of the lung lobe fissures by the DTCWT to reduce the surgical planning time. Three phases are in the proposed method and are implemented as explained below. The region of fissure is identified in the first phase. The lobar fissures are identified and found oblique fissures are extracted in the second phase. The horizontal fissure is identified and extracted in the third phase.

## 2 Proposed Work and Methodology

The isotropic CT images are preprocessed to remove the unwanted noises present in the input images. The noise removal is done by using the mean filter. The filter size of $3 \times 3$ matrix is preferred. Each input point is replaced by the mean of the neighborhood points. The noise in CT input image uses the Gaussian distribution of the mean filter. This only balances the noise elimination and over hiding of the images. For finding the fissures in the isotropic CT image, the adaptive fissure sweep is employed. The lung section is divided from the context [3] by analyzing the histogram and the connected component labeling. The flow chart of fissure extraction is given in Fig. 2.

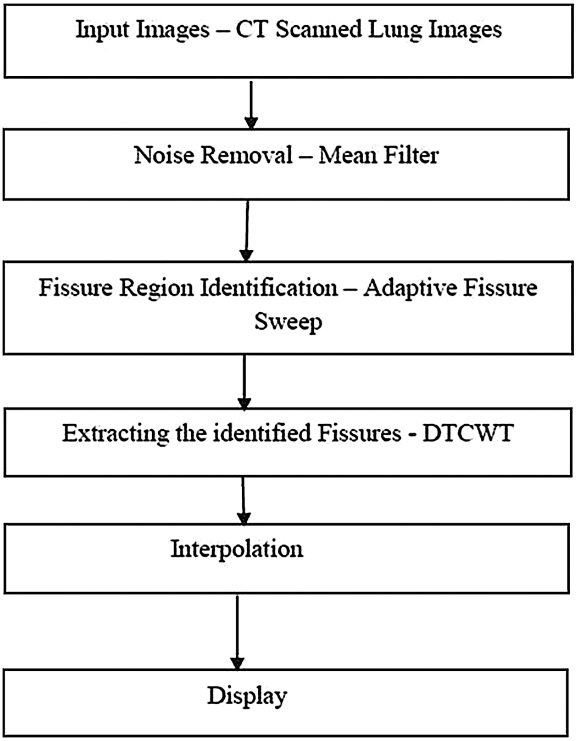

Input Images – CT Scanned Lung Images
Noise Removal – Mean Filter
Fissure Region Identification – Adaptive Fissure Sweep
Extracting the identified Fissures - DTCWT
Interpolation
Display

**Fig. 2.** Flowchart of fissure extraction

### 2.1 Noise Removal – Mean Filter [7]

The process before the adaptive fissure is preprocessing the input images in demand to remove the surplus noise present in the input image. The additional noise in the input image is removed by the mean filter. This will do the primary noise removal from the CT input image. The filter size of 3 × 3 is employed to remove the noise. The mean filter replaces each pixel of the input image by the weighted average of the neighbourhood pixels.

### 2.2 Fissure Region Identification Using Adaptive Fissure Sweep

The adaptive fissure sweep is the main step of the lobe segmentation process. This is used to find the fissure sections from the isotropic CT scanned lung images after attainment of input images. The lung area is partitioned from the background by using the histogram inquiry and the connected component labeling. The lung segmentation also involves the region growing, morphological operations and watershed algorithms etc. [4]. The removal of the fat and muscles surrounding the lungs is implemented by choosing the threshold value Tr which is based on the histogram.

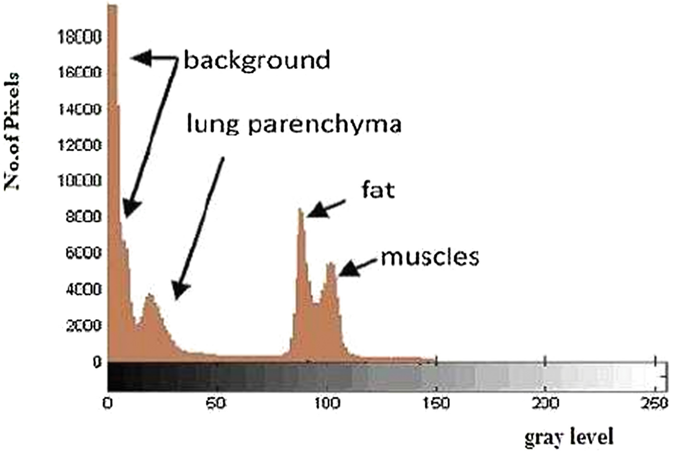


**Fig. 3.** CT lung image histogram

The Tr is calculated by the equation

$$\mathrm{Tr} = \frac{\mathrm{I}_{\mathrm{FM}-\mathrm{I}_{\mathrm{BL}}}}{2} + \mathrm{I}_{\mathrm{BL}} \tag{1}$$

where $\mathrm{I}_{\mathrm{FM}}$ is the mean pixel strength values of top analogous to the fat or muscles and is the mean pixel strength value of heights matching to the background or lung parenchyma respectively. The histogram of a computed tomography lung image [8] is shown in Fig. 3.

The two lungs are extracted by performing the bounding box and connected component labelling by the specified algorithm. Due to the presence of the bronchial and vascular tree, the extracted lung boundaries are irregular. By smearing the circular morphological closing operator the above said problem is rectified [4]. To smooth the lung boundaries, the filter size of 10 × 10 pixel is applied to collect the original shape of the lung. On each segmented lungs, the adaptive fissure sweep is achieved. This locates the fissure section within the partitioned lungs. The borders of the lung lobes are also identified. The morphological dilation operator is performed to enhance the vascular and bronchial trees [5]. These steps permit an enhanced framework for discovering the fissure section in the isotropic CT images.

## 2.3 Extracting Identified Fissures – DTCWT

The DTCWT computed the compound transform [5] using two separate Discrete Wavelet Transform decompositions of tree a and tree b which is presented in Fig. 4. DTCWT is implemented to remove the identified fissures from the isotropic CT images. If the filters used are different from each other, it is possible to have one DWT to create the real coefficient and the other one is considered as imaginary.

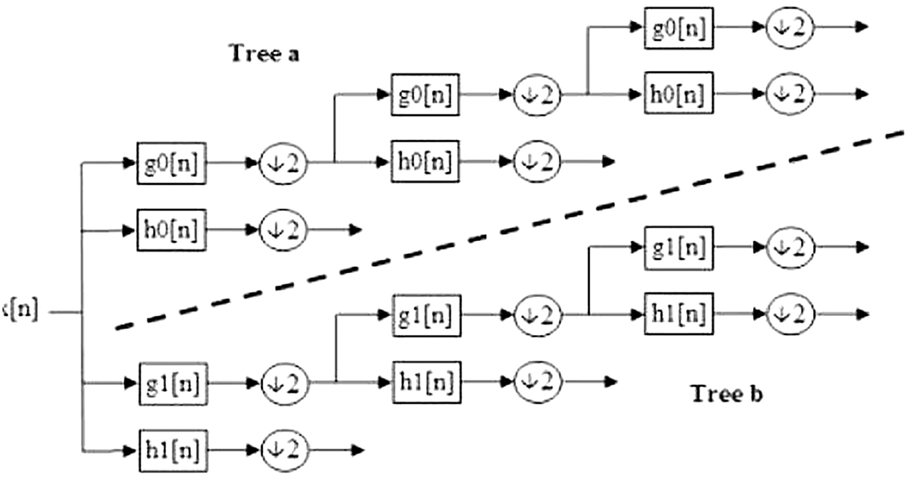


**Fig. 4.** Filter structure of DTCWT

The h0 is the real valued high pass and h1 is the real valued low pass filters respectively. Similarly, g0 and g1 is the imaginary tree. The important characteristics to be followed while designing the filters are:

i. The two trees are differing by half of the sample period in the low pass filter.
ii. Reconstructing the filters is the reverse process of analysis.
iii. Tree “a” filters are considered as the opposite of tree “b” filters.
iv. All filters are from the same orthonormal set.
v. Both the trees must have the same frequency response.

The static DWT uses a low pass and high pass filter to spoil the input image simultaneously. This will leads the formation of detailed coefficients of an input image. To convert the rows and columns, a 2D-DTCWT involves one dimensional DTCWT. The 2D – DTCWT [8] gives four sub images entails of three high pass filtered images such as: horizontal, vertical and diagonal and the low pass version of the original image, unlike the stationary 2D conventional DWT. This space invariance property [6] allows the procedure to find the fissure position and the curve using the detailed coefficient of the image.

Most of the lobar fissures look as if horizontally across the fissure sections so that the horizontal aspect of the sub image is used for the advance analysis. This is due to find the adaptive fissure sweep that familiarizes with the fissure sections which is along the fissure directions. The longest continuous lines crossing the fissure region is found by applying fissure search technique in the lobe segmentation algorithm [9]. This algorithm bearings point by point analysis and employing the anchor points automatically at remoteness of 5 points apart for identifying the fissures. The current fissure anchor points are compared with their matching part on a previous adjacent fissure. In two adjacent isotropic CT images, the fissure changes are very small. To define a precise fissure, the following criteria is used

$$\frac{1}{M}\sum\nolimits_{M}^{j-1} Z_{j,}\ 1 - Z_{j,}\ 2 \leq 3 \text{ pixels and } \mathrm{Z_j}\ 2 - Z_{j,}\ 1 \leq 9 \text{ pixels} \tag{2}$$

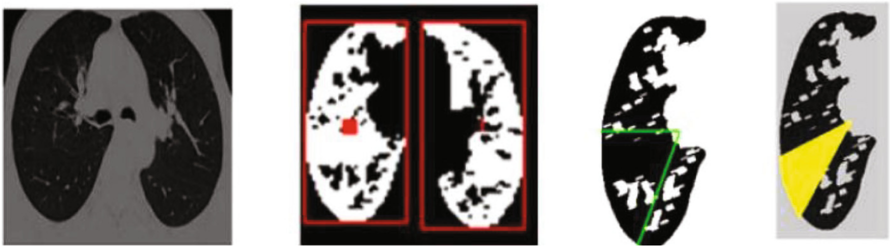
M represents the amount of anchor points used for a fissure and is the z-coordinate of the jth anchor point. The fissures lies between the adjacent CT images tend to change in the vertical direction. Hence z coordinate is used instead of Euclidian distance. The last three CT slices are reflected by applying the anchor points of this fissure to guide the fissure search in the next adjacent slice. This only finds the correct fissure. The identified fissures are discontinuous. So, the linear interpolation finds the continuous fissures. The average angle is given by

$$\Phi\text{average} = (\phi 1 + \phi 2 + \phi 3) \tag{3}$$

where Φj (j = 1, 2, 3) denotes the angle of the fissure segment between two adjacent anchor points. So the left and right oblique fissures are removed from the isotropic input CT images by this algorithm.

## 3 Results and Discussion

The slice of CT input image from the online database in Harvard University is shown in Fig. 5(a) and before the left and right lung segmentation by the bounding box, the result of connected component labeling is shown in Fig. 5(b). The outcome of fissure sweep and recognized fissure section in right lung is shown in Fig. 5(c) and (d).



(a) (b) (c) (d)

**Fig. 5.** (a) Original image (b) bounding box (c) fissure sweep (d) fissure region

The sequences of outputs from the fissure extraction algorithm after applying DTCWT are shown in Fig. 6(a)–(d).

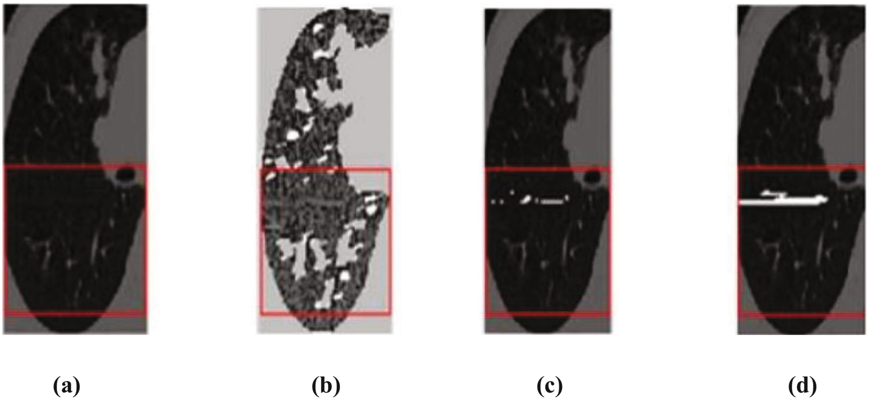
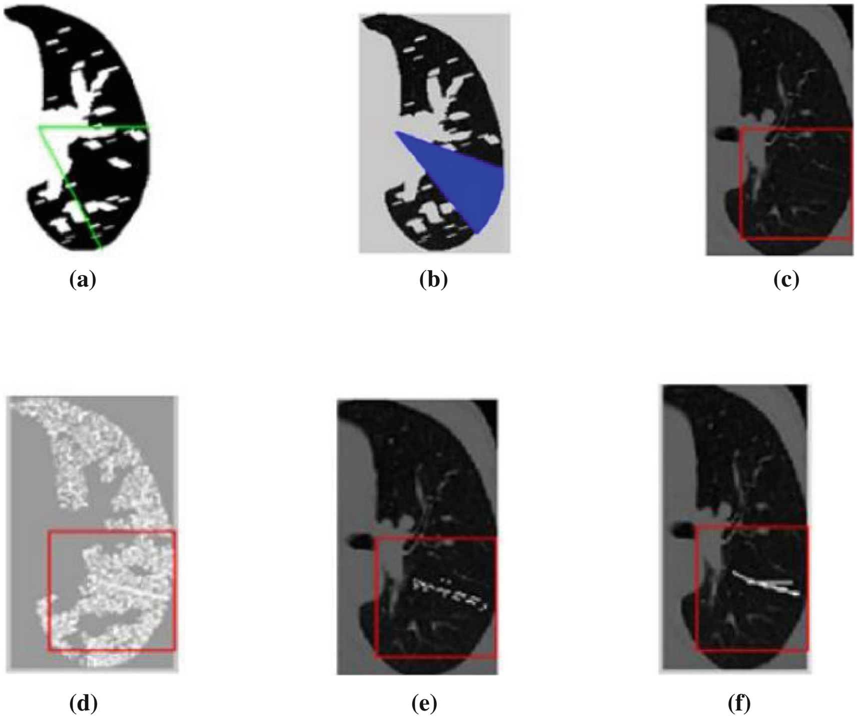

(a) (b) (c) (d)

**Fig. 6.** (a) Identified fissure (b) enhanced fissure (c) right oblique fissure (before interpolation) (d) right oblique fissure (after interpolation)

The sequences of outputs from the left lung are shown in Fig. 7(a)–(f).


(a) (b) (c)

(d) (e) (f)

**Fig. 7.** (a) Fissure sweep (b) fissure region (c) detected fissure (d) enhanced fissure (e) left oblique fissure (before Interpolation) (f) right oblique fissure (after interpolation)

The peak signal to noise ratio analysis of lung images from the LOLA database by using two methods are given in the Table 1. The analysis is performed with five lung images. From the table, it is identified that the proposed method has the highest PSNR ratio and the comparison of these three methods indicates that the proposed method is efficient and speedy. The lung images are taken from the LOLA database.

**Table 1.** PSNR Comparison with three methods

| Name of the image | PSNR ratio | | |
|---|---|---|---|
| | FLICM | FBEA-SPQT | Proposed method |
| Limgimagel | 64.73 | 68.34 | 71.94 |
| Lutigiinage2 | 68.55 | 72.65 | 74.75 |
| Littigimage3 | 64.27 | 67.26 | 69.54 |
| Lmigimage4 | 66.81 | 69.94 | 72.26 |
| LungimageS | 67.11 | 71.33 | 73.82 |

The PSNR analysis charts of different lung images with three methods are shown in Fig. 8.

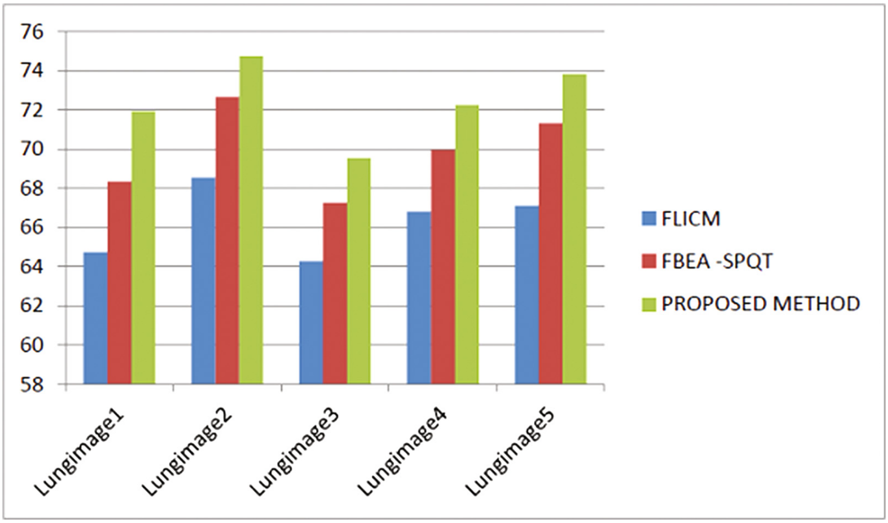


**Fig. 8.** PSNR analysis chart

## 4 Future Work and Conclusion

The lung lobe segmented by this algorithm is useful to identify the location of affected regions of CT lung image and helpful for the medical practitioner to identify the proper location of the affected lungs for the surgery. The proposed method is more accurate and speedy compared to the existing method. The horizontal fissures are identified with affected regions. The PSNR ratio of this proposed method is effective compared to the existing method. The advanced concept is applied in order to derive the accurate results of horizontal fissure in the future. The defined three phases are executed and the results are obtained with more accuracy. The concept consumes very less time to get the required results. This result is further used for the major surgery.

## References

1. Bharathi, N., Manikandan, T.: Lobar fissure extraction in ct lung image – an application to cancer identification. Int. J. Comput. Appl. **33**(6), 0975–8887 (2011)
2. Emedicinehealth. http://www.emedicinehealth.com/lungcancer/article_emt.htlunglobes
3. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision, 3rd edn. Thomson Learning, Boston (2008)
4. Anitha, S., Sridher, S.: Segmentation of lung lobes and nodules in CT images. Int. J. Sig. Process. Image Process. (SIPIJ), **1**, 1–12 (2010)
5. Haralick, R.M., Stenberg, S.R., Zhuang, X.: Image analysis using mathematical morphology. IEEE Trans. Pattern Anal. Mach. Intell. **4**, 532–550 (2007)
6. Kumar, S.N., Kavitha, V.: Automatic segmentation of lung lobes and fissure for surgical planning. In: Proceedings of ICETECT, pp. 546–550 (2011)
7. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall
8. https://www.google.com/search?q=Histgram+of+a+CT+Lung+image&tbm=isch&tbs=rimg
9. Bhadarani, A., Yu, R.: A Dual Tree Complex Wavelet with application in image denoising. In: IEEE International conference on Signal Processing and Communications, pp. 1203–1206 (2007)

# ASRD: Algorithm for Spliced Region Detection in Digital Image Forensics

A. Meenakshi Sundaram[1(✉)] and C. Nandini[2]

[1] School of Computing and Information Technology, REVA University,
Bangalore, India
meenakshisnarayana@gmail.com
[2] Department of Computer Science and Engineering,
Dayanand Sagar Academy of Technology and Management, Bangalore, India

**Abstract.** Image splicing is one of the most frequently exercised in the area of image forgery that is quite challenging to be identified. After reviewing existing techniques towards identification of spliced region, it was found that existing techniques are either computationally expensive or do not address the cumulative problem. Hence, this paper, a novel and simple algorithm is presented called as ASRD i.e. Algorithm for Spliced Region Detection. A simple statistical-based approach is presented that perform partitioned blocks followed by detection of various artifacts among the neighbor blocks. The algorithm then implicates a classification condition for tampered and non-tampered region to truly identify the spliced region. With an aid of histogram analysis, true positive score, true negative score, accuracy and computational performance, the proposed algorithm was found to excel better performance in detection of spliced region.

**Keywords:** Image splicing · Image forensics · Color filter array · Localization · Accuracy

## 1 Introduction

The utilization of image is consistently increasing in present era on multiple area e.g. domestic appliances, educational requirements, social networking, evidences in court cases, multimedia sharing, etc. There are various enterprise applications where image plays one of the most critical roles and the enterprise can suffer a collateral loss if such images or its contents are compromised by any means [1]. With an availability of modern image editing tools, it is now possible to re-create fake or forged image that area quite challenging to be identified as a real [2–4]. This phenomenon is called as image forgery [5]. Although, there are various types of attacks over an image, but splicing image is one of the frequently adopted practices to forge the image in social network or malicious spreading of fake propaganda. It merged two or more images or objects in one scene in such a way the final product looks unbelievably true. The fake image of Osama Bin Laden by Sun, UK Mail, and Telegraph is one live example of image splicing [6, 7]. Blitzer [8] has illustrated the underlying principle of image forensics where vivid description of many attacks can be studied. The core idea behind

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_10

image forensic is to tamper and manipulate the image and give a new look into the image in such a way that it is not feasible to be identified as forged [9, 10]. While doing so an attacker has the equal chances of getting themselves caught. Hence, it is necessary to remove all the possible traces by the attacker in order to increase more imperceptibility towards splicing. This paper introduces on such algorithm in order to identify the position compromised by image splicing attack. Section 2 discusses about the recent research work towards image splicing accompanied by briefing of problem identification in Sect. 3. The proposed algorithm is briefed in Sect. 4 followed by algorithm description in Sect. 5. Comparative analysis of performance is discussed in brief in Sect. 6 and finally the contribution of the paper is highlighted in Sect. 7.

## 2 Related Work

Our prior work has reviewed about the effectiveness of the existing techniques towards enhancing detection of image forgeries [11–13]. Our prior techniques has introduced techniques towards copy-move attack and retouching attacks [14, 15]. In this part of the study, our interest orients towards image splicing attacks. The significant work towards splicing attack was carried out by Cozzolino et al. [16] by deploying expectation-minimization algorithm and feature extraction carried out by statistical approach. Adoption of real-world image was seen in the work of Zampoglou et al. [17] towards image splicing detection. Various optimization techniques e.g. singular vector decomposition, discrete cosine transformation, support vector machine etc. has been found to be adopted in various research work like that of Amerini et al. [18] and Moghaddasi et al. [19].

The focus of such techniques was mainly to find the exact position of the spliced region. Markov modeling is another technique for feature extraction to be used for identifying the underlying image structure by Su et al. [20]. Usage of Markov model for spliced region detection in colored image was seen in the work of Han et al. [21]. Local binary pattern is also used for feature extraction for identifying the splicing features found in the work carried out by Zhang et al. [22]. Usage of local descriptors for spliced region detection was also witnessed in work of Saleh et al. [23]. Similar type of scheme was also used by Zhao et al. [24] where the abnormal channels are explored from color channels considering chrominance feature. Usage of probability matrix along with the statistical approach was also found to be providing more information about the spliced region. Application of descriptive statistical measures e.g. skewness and kurtosis was seen in the work of Pan et al. [25] for spliced region detection. Adoption of cryptographic encryption was seen in the study of Niu et al. [26] in order to resist statistical attack. The authors have also used quaternary coding and chaos theory. The next section outlines the problems in existing system.

## 3 Problem Identification

A closer look into the existing techniques of detection of spliced region is found to use supervised learning techniques, complicated classification-based approaches, and principle component analysis. Unfortunately, such mechanism are computational expensive and are accompanied by degraded system reliability. Although there are various studies towards image splicing, but the direction is more into classification and less into capturing the unique correlated information among the image blocks. Usage of color filter array can solve the problem of forced blur mechanism while manually performing image splicing operation.

However, there are less number of attempts towards using statistical-based approach by considering the abnormalities in blurring effect which are either not removed while removal of traces by attacker or has significantly produced an artifacts. Therefore, the open issue of the research is to find cost-effective computational algorithm which has higher accuracy and computational less complex while identifying spliced region for a given forged image.

## 4 Algorithm for Spliced Region Detection (ASRD)

The proposed system targets to extract the spliced region for a given forged image. Figure 1 exhibits the methodology adopted for designing ASRD. The mechanism performs partitioning of multiple blocks for extracting significant attributes of every blocks and its correlation with neighbor blocks. This phenomenon significantly assists to investigate the tampering of any part of the image. The local attributes are extracted and the algorithm looks for artifacts that could be underlying in the image. As there are infinite pixels to be scanned and investigated hence statistical measures using probability theory are the most appropriate technique to do this job. The algorithm is also associated with a classification technique in order to recognize tampered and not tampered zone.

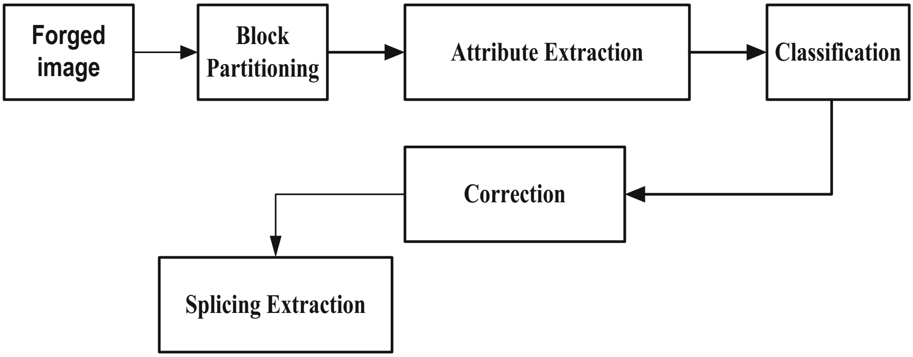


**Fig. 1.** Schema of proposed ARDS

## 5 Algorithm Implementation

The algorithm is mainly responsible for identifying the spliced region for a given image. The algorithm takes the input of $B_{i\ j}$ (block of image), $O_{i,j}$ (original Resolution image), $\rho_{i,j}$ (Gaussian Smoothening Function), $\sigma_{i,j}$ (Noise), $p$ (projection vector),

$C_m$ (Covariance matrix), $v_c/v_p$ (vector representing complete corruption and partial corruption respectively), $Th$ (Threshold), $\alpha_{class1}/\alpha_{class2}$ (classifier), etc., which after processing generates $S$ (Spliced Region). The steps of the algorithms are as shown below:

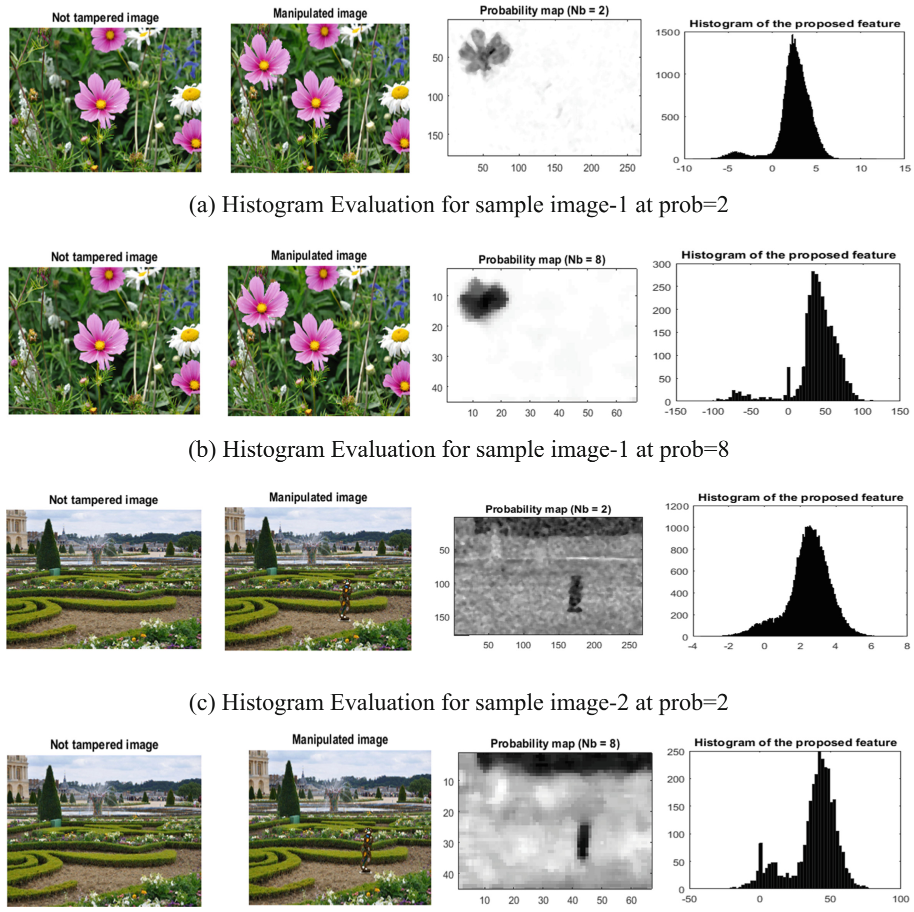
```
Algorithm for Spliced Area Detection
Input: $O_{i,j}, v_c, v_p, T$
Output: S
Start
1. $O_{i,j} \rightarrow$ digitize image
2. $B_{i\,j}=O_{i,j}*\rho_{i,j} + \sigma_{i,j}$
3. $p(B_{i\,j}) \rightarrow C_m^{-1}(v_c-v_p)$
4. If $f(x_{i,j}) \geq Th$
5.    $\alpha_{class1} \rightarrow$ Tamper Region
6. or else
7.    $\alpha_{class2} \rightarrow$ Non tamper Region
8. $S \rightarrow card(sort(\alpha_{class1}))$
9. spliced Region $\rightarrow bin$ (S), where bin=1 for $\alpha_{class1}$ and bin=[0, 256] for $\alpha_{class2}$
End
```

The algorithm initially converts the given image into grey scale B, which is then subjected to be partitioned by $B_{i,j}$ blocks. The first step of the algorithm is to empirically represent the tampered portion of the image as shown in Line-2. The computation of $\rho_{i,j}$ and $B_{i,j}$ can be carried out using deconvolution technique. The consecutive step of the algorithm is to identify the type of tampering followed by classifying the formation of spliced region. For this, the algorithm considers an attribute $\eta$ by integrating the dimensional parameter $\delta_1$ and standard deviation $\delta_2$. Empirically, it can be also represented as $\eta_{i,j} = [\delta_1\ \delta_2]^T$. The algorithm also computes the projection vector p as $p = [\ p(\delta_1)\ p(\delta_2)]^T$ in order to yield better formulation of an attribute $\eta_{i,j}$ using linear transformation where $\eta_{i,j} = p^T\ \eta_{i,j.}$ A covariance matrix $C_m$ is formulated for both complete ($C_c$) and partial corrupted ($C_p$) image i.e. $C_m = C_c - C_p$. The study considers the binary possibility of two types of regions i.e. (i) completely corrupted or spliced region and (ii) non tampered region. So, the algorithm represents its projection vector with respect to this binary classification as shown in Line-3. Forming a logical condition of tampered region, the $\alpha_{class1}$ is set for identification of both tampered and non-tampered spliced region (Line-4–7) using a particular threshold. The true positive and true negative parameters are then computed using $\alpha_{class1}$ and $\alpha_{class2}$ respectively. Depending upon the experimental value of $\alpha_{class1}$ and $\alpha_{class2,}$ the threshold value can be fine tuned. The studies also performs check for cardinality of such spliced region (Line-8) and then highlight the spliced region using the binary classifier *bin* (Line-9).

It will mean that as an outcome, the algorithm will make the entire image black with only the spliced region be explored as its natural or true color contents. Therefore, the proposed mechanism is able to identify spliced regions for number or any type of images corrupted by any degree of splicing operation.

## 6 Result Analysis

The analysis of the proposed study was carried out considering 1000 synthetic image dataset captured from SLR camera with varied range and resolution. Using existing image editing tool, they are also manipulated in order to obtain spliced image. Along with testing on synthetic dataset, the proposed system was also tested on standard datasets of Columbia Image Splicing database [27].



(a) Histogram Evaluation for sample image-1 at prob=2

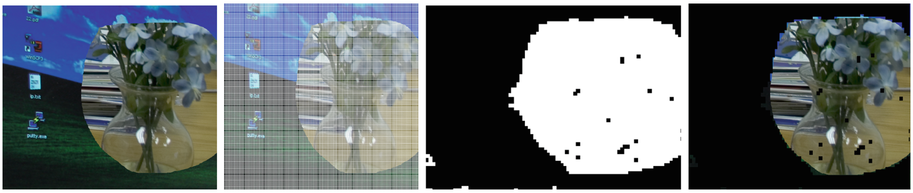
(b) Histogram Evaluation for sample image-1 at prob=8

(c) Histogram Evaluation for sample image-2 at prob=2

(d) Histogram Evaluation for sample image-1 at prob=8

**Fig. 2.** Visual outcomes of histogram analysis and probability in proposed system

The uncompressed spliced image dataset is used from Columbia Image Splicing database as shown in Fig. 2. Figure 2 (a) and (b) Shows sample-1 image with two different probability map (of value 2 and 8) shows different pattern of peaks in histogram. It interprets that increase in probability also increase better detection rate of spliced region. Probability maps are normal maps of gray scale image in order to yield binary feature of classification. The binarization was carried out after fine-tuning the threshold *Th* that was selected based on higher value of true positive and negative value.



(a) Spliced Image (b) Blocking (c) Corrected Image (d) Identified Spliced Region

**Fig. 3.** Visual outcomes of proposed system

The study outcome shown in Fig. 3 exhibits the visual outcomes of steps involved in processing. The spliced image is subjected for blocking operation that is further processed using binary classification technique illustrated in the algorithm to generate a binary image with tampered and non-tampered region. The elaborated discussion of the method used in implementation is as follows-The spliced region is varied from different ranges of sizes (smaller-bigger). The system then takes the spliced image as an input and multiple type of blocking operation is applied. The prime reason behind applying blocking operation is to ensure the detection performance. It also increases the feasibility of exploring the sensitive and critical area. Hence, it is recommended to apply smaller dimensions of the blocking operation in order to incorporate imperceptibility towards the spliced region. Moreover, maximized dimensions of the block partitioning can also be used for enhancing the difficulty level of an input spliced image. Once the finalization of the block partitioning operation is done then local level features are extracted. This operation is carried out using statistical-based approaches e.g. variance, mean, standard deviation, etc. Statistical-based features give more comprehensive information about the image and thereby over maximized benefits. Applying such forms of features extraction mechanism is quite deterministic in nature and therefore offers faster computation operations too. It is also cost effective in computational performance in contrast to any other statistical technique that uses inferential-based mechanism. The second advantage of this approach is its higher accuracy within inclusion of any recursive operation. Once the artifacts of the CFA are identified then the extraction of the outcome is carried out from the given image. Finally, the identification of the spliced region can be seen for the proposed system to retain its true color. For the purpose of an effective analysis, we compare outcome of proposed study ASRD with Ferrara et al. [28] and Han et al. [21] with respect to standard performance parameters of true positive, true negative, and accuracy. Figure 4 outlines the comparative performance analysis which takes the similar descriptive features discussed

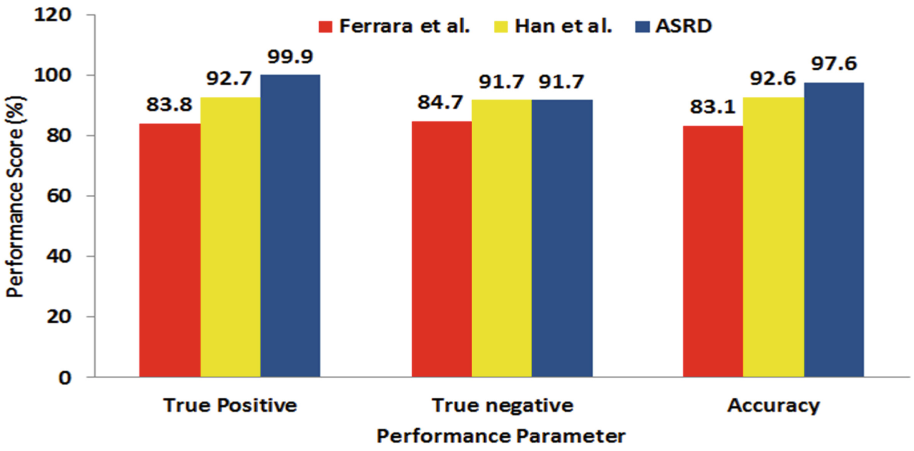


**Fig. 4.** Comparative performance analysis

above in order to extract the spliced region. Both the technique i.e. proposed and existing system considers its own flow of mechanism involved in detecting forged region in order to finally obtain the numerical outcome with respect to accuracy in detection process.

The outcome exhibits extensive identification of the spliced image as compared to Ferrara et al. [28] and Han et al. [21]. Though the difference is very marginal in terms of accuracy in detection factor, but proposed system has better computational capability in comparison to existing one. The complete algorithm processing time of ASRD was found to be 0.2765 s in core i7 machine while that of existing system was found to be approximately near to 1.2754 min. The memory complexity of the proposed system is also highly enhanced as it is free from any complex stochastic modeling like that of Han et al. [21] work. The performance of the proposed system with synthetic and standard dataset slightly differs in their outcome with 6.75%, which can be said to be within acceptable limits. Hence, the proposed study offers a robust and cost effective modeling for identification of regions within an image inflicted with image splicing operation. Apart from the accuracy, the response time of the proposed ASRD is found to be 75% of improvement as compared to the existing approaches of forged region detection in image processing.

## 7 Conclusion

The proposed study of ASRD has presented a technique that takes the forged image as an input in order to extract the precise region that has been maliciously tampered or corrupted with image splicing attack. The complete ideology of the proposed study is about uniformity among the neighborhood pixels with each in an original image. This uniformity is broken during image splicing in such a way that it is very difficult to perform identification of traces based on pixels. Hence, the proposed system presents a mechanism that performs partitioning of the blocks in order to obtain better granularity in the investigational findings. The study doesn't use any complex or iterative algorithms of optimization what can be seen in abundant in existing research techniques. This is where the proposed ASRD makes a different by introducing a very simple and cost effective algorithm for identifying the spliced region.

## References

1. Shih, F.Y.: Multimedia Security: Watermarking, Steganography, and Forensics. CRC Press, Boca Raton (2012)
2. Julliand, T., Nozick, V., Talbot, H.: Image noise and digital image forensics. In: Shi, Y.-Q., Kim, H.J., Pérez-González, F., Echizen, I. (eds.) IWDW 2015. LNCS, vol. 9569, pp. 3–17. Springer, Cham (2016). doi:10.1007/978-3-319-31960-5_1
3. Ding, F., Dong, W., Zhu, G., Shi, Y.-Q.: An advanced texture analysis method for image sharpening detection. In: Shi, Y.-Q., Kim, H.J., Pérez-González, F., Echizen, I. (eds.) IWDW 2015. LNCS, vol. 9569, pp. 72–82. Springer, Cham (2016). doi:10.1007/978-3-319-31960-5_7
4. Choi, C.-H., Lee, M.-J., Hyun, D.-K., Lee, H.-K.: Forged region detection for scanned images. Springer-Comput. Sci. Converg. **114**, 687–694 (2011)
5. Malviya, P., Naskar, R.: Digital forensic technique for double compression based JPEG image forgery detection. Springer-Inf. Syst. Secur. **8880**, 437–447 (2014)
6. Smith, S.: iMediaEthics' Top 10 Fake and Doctored Photo Stories. An online article of iMediaEthics 2016. http://www.imediaethics.org/imediaethics-top-10-fake-and-doctored-photo-stories/. Accessed 20 Oct
7. Vamosi, R.: Researcher: Bin Laden's beard is real, video is not. An online article of CNET. https://www.cnet.com/news/researcher-bin-ladens-beard-is-real-video-is-not/. Accessed 20 Oct 2016
8. Blitzer, H.L., Stein-Ferguson, K., Huang, J.: Understanding Forensic Digital Imaging. Academic Press, Cambridge (2010)
9. Stamm, M.C., Liu, K.J.R.: Forensic detection image manipulation using statistical intrinsic fingerprints. IEEE Trans. Inf. Forensics Secur. **5**(3), 492–506 (2010)
10. Weisi, L., Tao, D., Kacprzyk, J., Li, Z., Izquierdo, E.: Haohong Wang, Multimedia Analysis, Processing and Communications. Springer Science & Business Media, New York (2011)
11. Sundarm, A.M., Nandini, C.: Copy-move forgery detection- a survey. In: ICACCN-International Conference on Advanced Computing, Communication Networks, Chandigarh, 02–03 June 2011
12. Sundarm, A.M., Nandini, C.: Investigational study of image forensic applications, techniques and research directions. Int. J. Emerg. Technol. Adv. Eng. Certif. J. **4**(8), 1–9 (2014). https://www.ijetae.com. ISSN 2250-2459, ISO 9001:2008
13. Sundarm, A.M., Nandini, C.: Image retouching and it's detection-a survey. In: NCGCT-First National Conference on Green Computing Technologies, DSATM, Bangalore, 07 March 2015
14. Sundarm, A.M., Nandini, C.: Feature based image authentication using symmetric surround saliency mapping in image forensics. Int. J. Comput. Appl. **104**(13), 1–9 (2014)
15. Sundarm, A.M., Nandini, C.: CBFD: coherence based forgery detection technique in image forensics analysis. In: IEEE-ICERECT-2015-International Conference on Emerging Research in Electronics, Computer Science and Technology, 17–19 December 2015
16. Cozzolino, D., Poggi, G., Verdoliva, L.: Splicebuster: a new blind image splicing detector. In: IEEE International Workshop on Information Forensics and Security (2015)

17. Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Detecting image splicing in the wild (web). In: IEEE International Conference on Multimedia and Expo Workshop, pp. 1–6 (2015)
18. Amerini, I., Becarelli, R., Caldelli, R., Mastio, A.D.: Splicing forgeries localization through the use of first digit features. In: IEEE International Workshop on Information Forensics and Security (2014)
19. Moghaddasi, Z., Jalab, H.A., Md Noor, R.: SVD-based image splicing detection. In: International Conference on Information Technology and Multimedia (2014)
20. Su, B., Yuan, Q., Wang, S., Zhao, C., Li, S.: Enhanced state selection Markov model for image splicing detection. Springer-EURASIP J. Wirel. Commun. Netw. **2014**, 1–10 (2014)
21. Han, J.G., Park, T.H., Moon, Y.H., Eom, K.: Efficient Markov feature extraction method for image splicing detection using maximization and threshold expansion. J. Electron. Imaging **25**(2), 023031 (2016)
22. Zhang, Y., Zhao, C., Pi, Y., Li, S.: Revealing image splicing forgery using local binary patterns of DCT coefficients. In: Liang, Q., et al. (eds.) Springer Journals of Communications, Signal Processing, and Systems. LNEE, pp. 181–189. Springer, New York (2012). doi:10.1007/978-1-4614-5803-6_19
23. Saleh, S.Q., Hussain, M., Muhammad, G., Bebis, G.: Evaluation of image forgery detection using multi-scale weber local descriptors. In: Bebis, G., et al. (eds.) ISVC 2013. LNCS, vol. 8034, pp. 416–424. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41939-3_40
24. Zhao, X., Li, S., Wang, S., Li, J., Yang, K.: Optimal chroma-like channel design for passive color image splicing detection. Springer-EURASIP J. Adv. Signal Process. **2012**, 240 (2012)
25. Pan, X., Zhang, X., Lyu, S.: Exposing image splicing with inconsistent local noise variances. In: IEEE International Conference on Computational Photography, pp. 1–10 (2012)
26. Niu, H., Zhou, C., Wang, B., Zheng, X., Zhou, S.: Splicing model and hyper-chaotic system for image encryption. J. Electr. Eng. **67**(2), 78–86 (2016)
27. Columbia Image Splicing Detection Evaluation Dataset. http://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet.htm, Retrieved 06th February 2017
28. Ferrara, P., Bianchi, T., De Rosa, A., Piva, A.: Image forgery localization via fine-grained analysis of CFA artifacts. IEEE Trans. Inf. Forensics Secur. **7**(5), 1566–1577 (2012)

# Towards Data Storage for Online Analytical Antispam System – ASOLAP

Alexandr Vasilenko and Jan Tyrychtr(✉)

Faculty of Economics and Management,
Department of Information Technologies,
Czech University of Life Sciences in Prague, Prague, Czech Republic
{vasilenko, tyrychtr}@pef.czu.cz

**Abstract.** Junk mail is a major concern of Internet communication. It represents most of the received messages. To filter unsolicited bulk e-mails there is a large amount of human and financial resources and computing resources needed. One way to counter this problem is to maximize the information yield of the obtained unsolicited messages. This article aims to introduce the concept ASOLAP - the use of OLAP to store and analyze metadata of e-mail messages. We propose a conceptual data model and verify its quality. Based on the results, we recommend to use the design of the star schema that represents the potential for quality and efficient solution of the ASOLAP design.

**Keywords:** Unsolicited bulk e-mail · Spammer · OLAP · Analytical system · Conceptual schema

## 1 Introduction

Electronic communication via electronic mail is an important tool for information exchange. Despite the ever-growing social network, the mail remains the main means of electronic communication. This is taken advantage of by spammers who flood e-mail boxes with large numbers of unsolicited messages (Unsolicited Bulk E-mail – UBEs) [9, 11].

According to the workgroup Mail Anti-abuse around 80% of all e-mails sent via the Internet is spam. This figure oscillates between 75%–90% in the longterm. Fluctuation is affected by the scope of activities of spammers and also by measures against them. Eliminating a large botnet from active misuse can cause a drop from singles to tens of percentage points [5, 10, 11].

Blocking spam is a never-ending process because spammers are trying to find solutions against the ever-improving filters. It leads to a duel of measures and countermeasures. The actual antispam solution represents a collection of many methods of evaluation of spamicity of the message. The key is to constantly monitor incoming e-mail messages and the results of their evaluation is needed to be analyzed and modified.

Classification of anti-spam tools in professional and scientific literature respects their primarily its focus: the rules, enforcement of RFC (request for connect),

The original version of the book was revised. For detailed information please see Erratum. The erratum to the book is available at 10.1007/978-3-319-57141-6_53

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_11

authentication, and more. We consider this division to be valid, however at a lower level than is useful for the purposes of this article. Therefore, we classify spam tools into three groups according to their complexity:

**Elementary Tools**
These tools use a single component method to detect spam based on exactly one criterion. They serve as a sub-method in the context of combined instruments, or as a method applied to the initial analysis. Used detection methods are simple (comparison of IP addresses, quit detection, etc.) and do not require too many system resources [13]. Typical initial filtering can be considered anti-virus software [11].

**Advanced Tools**
The activity of tools with complex internal logic require more system resources, but show a better performance parameters. As an example, we can mention neural networks. This tool can, after a certain period (of learning) determine the spamicity very precisely. The disadvantage, however, are high demands on system resources and deployability is thus debatable [6, 12].

**Combined Tools**
The way to good performance of an anti-spam solution is a combination of selected tools. In this approach, it is necessary to monitor and adjust the weights to the partial methods to optimize the results of the evaluation. A typical example is SpamAssassin used for e-mail servers under Linux [14].

One of the ways to improve monitoring and evaluation of messages_ is represented by an improvement of analytical processes. Currently, it is common to evaluate the spamicity of the message according to the similarity of the hash (hash function nilsimsa E.G.). The actual message header contains also interesting data about the way from the source to the recipient.

The antispam issue is a multidiscipline that includes not only technological elements, but also theoretical assumptions, mathematical and statistical analysis and modeling. In practice, it is necessary to put several filtering methods in such a way as to maximize their effect while minimizing system requirements. The goal is to find the intersection between different instruments and combine their strengths so that they override the weaknesses, which could degrade the resulting solution.

We are convinced that this huge amount of information can be processed if we use methods from the field of Business Intelligence (BI) [7]. When analyzing large amounts of intercepted messages, we deem it appropriate to focus on the power of anti-spam tools. Based on past experience in this area it is necessary to accept the current limitation of performance characteristics of server solutions for e-mail centers.

The aim of this article is to design a conceptual model for use in Antispam Online Analytical System (ASOLAP design) with emphasis on performance characteristics and with an emphasis on optimizing the rules set by analyzing metadata of e-mail messages.

This could have a significant impact on the development of intelligent systems for effective corporate e-mail services.

## 2 Materials and Methods

In this paper, we use approaches on spamicity and vectorization of metadata UBEs and propose a conceptual schema for creating a data storage - here we use OLAP technology as a tool for processing large volumes of text data.

Typical uses OLAP _ in companies is for reporting, in-depth data analysis, forecasting and evaluation. Since the data is processed using OLAP the output from all internal information systems, the data volume from the biggest companies can be compared with the volume of data being produced by anti-spam and e-mail system [6].

### 2.1 Spamicity

For spamicity classification it is necessary to analyze maximum of available data [4]. These are available in the header, which has complete records about the message. Another source of data is the actual message content, which may be evaluated by partial methods, e.g. Bayesian filtering, where each word has its own score with representation in messages classified as spam. The second tool for the rating of content is the similarity hash, for example Nilsimsa. This, to some extent, eliminates partial changes that the spammer makes to make detection more difficult. Similarity hash eliminates this to some extent.

The last data source is an optional message attachment - usually with spam it contains malware. That is unchangeable, the hash function can analyze messages according to attachments.

Metadata e-mail messages we described by the following hierarchical structure:

- Return-path,
- Delivered-to, Received,
- Message-id,
- SMTP server,
- Subject,
- From,
- Nilsimsa hash,
- Links,
- Bayesian classifiers,
- DKIM,
- SPF,
- Attachment-hash.

Article [11] introduced the concept of classification of sets of unsolicited messages. As part of its evaluation work of spamicity using BI we use this concept. Identification of the sets represents of the spamicity a method to evaluate of the messages and a comprehensive view of the received message as a whole, not only to assess the report as an isolated entity. Characteristic is the comparison between individual messages made not only based on the aggregate value of the spamicity of the message, partial common characters with other already identified unsolicited messages.

Belonging to a set of defined metadata of e-mail messages. Each value determines the sub spam rating, the resulting vector [11], then in enables to evaluate the spamicity of the message comprehensively and not based on a simple comparison and of the sum of the limit values of the spamicity.

The utility of this concept is reinforced by binding spam messages to a botnet. That is usually not specialized in merely sending unsolicited messages, but also for other activities. One of them is the pursuit of a distributed attack on credentials, for example an attempt to guess a password for ssh access to the server.

We are seeing many such attacks. In an attempt to circumvent the tools to deny access for a single IP address (such as a tool fail2ban) there is a botnet used - each IP address has three attempts to log on, after the blockage it is relieved by another computer of the botnet [5].

## 2.2 Vectorisation of UBEs Metadata

Vectorization is a system of evaluation of messages which is not based on differentiating the messages according to the final score of the message but it is based on its vector – that is composed of individual parts. According to the concept of vectorization of metadata we describe individual tables by vectors which signify the spamicity of the message for the given dimensional table. This calculation is then modified by the change of weight of individual sub variables and then we modify the evaluation of individual messages based on appearance of chosen key characteristics found in the metadata of e-mail messages [11].

The proposed data scheme can be described by the following record:

$$V_s = v_1 + v_2 + v_3 + v_4 + v_5 + v_6 + v_7 + v_8, \tag{1}$$

where $V_s$ is a vector of spamicity of the given e-mail message and $v$ are individual sub vectors of evaluation:

- v1 the SMTP server used to send messages
- v2 the date of receipt
- v3 the message delivery time
- v4 the date of sending the message
- v5 is time to send a message
- v6 a hash attached file
- v7 the IP address of the sender node
- v8 is the subject of your message

Each set of messages has its own vector. To pertain to a certain set of messages is given by the level of toletance of the difference between individual vectors.

## 2.3 OLAP

In this paper, we adopt the formal apparatus of the data cube according to [1, 2]. Let us have 6-tuple $<D, M, A, f, V, g>$, where four components indicate properties of data cube. Those properties are:

- The set of n dimensions $D = \{d_1, d_2, \ldots, d_n\}$, where each $d_i$ is the name of dimension from particular domain $dom_{\mathrm{dim}(i)}$.
- The set of k measures $M = \{m_1, m_2, \ldots, m_k\}$, where each $m_i$ is the name of measure from particular domain $dom_{\mathrm{measure}(l)}$.
- The set of dimension names and measures is disjoint; i.e. $D \cap M = 0$.
- The set of t attributes $A = \{a_1, a_2, \ldots, a_t\}$, where each $a_i$ is the name of attribute from particular domain $dom_{\mathrm{attr}(r)}$.
- The one-to-many mapping $f : D \rightarrow A$ exists for every dimension and set of attributes. The mapping is such that attribute sets corresponding are pair wise disjoint, i.e. $\forall i, j, i \neq j, f(d_i) \cap f\left(d_j\right) = 0$.
- The set $V$ represents a set of values that were used to materialize data cube. Therefore every element $v_i \in V$ is $k$-tuple $< \mu_1, \mu_2, \ldots, \mu_k >$ , where $\mu_i$ is instance of $i$-th measure $m_i$.
- The $g$ represents a mapping $g : dom_{\mathrm{dim}(3)} \times dom_{\mathrm{dim}(2)} \times \ldots \times dom_{\mathrm{dim}(n)} \rightarrow V$. Thus, intuitively $g$ mapping indicates which values are associated with a particular ‘cell’. Cells are measures or values based on a set of dimensions.

### 2.4 OLAP Schemes

For the proposal of conceptual schemas, we use a star schema and a snowflake schema according to the rules of vectorization [3]. These schemas typically have one fact table and corresponding dimension tables. These are without a hierarchical structure for the schemas of the star type, schema of the snowflake includes a hierarchy of the dimensional tables.

### 2.5 Methods for Measuring the Quality of Data Warehouse

Design of a data warehouse is validated by the selected methodology. Based on the results of the assessment, data model design is accepted or rejected and possibly redesigned. Methodologies for evaluation of the design of data warehouses are currently several [12]. For the purposes of the article we use the evaluation according to the methodological approaches based on the clarity and comprehensiveness. These evaluation criteria were selected based on recommendations [4, 8].

We use the following evaluation criteria:

- NDT – number of dimension tables,
- NT – number of tables,
- NADT – number of attributes of dimension tables,
- NAFT – number of attributes plus the number of foreign keys,
- NA – number of attributes,
- NFK – number of foreign keys.

These metrics will help you measure the quality of OLAP data warehouse in the star or snowflake scheme according to the complexity of the scheme itself, where complexity is determined by the number of tables, attributes and foreign keys of the scheme. It is therefore through metrics assume that the quality is affected by the complexity and hence scheme with low complexity, it should be preferred to scheme with great complexity [1].

The last two metrics are very significant. Through the NFK metric we calculate the ratio of the number of foreign keys in the fact table to the total number of attributes. A high value for this metric penalizes the fact table that have a high number of foreign keys and several measures. Generally, in the dimension tables two attributes are usually sufficient: the primary key and descriptive attribute. Other descriptive attributes are often useless.

### 2.6 Creation of Prototype

Both prototype solutions are tested on real data (captured spam) and measure the time required to implement the selected view of data. Repeat measurements for elimination of outlying observation, which may be due to a temporary system overload by another task. The resulting average is then used as the third evaluating parameter.

We realize measurements with a measuring script in the VBA language (Visual Basic for Applications), which apart from including the launched code for the realization of looking at data also includes the measuring part. The output value is measured in seconds. The prototype of the OLAP is created using the ROLAP (Relational OLAP) technology [4] and Microsoft Power Pivot software.

## 3 Results

To verify the feasability of OLAP concept as a tool for storing and analyzing metadata of e-mail messages, we create two conceptual schemes. One based on a star schema, the latter conceived as a solution using a hierarchy dimensions.

### 3.1 Design of Conceptual Schemas

**Star Schema**

To verify the implementation options of ASOLAP as instruments for analyzing the metadata of e-mails, we designed a prototype solution to a multidimensional database. The primary function of the prototype is to verify the feasibility of this solution and performance analysis of this solution - its operation in real time to refine the rules of anti-spam tools. Source data of e-mail messages were imported and cleaned to simplify the model. This simplification reflects the purpose of the prototype - the evaluation the applicability of OLAP metadata for analysis of e-mail messages at the expense of full data analysis (Fig. 1).

**Snowflake Schema**

The advantage of the database snowflake schema is the conception of data in a standardized form, which makes the design of a data schema more natural, especially where it is proposed by administrators with experience in relational databases. The disadvantage is a lower pace of operations due to the higher number of sessions and dimensional tables.

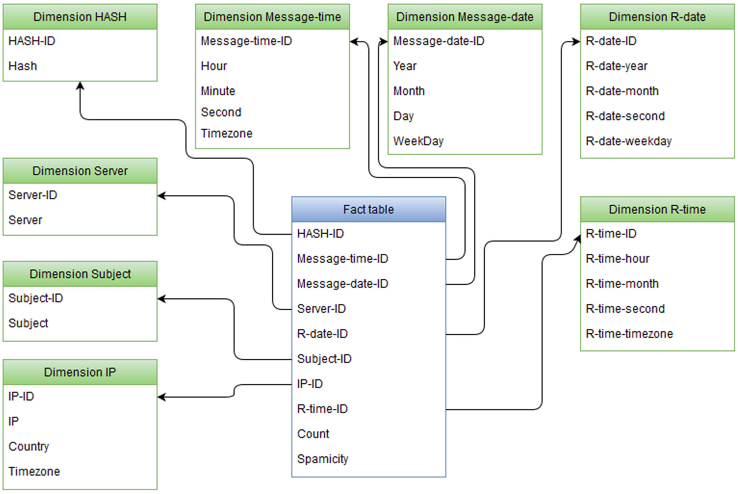


**Fig. 1.** Star schema of ASOLAP

Data snowflake schema requires changes compared to the previous solution in the adaptation of the dimensional tables and to establish new relations. At the same time the data is normalized and its volume is reduced. The result of our solution is then the following scheme (Fig. 2).

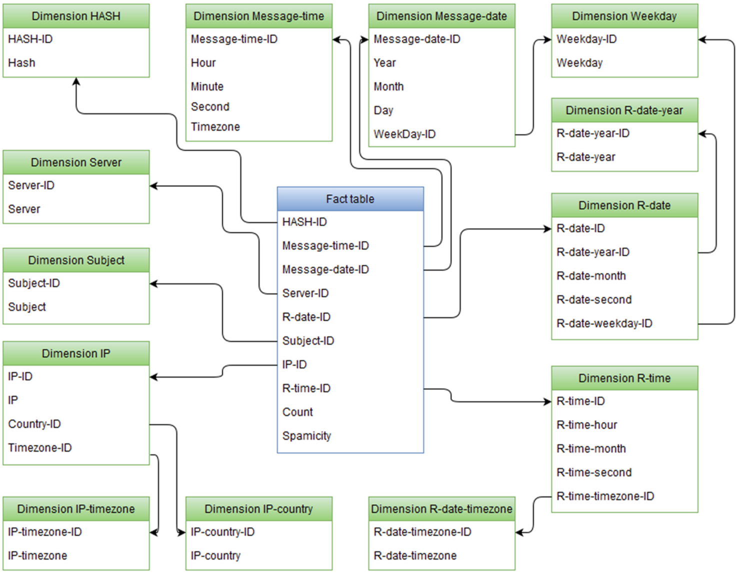


**Fig. 2.** Snowflake schema of ASOLAP

### 3.2 Quality Evaluation

The created conceptual schemes for ASOLAP are evaluated in terms of quality. In all variants of the comparison the star type schema is evaluated better. In its evaluation there is a significant difference in clarity (57:77). Based on these results, a star schema is suitable for the design of data storage of ASOLAP systems (Tables 1 and 2).

**Table 1.** Evaluation of the quality of the model

| Methods | Star | Snowflake |
|---|---|---|
| NFT | 1 | 1 |
| NDT | 8 | 13 |
| NFK | 9 | 9 |
| NMFT | 8 | 8 |
| Overall score | 26 | 31 |

**Table 2.** Evaluation of clarity of the model

| Methods | Star | Snowflake |
|---|---|---|
| NDT | 8 | 13 |
| NT | 9 | 14 |
| NADT | 30 | 40 |
| NAFT | 1 | 1 |
| NFK | 9 | 9 |
| Overall score | 57 | 77 |

### 3.3 Performance Evaluation

The performance measuring is based on the time required to show the same views of data. Within the performance testing was measured processing time views of the following entries:

1. The number of messages sent in each year from individual countries.
2. The number of messages sent using the detected STMP server.
3. The number of identified files sent from individual countries.
4. The number of messages sent in a single day each week in the year.
5. The number of messages from the identified IP addresses (Table 3).

Within the framework of performance, we conducted a measurement of speed. Both models received the same assignment of queries for data and with the help of a script the time required to build the view of the data according to the relevant query was measured. The results of the testing showed that the star data schema was faster in all five queries.

**Table 3.** Performance test results

| Operation | Star | Snowflake |
|---|---|---|
| The number of messages sent in each year from individual countries | 0, 35 s | 0, 41 s |
| The number of messages sent via detected SMTP servers | 2, 31 s | 2, 37 s |
| The number of identified files sent from individual countries | 1, 05 s | 1, 15 s |
| The number of messages sent each day in the week in the year | 1, 45 s | 1, 98 s |
| The number of messages from identified IP addresses | 3, 87 | 4, 25 s |
| Number of better times | 5 | 0 |
| Total time differences | 9, 03 s | 10, 16 s |

## 4 Conclusion

The design of conceptual models for the use in ASOLAP systems proposed for analyzing the metadata of e-mail messages made it possible to identify solutions that could help in the development of these complicated and complex systems for filtering spam. Based on our results the star schema design was chosen. This type of proposal will make the creation of quality and efficient data storage systems ASOLAP type possible. The resulting model and its application has been validated by altering the rules and changes in the order of priorities. For the administrator of the e-mail system it is easy to access the comprehensive data and statistics, according to which they decide on modifying or keeping the classifiers. Effective analysis of unsolicited messages from different perspectives (dimensions) will enable to increase the effectiveness of anti-spam solution and thereby reduce the burden on users' mailboxes. Such a solution is suitable for administrators of antispam engines and particularly in corporate environments that are overcrowded by junk e-mail.

**Acknowledgements.** The results and knowledge included herein have been obtained owing to support from the Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences in Prague, grant no. 20161019, “Economic value of analytical systems in agriculture”.

## References

1. Bellatreche, L., Cuzzocrea, A., Song, I.: Advances in data warehousing and OLAP in the big Data Era. Inf. Syst. **53**, 39–40 (2015)
2. Colliat, G.: OLAP, relational, and multidimensional database systems. ACM Sigmod Rec. **25**(3), 64–69 (1996)
3. Datta, A., Thomas, H.: The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. Decis. Support Syst. **27**(3), 289–301 (1999)
4. Di Tria, F., Lefons, E., Tangorra, F.: Hybrid methodology for data warehouse conceptual design by UML schemas. Inf. Softw. Technol. **54**(4), 360–379 (2012)
5. Kirubavathi, G., Anitha, R.: Botnet detection via mining of traffic flow characteristics. Comput. Electr. Eng. **50**, 91–101 (2016)

6. Chakraborty, M., Pal, S., Rahul Pramanik, C., Chowdary, R.: Recent developments in social spam detection and combating techniques: a survey. Inf. Process. Manag. **52**(6), 1053–1073 (2016)
7. Peters, M.D., Wieder, B., Sutton, S.G., Wakefield, J.: Business intelligence systems use in performance measurement capabilities: implications for enhanced competitive advantage. Int. J. Account. Inf. Syst. **21**, 1–17 (2016)
8. Gosain, A., Heena: Literature review of data model quality metrics of data warehouse. Procedia Comput. Sci. **48**, 236–243 (2015)
9. Spammer-X, Posluns, J., Sjouwerman, S.: Inside the SPAM Cartel. Syngress, Boston (2004). ISBN: 079-2502668603
10. Almeida, T.A., Yamakami, A.: Facing the spammers: a very effective approach to avoid junk e-mails. Expert Syst. Appl. **39**(7), 6557–6561 (2012)
11. Vasilenko, A., Ocenasek, V.: Spam as a problem for small agriculture business. Agris Online – Pap. Econ. Inform. **1**(8) (2013)
12. Zhang, X., Li, Y., Kotagiri, R., Lifang, W., Tari, Z., Cheriet, M.: KRNN: k Rare-class Nearest Neighbour classification. Pattern Recogn. **62**, 33–44 (2017)
13. Yevseyeva, I., Basto-Fernandes, V., Ruano-Ordás, D., Méndez, J.R.: Optimising anti-spam filters with evolutionary algorithms. Expert Syst. Appl. **40**(10), 4010–4021 (2013)
14. Meng, Y., Kwok, L.-F.: Adaptive blacklist-based packet filter with a statistic-based approach in network intrusion detection. J. Netw. Comput. Appl. **39**, 83–92 (2014)

# Development of Methodology for Entangled Quantum Calculations Modeling in the Area of Quantum Algorithms

Viktor Potapov(✉), Sergei Gushanskiy, Vyacheslav Guzik, and Maxim Polenov

Department of Computer Engineering, Southern Federal University, Taganrog, Russia
vitya-potapov@rambler.ru, {smgushanskiy, vfguzik, mypolenov}@sfedu.ru

**Abstract.** This paper assumes the description of the foundations of quantum information theory and the concept of quantum entanglement. Universal definition of a quantum algorithm was derived as well as the stages of its work. A methodology for modeling entangled quantum computing in the field of quantum algorithms is proposed. Such method is a complete the sequence of stages implementing of the universal quantum algorithm in terms of quantum computing. And the study of entanglement level influence on the operation of quantum algorithms is considered also.

**Keywords:** Quantum entanglement · Singular value decomposition · Schmidt's decomposition · Pauli matrices · Quantum algorithm · Entangled states

## 1 Introduction

The study of entanglement [1] of quantum-mechanical state systems is one of the most important areas of research in the theory of quantum information and quantum computing because of the confusion in the concept of the importance of applied algorithms and the relationship of this concept to other sections of quantum computing. As demonstrated by the history of this research development, deepening our understanding of the concept of entanglement leads to better understanding of the structure and logic of quantum mechanics and to a new look at the notion of physical condition.

Nowadays entanglement is considered to be one of the main differences between the quantum-mechanical systems from the classical ones. It makes the first ones interesting from the point of view of applications in matters of information processing and quantum communication. Some authors point to the involvement as the underlying cause of the acceleration of quantum-mechanical calculations compared to classical devices. Despite making great efforts to understanding of physical phenomena, a complete theory of entanglement has not currently been developed yet. So, the definition of universal measures of entanglement, applied not only to mixed, but also to the pure state; establishing of criteria entanglement quantum state; the study of general properties of multicomponent quantum systems are still remain unresolved.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_12

One of the most important properties of entanglement, which plays an prominent role in the subsequent discussion, is the invariance of confusion in relation to the local operations subsystems of the quantum system. In other words, the entangled state of two or more qubits can be neither created nor significantly changed (in the sense of confusion changes) by impact to individual qubits. Thus, all the conditions that may be obtained from a given state only through involvement of local operations can be identified in terms of entanglement value stored therein. Below a diagram [2] considering relations and interactions of the elements of quantum mechanics is shown in Fig. 1. Concepts related to the theory of quantum information and quantum computing and based on the notion of quantum entanglement are given in bold.

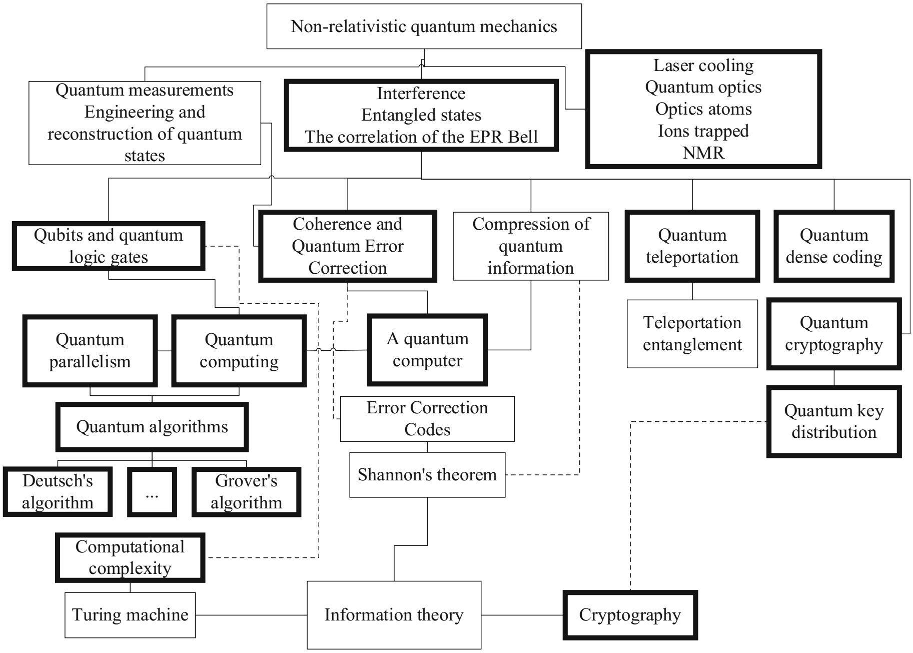

Non-relativistic quantum mechanics
Quantum measurements Engineering and reconstruction of quantum states
Interference Entangled states The correlation of the EPR Bell
Laser cooling Quantum optics Optics atoms Ions trapped NMR
Qubits and quantum logic gates
Coherence and Quantum Error Correction
Compression of quantum information
Quantum teleportation
Quantum dense coding
Quantum parallelism
Quantum computing
A quantum computer
Teleportation entanglement
Quantum cryptography
Quantum algorithms
Error Correction Codes
Quantum key distribution
Deutsch's algorithm
...
Grover's algorithm
Shannon's theorem
Computational complexity
Turing machine
Information theory
Cryptography

**Fig. 1.** The relationship of elements of quantum information theory

The question is how to describe the variety of such states. The solution of this question must necessarily precede the study of the relationship of these varieties with the measure of entanglement of quantum systems.

The field of quantum algorithms is constantly updated. Development of new quantum algorithm is a dynamic area, as evidenced by the set of quantum algorithms [3], it contains references to 45 algorithms and about 160 papers. Despite the fact that many quantum algorithms are designed for basic tasks (for example, determining the order of the ideal of finite ring, the computation of Boolean formulas and a sieve of Kuperberg) a number of quantum algorithms that solve applied problems has been developed. It is the task of cryptography (cryptographic compromise of the various systems and protocols to generate private keys), sample math problems on graphs and matrices, and they have a very large range of applications. However, the area of quantum computing transition from theory to practice in the process of developing, but we can assume the shape of a possible future quantum computer and an interface which can be used to interact with the quantum computing device.

## 2 The Concept of a Quantum Algorithm

Currently, there are a large number of quantum algorithms to solve a variety of tasks, but there is no universal definition of the quantum algorithm, which is not based on the properties of the specific quantum algorithm.

Figure 2 shows one cycle of a quantum resource access. Here, using the pulse of the classic computer, a set of qubits $S_k$ on universal basis is created, then a pure state $|0^n\rangle$ is prepared. Next are following the different unitary transformations $U_{b_l}[S_l]\ldots U_{b_1}[S_1]|0^n\rangle$ and measurement in the computational basis. The final step is the processing of the results of the measurement using classical tools – x.

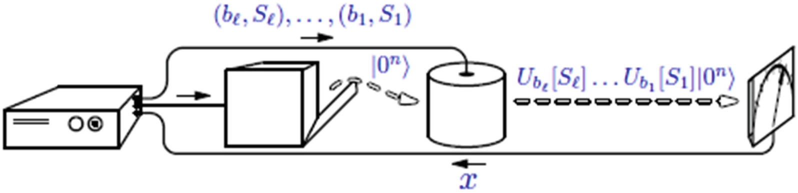

$(b_\ell, S_\ell), \ldots, (b_1, S_1)$
$|0^n\rangle$
$U_{b_\ell}[S_\ell]\ldots U_{b_1}[S_1]|0^n\rangle$
$x$

**Fig. 2.** Cycle of a quantum resource access

Thus, the probability of observing x is represented as follows:

$$\Pr(U_{b_l}[S_l]\ldots U_{b_1}[S_1]|0^n\rangle, x) = |\langle x|U_{b_l}[S_l]\ldots U_{b_1}[S_1]|0^n\rangle|^2 \tag{1}$$

The execution time of the action cycle is *T(l)*. But it should also be noted that such scheme could be improved. Only one cycle of access to quantum resource is enough and all the intermediate measurements could be simulated by the corresponding quantum circuits. To do this, after measuring a qubit it is necessary to use only operators

$$U : |b\rangle \otimes |\psi\rangle \mapsto |b\rangle \otimes U_b|\psi\rangle \tag{2}$$

Thus, let's give a definition of the quantum algorithm. Quantum algorithm $Q$ is a classic algorithm $A$, which can be implemented on classic computing devices and according to the input x, implements the quantum scheme description $C_x$ in universal finite basis with the operator $U_x$ on $n_x$ qubits and result register $S_x$.

The algorithm running time on input x is nothing but the sum of the running time of the algorithm $A$ and the size of circuit $C_x$.

There is also the probability of the result y in input x, and it is equal to:

$$\Pr(y|x) = \sum |\langle x|U_x|0^{n_x}\rangle|^2 \tag{3}$$

The algorithm computes the function f (x) with a probability of error ε, if

$$\Pr(y \neq f(x)|x) < \varepsilon \tag{4}$$

However, more common is the definition where the algorithm implements description of the entire scheme for all inputs of length n-th, and the input of the quantum circuit instead of $|0^n\rangle$ is supplied $|x\rangle$. As in the case of quantum probability calculations, quantum algorithms perform a rapid decrease a probability of error.

If you want to implement a quantum algorithm $Q$, which will calculate the *f(x)* with a probability of error ε < ½, then the algorithm $Q_s^{'}$ will run as follows:

1. Repeat an algorithm $Q$ $S$ times apart from each other;
2. Issue as a reasonable result the *y* value, which is obtained more often.

Anyway all quantum algorithms can be described using respective quantum circuits comprising, in turn, a certain set of gates. Therefore, when evaluating computational complexity it is reasonable to base on two basic measures of quantum circuit complexity, namely the size and depth.

The size of both quantum circuit and the classical one is the number of elements presented in the composition of this scheme. The size of the circuit, in turn, directly influences on the time of computation and the operation of the quantum algorithm on a serial device that performs elementary steps one after other.

The depth of the quantum circuit is the smallest possible number of layers (branches) [4], in which you can arrange the elements of the circuit. This measure of complexity is not as simple as it seems at first glance. The situation is complicated by the existence and quite often use of elements of quantum circuits (gates) that can use two, three, or even all the layers-branches of quantum circuit. These include: CNOT, Swap, CCNOT (Toffoli), CSWAP (Fredkin's gate), Grover operator, quantum Fourier transform and inverse quantum Fourier transform. It should be noted that single-qubit gates also occur in estimating the depth of the quantum circuit, but their value can be neglected because of their location and occupying only one branch.

This complexity measure directly affects the running time of computing device, which able to perform several different actions (parallel computation) simultaneously.

In turn, quantum circuits are not the only way to describe quantum algorithms. The process of evaluating the computational complexity depends on the computing models, such as the Markov algorithms and the Turing machine.

## 3 The General Structure of the Quantum Algorithm

Problems to be solved by means of quantum algorithms may be represented in the following form shown in Table 1.

The quantum computing basis are three operators acting on the quantum coherent states: superposition, entanglement and interference.

**Table 1.** The structure of the quantum algorithm task

| | |
|---|---|
| Input | Function $f : \{0,1\}^n \rightarrow \{0,1\}^m$ |
| Task | Find the specific property of the function f |

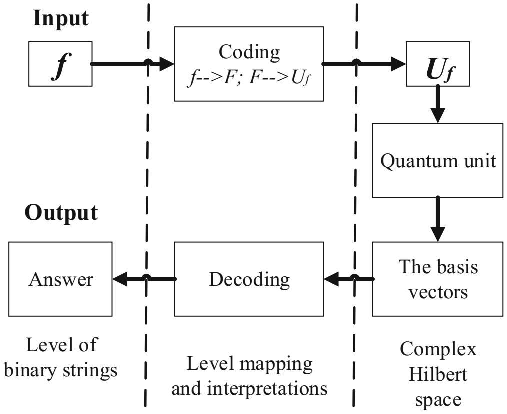


**Fig. 3.** Schematic diagram of the quantum algorithm

The operator of superposition is the tensor product of $n$ Hadamard operators $H$ and $m$ operators of the identical transformation $I$. The resulting operator acts on the first register to create their superposition and operates identically to the second register leaving it unchanged.

The operator $U_F$ is the entanglement operator. It is obtained by using of the encoding unit (Fig. 3). Its appearance depends on the properties of the original function $f$.

Considered algorithm sets a specific interference operator, which can be, for example, the quantum Fourier transform or Hadamard operator. The phenomenon of constructive/destructive interference is used as a tool of effective extraction of the results of calculations of quantum algorithms to improve the probability of measuring and extracting the desired solution in the process of designing different models of quantum algorithms. To increase the probability of extracting "successful" solutions constructive interference is applied, and for the reduction of "bad" decisions is applied destructive interference.

A schematic diagram of a quantum algorithm simulation on a classical computing device is shown in Fig. 3. Quantum unit here performs the alternate use of the quantum operator and the result measurement. It is run n times to obtain a set of basis vectors. Since the measurement operation is not determined then obtained basis vectors will be identical, and each of them will contain only the information needed to solve the problem.

The final stage of the of the quantum algorithm run is the interpretation of obtained basis vectors in order to get the correct answer to the initial problem with a certain probability.

# 4 Development of Methodology for Entangled Quantum Calculations Modeling

## 4.1 Entanglement Criterion

Entanglement is a special form of the quantum particles correlation, which has no classical analogs. In order to create the most entangled pair the Hadamard gate must impact on the first qubit, and then both qubits are impacted by CNOT gate [5]. Qubits are initially taken in a pure state. Let's suppose that there is a pure quantum state in space $H_A \otimes H_B$, $d_A = \dim(H_A)$, $d_B = \dim(H_B)$, $d = \min(d_A, d_B)$.

$$|\psi\rangle = \sum_{i=1}^{d_A} \sum_{j=1}^{d_B} a_{ij} * |i\rangle_A * |i\rangle_B \tag{5}$$

Note that no-entangled is a state that can be represented in the form $|\psi\rangle = |\psi\rangle_A \otimes |\psi\rangle_B$ and all remaining states are entangled.

Task. Is the state entangled:

$$|\phi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)?$$

Decision. The matrix corresponding to this quantum state will be

$$A = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

$rank(A) = 2$ therefore, the state is entangled.

## 4.2 Singular Decomposition of Matrices

Four Pauli matrices are the basis of all gates, and all other gates are obtained by the ordinary and tensor product of matrices, as well as multiplication of a factor. An important role in the entanglement of quantum states plays a singular value decomposition of the matrices (or SVD-decomposition, Singular Value Decomposition). Any complex matrix A of size m × n has the decomposition $A = U * S * V^*$, where U – a unitary matrix of m order, V – a unitary matrix of n order, S – diagonal matrix of m × n with non-negative real numbers $\{S_1, S_2, \ldots, S_n\}$, on the diagonal (these numbers are called the singular values of matrix A).

## 4.3 Schmidt Decomposition

Suppose we have a pure quantum state in the Hilbert space $H_A \otimes H_B$, $d_A = \dim(H_A)$, $d_B = \dim(H_B)$, $d = \min(d_A, d_B)$. A finite-dimensional Hilbert space is isomorphic to its conjugation, and for this reason there is a possibility to consider one instead of another. We replace the space $H_B$ with $H_B^*$.

This technique is intended for simulation of entangled quantum computation in quantum algorithms, as well as for the study of the influence of the entanglement level on the running of quantum algorithms.

Quantum algorithm $Q$ [6] is a classical algorithm $A$, which can be implemented on classic computing devices and according to the input $x$, the quantum description scheme $C_x$ implements a universal finite basis, which realizes $U_x$ operator on $n_x$ qubits and the description of result register $S_x$.

Let's propose the following principle of the method: starting with β = 0 (no entanglement) the running of a quantum algorithm is simulated, then β increases to a certain value and quantum algorithm is simulated again. To do this, you must synthesize the matrix of the operator in which you can adjust the degree of entanglement. Therefore, we use the following operator:

$$S = \exp\left(i\frac{\beta}{2}\sigma_1^{\otimes N}\right) = \frac{1}{\sqrt{1+\beta^2}}(\sigma_0 \otimes \sigma_0 + i\beta\sigma_1 \otimes \sigma_1) \tag{6}$$

where $\beta$ is the parameter determining the degree of qubit entanglement, which is set in the range $[0;\ \pi/2]$; $\sigma_1$ and $\sigma_2$ – the Pauli matrices.

This actions are repeated until the entanglement is become maximum ($\beta = \pi/2$). Quantum algorithm assumes influence on the qubit or qubits (quantum register) by unitary operators (gates). These operators are prepared beforehand: some of them are taken as they are, others are synthesized with the help of the tensor product and matrices multiplication. The tensor product is necessary for the preparation of operators that will be applied to the register. All gates and operators (except for entangling operator (6)) will be the same in all cycles of the algorithm, so they should be prepared before the start of the cycle, in order to reduce the simulation time. The inception of any quantum algorithm must start with the alignment probabilities of all states and their transfer to the basis states |0 > for the further implementation of the superposition state for the existing quantum register.

The algorithm (Fig. 4) includes a cycle to increase gradually the degree of entanglement from a small value to a maximum one and then to record the response of testing algorithm. At the beginning of each cycle it is necessary to create qubits with the degree of entanglement given by $\beta$ or construct the operator, with the help of which the qubits in the future can be confused with the degree of $\beta$, mentioned above (6).

It should be emphasized that this cycle will be executed at least one time, as there is no initial maximum degree of entanglement due to the decoherence of surrounding environment. Having created intricate qubits, we measure them in terms of the von Neumann entropy. You must do this immediately after the creation of entangled qubits while the qubit vector not changes as a result of processing. Next, create a vector of qubits, or qubit register which will be used for simulation. At this stage the preparation is completed, there are all the necessary gates (matrices) and a register with the required amount of qubits (vectors).

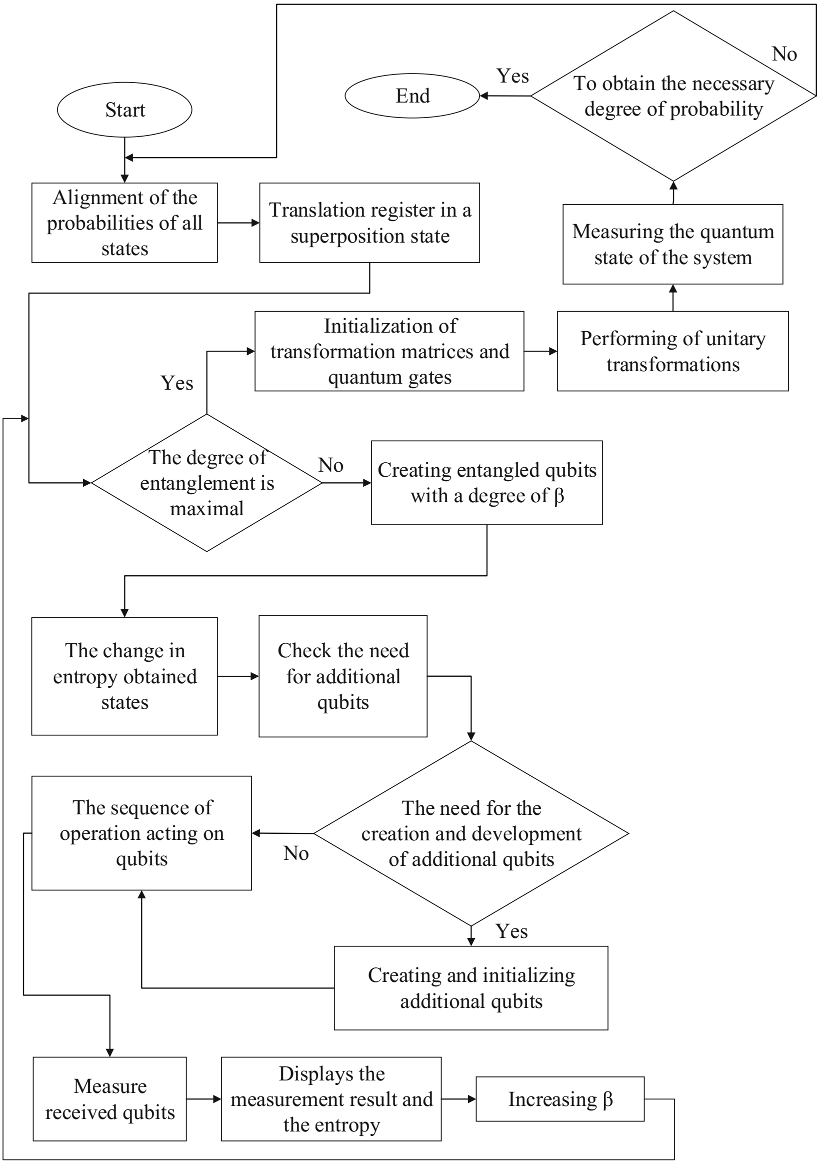

Start
Alignment of the probabilities of all states
Translation register in a superposition state
Initialization of transformation matrices and quantum gates
Performing of unitary transformations
Measuring the quantum state of the system
To obtain the necessary degree of probability
Yes
No
End
Yes
The degree of entanglement is maximal
No
Creating entangled qubits with a degree of β
The change in entropy obtained states
Check the need for additional qubits
The need for the creation and development of additional qubits
No
Yes
The sequence of operation acting on qubits
Creating and initializing additional qubits
Measure received qubits
Displays the measurement result and the entropy
Increasing β

**Fig. 4.** Method of modeling of entangled quantum computing

After the quantum algorithm was finished, qubit or qubits register is measured using a specially synthesized function, which is was described above. As a result of the decoherence instead of superposition we obtain the classical binary value with the probability specified in superposition. Then we save or output the degree of entanglement measured with the help of the von Neumann entropy. At the end of a cycle increment $\beta$, thus as a next step, the algorithm will be modeled with a greater degree of

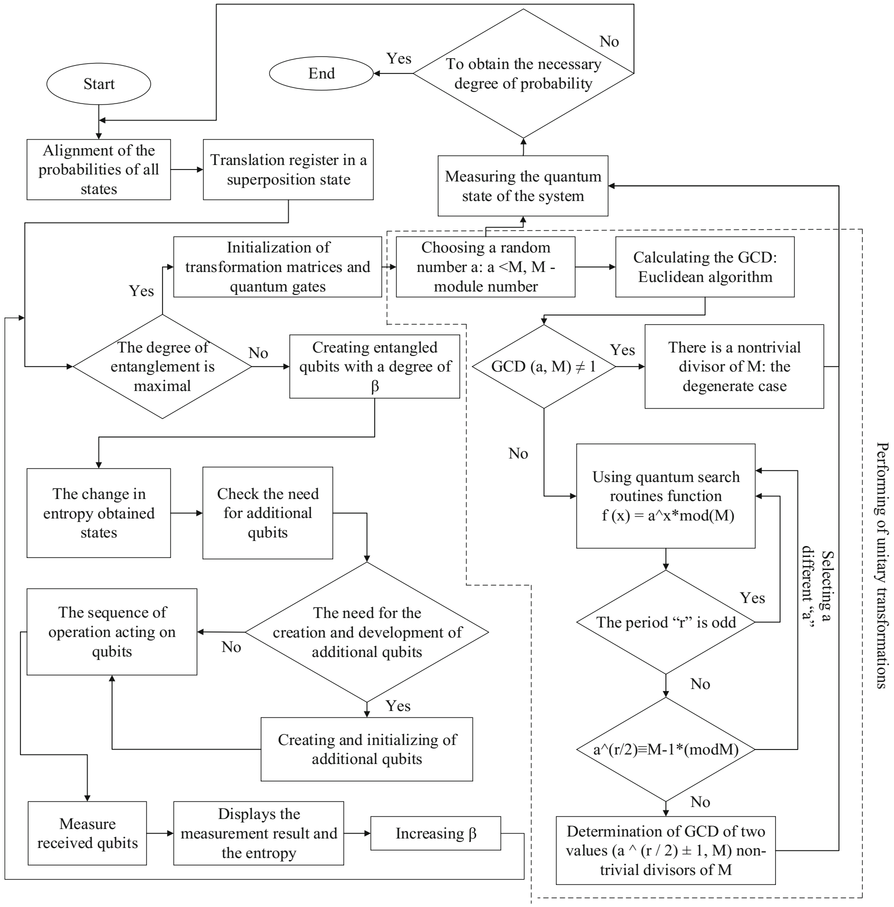

Start
End
Yes
No
To obtain the necessary degree of probability
Alignment of the probabilities of all states
Translation register in a superposition state
Measuring the quantum state of the system
Initialization of transformation matrices and quantum gates
Choosing a random number a: a <M, M - module number
Calculating the GCD: Euclidean algorithm
Yes
The degree of entanglement is maximal
No
Creating entangled qubits with a degree of β
GCD (a, M) ≠ 1
Yes
There is a nontrivial divisor of M: the degenerate case
No
Using quantum search routines function f (x) = a^x*mod(M)
The change in entropy obtained states
Check the need for additional qubits
Performing of unitary transformations
Selecting a different "a"
Yes
The period "r" is odd
The sequence of operation acting on qubits
No
The need for the creation and development of additional qubits
No
Yes
a^(r/2)≡M-1*(modM)
Creating and initializing of additional qubits
No
Measure received qubits
Displays the measurement result and the entropy
Increasing β
Determination of GCD of two values (a ^ (r / 2) ± 1, M) non-trivial divisors of M

**Fig. 5.** Method of modeling of entangled computing by implementing factoring Shor algorithm

entanglement. If the result is dependent on the random variables, such as in a model of quantum teleportation, the quantum algorithm should be repeated enough times with one level of entanglement, using the average result to eliminate the random factor, and only then to increase the degree of entanglement.

The Fig. 5 shows a simulation procedure of entangled quantum computing described above and the implementation of the algorithm Shor's factorization.

## 5 Conclusion

Modeling and implementation of entangled quantum computing in the field of quantum algorithms, based on suggested method, allows to:

- Predict and analyze the behavior of a quantum algorithm for the partial confusion that may arise under the influence of the environment on the quantum system.

Quantum systems cannot be fully separated from the environment, so this prediction is important in any algorithm with entanglement;

- Visually describe a universal method of realization of quantum algorithms based on varying degrees of entanglement;
- Find new ways to use partial for modeling any parameters of an executable task.

In the process of this research we analyzed the scheme of the elements of quantum information theory and the place of the concept of quantum entanglement. A methodology for modeling entangled quantum computing in the field of quantum algorithms, which is a complete the stage's sequence of the universal quantum algorithm. Study influence of entanglement level on its work.

**Acknowledgments.** This work was financially supported by the Ministry of Education and Science of the Russian Federation under Project No. 2.3928.2017 (Project part of State task).

## References

1. Guzik, V., Gushanskiy, S., Potapov, V.: Quantitative characteristics of the entanglement degree. Izvestiya SFedU. Eng. Sci. **3**, 76–86 (2016). SFedU Publishing, Taganrog (in Russian)
2. Kondrashin, M., Yakovlev, V.: Elements of Quantum Informatics Textbook. Moscow Engineering Physics Institute, Moscow (2004). (in Russian)
3. The Shortest Introduction to Quantum Computing. http://eax.me/quantum-computing-intro/. Accessed 23 Jan 2016
4. Guzik, V., Gushanskiy, S., Polenov, M., Potapov, V.: Models of a quantum computer, their characteristics and analysis. In: Proceedings of the 9th International Conference on Application of Information and Communication Technologies (AICT 2015), pp. 583–587. IEEE Press (2015)
5. Gushanskiy, S., Potapov, V.: Definition and implementation of the operators of quantum algorithms. Juvenis Scientia **2**, 38–40 (2016). (in Russian)
6. Guzik, V., Gushanskiy, S., Polenov, M., Potapov, V.: The concept and structure of the quantum algorithm. Inf. Commun. **1**, 13–18 (2016). (in Russian)

# Truly Parallel Model-Matching Algorithm in OpenCL

Tamás Fekete(✉) and Gergely Mezei

Budapest University of Technology and Economics, Budapest, Hungary
{fekete,gmezei}@aut.bme.hu

**Abstract.** The Model-driven Engineering (MDE) is coming into focus faster and faster nowadays because it can significantly simplify and accelerate the software development and maintenance processes. MDE can efficiently reduce resource requirements not only in development, but also in refactoring and maintenance tasks of complex software systems. There are several tools to support MDE. Although, these tools can deal with the average size of the currently applied domain models, the growing software systems can cause challenges in model manipulations. The growing size of systems can result in such a slow computation which cannot be accepted anymore. Therefore, more efficient model processing methods are needed. We are working on a complex, high performant model-transformation engine for MDE tools. Our solution can take the advantage of parallel computation available for example in modern GPUs. The engine is referred to as PaMMTE (Parallel Multiplatform Model-transformation Engine). In earlier publications, the architecture and functionality of our engine has been introduced and the functional correctness has also been proven. In this paper, we introduce a new pattern matching algorithm. The algorithm is truly parallel, it is scalable and more efficient than the previous version. Moreover, we analyze the current and the new pattern matching algorithms in general and the performance gain achieved. The new pattern matching algorithm can be effectively used not only in PaMMTE, but in any other cases, when high-performant pattern matching computation is required.

**Keywords:** PaMMTE · High-performant computation · Model-transformation · OpenCL framework · Software architecture · C++

## 1 Introduction

The Model-driven engineering (MDE) can efficiently simplify the software development which causes the sudden spreading of its usage in the software industry.

T. Fekete–This work is connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project. This project is supported by the New Széchényi Plan (Project ID: TÁMOP-4.2.1/B-09/1/KMR-2010-0002).

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_13

MDE works with models, which are not only created for presentation purposes anymore, but transformed, processed and often used directly or indirectly as the basis of the code generation. Therefore, it is an important and challenging part of MDE to find and apply efficient model-transformation methods. Several techniques exist; the graph rewriting-based transformation (referred as a graph transformation) is one of the most popular among them. Graph transformation is based on an NP complete problem (subgraph isomorphism) and may need serious amount of time depending on the size of the input model and the pattern to search for.

Motivated by the increasing requests for high performant model transformation engines, we analyzed the capability of existing tools. There are many studies and surveys (e.g. [1]) that collect and classify the model transformation tools, like GREAT, IncQuery or MOLA. While all of them are efficient and flexible to some extent, none of them is capable of using GPU-based parallel execution. We have decided to fill up this gap and create a new model-transformation engine supporting both usual features of model transformation and efficient GPU-based parallelism. The engine is referred to as PaMMTE (Parallel Multiplatform Model-transformation Engine). The core algorithm of a graph transformation engine lies in the pattern matching, this is the most computation intensive part. Therefore, currently we are focusing on this part. We have analyzed the working mechanism of our original matching algorithm and the architectural structure of GPUs. We have found that our original solution is not GPU-specific enough, although it has several steps applicable in parallel, we are still heavily relying on the CPU-based computation, which is the bottleneck from the performance's point of view. Therefore, we have created a completely new algorithm, which fits much more in the GPU-based world. In this paper, we present this new, truly parallel algorithm. Besides the details of the algorithm, we also present a short comparison of the original and the new algorithms.

The rest of the paper is organized as follows: In Sect. 2, the base conceptions are introduced which were assumed during the creation of the engine. In Sect. 3, for the sake of simplicity, a short overview is given about the base architecture of PaMMTE focusing on the part which is modified. In Sect. 4, main novelty of the current paper can be found in details. In Sect. 5, our theoretical conception is validated by measurements applied in a case study. In Sect. 6, we conclude and give some directions for the possible future research.

## 2 Related Work

On the market, numerous kinds of GPUs and other hardware elements can be found which have the ability to apply highly-parallel computation. Using a vendor or model-specific language and framework would need a tremendous effort and each user would need to provide the proper environment according to the available hardware. To avoid this, the OpenCL framework has been released in 2009. OpenCL is a platform independent framework which can be applied to handle the most widely used hardware uniformly (CPU, GPU, FPGA, DSP).

OpenCL is an interface defined by Khronos Group [10]. Each product vendor has its own implementation. In addition to move the computation into OpenCL devices, the computation capacity of the primary hardware (CPU) is less used producing thus a more balanced system in general.

At the beginning of our research, it was the one of center questions to be examined whether the usage of the OpenCL framework can be as good as using any other hardware specific environment. We studied this point carefully and also searched in the literature for other's results. In [8], we compared and evaluated several GPGPU-based solutions, OpenCL-based libraries and applications. Papers, like [6] pointed out that OpenCL framework can be effectively used in our research too and there are lots of optimization points which are probably different in case of variant hardware. It indicated us to collect the optimization opportunities. There are further examples in [5], which gives details how important the graph processing components are. [5] also focuses on mapping algorithms between the host and the GPU devices which is a big challenge in the effective usage of GPUs. They mapped 12 graph applications into the GPU device, studied the performance and suggested several approaches to accelerate the performance of the GPU-based algorithms.

There are further new studies which have influenced our current research: In [2], the k-Nearest Neighbor algorithm is implemented using OpenCL and CUDA. There is a big difference between the two implementations, the most emphasized is that CUDA is strongly hardware-dependent, while the OpenCL framework can be used on many hardware platforms. There are several measurements and comparisons between devices with a single CPU and devices with a CPU and a GPU. Furthermore, there is also a comparison between OpenCL and CUDA: in some measurements OpenCL seems to be perform better, but not in all cases. The thesis [3] compares several hardware and software solutions. One of the main reasons of the rapid spreading of highly-parallel computation (and the growing number of the computation units) is the expensive computation requirements in computer games. [3] gives an overview and compares not only the main competitors of the OpenCL framework (e.g. CUDA), and it also presents a wide benchmark. It highlights several solutions for parallel computation (e.g.: CUDA, OpenGL, DirectX, OpenMP). [4] provides the usage of the OpenCL with some C++ and STL related features (meta-programming) as part of the official Boost library. This part of the Boost library can also be used as a thin wrapper.

Taking all advantages (e.g. platform independence) and possible disadvantages (e.g. performance loss) into account finally we have decided to build our engine based on OpenCL. Because of the importance of performance, we use the OpenCL interface directly (not through a high level wrapper) and fine tune the performance keeping in mind that our final goal is to create a high-performant model transformation engine.

## 3 Architecture of PaMMTE

The architecture and working mechanisms of PaMMTE are complex. In [7], we elaborated them in details, however, in this paper, we give only a short overview, since the focus is on the new matching algorithm. The architecture of PaMMTE can be divided into three layers (Fig. 1):

**(i) Model-transformation Logic Layer (MTLogic Layer):** The model-transformation process is split up into three main model-transformation steps, all of them are implemented in the highest layer. Input model is read, processed and the output is also evaluated in the this layer. The three main steps of the model-transformation logic are the followings in the order of the first calls: pattern matching (PatternMatcher package), attribute processing (AttributeProcessor package) and finally the rewriting of the graph (ReWriter package). Note that: between the packages, in the highest layer, the model-transformation data (MTData) is passed again and again. MTData contains the processed input domain model, the pattern to be found and the results of the actual steps.

**(ii) Model-transformation Library Layer (MTLib Layer):** The middle layer contains the concrete pattern matching, attribute processing and model-rewriting algorithms which manage the core computations based on the OclAccessing Layer. Pattern-matching algorithm searches only for topological matching by using the symbol of the nodes in the input graph created from the input domain model. This is later extended and re-checked by the attribute processing step, where we check the attributes of the given nodes as well. The focus of the current study is the pattern-matching part of the engine. In MTLib, we introduced a common interface for pattern matching, attribute processing and model rewriting algorithms, thus each of them can be easily exchanged.

**(iii) OclAccessing Layer:** The lowest layer is a kind of abstraction layer. Other layers have no information about the type and the number of the currently used OpenCL devices, all of these are hidden from higher layers. Each OpenCL-based computation is managed via a general context provided by the OclAccessing Layer.

**The presented architecture has several advantages:** (i) The domain-related logic is implemented only on the highest level separated from the core algorithms (which are in the middle layer). Adding new model-transformation steps and changes in the hard coded configuration can be managed safely and easily. (ii) The core algorithms can be exchanged using a common interface and it can be replaced at run-time too. (iii) There is no hardware dependency, because the OclAccessing Layer provides a general context to the computation libraries. (iv) Implementation of the main interface of the modelPocessing package allows us to easily use new domain models. To achieve high-performant computation only C++14 is used in the implementation of PaMMTE and the only 3rd party dependency is the Boost library. To configure the engine, an XML file must be used following a predefined schema. The main development environment is the MS Visual Studio 2015 C++. During the development, we used elements from the Test-Driven Development (TDD) methodology to prove the correct functionality of the engine. In this paper, we focus on the improvement of the pattern-Matching package in order to create a generally usable truly pattern matching algorithm. We also compared two pattern matching solutions for OpenCL devices in [8]. In that research, we used the advantages of run-time information during

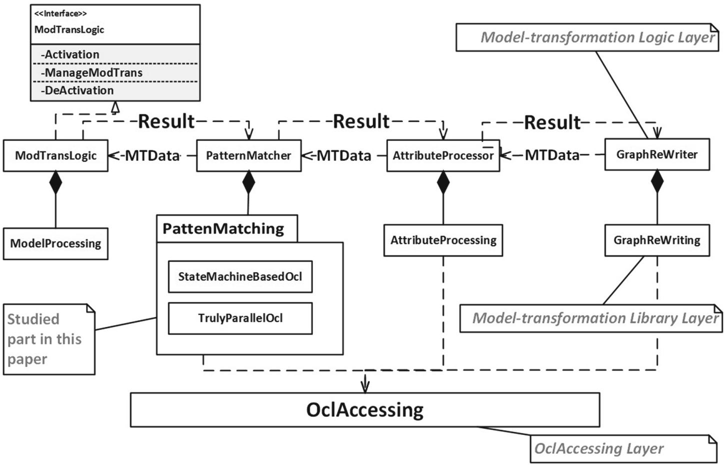


**Fig. 1.** The base architecture (packages) of PaMMTE.

building the kernel code. Now, the main goal is to achieve a truly parallel pattern matching to increase the performance of the whole model-transformation.

## 4 Towards the Truly Parallel Pattern Matching Algorithm

In [8], we introduced the importance of pattern matching algorithms and pointed out some open issues in pattern matching to be solved later. Since then, we managed to create a truly parallel pattern matching computation in our engine which is the main contribution of this paper. The main steps of both the old and the new algorithms are listed, as well as, the most important commonalities and differences.

### 4.1 Properties and Issues of the Old Algorithm

In the old algorithm, the OpenCL kernel was executed several times when result buffer overflow has occurred. We defined formulas to calculate the optimal size of the output result buffer and also designed and implemented a strategy, when the buffer is rarely used. The number of the threads equals to the number of the nodes in the input graph to be processed (each thread processes exactly one node). In case of buffer overflow, only those elements are re-used, which were not processed before, when the kernel is executed again. Avoiding to avoid processing one node two times does not significantly saves any time, because other threads must be waited. Although, in some cases, one node can have only a few neighbors, which results in a fast computation for those threads, there can be other threads with lots of neighbors. The kernel computation cannot be finished from the viewpoint of the host until each thread has finished the task assigned to it.

As it can be seen, the old algorithm works with parallel threads, but not in an efficient way. In some cases, only a few threads work. Another viewpoint is that, each thread has a complex inner state. The state has to store the parent node, the currently processed neighbor number, the deepest level, which means how far the actual node is from the pivot point and so on. In the current paper, we use the term of candidate multiple times with the following meaning: *candidate is part of the input graph which is supposed to match to a part of the pattern. Moreover, if the size of the candidate equals to the size of the pattern and they are still matching, the candidate is already a matching result.* Each thread tries to create a small candidate at the beginning and checks whether that candidate is matching or not. If the candidate is matching the thread takes a new neighbor to extend the size of the actual candidate and checks again the matching state. This process is applied until then the size of the matching candidate equals to the size of the pattern. To find results, in the implementation of the first kernel, there are two nested state machines (the first digs deeper in the graph while the second finds each neighbors at current level) and two function calls (to validate the candidate) which are not the best way to achieve optimal performance for a data oriented computation model such as used in OpenCL. In short, the old algorithm was executed semi-parallel.

### 4.2 The New Pattern Matching Algorithm

**Overview of the New Pattern Matching.** To evaluate the matching algorithm truly in parallel and avoid buffer size estimation, we created a new kernel source code. There are some lower (hardware) level assumptions which are considered in order to achieve the truly parallel functionality and increase performance: (i) Compare two numbers is fast on hardware level (in general, hardware computation units use gates instead of bit evaluation one-by-one to compare two numbers). (ii) The memory allocation is fast both on the host and on the OpenCL side. However, on host side, handling of memory fragmentation is required because of the big amount of data processed. The new approach is illustrated in Fig. 2. To find all results, the kernel is executed several times (two executions in this particular case). We use four buffers, but only these buffers are changed during the computation. Others, graph or pattern related buffers do not change during matching. The four buffers are the followings: (i) FH1 - first helper buffer, (ii) FB1 - first result candidate buffer, (iii) SB2 - second candidate buffer, (iv) SH2 - second helper buffer. The role of the buffers is explained later.

**Behaviour of the New Pattern Matching.** Since in the general case, the kernel is executed several times, let us suppose that the current loop number is N (kernel is already executed N-1 times successfully). Now, the following steps evaluated:

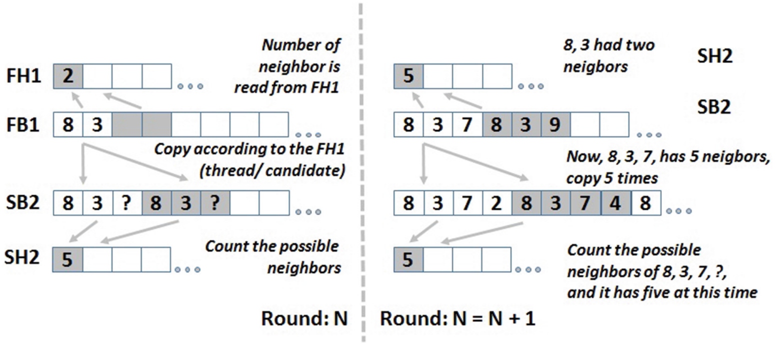

FH1
2
Number of neighbor is read from FH1
FB1
8 3
Copy according to the FH1 (thread/ candidate)
SB2
8 3 ? 8 3 ?
SH2
5
Count the possible neighbors
Round: N
8, 3 had two neigbors
SH2
5
SB2
8 3 7 8 3 9
Now, 8, 3, 7, has 5 neigbors, copy 5 times
8 3 7 2 8 3 7 4 8
5
Count the possible neighbors of 8, 3, 7, ?, and it has five at this time
Round: N = N + 1

**Fig. 2.** Buffers in the new algorithm.

**(i) Determine the size of the FH1 and FB1 buffers:** We store the data of M candidates here. The length of the candidates is N, the size of the firstCandidateBuffer (FB1) is N*M. The values of the elements of FirstHelperBuffer (FH1) denote how many new neighbor each candidate can have. Since, we need only one value for each candidate, the size of FH1 is M.
**(ii) Determine the size of the SB2 and SH2 buffers:** As far as FH1 is cumulated on the host, let C refer to the last value of the last element of the cumulated helper buffer (FH1). In this case, the size of the secondCandidateBuffer (SB2) is C*(N + 1). The meaning of secondHelperBuffer (SH2) is similar to that in the previous case: it shows how many new neighbors the candidates have. The size of SH2 is C size is C.
**(iii) Copy candidates from the FB1 to the SB2:** Each thread is responsible for exactly one candidate. The thread copies that candidate and adds the new neighbor. The thread checks whether the filled up candidate matches. If yes, it computes the number of the possible new neighbors and stores the number of neighbors in SH2.
**(iv) Change the pointers on the host side:** On the host side, the pointers of the FB1 and the SB2 are exchanged. Similarly, the SH2 is replaced with the FH1.
**(v) Prepare the new result buffers:** The first helper buffer is read from the global memory of the OpenCL device and cumulated on the host side. Based on the accumulation, we re-calculate the sizes of SB2 and SH2, moreover, the number of the threads received from the result of the cumulating are also recalculated. The SB2 and the SH2 are freed and new arrays are allocated with empty content.

**Details About the New Pattern Matching.** The kernel binary always works on the four buffers, it reads FH1 and FB1 and writes SH2 and SB2. The host manages two important steps before calling the kernel. Firstly, it cumulates the numbers in the first helper buffer, then it swaps the pointers of the first and second buffers. The kernel always works from the first buffers and saves the result to the second buffers: (i) The kernel copies the candidates from the first

**Table 1.** Average time values of several measurements are collected (only for the pattern matching).

| Platform | Intel (time) | Nvidia (time) |
|---|---|---|
| New OpenCL Kernel | 112 | 132 |
| Old OpenCL Kernel | 15761 | 15788 |
| Host version of the old one | 1264 | |

buffer to the second buffer and adds to the new neighbor using the helper buffer and the thread id (each thread uses the same formula to find the place of the old and the new neighbor candidate). The number of the threads equals to the number of new candidates. Each new thread knows its base candidate and copies the candidate from the first buffer to the second buffer (each thread copies the same number of elements). (ii) The thread knows which neighbor must be taken from the input graph to the new empty place. (iii) The thread verifies whether the new candidate is matching according to the pattern. In case of mismatching, the thread sets that the number of possible new neighbors to zero. If the new candidate is matching, the thread adds how many new neighbors must be checked in the next loop. Finally, the new candidate buffer is created.

## 5 Case Study

To test and evaluate the performance gain of the new algorithm, we measured the computation time in a case study. We selected the Internet Movie DataBase (IMDb) [9] as the target domain. IMDb is the largest film and TV show related database which is publicly available. It has approximately 3.3 million titles and 6.5 million personalities (actors, directors, etc.). IMDB contains information on several domain concepts, like movies (subtitle, creation time, and rate), actors (with movies they played in) and producers (with their movies). A simple example for a pattern to be searched is: *"Three actors playing in the same movie. The movie has an attribute showing that the movie is made in the USA. Furthermore, the first name of the director is Jack and at least one actress (besides the three actors) must play in the same movie."*. As far as the first logical step of the model-transformation works only with the symbols of the nodes, the other part of the rewriting rules are not considered now. Similarly, only the pattern matching part is measured in the current case study.

The test environment used in the case study is a simple notebook with the following configuration: Intel Core i7 HD Graphic 5500 and Nvidia Geforce GTX 950M with a Windows 10. In Table 1, the results are collected. Note that the results are the average of ten measurements. In this case study, we measured and compared three pattern matching algorithms: (i) the old algorithm, (ii) the old version executed on the host (CPU) and (iii) the new algorithm. Although, we do not intend to compare two different kinds of architectures/technologies, it also can be seen that even a relatively week GPU can result in faster computation

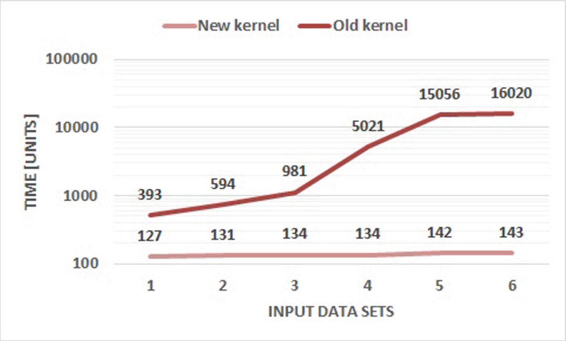


**Fig. 3.** Time results in case of different input models.

that a strong CPU. Among the algorithms, the first one uses the smallest input data and the simplest pattern. The last one uses the biggest input data and the most complex pattern.

In Fig. 3 the time of the execution is compared for the old and the new algorithms in case of different input domain models (the pattern to be searched is not changed). In case of the old algorithm, the computation time is directly proportional to the size of the input domain model. Bigger and more complex input graphs require more time to find each matching, because of the complex inner state machines and growing complexity. In contrast, the new algorithm is completed almost in the same time at each measurement. The reason of the almost constant time is the truly parallel behaviour of the new algorithm. However, it is suspicious that a remarkable portion of the time is caused by the constant time required by preprocessing steps (e.g. memory copy) and thus the real calculation needs increasing amount of time. We need a series of measurements in order to examine this question in detail. Note that this behavior does not change the fact that the new algorithm is several orders of magnitude faster than the old one.

## 6 Conclusion

The MDE approach is gaining more and more interest nowadays as we have to deal with bigger and bigger software systems. There are several tools to support model-based development however, existing tools does not support parallel execution natively. Our aim is to solve this issue. We have created an engine for model-transformation which is based on a new approach to achieve high-performant multiplatform computation. The base and the novelty of our engine is the usage of the OpenCL framework to significantly accelerate model-transformation. Pattern matching is one of the most important parts of a model-transformation engine. In the current research, we improved the pattern matching part of PaMMTE. The solution is presented, analyzed and validated by measurements using a case study. Other model-transformation steps like the attribute processing and the model rewriting are not changed in this paper.

By changing the pattern-matching algorithm, we could achieve a truly parallel model-transformation engine, which is deployed in PaMMTE, but it can also be used in any other similar cases. In the future, we are going to search for further case studies and to test the solution with other platforms to ensure that our solution works well with most of the OpenCL devices. Furthermore, we must study the advantages and the disadvantages of the algorithms.

## References

1. Jakumeit, E., Buchwald, S., Wagelaar, D., Dan, L., Hegedus, A., Herrmannsdorfer, M., Horn, T., Kalnina, E., Krause, C., Lano, K., et al.: A survey and comparison of transformation tools based on the transformation tool contest. Sci. Comput. Program. **85**, 41–99 (2014)
2. Masek, J., Burget, R., Povoda, L., Dutta, M.K.: Multi-GPU implementation of machine learning algorithm using CUDA and openCL. Int. J. Adv. Telecommun. Electrotechn. Sig. Syst. **5**(2), 101–107 (2016)
3. Sorman, T.: Comparison of technologies for general-purpose computing on graphics processing units (2016)
4. Szuppe, J.: Boost.Compute: a parallel computing library for C++ based on opencl. In: Proceedings of the 4th International Workshop on OpenCL, p. 15. ACM (2016)
5. Xu, Q., Jeon, H., Annavaram, M.: Graph processing on GPUs: where are the bottlenecks? In: 2014 IEEE International Symposium on Workload Characterization (IISWC), pp. 140–149 (2014)
6. Yan, X., Shi, X., Wang, L., Yang, H.: An openCL micro-benchmark suite for GPUs and CPUs. J. Supercomput. **69**(2), 693–713 (2014)
7. Fekete, T., Mezei, G.: Architectural challenges in creating a high-performant model-transformation engine. Subsequences. In: The 10TH Jubilee Conference of PhD Students in Computer Science, p. 20 (2016)
8. Fekete, T., Mezei, G.: Creating a GPGPU-accelerated framework for pattern matching using a case study. In: 24th High Performance Computing Symposium (HPC16), Pasadena, CA, USA (2016)
9. IMDb - Movies, TV and Celebrities - IMDb (2016). http://www.imdb.com/interfaces
10. OpenCL - The open standard for parallel programming of heterogeneous systems (2016). https://www.khronos.org/opencl

# A Novel Co-channel Deployment Algorithm Based on PCF in Multiple APs and High Density WLANs

Jianjun Lei(✉) and Jianhua Jiang

School of Computer Science and Technology,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, China
Leijj@cqupt.edu.cn, Jiangjianh@foxmail.com

**Abstract.** In Wireless Local Area Networks (WLANs), co-channel deployment is desirable for mitigating the collision and interference among different access points (APs). Therefore, it can bound channel access delay and improve network capacity. In this paper, we present an interference model for multiple APs co-channel deployment. And then we formulate the channel assignment problem into a time slot allocation problem and propose a co-channel assignment strategy, which includes two procedures: (1) Assigning the channel based on the vertex coloring algorithm. (2) Making extra polls to improve the channel assignment by applying time slot reservation strategy. The simulation results show that our proposed algorithm can improve the network throughput and transmission delay by 8% and 15% respectively compared with the TMCA scheme.

**Keywords:** WLANs · PCF · Time slot · Channel assignment · Co-channel deployment

## 1 Introduction

In recent years, with the development of wireless technologies, the demand of WLANs has risen and its deployment has become denser and denser. The wireless APs are placed everywhere in our daily life environment [1, 2]. A plenty of APs will cause serious collision, influencing clients' experience. Hence, eliminating interference among APs under the limited number of channels has become an important issue.

Typically, the clients for IEEE 802.11 operate on the distributed coordination function (DCF) mode, in which clients use the carrier sense medium access with collision avoidance (CSMA/CA) mechanism [3] to compete for the wireless medium to transmit data. Due to the contention nature and throughput fair property of CSMA/CA, it is difficult to provide guaranteed bandwidth and bounded access time for each client. The problem becomes even more challenging in high density and multi-AP WLANs. Even though such interference among neighboring APs can be mitigated by assigning them with different channels, which cannot be resolved completely for only limited channel resources are available.

This paper proposes a novel co-channel deployment algorithm to reduce the interference, which formulates the channel assignment problem into vertex coloring

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_14

problem. Furthermore, an improved channel assignment strategy, which makes extra poll mechanism for APs without collision, is presented to improve the channel utilization and guarantee the bandwidth demand, especially in high density WLANs.

The rest of this paper is organized as follows: we survey some related works in Sect. 2; in Sect. 3, the interference model and conflict constraint are introduced; the basic channel assignment strategy and the improved channel assignment strategy are presented in Sect. 4; after evaluating the performance of the proposed algorithm in Sect. 5, we make conclusion in Sect. 6.

## 2 Related Work

In WLANs, to degrade the interference, the researchers usually assign non-overlapping channels to neighboring APs. Achanta proposes the classic Least Congested Channel Search (LCCS) [4] algorithm. In the algorithm, APs scan each channel and detect the condition of data transmission in all channels. APs can obtain the load level in each channel and choose the channel in which the load is lighter and the interference is smaller. Since this method only assigns the channel statically on the AP side, the performance may drop sharply under the high density environment. In [5], the author proposes a kind of global channel allocation method based on graph theory. The main idea is to select the vertex according to the saturation and adjacent degree of vertices to color from the available color sets. A similar channel assignment scheme is proposed in [6], where the expected transmission delay due to the interference from a neighboring BSS is regarded as the weight of the edge. However, both of the methods do not consider the competition between different APs and load distributions. Thus, they cannot eliminate interference completely due to a limited number of channels available.

Other literatures focus on eliminating the interference by utilizing the PCF (Point Coordination Function). In [7], the author proposed a kind of collision-free client polling method to mitigate the conflict and interference, which polls the clients in a time slot manner. Though the interference is eliminated effectively, this method needs too much time slots to poll all clients so that the utilization of channel is degraded. In [8], a client polling framework, named MiFi, is proposed for multi-AP deployment, in which CFPs (Contention-free Periods) at all APs are synchronized and divided into time slots. Neighboring APs are assigned with different time slots, so as to poll all clients without collisions. Nevertheless, the constraints in the AP interference model are too much, suppressing the concurrent polls, and thus resulting in aggregated throughput decrease.

Yet, these algorithms mainly concentrate on allocating non-overlapping channels to APs while the coordination mechanism between APs is ignored. Hence, we propose a collision-free co-channel deployment algorithm, which consists of two procedures: (1) Conducting a basic channel assignment by vertex coloring algorithm that determines the minimum number of time slots. (2) Constructing an improved channel assignment by making extra polls for APs to improve the channel utilization. Additionally, since all APs in our algorithm use the same channel, we also can deploy multiple vertically overlapped networks in the same location, by assigning the non-overlapping channel to each network, which can greatly boost network capacity.

# 3 Interference Model and Conflict Constraint

## 3.1 Interference Model

In WLAN, all APs are operating in the PCF mode and there would not be any interference between two clients in the same AP. Two clients from neighboring APs would interfere with each other. Thus, $I_e$ is defined to denote the interference probability for every pair of APs. Assume that two APs of $AP_m$ and $AP_n$ have $S$ clients. Then, $I_e$ is estimated by Eq. (2).

$$I_e = \frac{\rho_1 + \rho_2 + \cdots + \rho_s}{S} \tag{1}$$

Equation (2) is subject to Eq. (3).

$$\rho_s = \begin{cases} 1, \; if \; \exists(i,j) \in \Phi_m \times \Phi_n \; s.t. \; i,j \; interfere \\ 0, \; otherwise \end{cases} \tag{2}$$

Where, $\rho_s$ describes the interference relationship. $\Phi_m$ and $\Phi_n$ are the sets of clients associating with $AP_m$ and $AP_n$ respectively. Note that the threshold $I_{th}$ can be adapted according to the network environment.

## 3.2 Conflict Constraint

Only one client in the same AP is allowed to transmit and receive data within a time slot. And, interfering clients from neighboring APs should be polled in different time slots to avoid transmission collisions. Hence, a variable $T_i$ is defined to denote the time slot assigned to client $i$ and then such conflict constraint can be expressed as follows.

$$1 \le T_i \le R, \qquad \forall i \in R \tag{3}$$

$$T_i \ne T_j, \; if \; I_e = 1, \qquad \forall i,j \in C \tag{4}$$

$$T_i \ne T_j, \; if \; A_m = A_n, \quad \forall i,j \in C, \; \forall m,n \in A \tag{5}$$

The conflict constraint of concurrent polling can also be described by a polling conflict graph $G = (V, E)$, where each vertex represents an AP and there is an edge between two vertexes if the interference probability $I_e$ exceeds the threshold $I_{th}$. Figure 1 gives an example of the polling conflict graph.

Note that the problem of polling conflict graph is equivalent to the vertex coloring problem. Furthermore, Coloring a vertex in the graph is equivalent to allocating time

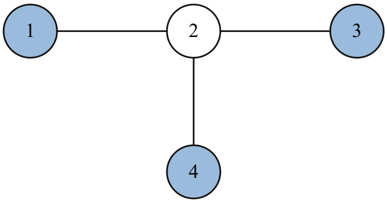


**Fig. 1.** An example of polling conflict graph

slot to the corresponding AP in the network. This is a *K-colorable* problem in graph theory, which is also a well-known NP-hard one.

## 4 Channel Assignment Algorithm

In this section, we propose a *co-channel assignment* (CCA) strategy based on vertex coloring, which includes a basic channel assignment and an improved channel assignment procedure.

### 4.1 Basic Channel Assignment

The *basic channel assignment* is aimed to reduce the interference between APs by allocating time slots to all APs in a collision-free manner based on PCF, with two components. The first one is the coloring algorithm given a graph $G = (V, E)$ and the number of colors *R*, which seeks out a feasible color assignment with minimal number of colors. The second component uses the coloring scheme to work out an efficient slot assignment, which is described as time slots assignment matrix.

To solve the vertex coloring problem, a mass of heuristic algorithms have been proposed in graph theory and many of them have been applied into practice, such as the maximal independent set (MIS) [9] and the EXTRACOL [10] which both can provide the optimal results. Thus, in this paper, the MIS algorithm will be applied on solving the problem of the vertex coloring graph to realize the basic channel assignment. The general procedure is described as follows.

```
Algorithm1: the basic Channel Assignment Procedure

Input:
Polling Conflict Graph $G=(V,E)$;
Time Slots Assignment Matrix $P$;
Output:
Time Slots Matrix $P'$;
Procedure:
Initiate the element of matrix $P$ to zero and time slot
$T_r = 0$
while( $|V| > q$ )
 $I_m = MIS()$ {Use MIS algorithm to find a maximal
independent set in $G$ }
end while
  Find the maximal independent sets as $\{I_1,\ldots,I_r\}$.
for $i = r$ down to 1
  Assign $r$ minimum available colors to $\{I_r\}$ and set $T_r = 1$ of
$\{I_r\}$.
end for
```

First, we construct the interference collision graph by using the proposed interference model. Second, the MIS algorithm is applied to solve the coloring problem repeatedly and the coloring scheme is recorded. After that, the time slots assignment matrix is established according to the coloring scheme and all APs poll its associated clients in the assigned time slots.

### 4.2 Improved Channel Assignment

Note that in the basic channel assignment, part of APs would be idle in some time slots. To utilize such idle time slots adequately and improve network performance, we propose an improved channel assignment that is called *time slots reservation* (TSR) scheme. Primarily, the AP can poll its associated clients in idle time slots by querying the polling list of all the neighboring APs. The extra polls must satisfy the conflict constraint among themselves. Therefore, we first define the variable $C_{sen}$ to denote the set of clients which are in the carrier sensing range of APs. And then, the general procedure is described as follows.

```
Algorithm2: the Improved Channel Assignment Procedure

Input:
The requirement of AP k to reserve the time slot.
Output:
A Boolean value;
Procedure:
Query the set N consisting of the neighboring APs of AP
k;
while( $m = 1 : |N|$ )
  Query the next polling client j of $N_m \left( N_m \in N \right)$;
  if(Client $j \notin C_{sen}$ ) then
    Reservation succeeds.
  else
    Reservation fails.
  end if
end while
```

First, idle APs in the next time slot query the set $N$ consisting of the adjacent APs and then query the polling list of all neighboring APs from the set $N$. After that, APs will judge whether the next polling client satisfies the conflict constraint. Accordingly, the next polling would not interfere with the neighboring AP polling.

The overhead of the proposed CCA strategy is negligible compared with data traffics in the network, for the time slot allocations for both basic channel assignment and improved channel assignment are updated only when interference or network topology changes.

# 5 Performance Evaluations

## 5.1 Simulation Methodology

In this section, we carry out the simulations by OPNET14.5, evaluate the proposed channel assignment strategy and compare it with the basic DCF method and the TMCA algorithm in [11].

In the simulation, the clients are randomly distributed in a 1000 × 800 $\mathrm{m}^2$ field and each client associates with an AP according to the largest RSSI. We set up the FTP server to simulate the real network traffic and set the uplink and downlink traffic in half respectively. Part of the experimental parameters are shown in Table 1.

The performance evaluation metrics include the delay and the throughput. The normalized throughput is also used in Sect. 5.3, which is computed by Eq. (6).

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{6}$$

Where, $X$ is the measurement data.$X_{max}$ and $X_{min}$ are the maximum and minimum of the measurement data respectively.

**Table 1.** Part of the experimental parameters

| Parameters | Value | Parameters | Value |
|---|---|---|---|
| ACK | 14 bytes | CTS | 14 bytes |
| RTS | 20 bytes | SIFS | 10 μs |
| DIFS | 50 μs | EIFS | 300 μs |

## 5.2 Simulation in Two APs Deployment

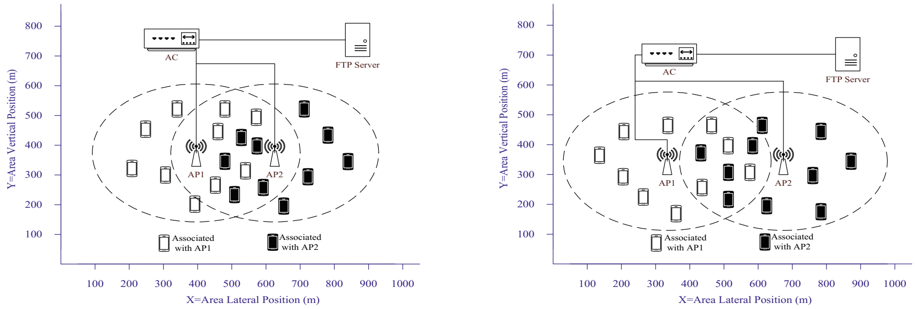
We evaluate CCA for transmission delay and average throughput with double APs deployed in the situation of tight coupling and loose coupling, which is shown in Fig. 2.

### 5.2.1 Transmission Delay

We first examine the transmission delay of the DCF and CCA strategy under various client densities. The simulation results are plotted in Fig. 3, where the number of clients increases from 5 to 30 for each AP. The transmission delay of two methods increases with the increasing of the number of clients. The reason is all clients have pending traffic and a larger number of clients lead to higher collision probability. Notably, our proposed strategy performs better compared with the DCF scheme both in tight or loose coupling deployment. The delay of CCA strategy is reduced by 45% in tight coupling and 32% in loose coupling respectively.

### 5.2.2 Network Throughput

Figure 4 shows the average network throughput changes as the number of the clients associating with each AP changes. When the number of clients for each AP is not



(a) Tight coupling (b) Loose coupling

**Fig. 2.** Network topology with double APs

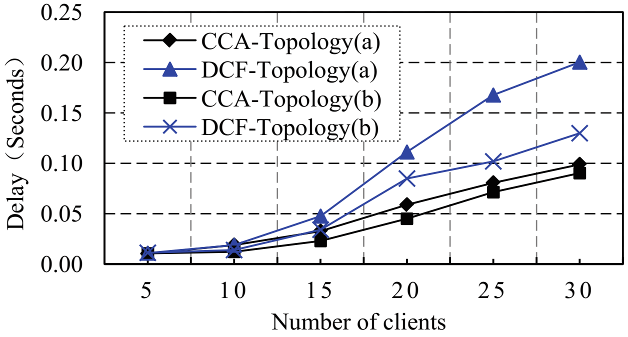


**Fig. 3.** Transmission delay under various client densities

beyond 20, the DCF method obtains better throughput compared with the CCA strategy, which may be due to little collision when the network density is low. However, the network throughput of DCF method begins to decrease as more clients compete for the channel. Our CCA strategy still keep high throughput in high density deployment.

## 5.3 Simulation in Multiple APs Deployment

We also evaluate CCA and TMCA algorithms in another two different network topologies, which is shown in Fig. 5.

### 5.3.1 Transmission Delay

Figure 6 shows the transmission delay of the CCA and TMCA algorithms under the topology depicted in Fig. 6(a) and (b). The CCA algorithm performs better and reduces transmission delay by 15% in delay compared with the TMCA algorithm. The reason is that TMCA algorithm uses binary back-off algorithm to avoid conflict, which cannot guarantee the clients to access the channel efficiently in high density deployment.

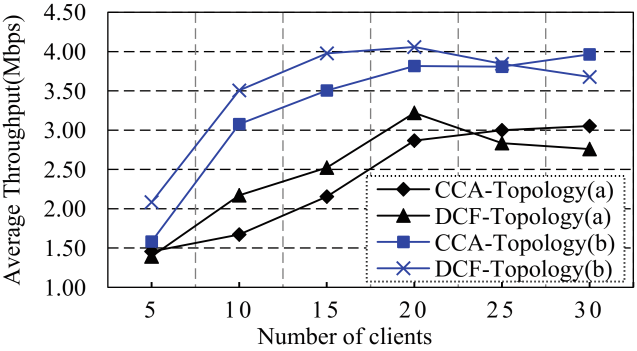


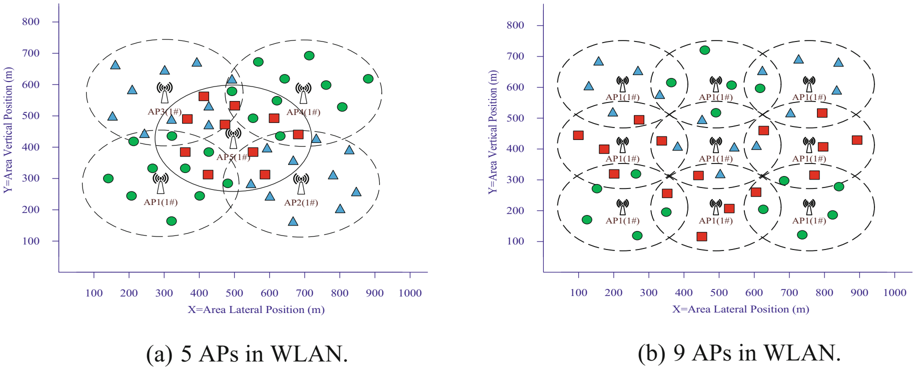
**Fig. 4.** Average network throughput under various client densities

Nevertheless, CCA strategy operates coordinately to control the channel access among APs so as to provide bounded access delay.

### 5.3.2 Network Throughput

We evaluate the normalized throughput using the proposed CCA strategy and TMCA algorithm under various client densities. As shown in Fig. 6, the number of clients increases from 5 to 30 for each AP.

Figure 7 shows the normalized throughput of the CCA and TMCA algorithms. The throughput of TMCA algorithm is slightly higher than that of the CCA strategy when the number of clients is less than 20. But as the number of client increases, the throughput of TMCA algorithm decreases while that of the CCA strategy always stays at high level and achieves 8% improvement compared with TMCA. The main reason is that TMCA algorithm only uses three non-overlapped channels available which will cause interference and conflict in high client density.



(a) 5 APs in WLAN.

(b) 9 APs in WLAN.

**Fig. 5.** Network topology with 5 and 9 APs respectively.

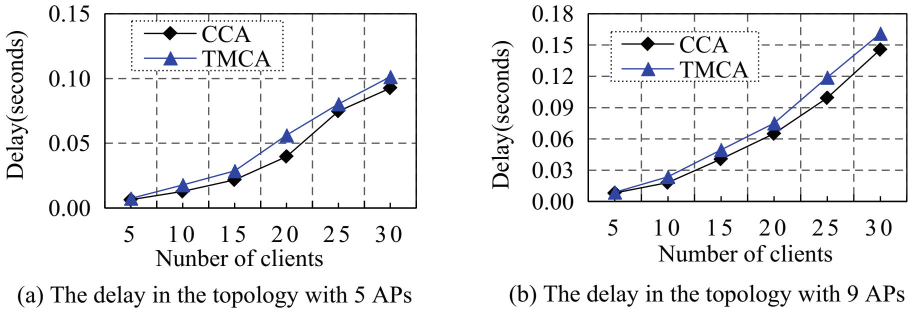


**Fig. 6.** Transmission delay in 5 and 9 APs deployment under various client densities.

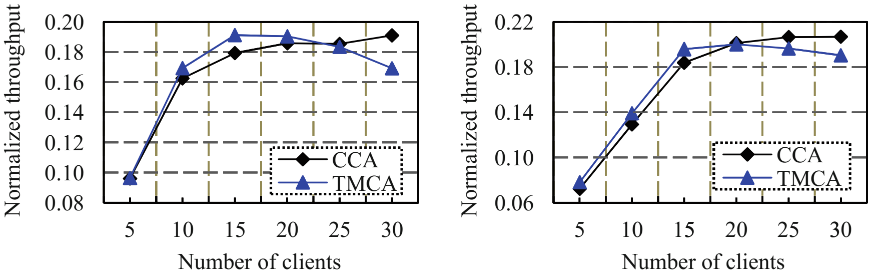


**Fig. 7.** Normalized throughput in 5 or 9 APs deployment topology of Fig. 6(a) and (b) respectively under various client densities.

## 6 Conclusions

In this paper, we have studied co-channel multiple APs deployment in PCF-based WLANs, which can provide high bandwidth and bounded access delay. We give the interference model and formulate the problem of channel assignment into a vertex coloring problem. Meanwhile, two strategies are used: (1) assigning the channel based on the vertex coloring approach; (2) making extra polls to improve the channel assignment by applying time slot reservation. Finally, we evaluate the proposed algorithm under different client density, the simulation results show that our proposed algorithm can obtain good throughput and low transmission delay.

**Acknowledgments.** This research was supported by Program for Innovation Team Building at Institutions of Higher Education in Chongqing (CXTDX201601021), the National Science Foundation of Chongqing (cstc2014kjrc-qnrc40002) and Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJ1500439).

## References

1. Fiehe, S., Riihijvi, J., Mahonen, P.: Experimental study on performance of IEEE 802.11n and impact of interferers on the 2.4 GHz ISM band. In: ACM IWCMC, pp. 47–51 (2010)
2. Gong, D., Yang, Y.: AP association in 802.11n WLANs with heterogeneous clients. In: IEEE INFOCOM, pp. 1440–1448 (2012)
3. Cuzanauskas, T., Anskaitis, A.: Multi-polling game for IEEE 802.11 networks. In: 2015 IEEE 3rd Workshop on Advances in Information Electronic and Electrical Engineering (AIEEE), pp. 1–5 (2015)
4. Achanta, M.: Method and apparatus for least congested channel scan for wireless access points. USA, 20060072602 (2006)
5. Zhang, H., Ji, H., Ge, W.: Channel assignment with fairness for multi-AP WLAN based on distributed coordination. In: Wireless Communications and Networking Conference (WCNC) 2011 IEEE, Cancun, Mexico, pp. 392–397 (2011)
6. Liu, Y., Wu, W., Wang, B., He, T., Yi, S., Xia, Y.: Measurement-based channel management in WLANs. In: IEEE WCNC, pp. 1–6 (2010)
7. Gong, D., Yang, Y., Li, H.: High-throughput collision-free client polling in multi-AP WLANs. In: Global Telecommunications Conference (GLOBECOM 2011), Houston, Texas (2011)
8. Bejerano, Y., Bhatia, R.S.: MiFi: a framework for fairness and QoS assurance for current IEEE 802.11 networks with multiple access points. IEEE/ACM Trans. Netw. **14**, 849–862 (2006)
9. Gualandi, S., Malucelli, F.: Exact solution of graph coloring problems via constraint programming and column generation. J. INFORMS Comput. **24**, 81–100 (2012)
10. Wu, Q., Hao, J.-K.: Coloring large graphs based on independent set extraction. Comput. Oper. Res. **39**, 283–290 (2012)
11. Gong, D., Zhao, M.,Yang, Y.: Channel assignment in multi-rate 802.11n WLANs. In: Wireless Communications and Networking Conference (WCNC), Shanghai, pp. 392–397 (2013)

# Determination of Optimal Cluster Number in Connection to SCADA

Jan Vávra(✉) and Martin Hromada

Faculty of Applied Informatics, Tomas Bata University in Zlin,
Nad Stráněmi 4511, Zlín, Czech Republic
{jvavra,hromada}@fai.utb.cz

**Abstract.** The recent evolution of cyber-attacks creates eminent pressure on information and communication systems. The increasing number of cyber-attacks and their sophistication have resulted in needs of the new type of cyber defense. The anomaly detection in relation to intrusion detection system (IDS) in connection with standard cyber defense technologies may be the answer to contemporary development in cyber security. Moreover, unsupervised anomaly detection based on K-means algorithm is broadly examined by a considerable number of researchers. Therefore, the algorithm is a solid selection in relation to intrusion detection system. However, one of the problems is to determine a proper number of cluster for the K-means. Nonetheless, there are methods to determine the optimal number of clusters. The aim of the article is to determine the number of clusters in relation to Supervisory Control and Data Acquisition system.

**Keywords:** Cyber security · Clusters · Anomaly detection · Supervisory control · Data acquisition

## 1 Introduction

The contemporary world has become unpredictable. The development in almost every area of human society is experiencing a rapid evolution. In the past years, the computer systems and networks passed through the fundamental change. Therefore, the information and communication system is one of the most dynamically developing sector. Moreover, the ICT has become highly interconnected with critical information infrastructure. Once completely isolated systems became open via the interconnection with business ICT. Thus, the main subgroup of industrial control system, the Supervisory Control and Data Acquisition (SCADA) systems, has become more open and less secured.

SCADA systems are often a target of sophisticated cyber-attacks known as Advanced Persistent Threat (APT). Moreover, the protective measures are often ineffective. The malware uses a variety of sophisticated techniques and is often focused on a predefined system which allows it hidden influence in the affected system. The main representatives of the APT are Duqu and Flame [1]. These examples of malware can successfully infect a variety of well-protected systems and collect important data for quite a long time. It is important to note that APT behaves as a non-lethal weapon,

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_15

which is created only for collecting information. However, Stuxnet was the first software, which can be referred as a cyber-weapon. Its core was similar to APT. However, this malware was accompanied by an executive part, which was intended to attack specific physical devices. Thus, a considerable number of security measures is inadequate. One of the way how to build reliable SCADA cyber defense seems to be IDS based on anomaly detection.

A comprehensive picture of SCADA cyber incidents is monitored and evaluated by "Industrial Security Incidents Database". The trend of SCADA cyber security is clear. The total number of the incidents is increasing since 1982. Moreover, the most affected sectors are transport, energy, oil and chemical industries. Therefore, it shows that current concept of cyber security has become insufficient. In this paper, we present the techniques designed to evaluate the appropriate number of clusters in relation to SCADA datasets. The results are examined by simple K-means algorithm.

Anomaly detection is one of the main subgroups of IDS detection methodology. The detection and identification of an anomaly can be critical for every system. The needs of IDS is discussed by Horkan [2]. He concluded that the IDS going to be an essential part of the SCADA systems. Moreover, Pollet [3] predicted increasing dependency of the SCADA systems on IT; therefore, the percentage of industrial companies utilizing the IDS will rapidly grow. The implementation of the IDS in the SCADA systems has been widely investigated (Cheung et al. [4], Verba and Milvich [5], Valli [6], Carcano et al. [7], Zhu and Sastry [8], Yang et al. [9], Maglaras and Jiang [10], Marton et al. [11]). However, these studies are mainly focused on supervised and semi-supervised anomaly detection. The deployment and operation of such IDS are restricted. Tomlin and Tomlin [12] concluded that "the unsupervised anomaly detection provides benefits over misuse detection and supervised anomaly detection. Unsupervised methods do not require prior knowledge of states or training data; thus, potentially detecting new attacks without any record of prior attacks or possible normal states." Moreover, Chiang and Chiang [13] suggest the importance of determining the right number of clusters for K-means in order to positively affect the results. Furthermore, Yang et al., [14] suggest that the current state of IDS is not entirely adapted to be widely deployed in the SCADA systems; accordingly, future research is needed [15].

The rest of the article is organized as follows. Section 2 represents basic information about SCADA systems. The detection techniques are analyzed and specified in Sect. 3. Section 4 gives an elementary overview of the methods to determine the number of clusters. The next Sects. 5 and 6 include methods and results. Finally, Sect. 7 provides the conclusion of the article.

## 2 Supervisory Control and Data Acquisition System

Supervisory Control and Data Acquisition System (SCADA) is highly centralized system which uses ICT to management, control and monitoring industrial processes. This industrial automation control system is the main subgroup of industrial control systems (IDS). The second subgroup is characterized as a geographically dependent automation control system known as Distributed Control System (DCS) [16].

Protection of Critical Information Infrastructure (CII) become a dominant issue in modern history. Contemporary society has become crucially dependent on ICT. Furthermore, SCADA systems and CII are the most important ICT systems in terms of state, environment, and population. SCADA systems can be implemented in power plants, transportation systems, dams, water management, oil production, manufacturing facilities, chemicals and, gas distribution, etc. Thus, the disruption of SCADA services may become a serious threat to life, health and basic needs of population. Therefore, we can describe SCADA system as a very important target which is highly prioritized for cyber-attacks.

## 3 Detection Methodologies

Detection methodologies have become essential for the current concept of cyber security. Moreover, detection methodologies are a major part of cyber security tools like IDS, antivirus, and firewall. They are dedicated to separate the attacker from the secured system. The detection of a cyber-attack is a primal issue of every cyber security tool. Moreover, the malicious behavior is detected based on signature or anomaly, which leads to the implementation of cyber security measures.

### 3.1 Signature Based Detection

The signature based detection is a retroactive detection methodology which is commonly used. Furthermore, this methodology is considerably effective against well-known cyber-attacks and, therefore, has a low rate of false positive detection. In the first phase, the cyber-attack must be recorded and analyzed. It is followed by the creation of signature or rule which is added to a signature database. Therefore, the signature based detection is only as good as its signature database. It can be relatively easily deceived by unknown threats like zero-day attacks or the modifications of already known cyber-attacks.

### 3.2 Anomaly-Based Detection

The anomaly-based detection is a complex and progressive methodology of detecting malicious behavior in ICT. Every deviation from normal behavior may indicate a cyber-attack against the protected system. Therefore, this methodology is able to detect new or unusual traffic that can be potentially new cyber-attack. These anomaly patterns are often referred to as outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants in different application domains [17].

Anomaly detection is used not only in cyber security but also in calling card and telecommunications fraud, false advertising, medical problems, insider threat, malicious cargo shipments, image processing, auto insurance fraud, accounting inefficiencies, auction fraud, email and web spam, tax evasion, data center monitoring, image/video surveillance, etc. [18].

### 3.3 Supervised Anomaly Detection

Supervised anomaly detection is critically dependent on labeled data which provides additional information. Moreover, it shows if the instances in the dataset are normal or anomalous. In the case of supervised anomaly detection, the whole dataset is labeled (normal data and anomalies). The predictive model based on training dataset is usually created in order to distinguish between normal and anomaly data. Moreover, the model is often based on classification techniques. The classifier must be able to manage imbalanced class distributions. Furthermore, supervised anomaly detection has high accuracy in relation with known anomalies.

### 3.4 Semi-supervised Anomaly Detection

Semi-supervised anomaly detection is based only on normal harmless operations. Therefore, the anomalies are contained only in test dataset, which is used for evaluation of classification model. Thus, the classification model is based only on normal behavior in order to identify anomalies in the test dataset. However, the model could have a high false positive rate. Thus, new legit data can be recognized as anomalies.

### 3.5 Unsupervised Anomaly Detection

Unsupervised anomaly detection is a methodology which does not need any training data or labeled data. Moreover, the main idea is based on assumption that normal behavior of the system is far more common than anomalies [17]. It is important to note that there is no difference between training and test dataset. The dataset is evaluated and sorted by distances and densities between data points. Thus, the anomalies can be separated from normal behavior [19]. In general, the unsupervised technique is more flexible than other detection techniques. Furthermore, it relies on data separation in order to assembly similar objects into subsets.

## 4 Methods to Determine the Number of Clusters

### 4.1 The Rule of Thumb

The Rule of thumb is very simple method how to determine the number of clusters. Moreover. It can be calculated by Eq. 1.

$$k = \sqrt{\frac{n}{2}} \tag{1}$$

Where n is total number of elements in dataset. However, this method is not suitable for each case. Thus, the method overestimates the number of clusters in some cases.

### 4.2 The Elbow Method

One of the oldest and verified method how to determine the number of clusters. The results are presented in graphical form; precisely speaking in graphs. Each point in the graph represents a number of clusters. Moreover, the number of clusters starts at K = 1 and gradually increasing the number by one cluster. Within groups sum of squares represents the sum of the squared deviations from the individual scores. The main goal is to compute a sum of squared error (SSE) for each cluster representation. SSE is defined as the sum of the squared distance between each member of a cluster and its cluster centroid. Furthermore, with the increasing number of clusters the SSE is decreasing. Thus, within groups sum of squares is calculated within each group which is presented in the graph. The optimal number of the clusters is chosen according to "the elbow". It is the spot in the graph where the increasing number of clusters do not provide the required decrease of within groups sum of squares [20]. The sum of squared distance between each member of a cluster is presented in Eq. 2.

$$\mathrm{W} = \sum\nolimits_{\mathrm{r}=1}^{\mathrm{t}} \frac{1}{\mathrm{p_r}} \mathrm{S_r} \tag{2}$$

Where t is total number of clusters and $p_r$ is number of points in a cluster. Moreover, $S_r$ is sum of the squared distances between each member of the cluster.

### 4.3 The Silhouette Method

Silhouette method is a commonly used technique used for evaluation of consistency within the dataset. The method provides addition information about how appropriately is data element assigned to its cluster. Hence determine which data elements lie within the cluster or merely hold an intermediate position. The average silhouette width provides an evaluation of clustering validity, and might be used to select an 'appropriate' number of clusters [21]. For each object is calculated average distance $a_i$ to all other object within a cluster and for every object outside the cluster is calculated value $b_i$ which represents minimum average distance to all object in the cluster. Consequently, the Silhouette value for the object i is calculated according to Eq. 3.

$$\mathrm{s_i} = \frac{(\mathrm{b_i} - \mathrm{a_i})}{\max(\mathrm{a_i}, \mathrm{b_i})} \tag{3}$$

## 5 Methods

The article is focused on determination of K number in relation to SCADA dataset. The results are verified via K-means algorithm in order to evaluate the deployment of the algorithm in SCADA environment. Furthermore, the Mississippi State University and Oak Ridge National Laboratory SCADA dataset was used.

The dataset consisting of 37 power system event scenarios. Moreover, the dataset is divided into natural events (8), no events (1) and attack events (28). The natural events include natural faults and maintenance. Furthermore, attack events include data injection attacks, remote tripping command injection attacks, relay setting change attacks [22]. However, for the purpose of the article one cyber-attack was chosen, early state of The Relay Setting Change.

The test bed consists of power generators G1, G2, Intelligent Electronic Devices (IED) R1–R4, and breakers BR1–BR4 for field level. Furthermore, the Phasor Data Concentrators (PDC), Syslog server, Snort, and control panel represents supervision and control level. The 129 features are contained in the dataset. Moreover, 29 features for each of the four phasor measurement units (PMU) [22]. Furthermore, each PMU is evaluated separately. Therefore, there are four subsets which represent data from each PMU (dataset 1–4) (Fig. 1).

The research was performed on a specific hardware configuration. The computer specification is shown in Table 1.

For the purpose of evaluation of the proposed approach, we use two assessment criteria - true positive rate (TPR), false positive rate (FPR).

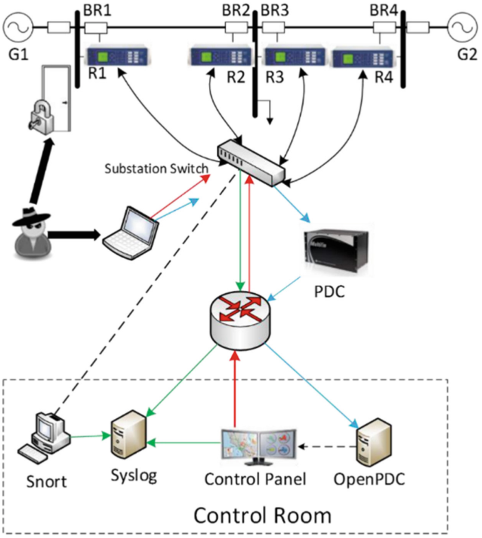


**Fig. 1.** Power system test bed [6]

**Table 1.** Hardware specification

| Hardware | Values |
|---|---|
| Procesor | Intel Core i7-4710HQ @ 2.50 GHz |
| RAM | 8 GB |
| Graphic processing unit | NVIDIA - GeForce GTX 860 M |
| Hard disk | SSD – 250 GB |
| Operating system | Windows 10–64 bit |

- True positive rate - TPR is also known as a sensitivity. Moreover, it is a statistical distribution of positive identification. Each element is evaluated and categorized into predefined class. In the case of IDS, the TPR represents correctly identified normal communication without anomalies. Moreover, the higher value represents better detection capabilities. The TPR is calculated according to Eq. 4.

  $$TPR = \frac{TP}{TP + FN} \tag{4}$$

  Where TP (True positive) represents correctly identified normal behavior of the system and FN (False negative) represents incorrectly rejected normal behavior of the system.
- False positive rate - FPR is also known as a false alarm rate. Moreover, it is a statistical distribution of false positive identification. In the case of IDS, the FPR represents incorrectly identified communication under cyber-attacks. The misclassification leads to a false alarm which is the critical threat to availability of the system. It is important to note that the availability of the services is the most important cyber security criterion for SCADA systems. Furthermore, the lower value represents the better anomaly detection capabilities. The FPR is calculated according to Eq. 5.

  $$FPR = \frac{FP}{FP + TN} \tag{5}$$

  Where FP (False positive) represents incorrectly identified normal behavior of the system and TN (True negative) represents correctly rejected normal behavior of the system.

## 6 Results

The article is created in order to specify the appropriate number of clusters in relation to SCADA systems. The determination of the optimal number K in K-means algorithm is the main goal of the research. Therefore, three techniques for determining the number of optimal clusters were used (the Rule of thumb, the Elbow method, and the Silhouette method). The research is based on the basic idea that normal behavior of the system is far more common than anomalies.

The Rule of thumb is very simple method how to determination of the optimal number of clusters. The total number of cluster is calculated for each dataset. The value is determined for: dataset 1–11 clusters, dataset 2–13 clusters, dataset 3–14 clusters, dataset 4–13 clusters. The results show disadvantage of the method which is caused by the relatively high number of clusters based on the dimension of the dataset. Furthermore, each dataset was evaluated by the simple K-means algorithm in order to specify true positive rate and false positive rate. The overall results are shown in Table 2. Furthermore, the overall best result represents dataset 4 with high TPR and the lowest FPR.

**Table 2.** An evaluation of the rule of thumb method

| Datasets | TPR | FPR |
|---|---|---|
| Dataset 1 | 1 | 0.6682 |
| Dataset 2 | 1 | 0.4712 |
| Dataset 3 | 1 | 0.375 |
| Dataset 4 | 1 | 0.3115 |

The Elbow method was chosen as the second technique for determining the number of optimal clusters. Moreover, each dataset was evaluated by the technique. According to the graph in Fig. 2, the total number of clusters is determined: dataset 1–4 clusters, dataset 2–2 clusters, dataset 3–3 clusters, dataset 4–3 clusters.

The True Positive Rate and False Positive Rate for four clusters are shown in Table 3. Moreover, The FPR for the second dataset is not calculated due to an absence of false positive cases. Furthermore, the overall best result represents dataset 3 with high TPR and the lowest FPR.

The Silhouette method was chosen as a third algorithm for determining an optimal number of clusters. According to the datasets, the optimal number of clusters was calculated on 2 clusters. Thus, the TPR and FPR is shown in Table 4 for each dataset.

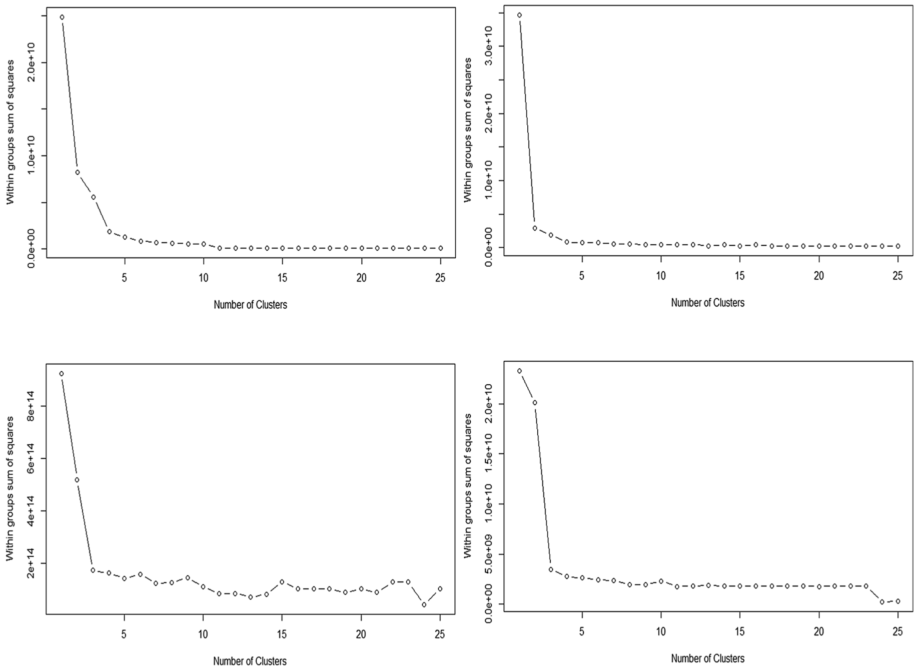


**Fig. 2.** The Elbow method graphs (top left corner - dataset 1, top right corner - dataset 2, bottom left corner - dataset 3, bottom right corner - dataset 4)

**Table 3.** An evaluation of the Elbow method

| Datasets | TPR | FPR |
|---|---|---|
| Dataset 1 | 1 | 0.5310 |
| Dataset 2 | 0.5457 | x |
| Dataset 3 | 0.8701 | 0.1743 |
| Dataset 4 | 0.707 | 0.3333 |

**Table 4.** An evaluation of the Silhouette method

| Datasets | TPR | FPR |
|---|---|---|
| Dataset 1 | 1 | 0.4729 |
| Dataset 2 | 0.5457 | x |
| Dataset 3 | 0.8701 | 0.1743 |
| Dataset 4 | 0.707 | 0.3333 |

Furthermore, the overall best result is represented by the same dataset (dataset 3) as for the Elbow method.

The final part of the research is to calculate TPR and FPR for clusters from 2 to 14 in order to evaluate and compare all methods for determining the number of clusters. The graph in Fig. 3 shows a comprehensive comparison between clusters. Moreover, it is focused on TPR. We can conclude that the major trend is continually increasing based on the increasing number of clusters.

The Fig. 4 describes progress of FPR in relation to number of clusters. Each dataset is evaluated by simple K-means algorithm according to a various number of clusters (from 2 to 14). However, dataset 2 has not define FPR for 2 and 3 clusters.

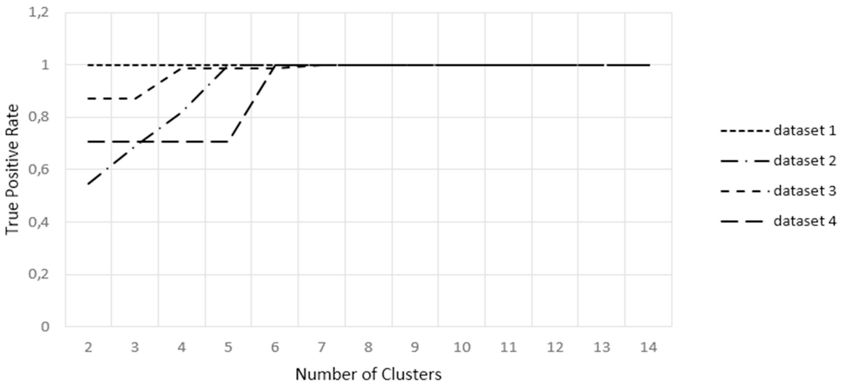


**Fig. 3.** The True Positive Rate distribution for selected clusters

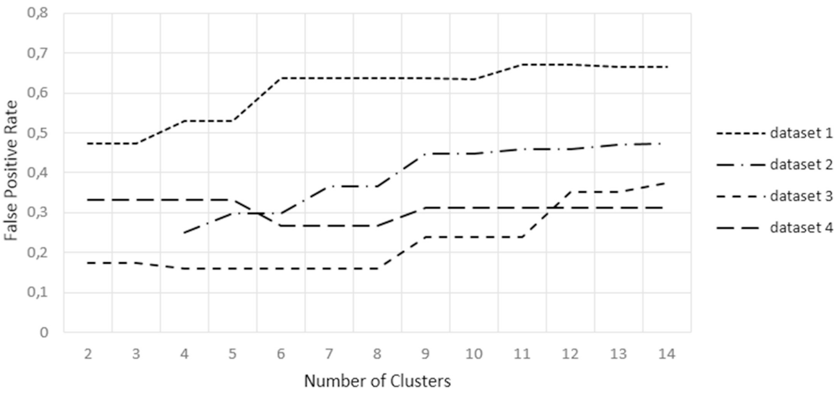


**Fig. 4.** The False Positive Rate distribution for selected clusters

## 7 Conclusion

The presented research is aimed at the significant part of cyber security in connection with IDS and SCADA systems. The unsupervised anomaly detection is examined. Therefore, methods used for determination of an optimal number of clusters were examined. The results are consistent with earlier studies conducted with anomaly detection (Tomlin and Farnam [12], Chiang and Mirkin [13], Yang et al. [14]).

The selected methods (the Rule of thumb, the Elbow method, and the Silhouette method) are commonly used in order to determine the optimal number of cluster for clustering algorithms like simple K-means. We can conclude that the overall results indicate relatively high TPR in almost every case with a considerable number of clusters. However, the FPR is unacceptable high in almost every case. We can conclude that the overall results indicate relatively high TPR in almost every case with a considerable number of clusters. However, the FPR is unacceptable high in almost every case. Furthermore, it should be noted that FPR is the most critical criterion in relation to SCADA systems due to availability which is the most important for the SCADA systems. Therefore, the best result of FPR indicates the Silhouette method. The last section of the research was dedicated to determinate an optimal number of clusters from range (2–14). The result shows that all datasets reach their TPR maximal values in a different number of clusters (dataset 1–2 clusters, dataset 2–5 clusters, dataset 3–7 clusters, dataset 4–6 clusters) as can be seen in Fig. 3. On the other hand, the representation of FPR is shown in Fig. 4. The best results are represented by a number of clusters (dataset 1–2 or 3, dataset 2–4, dataset 3–7 or 8, dataset 4 - from 6 to 8).

The overall results showed that the traditional methods of determination of cluster number are in some cases ineffective for the purpose of anomaly detection. The overall best method seems to be the Silhouette method. However, even the best method does not fit the needs of anomaly detection. It should be noted that among all results, there is no result with the sufficiently low value of FPR. Therefore, the future research is required in order to improve SCADA cyber security.

**Acknowledgments.** This work was funded by the Internal Grant Agency (IGA/FAI/2017/003) and supported by the project ev. no. VI20152019049 “RESILIENCE 2015: Dynamic Resilience Evaluation of Interrelated Critical Infrastructure Subsystems”, supported by the Ministry of the Interior of the Czech Republic in the years 2015–2019 and also supported by the research project VI20172019054 “An analytical software module for the real-time resilience evaluation from point of the converged security”, supported by the Ministry of the Interior of the Czech Republic in the years 2017–2019. Moreover, this work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014) and also by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

## References

1. Vávra, J., Hromada, M.: An evaluation of cyber threats to industrial control systems. In: The ICMT 2015 Conference Proceeding, 19–21 May 2015, Brno, pp. 369–373 (2015). ISBN 978-80-7231-976-3
2. Horkan, M.: Challenges for IDS/IPS deployment in industrial control systems (2015)
3. Pollet, J.: SCADA 2017: the future of SCADA security. Red Tiger Security (2017)
4. Cheung, S., Dutertre, B., Fong, M., Lindqvist, U., Skinner, K., Valdes, A.: Using model-based intrusion detection for SCADA networks. In: Proceedings of the SCADA Security Scientific Symposium, vol. 46, pp. 1–12 (2007)
5. Verba, J., Milvich, M.: Idaho national laboratory supervisory control and data acquisition intrusion detection system (SCADA IDS). In: 2008 IEEE Conference on Technologies for Homeland Security, pp. 469–473. IEEE (2008)
6. Valli, C.: SCADA forensics with snort IDS. In: Proceedings of WORLDCOMP 2009, Security and Management, Las Vegas, USA, pp. 618–621 (2009)
7. Carcano, A., Fovino, I.N., Masera, M., Trombetta, A.: State-based network intrusion detection systems for SCADA protocols: a proof of concept. In: Rome, E., Bloomfield, R. (eds.) CRITIS 2009. LNCS, vol. 6027, pp. 138–150. Springer, Heidelberg (2010). doi:10.1007/978-3-642-14379-3_12
8. Zhu, B., Sastry, S.: Intrusion detection and resilient control for SCADA systems. In: Securing Critical Infrastructures and Critical Control Systems: Approaches for Threat Protection: Approaches for Threat Protection, p. 352 (2012)
9. Yang, Y., McLaughlin, K., Littler, T., Sezer, S., Wang, H.F.: Rule-based intrusion detection system for SCADA networks. In: Renewable Power Generation Conference (RPG 2013), 2nd IET, pp. 1–4. IET (2013)
10. Maglaras, L.A., Jiang, J.: Intrusion detection in scada systems using machine learning techniques. In: Science and Information Conference (SAI), pp. 626–631. IEEE (2014)
11. Marton, I., Sánchez, I.A., Carlos, S., Martorella, S.: Application of data driven methods for condition monitoring maintenance. Chem. Eng. Trans. **33**, 301–306 (2013)
12. Tomlin Jr., L., Farnam, M.R.: A clustering approach to industrial network intrusion detection (2016)
13. Chiang, M.M.T., Mirkin, B.: Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. J. Classif. **27**(1), 3–40 (2010)
14. Yang, Y., McLaughlin, K., Sezer, S., Littler, T., Im, E.G., Pranggono, B., Wang, H.F.: Multiattribute SCADA-specific intrusion detection system for power networks. IEEE Trans. Power Deliv. **29**(3), 1092–1102 (2014)

15. Vávra, J., Hromada, M.: Comparison of the intrusion detection system rules in relation with the SCADA systems. In: Silhavy, R., Senkerik, R., Oplatkova, Z., Silhavy, P., Prokopova, Z. (eds.) Software Engineering Perspectives and Application in Intelligent Systems. AISC, vol. 465, pp. 159–169. Springer, Cham (2010). doi:10.1007/978-3-319-33622-0_15
16. Macaulay, T., Singer, B.: Cybersecurity for Industrial Control Systems: SCADA, DCS, PLC, HMI, and SIS. 193 p. CRC Press, Boca Raton (2012). ISBN 14-398-0196-7
17. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. (CSUR) **41**(3), 15 (2009)
18. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data Min. Knowl. Discov. **29**(3), 626–688 (2015)
19. Goldstein, M., Uchida, S.: A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS One **11**(4), e0152173 (2016)
20. Peeples, M.A.: R script for K-means cluster analysis (2011)
21. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
22. Pan, S., Morris, T., Adhikari, U.: A specification-based intrusion detection framework for cyber-physical environment in electric power system. Int. J. Netw. Secur. (IJNS) **17**(2), 174–188 (2015)

# Compensation Model of Multi-attribute Decision Making and Its Application to N-Version Software Choice

Denis Vladimirovich Gruzenkin[1(✉)], Galina Viktorovna Grishina[1],
Mustafa Seçkin Durmuş[2], Ilker Üstoğlu[3],
and Roman Yurievich Tsarev[1]

[1] Siberian Federal University, Krasnoyarsk, Russia
gruzenkin.denis@good-look.su, ggv-09@inbox.ru,
tsarev.sfu@mail.ru
[2] Pamukkale University, Krasnoyarsk, Turkey
msdurmus@pau.edu.tr
[3] Yildiz Technical University, Istanbul, Turkey
ustoglu@yildiz.edu.tr

**Abstract.** Multi-attribute decision making deals with discrete finite set of alternatives. The solution to the multi-attribute decision making problem is the choice of an alternative from the set of all possible alternatives on the base of usually contradicting attributes. In this paper, a new multi-attribute decision making model is presented. The proposed model develops a linear compensatory process for the interconnected attributes. It concerns the overall ranking of the alternatives based on the attribute-wise ranking as well as the interaction and the combination of the attributes. The compensation model of multi-attribute decision making is applied to N-version software selection. N-version programming is one of the well-known software development approach which ensures high dependability and fault tolerance of software. However, the problem of extra resource involvement arises since the N-version programming stipulates program redundancy. A set of characteristics/attributes have to be considered when choosing an optimal variant of N-version software. The proposed compensation model of multi-attribute decision making provides a solution to this problem. Additionally, a case study on choosing N-version software for a real-life information and control system problem is provided to verify the correctness of our model.

**Keywords:** Multi-attribute decision making · Compensation model · Dependability · N-version software

## 1 Introduction

Modern control systems are characterized mainly by the ability of their data processing power where the control functions are executed by software. The main reason for this situation is that the control processes and the complex calculations of a huge amount of data cannot be carried out only by means of hardware [1].

The original version of the book was revised. For detailed information please see Erratum. The erratum to the book is available at 10.1007/978-3-319-57141-6_53

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_16

The field of control systems application defines the requirements of the dependability of these systems [2]. There is a number of areas where the failure of control system can result in severe financial and economic losses or can harm human health and even cause death [3]. Since, the data processing and control is executed by means of software, then its dependability characteristics directly define the dependability of a control system as a whole [4].

One of the most well-established methodologies ensuring high level of dependability and fault tolerance of software is N-version programming [5]. This methodology is based on the principle of program redundancy which allows to increase significantly the dependability of software for control systems [6, 7].

A large number of N-version software modules, redundant versions, and also some restrictions of the real-world problems such as cost, execution time, memory requirements, dependability properties compose a decision making problem [7]. The problem is to determine the best variation of N-version software for a control system by taking into account a number of criteria [8]. The solution to this problem can be obtained by applying a multi-attribute decision making method [9].

This paper considers the compensation model of multi-attribute decision making which allows to perform an overall alternative ranking based on the attribute-wise ranking as well as the interaction and the combination of the attributes. The proposed compensation model is applied to determine the best variation of N-version software for a control system.

## 2 Compensation Model of Multi-attribute Decision Making Providing Overall Ranking Alternatives on the Base of Attribute-Wise Ranking

The compensation model of multi-attribute decision making suggests alternatives ranking according to their order of preference. The overall ranking of the alternatives is based on the attribute-wise ranking. The alternative that is assigned the first rank is the best one. Ranking of the alternatives by considering just the order of their preferences allows us to avoid scaling of the quality-type attributes. This process uses ordinal input data rather than cardinal ones [9, 10].

The model describes a linear compensatory process for the attribute interaction and combination. The overall ranking of the alternatives can be obtained by the attribute-wise ranking where the interaction of the attributes is ignored [9]:

$$k_i = \sum\nolimits_{j=1}^{n} k_{ij};\ i = 1, 2, \ldots, m \tag{1}$$

where

$n$ – is the number of the attributes,
$m$ – is the number of the alternatives,
$k$ – is the number of ranks ($k = m$),
$k_i$ – is the overall rank of the alternative $i$,
$k_{ij}$ – is the rank of $i^{\text{th}}$ alternative on the $j^{\text{th}}$ attribute.

However, it is important to consider this dependence for the majority of decision making problems [11–13]. According to this, the compensation model of multi-attribute decision making has been developed. In this case, the idea of compensation consists of accounting the interdependence between the attributes: the change of a value of one of them leads to the change of values of some other attributes.

Let us define the matrix π as a square nonnegative matrix $m \times m$ where the element $\pi_{ik}$ represents the number (or the frequency) of ranking of an alternative $A_i$ the $k^{\text{th}}$ attribute-wise ranking. The matrix π is based on the matrix of the attribute-wise ranking $D$ of the alternatives.

$$\pi_{ij} = \sum\nolimits_{l=1}^{n} I(D_{jl}^{i})w_l;\ i = 1,2,\ldots,m;\ j = 1,2,\ldots,k, \tag{2}$$

$$I\left(D_{jl}^{i}\right) = \begin{cases} 1, \ \mathit{if}\ D_{jl}^{i} = i \\ 0, \ \mathit{if}\ D_{jl}^{i} \neq i \end{cases} \tag{3}$$

where

$I(x)$ – is the indicator function; $w_l$ – is the weight coefficient of the attribute $l$.

In the case of different weight coefficients the elements of the matrix π represent the sum of attributes weights of the appropriate rank. The weight coefficients are supposed to be normed.

It is obvious that the element $\pi_{ik}$ defines the contribution of the alternative $A_i$ into the overall ranking. The more $\pi_{ik}$ value the more alternative $A_i$ deserves to be assigned to rank $k$.

Let us define a permutation matrix $Z$ as a square matrix $m \times m$ whose elements are $Z_{ik} = 1$, if the alternative $A_i$ is assigned to the overall rank $k$, and $Z_{ik} = 0$ otherwise. The objective function can be expressed as follows:

$$\max_{Zij} \sum\nolimits_{i=1}^{m} \sum\nolimits_{j=1}^{k} \pi_{ij} Z_{ij} \tag{4}$$

subject to:

$$\sum\nolimits_{j=1}^{k} Z_{ij} = 1;\ i = 1,2,\ldots,m, \tag{5}$$

$$\sum\nolimits_{i=1}^{m} Z_{ij} = 1;\ j = 1,2,\ldots,k. \tag{6}$$

The conditions mean that the alternative $A_i$ can be assigned only to one rank, and the rank $k$ can be assigned to only one alternative.

The optimal permutation matrix representing the solution to the linear programming problem mentioned above is designated as $Z^*$. Then, the preference order can be achieved by multiplying the matrix π by the matrix $Z^*$.

# 3 A Case Study on the Application of the Compensation Model of Multi-attribute Decision Making

In this section, an application of the proposed compensation model of multi-attribute decision making is considered. In the first example the optimal solution is to be found out of three alternatives considering three attributes. The attribute-wise ranking of the alternatives is shown in Table 1.

**Table 1.** The attribute-wise ranking of the alternatives

| Attribute | | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|---|
| Rank | 1 | $A_1$ | $A_1$ | $A_2$ |
| | 2 | $A_2$ | $A_3$ | $A_1$ |
| | 3 | $A_3$ | $A_2$ | $A_3$ |

The attribute-wise ranking of the attributes is represented by the matrix $D$. The indexes of the alternatives shown in Table 1 are considered to be the elements of the matrix $D$. Therefore, the matrix $D$ can be written as follows:

$$D = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \\ 3 & 2 & 3 \end{bmatrix}.$$

The matrix $\pi$ can be developed on the basis of the matrix $D$ whose elements are presented by the number of alternative assignments of each rank. The first alternative is assigned to the first rank twice, the second rank once, and the first alternative is not assigned the third rank. The first line of the matrix $\pi$ reflects the first alternative assignments:

$$\pi = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}.$$

Applying the attributes weight coefficients $w_1$ = 0.2, $w_2$ = 0.4, $w_3$ = 0.4, the elements of the matrix $\pi$ can be rewritten as follows:

$$\pi = \begin{bmatrix} 0.2+0.4 & 0.4 & 0 \\ 0.4 & 0.2 & 0.4 \\ 0 & 0.4 & 0.2+0.4 \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.4 & 0.2 & 0.4 \\ 0 & 0.4 & 0.6 \end{bmatrix}.$$

The optimal permutation matrix $Z^*$ is as follows:

$$Z^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The most left-top element of matrix $Z^*$ reflects assignment of the first alternative (the first column) to the first rank (the first line). The preference order is obtained by multiplying the matrix π by the matrix $Z^*$ is as follows:

$$A_1 \succ A_2 \succ A_3.$$

Thus, the best alternative is alternative $A_1$.

Let us consider another example. The values of the attributes of three alternatives are shown in Table 2.

The attribute-wise ranking of the alternatives is shown in Table 3.

The overall rank can be obtained on the basis of the attribute-wise ranking (1):

$$\begin{aligned} k_1 &= 1+3+3+1 = 8, \\ k_2 &= 2+2+2+2 = 8, \\ k_3 &= 3+1+1+3 = 8, \end{aligned}$$

where $k_i$ is a rank of the alternative $A_i$.

The result provides a decision maker with no information.

Applying the proposed compensation model of multi-attribute decision making the matrix π can be obtained as follows:

$$\pi = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \end{bmatrix}.$$

**Table 2.** The values of the attributes

| Alternative | Cost ($) | Reliability | Processor time (msec) | RAM volume (Gb) |
|---|---|---|---|---|
| $A_1$ | 1000 | 0.90 | 7 | 2 |
| $A_2$ | 1500 | 0.95 | 5 | 4 |
| $A_3$ | 2000 | 0.99 | 2 | 8 |

**Table 3.** Attribute-wise ranking of the alternatives

| Rank | Cost ($) | Reliability | Processor time (msec) | RAM volume (Gb) |
|---|---|---|---|---|
| 1 | $A_1$ | $A_3$ | $A_3$ | $A_1$ |
| 2 | $A_2$ | $A_2$ | $A_2$ | $A_2$ |
| 3 | $A_3$ | $A_1$ | $A_1$ | $A_3$ |

In this case, two instances of the matrix $Z^*$ are relevant:

$$Z_1^* = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \; and \; Z_2^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$A_1, A_3 \succ A_2.$$

The compensation model demonstrates that the alternatives $A_1$ and $A_3$ are more preferable than the alternative $A_2$.

Let us define the weight coefficients reflecting the importance of attributes:

$$\omega_{cost} = 0.2,$$
$$\omega_{rel} = 0.5,$$
$$\omega_{time} = 0.1,$$
$$\omega_{vol} = 0.2.$$

Then matrixes $\pi$ and $Z^*$ are as follows:

$$\pi = \begin{bmatrix} 0.4 & 0 & 0.6 \\ 0 & 1 & 0 \\ 0.6 & 0 & 0.4 \end{bmatrix},$$
$$Z^* = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$
$$A_3 \succ A_2 \succ A_1.$$

In this case, the most preferable alternative is the alternative $A_3$.

## 4 Choice of the N-Version Software for an Information and Control System

The compensation model of the multi-attribute decision making model was applied to solve a real-life problem for the choice of N-version software for an information and control system used in a Krasnoyarsk transport company.

An analysis of the existing information and control system of the company was performed prior to the modernization of its software. The company provides a user with a workplace equipped with a computer which continuous functioning was supported by an uninterruptible power supply (UPS).

The core of the system is a server with an individual UPS of extended capacity and some network equipment. Besides, the information and control system includes a rented virtual server which is placed remotely on the Internet. In case of the primary server failure, the requests from clients are redirected to this backup server. Interacting with controllers, both of these servers increase external data traffic doubling network load. As a result, it negatively affects the productivity of the whole information and control system.

Only the primary server interacts with controllers. At the same time, it exchanges data with a backup server which continuously updates its status. If data exchange with the primary server stops, the backup server supposes that the primary server has failed and takes over its functions. After the failure of the primary server is fixed, it restores and switches on again. The primary server reads the current state from the backup server and retakes over its functions as a primary server.

The company's information and control system can be considered rather unreliable as data duplication occurs for the data transferred on a network, i.e. server failure will result in malfunctioning of the regulated subsystems. Thus, for instance, the occurrence of a failure in at least one transmission channel of the system will lead to an inappropriate control of the operations execution and can result in essential financial losses.

The increase of the dependability and implementation of the fault tolerance of software of the information and control system is applied on programming level according to the N-version methodology. There is a set of possible variants for N-version software to implement. The problem is to choose an optimal alternative that is a variant of N-version software for applying to the information and control system. The solution to this multi-attribute decision making problem is to be obtained on the basis of the attribute values. The attribute values are shown in Table 4.

**Table 4.** Variants of the N-version software

| Attribute | | Cost | Time | Vol | Reliability | MTTF |
|---|---|---|---|---|---|---|
| Alternative | $A_1$ | 11000 | 850 | 15 | 0.99911 | 5.11 |
| | $A_2$ | 12000 | 400 | 11 | 0.99934 | 3.89 |
| | $A_3$ | 14000 | 400 | 15 | 0.99956 | 4.65 |
| | $A_4$ | 14000 | 300 | 14 | 0.99921 | 4.35 |
| | $A_5$ | 15000 | 700 | 12 | 0.99965 | 3.77 |
| | $A_6$ | 18000 | 500 | 10 | 0.99943 | 5.34 |
| | $A_7$ | 19000 | 600 | 7 | 0.99932 | 4.54 |
| | $A_8$ | 21000 | 500 | 11 | 0.99984 | 3.56 |
| | $A_9$ | 24000 | 300 | 9 | 0.99977 | 5.66 |

Notation:

*Cost* – Cost of the N-version software, $
*Time* – Processor time, sec.
*Vol* – Volume of RAM required, Gb
*Reliability* – Reliability of the N-version software
*MTTF* – Mean time to failure, months

The importance of an attribute is reflected by its weight:

$$\begin{aligned} \omega_{Cost} &= 0.20; \\ \omega_{Time} &= 0.15; \\ \omega_{Vol} &= 0.10; \\ \omega_{Reliability} &= 0.30; \\ \omega_{MTTF} &= 0.25. \end{aligned}$$

The result of the calculations according to the proposed compensation model of multi-attribute decision making is as follows:

$$A_8 \succ A_9 \succ A_2 \succ A_3 \succ A_6 \succ A_4 \succ A_7 \succ A_5 \succ A_1.$$

Thus, the optimal solution to the given multi-attribute decision making problem is the alternative $A_8$, that represents a certain variants of the N-version software implementation.

The reliability of the chosen N-version software for the information and control system is 99,984% that is 0,72% more than the previous value of reliability. The economic effect from modernization of the information and control system software was assessed by means of Advisor Client & Server Model [14, 15]. The results of economic assessment show that even with some extra modification costs the total expenses in case of system failure are reduced for $27.981,35.

## 5 Conclusion

Redundant software such as N-version software requires more resources than classic one-version software. Therefore, software designers and developers face the problem of compromise between benefits and losses. To increase the dependability of the software and avoid extra expenses at the same time, they have to deal with multi-attribute decision making. The proposed compensation model in this paper allows to solve this multi-attribute decision making problem.

In this paper we present the results of an application of the compensation model of multi-attribute decision making to choose the optimal variant of the N-version software for an information and control system. The results demonstrate the ability of the proposed model to bring us to the solution to the given multi-attribute decision making problem.

The compensation model of the multi-attribute decision making allows to perform the overall ranking of the alternatives in the order of their preferences. The overall ranking is based on the attribute-wise ranking. The proposed model takes into account the interaction and the combination of the attributes. The compensation model concerns ordinal preferences rather than cardinal ones. It allows to avoid problems related to the scaling of the attribute values and make it possible to consider both quantity- and quality-type attributes. The only input data for the model is the attribute-wise ranking of the alternatives. The compensation model of multi-attribute decision making describes a linear compensatory process. Therefore, the solution can be obtained by means of an available mathematical package.

The proposed multi-attribute decision making model was also applied to solve a real-life problem to verify its correctness. The optimal variant of the N-version software has been selected with assistance of the model.

Moreover, the developed compensation model of the multi-attribute decision making can be applied to solve a decision making problem in the discrete finite space of alternatives where the decision is based on the attribute values of the alternatives.

## References

1. Tsai, W.-T., Zhou, X., Paul, R.A., Chen, Y., Bai, X.: A coverage relationship model for test case selection and ranking for multi-version software. In: Zhang, L.-J., Paul, R., Dong, J. (eds.) High Assurance Services Computing, pp. 285–311. Springer, Heidelberg (2009)
2. Eckhardt, D.E., Lee, L.D.: A theoretical basis for the analysis of multiversion software subject to coincident errors. IEEE Trans. Softw. Eng. 1511–1517 (1985)
3. Guo, P., Liu, X., Yin, Q.: Methodology for reliability evaluation of N-version programming software fault tolerance system. In: Proceedings of the IEEE Computer Science and Software Engineering International Conference, pp. 654–657 (2008)
4. Zuzana, K.: Software reliability models. In: Proceedings of the IEEE Radioelektronika, 2007 17th International Conference (2007)
5. Avizienis, A., Chen, L.: On the implementation of N-version programming for software fault-tolerance during program execution. In: Proceedings of IEEE Computer Society International Conference on Computers, Software and Applications Conference, COMP-SAC, pp. 149–155 (1977)
6. Gruzenkin, D.V., Tsarev, R.Y., Pupkov, A.N.: Technique of selecting multiversion software system structure with minimum simultaneous module version usage. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Silhavy, P., Prokopova, Z. (eds.) Software Engineering Perspectives and Application in Intelligent Systems. AISC, vol. 465, pp. 375–386. Springer, Cham (2016). doi:10.1007/978-3-319-33622-0_34
7. Kulyagin, V.A., Tsarev, R.Y., Prokopenko, A.V., Nikiforov, A.Y., Kovalev, I.V.: N-version design of fault-tolerant control software for communications satellite system. In: International Siberian Conference on Control and Communications (SIBCON), pp. 1–5 (2015)
8. Sklyar, V., Karchenko, V.: A method of multiversion technologies choice on development of fault-tolerant software systems. In: Workshop on Methods, Models and Tools for Fault Tolerance, pp. 148–157 (2007)
9. Hwang, C.-L., Yoon, K.: Methods for Multiple Attribute Decision Making. Multiple Attribute Decision Making. Springer, Heidelberg (1981). pp. 58–191
10. Bernardo, J.J., Blim, J.M.: A programming model of consumer choice among multi-attributed Brands. J. Consum. Res. **4**, 111–118 (1977)
11. Saidi Mehrabad, M., Fathian Brojeny, M.: The development of an expert system for effective selection and appointment of the jobs applicants in human resource management. Comput. Ind. Eng. **53**(2), 306–312 (2007)
12. Işıklar, G., Büyüközkan, G.: Using a multi-criteria decision making approach to evaluate mobile phone alternatives. Comput. Stand. Interfaces **29**, 265–274 (2007)

13. Xu, Z.: Uncertain Multi-Attribute Decision Making. Methods and Applications. Springer, Heidelberg, New York, Dordrecht, London (2015)
14. Schwalbe, K.: Information Technology Project Management. Cengage Learning, Australia, Brazil, Mexico, Singapore, United Kingdom, United States (2015)
15. Chalutz Ben-Gal, H., Tzafrir, S.S.: Consultant-client relationship: one of the secrets to effective organizational change. J. Organ. Change Manag. **24**(5), 662–679 (2011)

# Empirical Testing of Bends in Workflow Diagrams by Eye-Tracking Method

Zdena Dobesova(✉)

Department of Geoinformatics, Faculty of Science, Palacký University,
17. listopadu 50, 779 00 Olomouc, Czech Republic
zdena.dobesova@upol.cz

**Abstract.** Workflow diagrams consist of nodes and connectors to express the steps of processing in the form of a visual program. The graphical vocabulary and the layout of the diagram have an influence on the user cognition of diagram. The aesthetic aspects also have an impact on users understanding. One aesthetic recommendation – "minimize beds in edge" was tested in workflow diagrams from ArcGIS ModelBuilder. Eye-tracking measuring in the laboratory was prepared for objective empirical testing. Five couples of diagrams with and without orthogonally bends were showed to 26 respondents. The user executed specific tasks above diagrams. Eye-tracking measuring brought interesting objective results. Eye-tracking metrics affirm that diagrams with orthogonal bends on connector lines have an average higher number of fixations, longer length of scanpath, shorter average time of fixation and longer duration time. The result is that the using of straight lines brings effective cognition of workflow diagrams in case of spatial data processing in geographic information system (GIS).

**Keywords:** Workflow · Human-Computer Interaction · Visual programming language · Eye-tracking · Aesthetic · Cognition · Geographic information system

## 1 Introduction

Workflow diagrams are used for graphical expression of steps of the process (algorithm). In the area of geographical information systems, the workflow diagram designs the processing of spatial data. Different graphical editors are available in geographical information systems (GIS) for design workflows. These types of GIS software and their workflow editors exist:

- ArcGIS for Desktop (editor ModelBuilder),
- Erdas Imagine (editors Model Maker and Spatial Model Editor),
- IDRISI (editor Macro Modeler),
- AutoCAD Map 3D (editor Workflow Designer).

Moreover, two open source GIS software have workflow editor:

- QGIS (editor Processing Modeler),
- GRASS GIS (editor Graphical Modeler).

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_17

The overview and description of these graphical editors are in the article [1].

All workflows diagrams belong to the group named visual programming languages [2]. Visual programs are easier understood than textual programs. The workflow editors are possible to describe from the point of the method of design, the amount of functionality or describe the symbols from their graphical vocabulary. The phase of diagram design and utilization belongs to the Human-Computer Interaction (HCI) research area. The graphical vocabulary (notation) is important from the point of user perception and cognition. Moody in his theory “Physics of Notations” stated that is necessary to use cognitively effective visual notations [3]. Cognitively effective means optimized for processing by the human mind.

A research group at Department of Geoinformatics of Palacky University has made an effort to evaluate visual programming languages in the area of GIS software mentioned above. The aim is finding the level of cognitive effectiveness and aesthetics of visual vocabularies and diagrams subsequently. For evaluation were used the theory of Physics of Notations, the rules of aesthetics and empirical testing by the eye-tracking equipment in the laboratory. Several tests for various diagramming language from GIS software experimented from 2014 to 2016. Application of these methods in the area of HCI discipline brings improvements and recommendations for user design of workflow diagrams. The article describes the result of the empirical test for diagrams with straight and orthogonally curved connector lines for a set of workflow diagrams from ArcGIS ModelBuilder.

## 2 Methods and Materials

Theory “Physics of Notations” defines nine principles for evaluation and design of cognitively effective visual notations [3]. One of the principles is “Principle of Cognitive Interaction”. This principle states that it is necessary to include explicit mechanisms to support the integration of information from different diagrams. In the phase of design or reading diagrams is a demand on simple navigation and transitions between diagrams. Connector lines in one separate diagram are also important for simple navigation in the diagram. Connector lines help in wayfinding and contribute to answering to a set of questions:

- *Orientation:* Where am I?
- *Route choice:* Where can I go?
- *Route monitoring:* Am I on the right path?
- *Destination recognition:* Am I there yet?

Additionally, the set of aesthetic rules and recommendations for diagram design is mentioned in literature:

- *Minimize bends in the edge* (the total number of bends in polyline edges should be minimized) [4, 5]
- *Minimize edge crossing* (the number of edge crossing in drawing should be minimized) [6]

- *Maximize minimum angle* (the minimum angle between edges extending from a node should be maximized) [7, 8]
- *Orthogonality* (fix nodes and edges to an orthogonal grid) [4, 9]
- *Symmetry* (where possible, a symmetrical view of the graph should be displayed) [10]
- *Good continuity* (minimize angular deviation from straight line of one bended edges or two followed edges connecting two nodes) [11]

Aesthetic rules concerns both to the connector lines (edges) both the layout and arrangement of symbols in the diagram. Some rules have a positive or negative influence on other rules. Empirical study Cognitive Measurement of Graph Aesthetics [11] verified the aesthetic rules. The respondents tried to find the shortest path above diagrams. Their testing proved that response time depends on the number of edge crossing and continuity of graph. Good continuity will be more readily received if nodes in the diagram are not in a zigzag pattern but form a smooth continuous sequence. Also, zigzag connecting lines are perceived worse.

In this type of studies are used "comprehension tasks" to measure response time and correctness of user answers [12–14]. The set of diagrams or pictures (maps) is often used for evaluation of usability of visualization methods in cartography and GIS [15, 16]. In our research, we tried to empirically verify the influence of orthogonal bends in connector lines to the effective cognition. We prepared the workflow diagrams and comprehension tasks (questions) for experimental test.

### 2.1 Workflow Modeling in ArcGIS ModelBuilder

ArcGIS (producer Esri) has an embedded graphical editor called ModelBuilder to create and execute the steps of spatial data processes. The workflow diagram is called model process in this editor. The design of flow is very easy, only by drag and drops the spatial functions (tools) to the canvas. The functions are represented by the yellow rectangle symbol, and blue/green ovals represent data. Moreover, the orange hexagon expresses the iterator for the construction of cycle. The connectors between symbols are black lines ended by an arrow that expresses the orientation of flow. The workflow is expressed as a fluent chain of input data, functions, and output data. The basic graphical vocabulary is described in the documentation [17]. The basic evaluation according to theory Physics of Notation was made in previous work [18].

The basic setting of diagram properties allows the automatic change of orientation of diagram. There is also an option to set the connection routing type. The user can switch between "Orthogonal routing" and "Straight routing" of connectors. The whole diagram is automatically redrawn according to the selected option.

Straight routing is the default. Form of workflow diagram with "Orthogonal routing" is in Fig. 1.

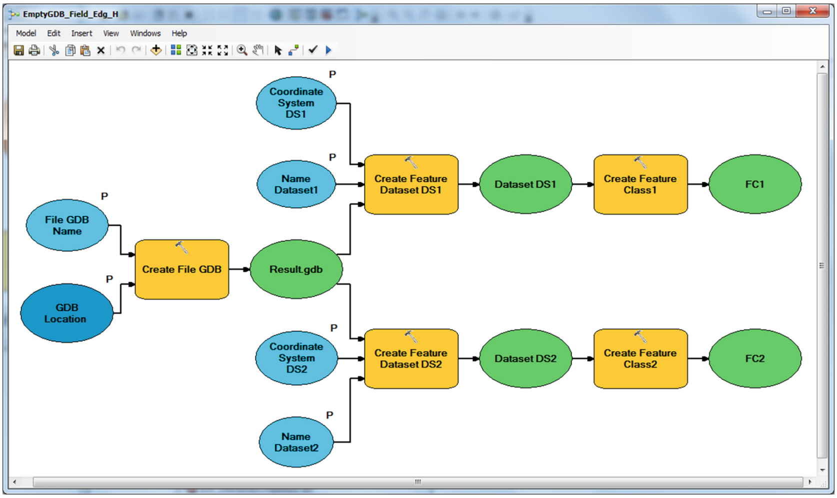


**Fig. 1.** Interface of ModelBuilder with workflow diagram

## 2.2 Eye-Tracking Testing

The eye-tracking measurement was used for evaluation of cognition of workflow diagrams. The test consists of 22 workflows diagrams from ModelBuilder. We tested several diagrams with various complexities, different arrangements of symbols and orientation of flow (vertical and horizontal directions), with a change of colors and also with straight and bend connector lines.

The respondents were the students of the second grade of bachelor study Geoinformatics at the end of the semester. They had the subject “Programming 2” where the design of workflow models in ModelBuilder was explained and detailed practiced in four lectures. Also, students accomplished four home works with the construction of complexity diagrams. The group of respondents was assumed as skilled users. The total number of respondents was 27. One of then was excluded due to bad calibration of gaze. The group consists of 6 women and 20 men finally, with age from 22 to 25. The age of respondents was from 20 to 25. The test proceeded in May of 2016.

The testing was run at an eye-tracking laboratory in the Department of Geoinformatics at the Palacky University in Olomouc (Czech Republic). For the experiment, we used eye-tracker SMI RED 250 with software SMI Experiment Suite 360°. To define the test, we used SMI Experiment Center program; to visualize the results we used SMI BeGaze. The evaluation was also done in software Ogama 4.5. The size of the monitor to record eye movement was 1920 × 1080 pixels for displaying diagrams. The sampling frequency was 250 Hz.

## 2.3 Diagram Stimulus

The term stimulus is used in the process of eye-tracking testing [19]. The stimulus could be any picture, photo, map or drawing like graph or diagram. In the case of testing of workflow models the series of 20 various diagram was prepared as an experiment. The diagrams were presented individually on the screen in random order to prevent “learning effect” [20].

Each stimulus was accompanied by the special task to record the understanding, comprehension and cognition of diagram. The users solved the task by finding and clicking on the correct symbol(s) in the diagram.

The eye-tracker collected the position of gaze above stimulus. From the raw data, the position of eye fixations and the scanpath (the path between eye positions) were calculated by OGAMA software. The response time and total time of each user were also measured. Two or more correct answers (symbols) exist in some diagrams (depends on the task). All mouse clicks were recorded. Moreover, other numeric characteristics (metrics) from eye-tracking data were calculated. They are the total length of scanpath, the average time of fixation, frequency of fixation per second, fixations/saccade ratio, average saccade length, path velocity in pixel per second and others. Aggregation of respondent scanpaths brings clear evidence of reading patterns. The orientation of and continuity of reading patterns follow mainly the orientation of connector lines [21].

Ten diagrams were present in the eye-tracking experiment for the testing of the influence of bends in connector lines. All ten diagrams consist of five couples of the same diagrams. The functionality of the diagrams was the same for each couple. Also, the tasks solved above a couple of diagrams were the same. Examples of one couple are in Figs. 2 and 3. The first is with straight connector lines and the second is with orthogonal bends on connector lines. The question solved above the diagrams was “*Mark input data of function Select Layer By Location.*” The places with correct answers are marked by a red dot in Figs. 2 and 3.

The respondent had to find the yellow rectangle with function “*Select Layer By Location*” firstly. After that, the gaze moved to the green ovals and marked them as answers by mouse clicks. The connectors are straight between the yellow symbol of function and green ovals (Fig. 2) or with two orthogonal bends (Fig. 3). In fact, the lines with orthogonal bend are longer than straight lines.

The research task was if the variant shapes of connectors have an influence on any eye-tracking metrics. The hypothesis was that bended lines are worse for reading and aesthetics perceive. Firstly the scanpath of individual respondents was explored. The fixations are mainly on the color symbols in the ModelBuilder workflow diagrams.

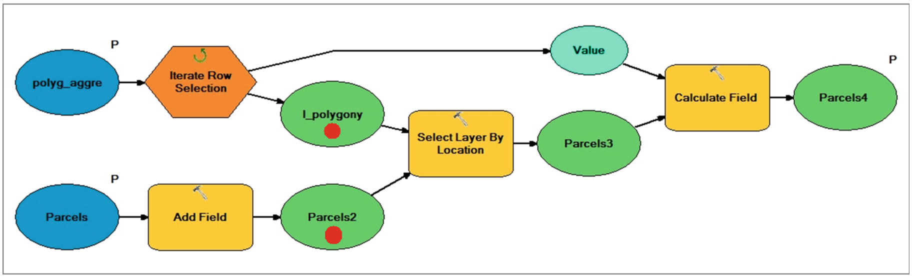


**Fig. 2.** Workflow diagram with the straight connector lines

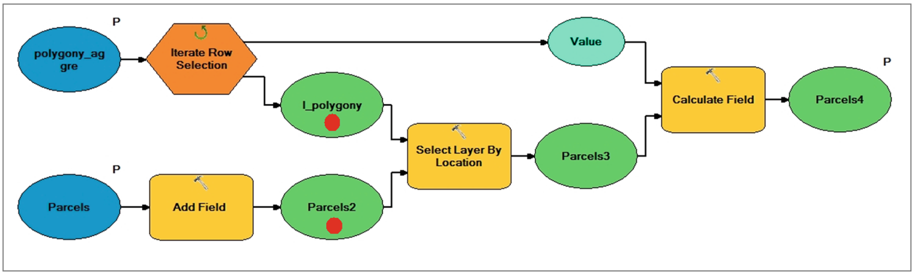


**Fig. 3.** Workflow diagram with bends in line connectors

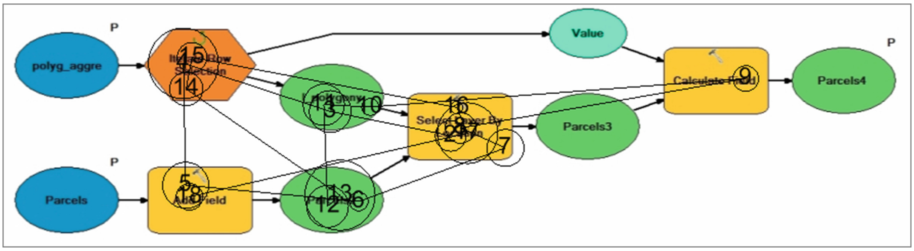


**Fig. 4.** Scanpath of one respondent with order of fixations

Connectors have nearly no fixations. The fixations are presented by black circles where inner number expresses the order of fixation, and the diameter expresses the duration of fixation (bigger has longer time). Black lines are eye quick movements of the eye between fixations (Fig. 4.). The connector line has mainly influence on the orientation of reading. The scanpath exposes that the user gazes skips between the yellow symbol and green ovals several time forward and backward. Left part and right part of the diagram are not nearly explored by respondents (no fixations are there).

Statistics evaluation of measured eye-tracking metrics was calculated after individual exploration of user recorded scanpaths. The score of correct and bad answers was assessed. All answers were correct for all five couples and 26 respondents. The shape of connector lines does not have negative influence to correct answers.

The Shapiro–Wilk test was used to verify the normality of eye-tracking data. The hypothesis of the normal distribution of data was not proving. Subsequently, the non-parametric tests were used. Non-parametric Mann–Whitney U test examined corresponding couples of diagrams. This test verifies null hypothesis $H_0$: The distributions of both populations are equal.

The calculated metrics are in Table 1. Values for B means diagram with orthogonal bends; S means straight line connectors. The average time of response (duration time) is shorter for all diagrams with straight lines in comparison with the same diagrams matched in couples. An average number of fixations is greater for diagrams with the orthogonal bends in line connector than for diagram with straight lines. Also, shorter

**Table 1.** Average value of eye-tracking metrics for orthogonal bend (B) and straight (S) lines

| Diagrams | Type of lines | Duration time [s] | Number of fixations | Length of scanpath [px] | Avg. time of fixation [ms] |
|---|---|---|---|---|---|
| Couple 1 | B | 10 | 25 | 3 373 | 223 |
| | S | 9 | 22 | 3 219 | 241 |
| Couple 2 | B | 11 | 30 | 7 104 | 197 |
| | S | 10 | 29 | 6 425 | 205 |
| Couple 3 | B | 16 | 44 | 9 449 | 215 |
| | S | 14 | 37 | 9 136 | 217 |
| Couple 4 | B | 11 | 26 | 3 813 | 211 |
| | S | 10 | 25 | 3 593 | 231 |
| Couple 5 | B | 11 | 29 | 3 695 | 224 |
| | S | 9 | 19 | 2 459 | 247 |

average scanpaths is for straight lines. The most interesting result is the average time of fixation. The straight lines have a longer time of average time fixation.

We assumed that the gaze is sputtered in the case of orthogonal bends. There are longer response time, longer scanpath and bigger count of total fixations and more repetitive gaze movements. The statistical evaluation does not validate the statistical significance of compared metrics. A significant difference has only for the last couple of diagrams where the diagrams have the vertical orientation. The difference was significant for duration time metric.

Subsequently, the number of fixations was calculated only for Area Of Interest (blue rectangle AOI) nearly to the place of the correct answer (Fig. 5). The green ovals (express input data) and yellow rectangle (with mentioned function in question) were incorporated to AOI together with the lines. These two connector lines have an influence on the number of fixation. The AOI with straight lines has 281 fixations (total for 26 respondents) (Fig. 5 left). In the second case, the AOI has 319 fixations also for 26 respondents (Fig. 5 right).

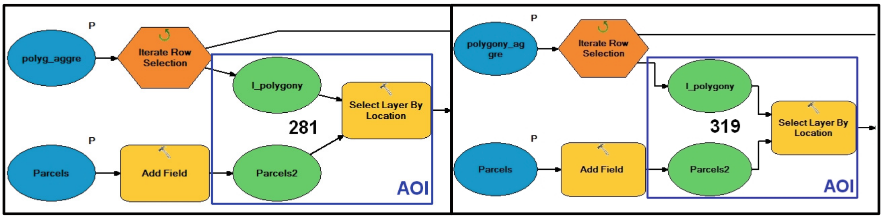


**Fig. 5.** Comparison of number of fixations in the same area of interest in both diagrams

# 3 Results

The eye-tracking testing empirically verified the aesthetic rule that state "*Minimize bends in the edge*". Five couples of various diagrams were tested with two modifications with orthogonal bends and without bends. The functionality was the same in couples, and the same comprehension task was assigned. The evaluated eye-tracking metrics prove that straight lines have in average:

- Lower number of fixation (also in AOI near correct place of answer)
- Shorter scanpath
- Longer average time of fixation
- Shorter total time of response (for one couple is statistically significant)

All these eye-tracking metrics were worse in the case of the orthogonal bend lines. In the case of longer average time of fixation, we assume that the respondent reading is calmer for straight lines than in case bends on lines. The orthogonally bended lines disturb the reading and user gaze skips several times between symbols with very short fixations. The presented testing of bends in workflows diagrams from ModelBuilder is valid for design any diagram in ModelBuilder. The finding also supports the validity of the aesthetic rule "*Minimize bends in the edge*" in general. The result is sustained not only by total time of response but by other eye-tracking metrics as number of fixations and scanpath length.

# 4 Discussions

The default option "Straight routing" of the connector is better than orthogonal routing. The default setting helps to design better aesthetic diagrams than diagram with bends on connector lines by the user in practice. We do not advice to users the intentional switching to a worse type of connectors. Our eye-tracking testing verified one rule from the set of aesthetic principles.

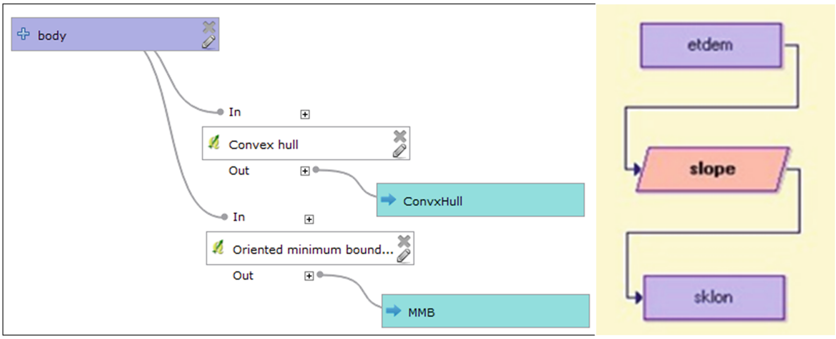


**Fig. 6.** Workflow with curved lines from QGIS Processing Modeler (left) and orthogonally curved lines in IDRIS Macro Modeler (right)

The result is also applicable to any other diagramming language in GIS and other IT diagramming fields. However, some GIS software does not support the variability of the shape of connectors. E.g. editor Processing Modeler for QGIS software use curved connector lines, and there is not possible to change to another shape. The connector lines are too long and space consuming of canvas (Fig. 6 left). Another example is IDRISI Macro Modeler in the area of GIS software. The lines in vertical (top-down) orientation are automatically orthogonally curved (Fig. 6 right). There is no possibility of user change. The aesthetic and cognitive quality of user workflow diagrams are under the influence of capabilities of the graphical editors and their limitations. The recommendation is: Draw or change to straight lines if there is an option in graphical editors for workflow diagram design.

**Acknowledgment.** This article has been created with the support of the Operational Program Education for Competitiveness – European Social Fund (project CZ.1.07/2.3.00/20.0170 Ministry of Education, Youth, and Sports of the Czech Republic).

## References

1. Dobesova, Z.: Data flow diagrams in geographic information systems: a survey. In: Proceeding of 14th SGEM GeoConference on Informatics, Geoinformatics and Remote Sensing, pp. 705–712. STEF92 Technology Ltd., Sofia (2014)
2. Boshernitsan, M., Downes, M.S.: Visual programming languages: a survey. EECS Department, University of California, Berkeley (2004)
3. Moody, D.L.: The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering. IEEE Trans. Softw. Eng. **35**, 756–779 (2009)
4. Tamassia, R.: On embedding a graph in the grid with the minimum number of bends. SIAM J. Comput. **16**, 421–444 (1987)
5. Battista, G.D., Eades, P., Tamassia, R., Tollis, I.G.: Graph Drawing: Algorithms for the Visualization of Graphs. Prentice Hall PTR, Upper Saddle River (1998)
6. Reingold, E.M., Tilford, J.S.: Tidier drawings of trees. IEEE Trans. Softw. Eng. **7**, 223–228 (1981)
7. Coleman, M.K., Parker, D.S.: Aesthetics-based graph layout for human consumption. Softw.: Pract. Exp. **26**, 1415–1438 (1996)
8. Gutwenger, C., Mutzel, P.: Planar polyline drawings with good angular resolution. In: Whitesides, S.H. (ed.) GD 1998. LNCS, vol. 1547, pp. 167–182. Springer, Heidelberg (1998). doi:10.1007/3-540-37623-2_13
9. Papakostas, A., Tollis, I.G.: Efficient orthogonal drawings of high degree graphs. Algorithmica **26**, 100–125 (2000)
10. Eades, P.: A heuristic for graph drawing. Congr. Numerantium **42**, 149–160 (1984)
11. Ware, C., Purchase, H., Colpoys, L., McGill, M.: Cognitive measurements of graph aesthetics. Inf. Vis. **1**, 103–110 (2002)
12. Figl, K., Mendling, J., Strembeck, M.: The influence of notational deficiencies on process model comprehension. J. Assoc. Inf. Syst. **14**, 312–338 (2013)
13. Störrle, H., Baltsen, N., Christoffersen, H., Maier, A.M.: On the impact of diagram layout: how are models actually read? In: MODELS (2014)

14. Störrle, H.: On the impact of layout quality to understanding UML diagrams: diagram type and expertise. In: IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp. 49–56 (2012)
15. Pődör, A.: Usability study on different visualization methods of crime maps. Int. J. Geoinform. **11**, 15–22 (2015)
16. Sedlák, P., Komárková, J., Hub, M., Struška, S., Pásler, M.: Usability evaluation methods for spatial information visualisation case study: evaluation of tourist maps. In: ICSOFT-EA 2015 - 10th International Conference on Software Engineering and Applications, Proceedings; Part of 10th International Joint Conference on Software Technologies, ICSOFT 2015, pp. 419–425 (2015)
17. Esri: What is ModelBuilder? http://desktop.arcgis.com/en/arcmap/10.3/analyze/modelbuilder/what-is-modelbuilder.htm
18. Dobesova, Z.: Using the physics of notations to analyse ModelBuilder diagrams. In: Proceeding of 13th International Multidisciplinary Scientific GeoConference, STEF 1992, pp. 595–602. Technology Ltd., Sofia (2013)
19. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye Tracking: A Comprehensive Guide to Methods and Measures. Oxford University Press, Oxford (2011)
20. Martin, D.W.: Doing Psychology Experiments. Wadsworth Cengage Learning, Belmont (2008)
21. Dobesova, Z.: Student reading strategies of GIS workflow diagrams. J. Adv. Soc. Sci. Educ. Humanit. Res. **70**, 319–325 (2016)

# 3D Models to Educated Museum Interactive Exhibition with Computing Techniques

Hao Jiang[1(✉)], Xiao-Li Liu[1], Xiang Peng[1], Ming-Xi Tang[2], Dong He[1], Hai-Long Chen[1], Kai-Bing Xiang[3], and Bo Man[4]

[1] Key Laboratory of Optoelectronic Devices
and Systems of Ministry of Education and Guangdong Province,
College of Optoelectronic Engineering, Shenzhen University,
Shenzhen, People's Republic of China
lxl@szu.edu.cn

[2] School of Design, The Hong Kong Polytechnic University,
Kowloon, Hong Kong
celestehao@gmail.com

[3] ESUN Ltd., Shenzhen, China

[4] Delta Dental, Alpharetta, USA

**Abstract.** The use of 3D scanner for the digitalization of 3D objects of cultural heritage is now an established approach. This paper aims at contributing to a better understanding of the emergence of new ideas in exhibition and education with such an approach. Researcher takes advantage of 3D optical technology cooperate with digital museum exhibition, the procedure of 3D scanning has been demonstrated in this research paper. Then, we create a mobile application which concentrate interactive design, computer technique in image tracking algorithm, and virtual model with 3D optical technology. The aim of research in virtual museum opens up a new area for the museum education and interaction. The results we obtain allow us to devise methods for enhancing creativity.

**Keywords:** 3D optical scanning · Interactive museum · Computer technique · Education

## 1 Introduction

Physical museums have been a common medium for presenting history and culture to the public for a long time. The visitors to the museum are able to see the real exhibits and explore the knowledge related to the contexts. However, many museums have not enough spaces and facilities to represent the culture heritages in different views of angles (De Backer et al. 2014). Therefore, digitalization and virtualization are gradually changing the museums settings (Padilla-Meléndez and del Águila-Obra 2013). Digital museums with the technology of virtual reality can bring the exhibition of culture relics into a new generation. The digitalization of culture heritages can enhance and complement the museum experience through interactivity and personalization of the contents (Ikei et al. 2013) with 3D scanning and 3D printing techniques (Douma et al. 2010). In this paper, we present a structured light scanning technology used for 3D

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_18

reconstruct of culture relics. An interactive platform has been generated in an iOS system using this technology.

## 2 Literature on Digital Museum

A digital museum is also called a virtual museum which can perform as the digital footprint of a physical museum (Styliani et al. 2009). Comparing with a traditional museum, a digital museum can be designed for a variety types of exhibition without the location and time limitation. In the early years of 90s, many museums around the world considered to break through the limiting stereotype, and tried to find new ways for culture communication.

WebExhibits founded in 1999 by Michael Douma at IDEA (Franklin 1999) is an interactive and cross-curricular virtual museum of science, humanities, and culture for K-12 and higher. It takes advantage of virtual experiments, and hand-on activities to guide visitors to formulate questions, and to think topics from different point of angles (Pipes 2003). WebExhibits taps into experts' knowledge and technology's potential to create a rich, immersive experience that incorporates narratives, descriptions, maps, photos, video, and audio. In addition to spurring the informal learning process, WebExhibits also supports structured educational efforts. Nine exhibits can run online, such as "The Causes of Color" used to explore why we see the brilliant colors of butterflies like Blue Morpho. Except for the exhibits, a link to teaching resources includes a toolbox of tips for using the exhibits for K-12 and higher (Melber and Hunter 2009).

The Virtual Museum of Canada (VMC) is a directory of over 3,000 Canadian heritage institutions and a database of over 600 virtual exhibits. It brings together Canada's museums regardless of size of geographical location (VMC 2014). The VMC includes virtual exhibits, educational resources for teachers, and over 900,000 images. All the resources are bilingual which are English and French. It provides an online environment for Canadian communities to tell their stories and preserve their history (VMC 2014). The other offering is the Community Memories Program, which is an investment program designed for smaller Canadian community museums, to allow them to create online exhibits about their history. Many others museums were now running online after a decade of development. Google art project is one of online platforms which allowed users to access high-resolution images of artwork in the collaborative museums (Valvo 2012). The project was completed in cooperation with 17 international museums, and launched on 2011. This platform enables users to virtually tour partner museum's galleries. The users could explore physical and contextual information about artworks, and compile their own virtual collections (Finkel 2012). There are three main components of the site including virtually gallery tour, artwork view, and creating an artwork collection in the first generation (Coombs and Ahmed 1973). When it is developed into a mature period, the features were enhanced for exploring and discovering video and audio contents in education. In the second generation of art project, Google updated searching capabilities for users to find the artwork from different categories to fit their parameters of interests (Berwick 2011). Some of partner museums design a virtual tour of their galleries. Users are able to walk through the museum with audio guide when visiting. In addition, several educational tools and resources were used to support teachers and students. Google created a multitude of educational videos embedded on the webpage or available through YouTube channel. And two pages named “Look like an expert” and “DIY” provide visitors lots of teaching resources and toolkits for learning (Pack 2011).

In order to realize this art project plan, Google team takes advantage of the existing technologies, such as Google street view and Picasa, which are the new tools built specifically for the art project. The Google street view camera can capture 360 degree images as it moves through the location (Fig. 1). Generally, the camera sits atop a car to capture street view images, and the art project camera was installed on an indoor trolley. The team created an indoor-version of the google street view, i.e., 360-degree camera system, to capture gallery images by pushing the camera “trolley” through a museum. The two professional panoramic heads LAUSS RODEON VR Head HD and CLAUSS VR Head ST (Proctor 2011) are used to take high resolution photos of the artworks within a gallery. Only this technology allowed to achieve the excellent attention to details with the highest resolution. Every partner museum selected one artwork to be captured at ultra-high resolution with approximately 1,000 times more details than the average digital camera (Davis 2012). The street view technique is a great help for artworks like painting and sculpture.

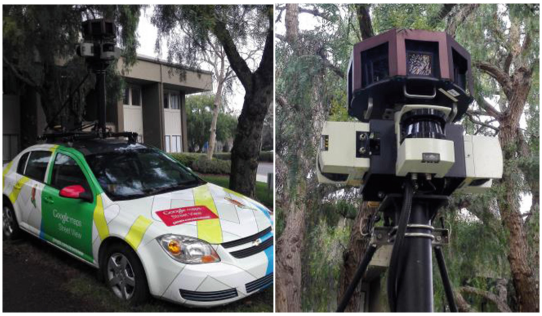


**Fig. 1.** Google street camera

However, there are many others culture relics in the museums in China which included Chinese Bronzes, golden and silver artifacts with complex textures (Mediati 2011). The 3D imaging technique can help solving the problem of complex texture capturing in order to realize full angels display. In the research presented in this paper, the researchers cooperated with Shenzhen museum in China, to explore innovative 3D optical scanning technology applied in culture relic’s protection and museum education.

## 3 Case Study with Shenzhen Museum

Shenzhen Museum is a comprehensive museum that has become an important cultural facility of Shenzhen, giving full play to its function of cultural relic collection, propaganda, education, and scientific research. Up till now, there are more than 20,000 pieces of cultural relics, such as the specimens of paleobioligical fossils dated 100 million years ago, ancient history and art treasures, showing 5000 years of civilization of China. It also has important historic materials about the development history of modern and contemporary Shenzhen. These precious cultural relics are the important cultural foundation of Shenzhen, as a city created only some 30 years ago. As an education base for the youngsters, Shenzhen Museum has kept on perfecting its permanent exhibitions since the very beginning. The permanent exhibitions used full and accurate materials to tell visitors the history of development process in Shenzhen region, and to introduce how our ancestors worked, struggled, and created many wonders with their wisdoms. Aim to enrich children's natural scientific knowledge, in the exhibitions of Marine Organisms and Wild Animal Specimens, Children can find out the natural resources around them from different shapes of animal specimens, which educates people to love the nature and to protect the natural environment. The museum also introduces many precious ancient cultural relics and many different schools of art from home and abroad to exhibit here at irregular intervals. As a research organization of social science, Shenzhen Museum has published some monographs on professional area, and all the scientific research results have attracted great attention of researchers around the world. Therefore, Shenzhen Museum has become an important research organ for the study of Shenzhen history and the history of Hakka culture. The case study we conducted is a collaboration with Shenzhen museum in relation to the exhibition of bronze relics. The museum intends to create an online exhibition to explore and learn the relics in virtual environment. This platform aims at educating users to study the history, material, and texture of the relics. The main technologies of platform include 3D scanning and imaging tracking.

## 4 The Process and Device of the 3D Scanning System

Culture heritages are precious items in museum exhibitions. Generally, each heritage has its own protector to prevent corrosion damage. However, it is inconvenient to get the 3D measurement data of items without physical contact with conventional methods. Non-contact 3D scanning technology could analyze a 3D object to collect data on its shape and color, and the collected data can be used to construct digital 3D models. As we know, cultural relics placed in the open air for a long time will be subject to a certain level of damage. Therefore, shortening the time of data collection will be the main issue that needs to be fixed.

Structured-light 3D scanner is a scanning system device for measuring the 3D shape of an object using projected light patterns and a camera system. The data taken this way has low noises and can be ordered. The system can be adjusted to take very small pictures for smaller parts or large photos for larger parts.

In this research, the structure light scanning system has been used, which is named "3D digital platform of culture relics based on binocular stereo vision and temporal phase unwrapping technology." This platform has been specially designed for the precious and vulnerable culture relics. The strength of optical 3D imaging is that it can avoid multiple moving objects and 3D scanning platform around the relics in 360 degrees, to capture data with a texture through the annular soft track. The lifting frame

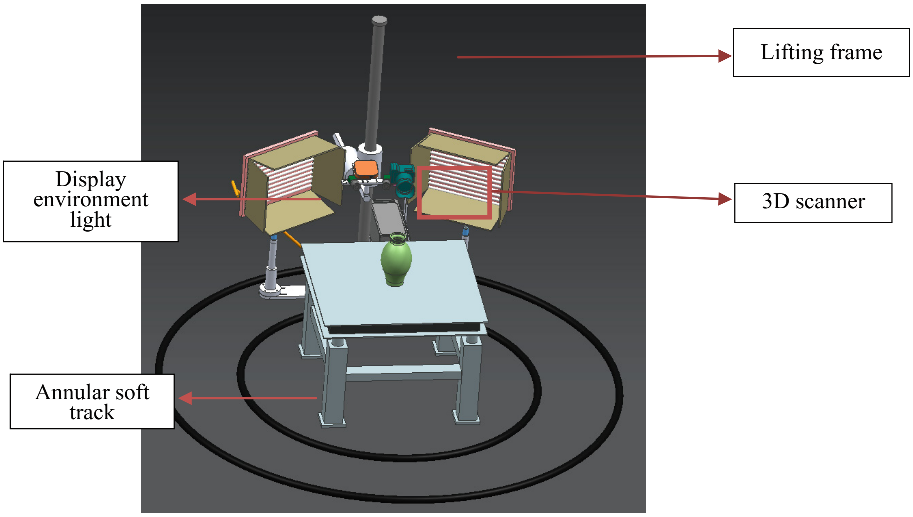


**Fig. 2.** Scanning platform_1

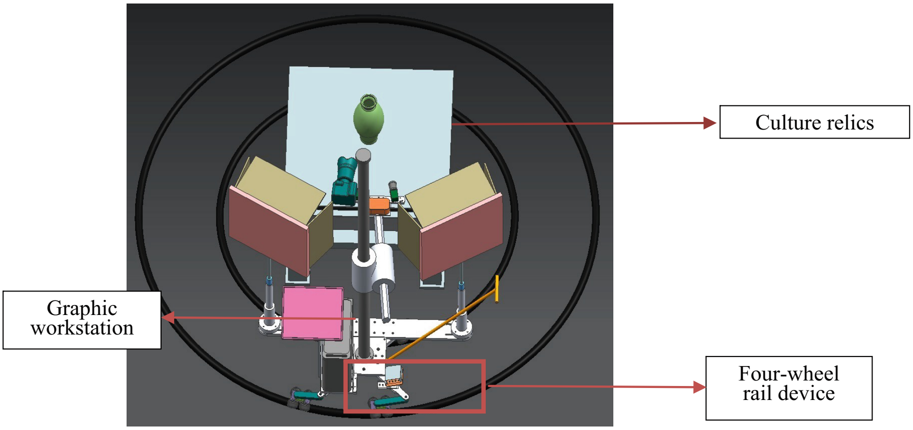


**Fig. 3.** Scanning platform_2

can be adjusted for the scanner to facilitate relics scanning in different point of views, in order to minimize the frequency of flip relics, and to protect the culture relics as much as possible. There are two high resolution CCDs, a projection device, a digital single lens reflex and ambient light system, to construct this scanning platform (Figs. 2 and 3).

This system can obtain the detailed data of an object quickly, and capture texture data under a uniform illumination condition control. Then, this is combined with self-developed imaging and modelling software to realize functions of 3D reconstruction, detailed data matching and integration, texture data mapping and integration. The accuracy of scanning field can be adjusted according to the requirements.

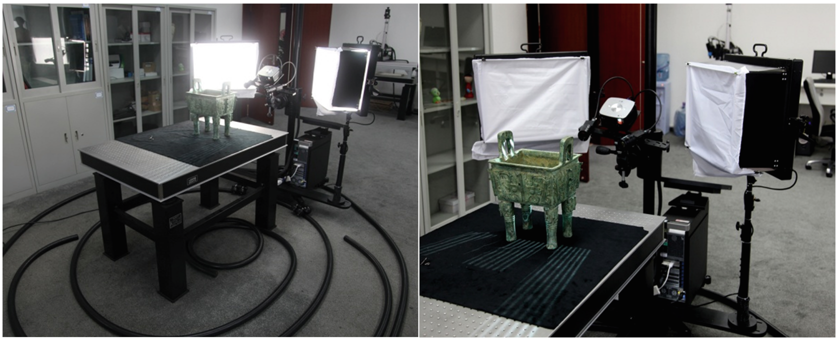


**Fig. 4.** The real scanning platform

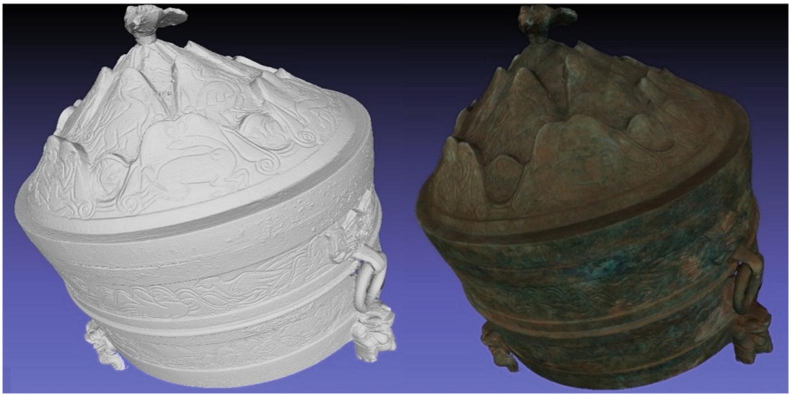


**Fig. 5.** The 3D reconstruction of relics

The minimum field of view is 10 mm × 10 mm, the accuracy of single-sided model is 0.01 mm, and the dot pitch of model is about 0.05 mm. In addition, two professional flashlights are able to create a uniform soft light filed to improve the quality of texture which is shot by SLR camera. The graphic workstation in high performance is used to calculate 3D scanning results quickly.

Generally, the speed of single-sided scanning and calculation is 30 s/f (Fig. 4). This system used for the 3D digital museum of culture relics in Shenzhen museum (Fig. 5).

## 5 Image Tracking Algorithms

In order to track the image and character in virtual environment, the open source Vuforia tools have been used for the environment development. Vuforia SDK is the main toolkit for imaging detecting and tracking development. It is an augmented reality software development kit for smart and similar mobile device that enables the executes AR function into real time video camera which are obtained in these kinds of devices.

The toolkit uses Computer Vision technology to recognize and track planar images (Image Targets) and simple 3D objects, such as boxes, in real-time. This image registration capability enables developers to position and orient virtual objects, such as 3D models and other media, in relation to real world images when these are viewed through the camera of a mobile device. The virtual object then tracks the position and orientation of the image in real-time so that the viewer's perspective on the object corresponds with their perspective on the Image Target, so that it appears that the virtual object is a part of the real world scene. The Vuforia SDK supports a variety of 2D and 3D target types including 'markerless' Image Targets, 3D Multi-Target configurations, and a form of addressable Fiduciary Marker known as a Frame Marker. Additional features of the SDK include localized Occlusion Detection using 'Virtual Buttons', runtime image target selection, and the ability to create and reconfigure target sets programmatically at runtime. Vuforia provides Application Programming Interfaces (API) in C++, Java, Objective-C, and the Net languages through an extension to the Unity game engine. In this way, the SDK supports both native development for iOS and Android while also enabling the development of AR applications in Unity that are easily portable to both platforms. AR applications developed using Vuforia are therefore compatible with a broad range of mobile devices including the iPhone, and iPad with iOS and Android phones and tablets running Android OS version 2.2 or greater and an ARMv6 or 7 processor with FPU (Floating Point Unit) processing capabilities. Figure 6 presents two example of working process and final results with tracking algorithm. There are several recognizing points which will be generated on both images and characters, and the information will appear on monitor when camera faces to the objects.

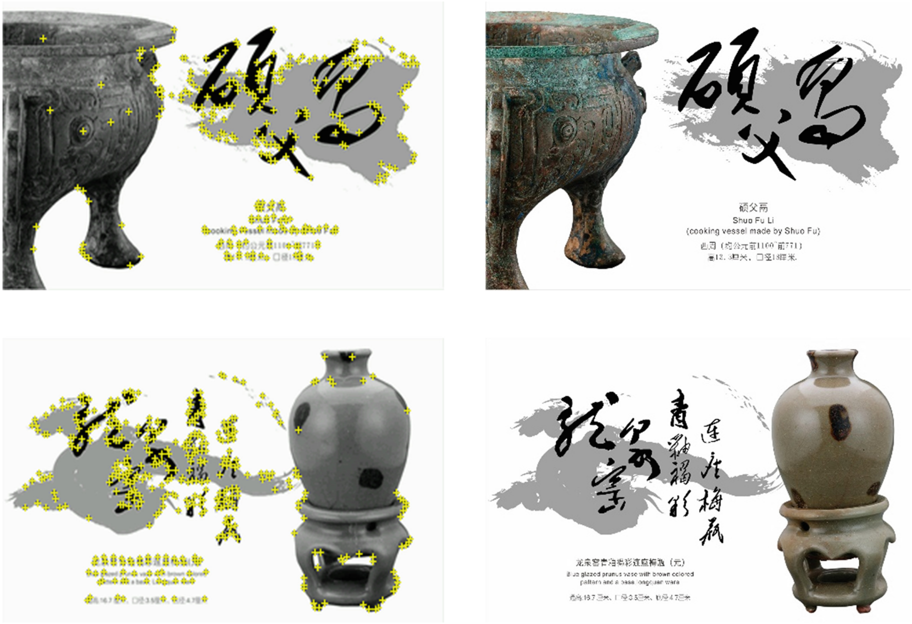
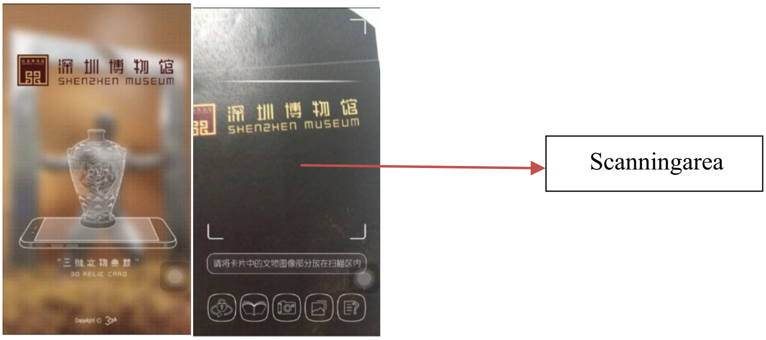

硕父鬲
硕父鬲
Shuo Fu Li
(cooking vessel made by Shuo Fu)
西周（约公元前1100~前771）
高13.3厘米，口径18厘米.
龙泉窑
青釉褐彩
莲座梅瓶
龙泉窑青釉褐彩莲座梅瓶（元）
Blue glazed prunus vase with brown colored
pattern and a base,longquan ware
通高16.7厘米、口径3.5厘米、底径4.7厘米

**Fig. 6.** Imaging tracking realization with Vuforia

# 6 The Demonstration of Digital Museum Exhibition

There are five functions buttons in the bottom of interface from left to right, which are object locking, introduction, photo taking, background selection, and help (Fig. 7). The object locking function is used to stop the autorotation of object for details viewing. The introduction function with background music and guide audio provides the information of relics for learning. The users are able to use photo taking function to save picture to their mobile phones, and they could change the background of the pictures if they want through background selection function. The middle of interface is scanning area for the objects, both characters and image are available for this function.

There is a demonstration of a relic named Celadon tripod with overhead handle (Fig. 8). The model will be display on the screen which is scanned from a postcard. The information of this relic which includes dynasty, dimension, and appearance will be appeared on the screen as well.



请将卡片中的文物图像部分放在扫描区内

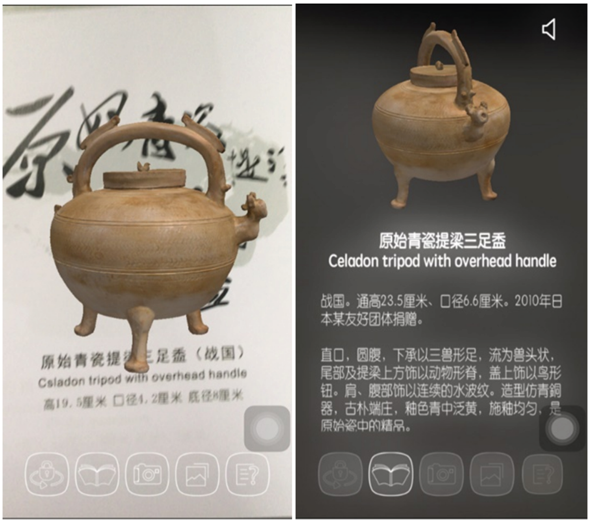
**Fig. 7.** The main interface of application

## 6.1 Results and Feedback

There are 200 people participated in the survey, and the effective questionnaires are 185. Researcher prepared six questions for participants. Each question has five options which are 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree.

Q1. I would like install the application through two-dimension code scanning during Museum tour.
Q2. The interface attract my attention.
Q3. The application easy to use.
Q4. I like the interaction mode of this application.
Q5. I think this application could help museum education and culture.
Q6. I will recommend this application to others.

From the data analyzation which is presented in Fig. 9, the most of people think this application would help them in museum education and culture learning, the


原始青瓷提梁三足盉（战国）
Csladon tripod with overhead handle
高19.5厘米 口径4.2厘米 底径8厘米

原始青瓷提梁三足盉
Celadon tripod with overhead handle

战国。通高23.5厘米、口径6.6厘米。2010年日本某友好团体捐赠。

直口，圆腹，下承以三兽形足，流为兽头状，尾部及提梁上方饰以动物形脊，盖上饰以鸟形钮。肩、腹部饰以连续的水波纹。造型仿青銅器，古朴端庄，釉色青中泛黄，施釉均匀，是原始瓷中的精品。

**Fig. 8.** Celadon tripod with overhead handle

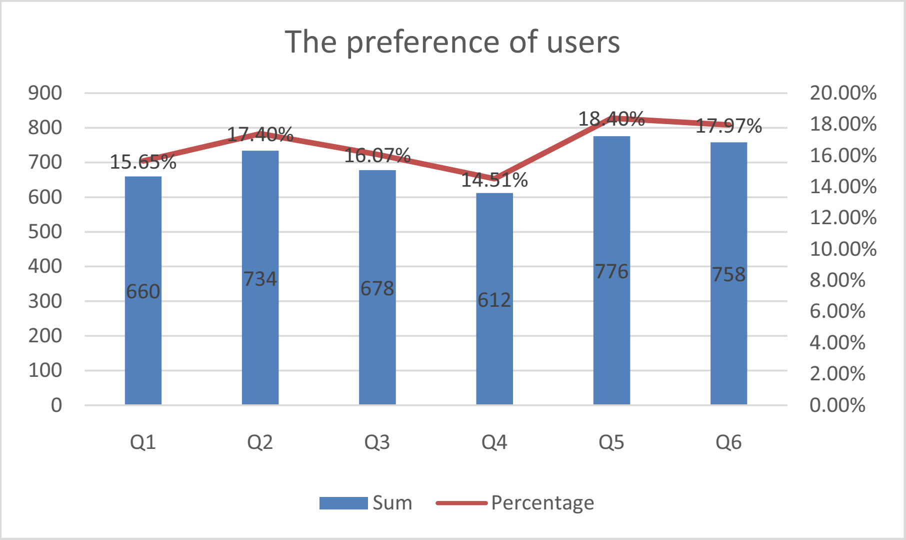


**Fig. 9.** The data of feedback

percentage is 18.40% which is means they strongly agree of this option. And they will recommend their friends, colleagues, and family member to install this mobile application. And the most important factors lead them to use this application is because of the interface attracted their attention. In additional, researcher noticed that the lowest percentage which is 16.07% presented a part of users feel the application difficult to use. Many of users think they cannot rotate virtual model very skillfully, and another problem is audio guide unable to pause during play. And some foreigners prefer English version of this application. Researcher will continue to optimize and update the application in future works. In any case, the results are in line with our expectations.

## 7 Conclusions

In this research study, the combination of 3D scanning and image tracking technologies to create an innovative non-formal museum education paradigm is used. It is different from traditional digital museum exhibitions which are always dependent on Web and Internet techniques (Falk and Dierking 2000). Many museums provide interactive monitors for guests to use during the tour. However, the web technique is unable to realize simulation under virtual environment with portable devices. Therefore, an application for mobile phone provides the opportunities for all the people who are interested in understanding the culture of history relics. All the digital resource can be downloaded from the official website of the museum to support learning more effectively. This type of experience created in the virtual museum opens up a new area for the museum education and interaction (Ansbacher 1998). It will be rapidly becoming possible to create interactive education and culture patterns using high quality 3D modeling.

**Acknowledgements.** This work has been partially supported by Esun. Ltd, Shenzhen, and Shenzhen Museum, by the Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Optoelectronic Engineering, Shenzhen University.

**Funding Sources.** This work was supported by the research project 2016M590801 sponsored Chinese Post-doctoral Science Funding.

## References

Ansbacher, T.: Learning in the museum [Review of the book learning in the museum]. Curator **41**(4), 285–290 (1998)

Berwick, C.: Up Close and personal with Google art project. Art Am. **99**(4) (2011)

Coombs, P.H., Ahmed, H.: New Paths to Learning: For Rural Children and Youth, pp. 71–72. International Council for Educational Development Publications, New York (1973)

Davis, J.: Google Art Project: Behind the Scenes. Tate Blogs. Tate Britain (2012). Accessed 24 Mar 2012

De Backer, F., Peeters, J., Buffel, T., Kindekens, A., Reina, V.R., Elias, W., Lombaerts, K.: An integrative approach for visual arts mediation in museums. Proced. – Soc. Behav. Sci. **143**, 743–749 (2014)

Douma, M., et al.: Concept maps for on-line exhibits: using SpicyNodes. In: Trant, J., Bearman, D. (eds.) Museums and the Web 2010: Proceedings. Archives and Museum Informatics, Toronto (2010)

Falk, J.H., Dierking, L.D.: Learning from Museums: Visitor Experiences and the Making of Meaning. AltaMira Press, Walnut Creek (2000)

Finkel, J.: LACMA, Getty among 134 museums joining Google's art site. LA Times (2012). Accessed 6 Apr 2012

Franklin, J.L.: Show us what you've got. In: The Boston Globe, 2 September 1999, p. 2. West Weekly (1999)

Ikei, Y., Abe, K., Masuda, Y., Okuya, Y., Amemiya, T., Hirota, K.: Virtual experience system for a digital museum. In: Yamamoto, S. (ed.) HIMI 2013. LNCS, vol. 8018, pp. 203–209. Springer, Heidelberg (2013). doi:10.1007/978-3-642-39226-9_23

Leah, M., Melber, A.H.: Integrating Language Arts and Social Studies: 25 Strategies for K-8 Inquiry, p. 113. SAGE (2009)

Mediati, N.: An extension of Google street view enables interactive. Web-based Virt. Mus. Tours **29**(4) (2011) (PC World)

Pack, T.: The Google Art Project is a Sight to Behold. Inf. Today **28**(5) (2011)

Padilla-Meléndez, A., del Águila-Obra, A.R.: Web and social media usage by museums: online value creation. Int. J. Inf. Manag. **33**(5), 892–898 (2013)

Pipes, A.: Foundations of Art + Design, vol. 264. Laurence King Publishing, London (2003)

Proctor, N.: The Google art project: a new generation of museums on the web? Curator: Mus. J. **52**(2) (2011)

Styliani, S., Fotis, L., Kostas, K., Petros, P.: Virtual museums, a survey and some issues for consideration. J. Cult. Herit. **10**(4), 520–528 (2009)

Valvo, M.: Google Goes Global with Expanded Art Project. Press Release. Google Art Project (2012)

Virtual Museum of Canada (2014). http://www.museevirtuel-virtualmuseum.ca/. Accessed 16 May 2018

# A Framework for Image Synchronization from Mobile NoSQL Database to Server-Side SQL Database

Abu Zarin Zulkafli[1](✉), Shuib Basri[1], Rohiza Ahmad[1],
and Abdullahi Abubakar Imam[2]

[1] Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Perak, Malaysia
{abu.zarin_g03382, shuib_basri,
rohiza_ahmad}@utp.edu.my
[2] Computer Science Department, Ahmadu Bello University, Zaria, Nigeria
aiabubakar@abu.edu.ng

**Abstract.** Last decades, software systems use SQL database to handle structured data between server and clients that hold text and integer value in column-store format. Nowadays, emergence of media entertainment had introduced unstructured data like images that acquires big memory allocation to be stored in database. Hence, some of software and mobile application developers tend to implement schema-less databases, which known as NoSQL. They handle synchronization process for unstructured data in a simple and fast manner. However, dissimilar database structure and framework make software developers tend to implement only either one type of databases: SQL or NoSQL, but not both. No interaction between SQL and NoSQL databases will prevent data sharing and accessibility. Hence, one-way image synchronization framework for heterogeneous databases in mobile environment has been developed and new sync gateway is introduced to synchronize images from client NoSQL database to server SQL database regardless of different schema and database structure.

**Keywords:** RDBMS · SQL · NoSQL · Mobile database · Server · Synchronization · Image

## 1 Introduction

Synchronization is an activity of transferring data between two entities, which aims to have same data in both entities. According to [1], it is either transferring data from server to client or vice versa. Synchronization consist of two operations which are upload and download, and upload operation is executed first and followed by download operation to get equivalence data for both interacted database [2]. Synchronization methods depend on type of database management system (DBMS) supports such as trigger, timestamp, meta-stored procedure which bring varieties of synchronization result performance. Relational Database management system (RDBMS) also known as SQL database is the developers' choice for almost 30 years back as the emergence of webs application development especially in late 1990's to handle simple structured data like text and numeric value. However in this 21st century, clients tend to operate the

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_19

system using mobile devices and emergence of media had introduced new database management system: NoSQL, Schema-less database that able to handle better unstructured data which handle big size of media data and high processing power to perform synchronization.

According to [3], the emergence of advance applications that extend its functionality and requirements, resulting the SQL database unable to fully fulfil the requirements in distributed environment. Besides, implementation of SQL database type alone unable to perform very well in a distributed system and cannot provide well availability of data [4]. This is because SQL database does not scale out very well in a distributed system. Hence SQL database need to operate together with NoSQL database, which according [5], that NoSQL database provide simpler structure and faster database management in distributed environment to bring high synchronization performance to server and mobile device client as well.

The paper is organized as follows, section literature review will explain the existing synchronization solutions, problems and NoSQL capabilities as well. Methodology section will discuss proposed new sync gateway integrated components to make possible synchronization between dissimilar databases type to interact. Besides, proposed framework for one-way image synchronization from NoSQL mobile client database to SQL server database is presented. Lastly, conclusion and future works recommendation.

## 2 Literature Reviews

According to [6], more than 80% of potential business information nowadays are in unstructured data form. Unstructured data consists of various formats like image, video, document etc. This unstructured data provides burdensome to synchronize compare to structure data in mobile environment that have limitations in terms of processing power, network bandwidth and battery capacity [7]. For example, unstructured data like images and documents need to go through binary diff, and data compression to reduce network consumption in order for mobile client to synchronize with Dropbox [8]. According to Li et al. [8], some of the synchronization method in cloud storage application facing traffic overuse problem that create network overhead. Therefore, they proposed Update-batched delayed synchronization (UDS) mechanism, which aims to reduce the overhead and preserving the rapid file synchronization from client's file storage to cloud storage application. This shows that unstructured data need to have extra procedure and additional mechanism to reduce mobile client's network data usage and transfer speed. Hence, image synchronization approaches need to be efficient in server-mobile client environment by reducing the burden of computing capacity in mobile client side. As NoSQL database characteristics are able to lessen processing time [3] simpler and faster [5] and schema-less, we believed by implementing NoSQL database in client side could help in reducing clients' mobile devices computing resources and synchronization process time. Hence, this framework proposed the implementation of NoSQL database type in mobile clients, in order to conduct experimentation in terms of transfer speed and memory consumption performance comparison with mobile client's SQL database. This is to justify the claimed made by NoSQL database about its performance through experimentation.

What are the benefits if the proposed framework for image synchronization between heterogeneous databases? As stated by [9], when heterogeneous databases are able to communicate with each other, huge advantages for the engineering, manufacturing and business operations as efficiency improvement in data sharing and availability. Thus, with this proposed framework, there will be an efficient data sharing environment whereas images are able to transfer between dissimilar database structures which brings high impact in engineering and manufacturing activities and business operation for mobile business system. Images can be shared across the platform, thus create new potential in manufacturing activities and business operation.

However, many synchronization approaches either in server and client side are dependent only on proprietary implementation of DBMS which is vendor specific. In other words, only mobile database and server side database that have identical type of database structure, framework and programming language are able to synchronize without restrictions and conflicts [1]. Different type of vendors, devices operating system and different database management, make synchronization process are facing problems as having different framework, data model and synchronization method as well. In addition, every new release of DBMS, operating system or even new version of existing DBMS, there is unclear gateways such as framework and data model that are provided by anyone to support synchronization process between heterogeneous databases that resulting in difficulties of reconciliation between databases [9].

Sometimes, programmers even need to modify existing applications for synchronization process for reconciliation of dissimilar databases [10]. Reconciliation of data synchronization method of these dissimilar databases could be perform by manual query programming between SQL and NoSQL language [11]. Thus, this had proved establishment of connection for dissimilar database is possible.

### 2.1 Why NoSQL Is Introduced?

Web 2.0 applications store the attributes in numerous tables like text, comment, image, video and source code. To support easy schema evaluation, underlying databases have to be flexible since web applications are very agile and NoSQL offers that flexibility [12]. SQL database is not able to add and remove features when it comes to system unavailability. In addition, NoSQL manages to handle massive data set by indexing arbitrarily and at the same time enabling a large amount of concurrent user request [13].

In addition, according to [14], to support mobility, low bandwidth variability and heterogeneous networks and security, the new database processing schemes need to be introduced because old type of databases schemes are not suitable to be applied more to provide effective synchronization process. Mobile databases need to be synchronized with server-side database in order to get updated data version, avoid data inconsistencies and maintain integrity of data. Thus, new efficient synchronization method and database type choices in mobile devices are extremely important in producing great synchronization process performance.

As stated by [15], NoSQL databases able to extend its performance and scalability compared to SQL by abandoning atomicity, consistent, durable, isolated (ACID) properties. This statement also supported by [16], which he properly explained that NoSQL are introduced to achieve more availability and scalability, which SQL unable to achieve as better as NoSQL, especially when handling with big data. [17] stated that SQL databases unable to provide enough computational and storage allocation when handling massive data. [16], stated that NoSQL is becoming popular approach in handling the data as the its enhance scalability, and practicing non-relation data stores and some of the database like Cassandra and Voldemort had proven that NoSQL able to produce cost effectiveness approach when handling the data in databases by using distributed key-value stores [18].

In conclusion, this paper will be focusing on implementing NoSQL database in mobile client sides, which to reduce the computing and memory resources of mobile device by adopting NoSQL characteristics which is simpler deployment and schema-less. The synchronization process to SQL server side will follow proposed sync gateway and framework thus will adopt some previous synchronization techniques between SQL server and SQL client like timestamp, triggered and meta-stored procedure by introducing new technique for heterogeneous databases able to synchronize successfully.

## 2.2 Existing Approach

Suitable development toolkits used to create mobile applications and database connection are in need to create communication links between server and mobile client. For example, [19] approach to use Android Cloud to Device Messaging (C2DM) is already deprecated on October 2015 hence new enhancement or migration of project need to use another toolkits like Google Cloud Messaging (GCM). Empirical study has been done to study the use of tools used to implement mobile application and database connection. Table 1 shows the choice of researchers in setting up the environment to implement database and database connectors from mobile devices to server as well.

**Table 1.** Database connections and development kits

| Author | DB connection | Development kits |
|---|---|---|
| Choi et al. (2010) | Java Database Connection (JDBC) | Android |
| Ajila et al. (2011) | MS sync services | .NET framework |
| Balakumar et al. (2012) | Java Database Connection (JDBC) | Android |
| Sedivy et al. (2012) | MCSync API, GAE's datastore | C2DM |
| Ramya et al. (2012) | Java Database Connection (JDBC), SSH | Android, GPRS |
| Alhaj et al. (2013) | HTTP | Android |
| Sethia et al. (2014) | HTTP | MySQL |
| Gupta et al. (2014) | Java ADT eclipse IDE | Android XML |
| Zaia et al. (2014) | HTTP | MSqlite |
| Imam et al. (2015) | HTTP | Android |

Most of the researchers used Java language to develop mobile application and connecting to the database. [7, 20, 21], use Android platform and most of them use JDBC to connect to the database. According to [19], HTTP is also popular approach in connecting databases as it is common practice and follow popular REpresentional State Transfer (REST) architecture. REST architecture able to create simple and extensible web-service, HTTP was implemented by [20, 22].

From the study above, it also shown that SQL databases are being used in almost all both server and client databases. MySQL, Microsoft SQL server and SQLite being selected as databases to store and transfer the data. Unlike [19], they implement two databases, SQLite as SQL database to interact with Android client, while BigTable as NoSQL database to serve Web clients by using Google App Engine's datastore as mediator. According to [23], they are implemented 3 type of databases in distributed environment consist of MySQL database on server side and SQLite and XML database in mobile client side. XML database is used by Imam et al. [23] to create a generic data model for data synchronization between heterogeneous databases by using JSON language as the intermediary compound to execute transaction process of structured and semi-structured data. Above all, recent researches like [19, 23] showed pattern towards reconciliation of different type of databases to adapt the heterogeneity environment as cause from variety of vendors, DBMS, and operating system nowadays.

### 2.3 Existing SQL Server to SQL Mobile Client Synchronization Solution

Study by [24] introduced Synchronization Algorithm Based on Message Digest (SAMD), to synchronize business data between Server and mobile client. There are 3 synchronization stages, which 1st and 2nd synchronization are conducted between Database Server Data Table (DSDT) and its message digest table Database Server Message Digest Table (DSMDT). The 3rd between Mobile Client Data Table (MCDT) and Mobile Client Message Data Table (MCMDT) (Fig. 1).

To reduce the computing power consumption of mobile side, the MCMDT is located inside database server and MCMDT has smaller size compared to DSMDT as every mobile devices has its own mobile id (Mid) and its own MCDMT. Any changes in DSDT and MCDT will be flagged (F) as '1' to state that the row need to be updated and will be synchronized with DSMDT and MCMDT to get the latest data. The 3rd synchronization will be conducted between DSDT and MCDT (only selected Mid). This process will update any changes in DSDT will be updated to MCDT and vice versa. However, synchronization conflicts occurs when both data tables have been changed (insertion, deletion, modification), thus the researcher introduced 4 Cases Analysis to prioritize which type of data changes will be updated in both databases.

The approach using message digest able to reduce size of data and creating MCMDT in server side has achieved the objective to lesser the mobile devices processor burden during synchronization. This approach is for SQL database in both sides and the use of message digest is important to reduce the size of data by encoded it to using md5.

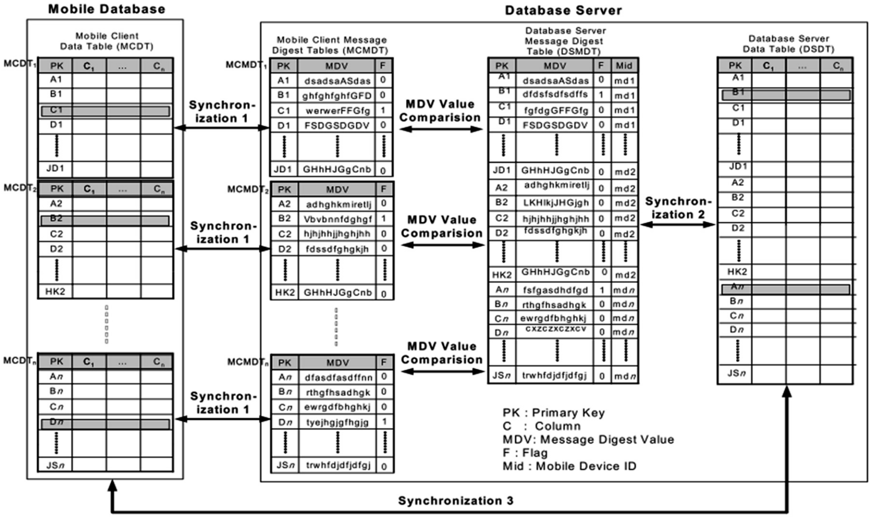


**Fig. 1.** Synchronization Algorithm Based on Message Digest (SAMD) for mobile DB [24]

## 3 Methodology

This framework will execute synchronously, which allocate one specific time to this synchronization thread on the mobile processor. Mobile processor will pause all other process until the synchronization process is completed. Server side consists of RDBMS which is MySQL DB that carries SQL data model behavior, which is table form database type whereas mobile client side consists of NoSQL Couchbase Lite DB, that stores data in schema-less JSON Native Document database. Components to represent the two type of database characteristics are chosen in Fig. 2 below.

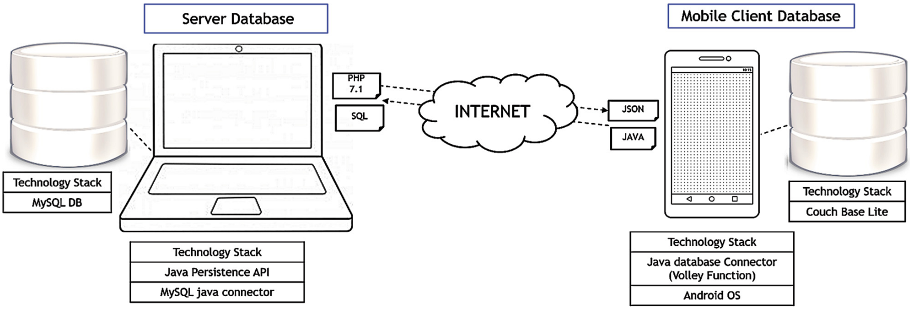


**Fig. 2.** System architecture

From Fig. 2, synchronization process of images needs wireless internet connectivity. Volley function in Android library is used in sync gateway to create connection between these two data-bases. Base64 is used to encode binary files such as images within scripts, to avoid depending on external files. Hence the image are able to save in databases in encoded forms thus its lighter and faster to perform image transaction, reducing the time for apps and server have to load will naturally make that apps faster. Base-64 primary purpose is to make it efficient to store large binary data in a document. Binary data stored in JSON has to be base64-encoded into a string, which inflates its size by 33% [25]. Sync Gateway for dissimilar databases and image synchronization framework from NoSQL client DB to SQL server DB framework will be discussed in next section.

### 3.1 Novel Sync Gateway

Sync gateway for dissimilar databases type will be focused on the compatibility of communication protocol between SQL and NoSQL database. This sync gateway consist of several class functions. This class function will manage image data structure, insertion and deletion of images and also information for databases connectivity and error handling. Figure 3 shows components in proposed sync gateway between MySQL – Couchbase lite db.

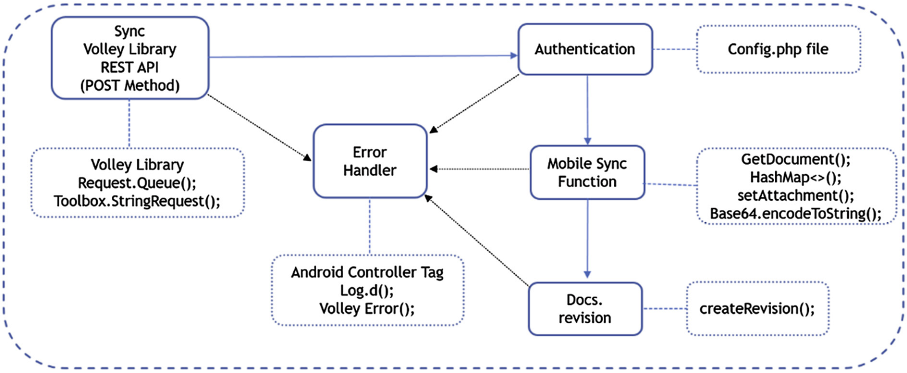


**Fig. 3.** Sync gateway architecture & components for MySQL - couchbase lite

From Fig. 3, Volley function is use to pass its request objects in paramater. The parameter consists of 3 elements which are POST method, server URL, and Listener to do error checking while synchronizing the image to server. Request to sync client DB with server DB will be put in Requestque(). Requestqueue() will handle multiple request, where when one request is added, it is picked up by the cache thread and triage and will be place on network queue.

This volley function manages the dispatching parsed response back to the main thread. Thus, the first request in requestqueue() will perform POST transaction, parses the response on the worker thread, write the respond to cache, and post the parsed response back to the main thread for delivery to server. This volley requires internet permission in client android manifest. Images information which consist of ID, name, size, base-64 encoded image content will the be put in POST parameter upon delivery data to server through the pipeline.

Mobile Sync function components consist of two major functions that carry insertion and deletion of images activities in client's side. Image is either capture using phone camera or from image gallery. The image is then given their own image ID number and encoded to string using Base64.encodeToString() function before saving into couchbase lite. db.Manager() is created to handle all transaction insertion and deletion activities, images sync upload status either 'True' or 'False' in couchbase lite db. GetDocument() function is use to retrieve Couchbase's JSON document that stores encoded image. HashMap <> is use to map keys to values which implements hash function to compute an image index into an array of buckets or slots. This means image ID is an index key that maps into their 64-encoded image string values.

Image in client database is deleted by following several procedures. Firstly, get.-document(ID) function will retrieve image ID that selected by client to delete it from client database. Then, function deleteFromserver() will be executed that pass the image name in POST parameter located in Volley function. Image name will be parsed into PHP file in the host server that will delete the image in the server that have the same image name in POST parameter.

Error Handler component has major responsibility in tracing error in android java classes. The error in Volley function while estabilishing connection to server, will be traced in the log cat in android studio. Besides, error while inserting new images and deleting existing images in client DB is notified to client through log cat and android 'toast' in mobile device's screen. This will track synchronization activities either the functions executed in normal manner or having problem.

## 3.2 Image Synchronization from Client to Server

This image synchronization framework is one way direction, from client to server. An android apps was developed to retrieve images from camera or SD card and save it to the client NoSQL database before synchronizing to SQL server. Then, images status either already synchronized or not to the server are based on two image status in client DB, 'True' for already synchronized to server, and 'False' is not synchronize yet. Only images with 'False' status will be synchronized to server once apps synced button is pressed. Figure 4 below shows the overall view of framework for image synchronization.

Mobile client sends request to synchronize with server. Once request is received, the images with 'false' status will be transferred to server and store the encoded image in SQL row and column (table) in Server Database. To view the image, PHP file is use to execute Base64_decode function to encoded images. Once the synchronization is completed, notification will be sent out to client devices. Figure 5 shows synchronization of images activities according to the proposed framework.

Android Logcat will play a major roles as it will check images status sync and any deletion of images occurs inside client database. Its means image ID of deleted images and images 'False' status will be parsed into POST parameter. Any failure of images

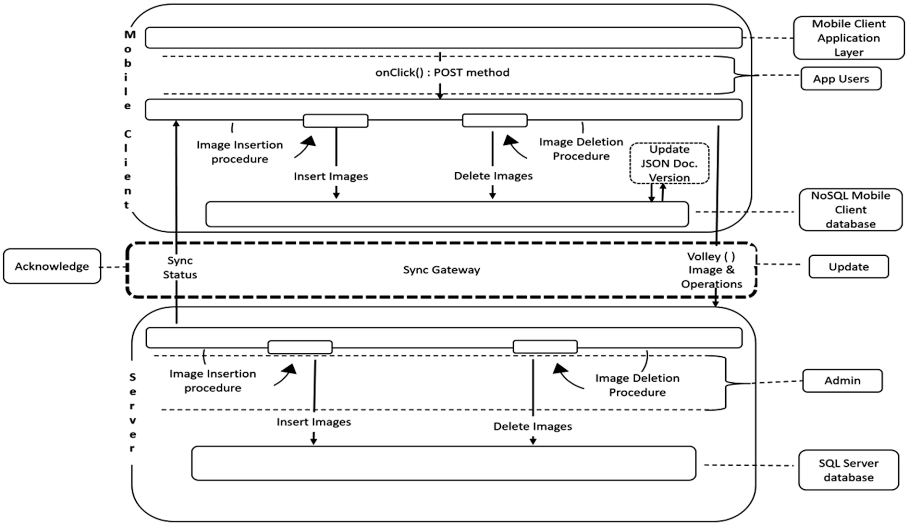


**Fig. 4.** ImgSync_Out framework

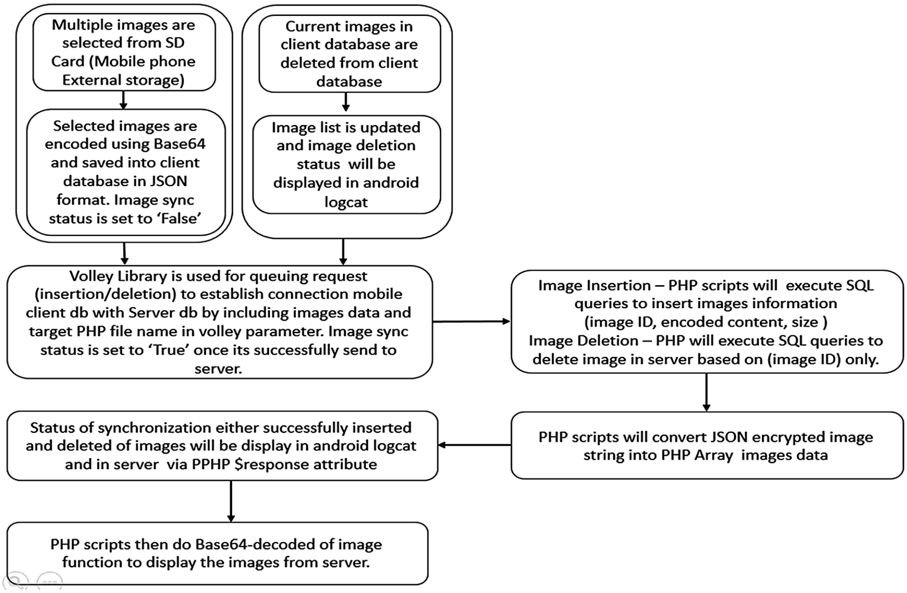

Multiple images are selected from SD Card (Mobile phone External storage)
Selected images are encoded using Base64 and saved into client database in JSON format. Image sync status is set to 'False'
Current images in client database are deleted from client database
Image list is updated and image deletion status will be displayed in android logcat
Volley Library is used for queuing request (insertion/deletion) to establish connection mobile client db with Server db by including images data and target PHP file name in volley parameter. Image sync status is set to 'True' once its successfully send to server.
Image Insertion – PHP scripts will execute SQL queries to insert images information (image ID, encoded content, size )
Image Deletion – PHP will execute SQL queries to delete image in server based on (image ID) only.
PHP scripts will convert JSON encrypted image string into PHP Array images data
Status of synchronization either successfully inserted and deleted of images will be display in android logcat and in server via PPHP $response attribute
PHP scripts then do Base64-decoded of image function to display the images from server.

**Fig. 5.** ImgSync_Out activity flow

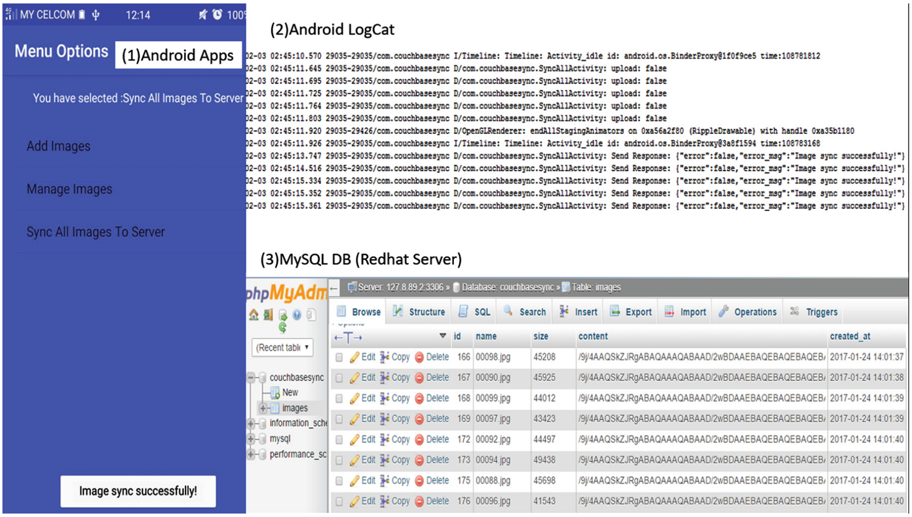
synchronization to server, images sync status will not update to 'True' and images ID of deleted images will be traced in error logcat tray and images deletion resynchronization request will be put back on Requestque() volley function.

This image synchronization framework was validated following [26] approach, through working prototype which shows added functionalities such as Sync Function and Sync Volley to prove its effective deployment. Below are details of software used:- (Table 2)

**Table 2.** Software used for prototype development

| Software | Version (V) | Type |
|---|---|---|
| Android studio | V 2.2.3 | Open source |
| Android SDK | API 25 | Open source |
| Red hat cloud | MySQL V 5.5.52 | Free host/server |
| Couchbase lite DB | V 1.3.2-4 | Open source |

Images dataset are chosen from Digital images in LabelMe, by MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) [27], secondary data source categories. Prototype deployment shows synchronization processes follows proposed framework accordingly. Images dataset was successfully synchronized to server. Images sync status is being observed before and after synchronization process and produce error-free status. Connection using Volley function from Couchbase Lite to MySQL server was recorded behavior and it brings no network breakage within stable Wi-Fi connection. Future experiment will be focusing on comparing the performance of the proposed framework with existing SQlite-MysSQL framework and it will discussed further below in future works section. Figure 6 shows working prototype during image synchronization, where ‘Toast’ message ‘Image sync successfully’ appear on client’s Android mobile screen after image upload status is successfully change from ‘False’ to ‘True’. ‘True’ status indicates images are successfully synchronize to MySQL server.



You have selected :Sync All Images To Server

Image sync successfully!

```
02-03 02:45:10.570 29035-29035/com.couchbasesync I/Timeline: Timeline: Activity_idle id: android.os.BinderProxy@1f0f9ce5 time:108781812
02-03 02:45:11.645 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: upload: false
02-03 02:45:11.695 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: upload: false
02-03 02:45:11.725 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: upload: false
02-03 02:45:11.764 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: upload: false
02-03 02:45:11.803 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: upload: false
02-03 02:45:11.920 29035-29426/com.couchbasesync D/OpenGLRenderer: endAllStagingAnimators on 0xa56a2f80 (RippleDrawable) with handle 0xa35b1180
02-03 02:45:11.926 29035-29035/com.couchbasesync I/Timeline: Timeline: Activity_idle id: android.os.BinderProxy@3a8f1594 time:108783168
02-03 02:45:13.747 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: Send Response: {"error":false,"error_msg":"Image sync successfully!"}
02-03 02:45:14.516 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: Send Response: {"error":false,"error_msg":"Image sync successfully!"}
02-03 02:45:15.334 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: Send Response: {"error":false,"error_msg":"Image sync successfully!"}
02-03 02:45:15.352 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: Send Response: {"error":false,"error_msg":"Image sync successfully!"}
02-03 02:45:15.361 29035-29035/com.couchbasesync D/com.couchbasesync.SyncAllActivity: Send Response: {"error":false,"error_msg":"Image sync successfully!"}
```

| ←T→ ▾ | id | name | size | content | created_at |
|---|---|---|---|---|---|
| □ Edit Copy Delete | 166 | 00098.jpg | 45208 | /9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAEBAQEBAQEBAQEB | 2017-01-24 14:01:37 |
| □ Edit Copy Delete | 167 | 00090.jpg | 45925 | /9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAEBAQEBAQEBAQEB | 2017-01-24 14:01:38 |
| □ Edit Copy Delete | 168 | 00099.jpg | 44012 | /9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAEBAQEBAQEBAQEB | 2017-01-24 14:01:39 |
| □ Edit Copy Delete | 169 | 00097.jpg | 43423 | /9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAEBAQEBAQEBAQEB | 2017-01-24 14:01:39 |
| □ Edit Copy Delete | 172 | 00092.jpg | 44497 | /9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAEBAQEBAQEBAQEB | 2017-01-24 14:01:40 |
| □ Edit Copy Delete | 173 | 00094.jpg | 49438 | /9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAEBAQEBAQEBAQEB | 2017-01-24 14:01:40 |
| □ Edit Copy Delete | 175 | 00088.jpg | 45698 | /9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAEBAQEBAQEBAQEB | 2017-01-24 14:01:40 |
| □ Edit Copy Delete | 176 | 00096.jpg | 41543 | /9j/4AAQSkZJRgABAQAAAQABAAD/2wBDAAEBAQEBAQEBAQEB | 2017-01-24 14:01:40 |

**Fig. 6.** Working prototype implementation on android apps, LogCat and RedHat server.

## 4 Conclusion and Future Work

We have presented the novel sync gateway to establish connection between heterogeneous databases structures to make image synchronization from NoSQL client DB to SQL server DB successfully executed. Proposed framework had implemented JSON language able to handle image transaction in client DB and parse the images data to PHP file in server for synchronization process. Thus, JSON language able to solve problem of limitation of NoSQL client DB that are unable to use general SQL query languages to send image data. From this research, JSON language able to be intermediator compound of NoSQL DB to send the images data to SQL server without conflicts.

This framework solution achieved to provide alternative for client database to store image schema-lessly which is in Document Type database instead in SQL table form. Installation of NoSQL data-base in client side will help speed up the development phase of database as it requires no schema definition. In addition, by using horizontally indexing scalability in NoSQL client DB, it is believed that it will be able to reduce the workload of client mobile devices which does not need to use SQL query language to check the image ID synchronization status in the SQL table data column by column, but only sorting by image upload status.

Future work is to conduct experiment for performance comparison between proposed framework and existing SQL server-SQL client framework in terms of capability to handle bigger size of image during synchronization. In addition, by using validation tools, T-Test and Chi2 analysis methods, comparison of image transmission speed during synchronization and computing power resource of mobile is analyzed to validate the performance of proposed framework.

**Acknowledgments.** This research is fully supported by Ministry of Education Malaysia (MOE) under Fundamental Research Grant Scheme (FRGS) for extensive research on Software Engineering. Any opinions, findings, and conclusions stated in this paper are those of authors and do not necessarily reflect those of the MOE.

## References

1. Domingos, J., Simões, N., Pereira, P., Silva, C., Marcelino, L.: Database synchronization model for mobile devices. In: Iberian Conference on Information Systems and Technologies (2014)
2. Zaia, G.P., Ronaldo, C.M., Messias, C., Eduardo, R.G., Olivete, C.J.: MySQLite Sync Middleware for stored data synchronization in mobile devices and DBMSs. In: Proceedings Latin American Computing Conference, pp. 1–7 (2014)
3. Strauch, C.: NoSQL Databases (2011). http://www.christofstrauch.de/nosqldbs.pdf/. Accessed 15 Nov 2014
4. Maan, P.K.: Database for unstructured, semi structured data - NoSQL focus on availability. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET), pp. 466–469 (2015)
5. Nasholm, P.: Extracting data from NoSQL databases: a step towards interactive visual analysis of NoSQL data. Master's thesis, University of Gothenburg, Sweden (2012)

6. Das, M.E., Suresh, S.: A synchronization algorithm of mobile database by using SAMD algorithm. In: International Conference Computing Control Engineering (2012)
7. Balakumar, V., Sakthidevi, I.: An efficient database synchronization algorithm for mobile devices based on secured message digest. In: International Conference on Computing, Electronics and Electrical Technologies, pp. 937–942 (2012)
8. Li, Z., Wilson, C., Jiang, Z., Liu, Y., Zhao, B., Jin, C., Zhang, Z.-L., Dai, Y.: Efficient batched synchronization in Dropbox-like cloud storage services. In: Eyers, D., Schwan, K. (eds.) Middleware 2013. LNCS, vol. 8275, pp. 307–327. Springer, Heidelberg (2013). doi:10.1007/978-3-642-45065-5_16
9. Thomas, G., Glenn, R.T., Chin-Wan, C., Edward, B., Carter, F., Marjorie, T., Stephen, F., Berl, H.: Heterogeneous distributed database systems for production use. ACM Comput. Surv. **22**(3), 238–266 (1990)
10. Singh, R., Dutta, C.: A synchronization algorithm of mobile database for cloud computing. Int. J. Appl. Innov. Eng. Manag. (IJAIEM) **2**(3), 491–497 (2013)
11. Leavitt, N.: Will NoSQL databases live up to their promise? Computer **43**, 12–14 (2010)
12. Hecht, R., Jablonski, S.: NoSQL evaluation: a use case oriented survey. In: 2011 International Conference Cloud and Service Computing (CSC), pp. 336–341. IEEE (2011)
13. Konstantinou, I., Angelou, E., Boumpouka, C., Tsoumakos, D., Koziris, N.: On the elasticity of NoSQL databases over cloud management platforms. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 2385–2388 (2011)
14. Sharma, A., Kansal, V.: Replication management and optimistic replication challenges in mobile environment. Int. J. Database Manag. Syst. **3**, 81–99 (2011)
15. Cattell, R.: Scalable SQL and NoSQL data stores. In: ACM SIGMOD, vol. 1, pp. 12–27. ACM Digital Library, New York (2010)
16. Moniruzzaman, A.B., Hosaain, S.A.: NoSQL database: new era of databases for big data analytics - classification, characteristics and comparison. Int. J. Database Theory Appl. **6**, 1–14 (2013)
17. Abadi, D.J.: Data management in the cloud: limitations and opportunities. IEEE Data Eng. Bull. **32**, 3–12 (2009)
18. Stonebraker, M., Hong, J.: Saying good-bye to DBMSs, designing effective interfaces. Commun. ACM **52**(9), 12–13 (2009)
19. Sedivy, J., Barina, T., Morozan, I., Sandu, A.: MCSync – distributed, decentralized database for mobile devices. In: IEEE International Conference on Cloud Computing in Emerging Markets, pp. 1–6. IEEE Press (2012)
20. Alhaj, T.A., Taha, M.M., Alim, F.M.: Synchronization wireless algorithm based on message digest (SWAMD) for mobile device database. In: International Conference on Computing, Electrical and Electronic Engineering Synchronization, pp. 259–262 (2013)
21. Gopta, K., Kumar R., Loothra S.: Smartphone security and contact synchronization. In: 2014 Fourth International Conference on Communication Systems and Network Technologies, pp. 621–625 (2014)
22. Sethia, D., Mehta, S., Chodhary, A., Bhatt, K., Bhatnagar, K.: MRDMS-mobile replicated database management synchronization. In: International Conference Signal Processing Integrate Networks, pp. 624–631 (2014)
23. Abdullahi, A.I., Basri, S., Ahmad, R.: An efficient data synchronization model for heterogeneous mobile device databases and server side database. Unpublished Master's thesis, Universiti Teknologi PETRONAS, Perak, Malaysia (2016)
24. Choi, M.Y., Cho, E.A., Park, D.H., Bae, J.Y., Moon, C.J., Baik, D.K.: A synchronization algorithm of mobile database for ubiquitous computing. In: Fifth International Joint Conference on INC, IMS IDC, NCM, pp. 416–419 (2009)

25. Calhoun, D.: When to Base64 encode images (and When Not To) (2011). http://davidbcalhoun.com/2011/when-to-base64-encode-images-and-when-not-to/. Accessed 02 Dec 2016
26. Sinitsyn, A.: A synchronization framework for personal mobile servers. In: Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, p. 208 (2004)
27. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. Int. J. Comput. Vis. **77**(3), 157–173 (2008)

# Using Virtualization Technology for the User Authorization System

Maxim Polenov(✉), Vyacheslav Guzik, and Vladislav Lukyanov

Department of Computer Engineering,
Southern Federal University, Taganrog, Russia
{mypolenov,vfguzik}@sfedu.ru, sith@pochta.ru

**Abstract.** The paper deals with the use of virtualization technology for the development of software system that supports authorization of Wi-Fi networks users. The suggested approach is to use two databases of operators and users, implemented on MySQL, and two subsystems of registration and authorization to be placed on a virtual machine. The paper presents the organization and a general algorithm of the authorization system operation, as well as the load analysis of server with the authorization system. Here considered the graph of hardware resources loads used by the virtual machine. Using the hypervisor allows changing the configuration and provides the possibility to monitor the virtual machine, the necessary disk space, the memory consumption and CPU usage.

**Keywords:** Virtualization · Wireless network · Authorization system · Database · Virtual machine · Server

## 1 Introduction

Wireless technology allows users to provide the required mobility while establishing access to local networks and the Internet. There is often a need to provide a specific group of users with access to a wireless network, and the use of standard methods of protecting the access point by establishing a standard password is not an adequate solution, as the number of users can change, so it will be necessary to constantly change the shared password. There are many simple solutions to this problem, but they require to restrict the number of users due to the limited memory capacity of the device where the access point is placed. Therefore, there is a need for operation opportunities of these devices with external user databases deployed on the servers. In addition, system operators need to be able to edit the user database, and all of these devices should be deployed on one server.

It is worth to consider that for developing system it is assumed that the operators are ordinary users who have a basic knowledge of personal computers and then use solutions such as Radius authorization server would be unpractical. This solution should also be mobile, quickly deployable and easily configurable.

Based on the foregoing, we can formulate the following tasks for the development of the authorization system. It is necessary to develop a database that will store the user's identity. The subsystem, which will allow adding new users, should be developed and

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_20

connected to the database. Since the majority of ready-made solutions are not compatible with the database, it is necessary to create a system that will communicate with the access point using standard scripts of authorization and with users checking the data entered by them. Database of users should be isolated from them, as well as the registration subsystem itself. The entire system must be installed on a single server. It is also necessary to determine on which servers the proposed approach can be implemented.

## 2 Software Implementation of the Authorization System

For the implementation of the approach, the deployment of authorization system on a server with a hypervisor is proposed. This will solve the problem of the complex deployment on a single server. It is also necessary to create a database, which will contain data on system operators for their authorization.

To restrict the access to the subsystem of user registration, a configuration file has been created in the root of the web-server. The file allows the restriction of access to the system by the internal local address, specifying the IP-address of the operator of computers. Address filtering algorithm is shown in Fig. 1.

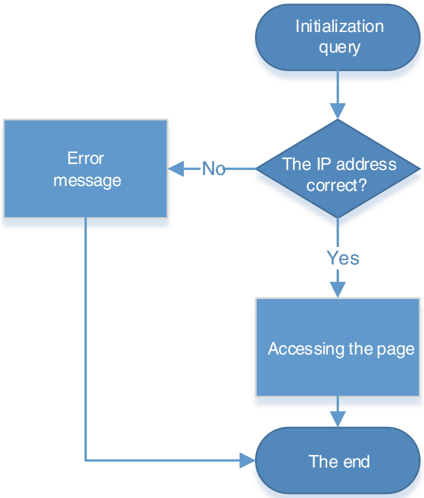


**Fig. 1.** Base algorithm of IP-filtering

A test version of the developed system has been installed on a server with a hypervisor. It was launched on the virtual machine (Table 1) working under the operational control of the Debian system where the web-server with PHP module was installed as well as the MySQL database and database administration tools.

While testing the system, hypervisor VMware vSphere [1] and the access point developed by Ubiquiti Networks [2] were used.

**Table 1.** Characteristics of the virtual machine

| Characteristics | Virtual machine |
|---|---|
| Installed OS | Debian |
| Installed components | Apache, MySQL, php, phpMyAdmin |
| The roles of the machine | The user database. The subsystem of users registration, The operators database, user authorization subsystem |
| Connected to local area network | Network 1,2 |

The database management system MySQL [3] has been used to create a database (DB) of users and operators. The PHP language [4] has been used to implement the subsystem of user authorization, configuration files for the database connection, the system scripts to check the entered data, and Shell-script [5] for transmitting of users' data, authorizing them on the access point, and providing them access to the Internet. The system administrator carried out database management using of PhpMyAdmin [6].

As a result, the authorization system consisting of two subsystems and two databases has been set up (Fig. 2).

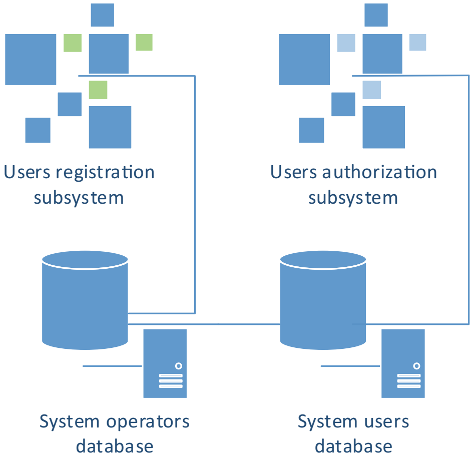

Users registration subsystem
Users authorization subsystem
System operators database
System users database

**Fig. 2.** The main components of the authorization system

In this system, the subsystem of users registration is connected to the users database to perform operations on it; as well as this, it is connected to the database of operators for their authorization and access to the Add User interface. In this case, user authorization subsystem is connected only to the user database and has limited access to it (search function), which allows the subsystem to compare the entries with the available database. Both databases are installed on a single workstation, but are logically in different local networks.

## 3 Hardware Implementation of the Authorization System

In order to study the proposed approach, the structure of hardware implementation of the authorization system, which is shown in Fig. 3, has been developed and tested.

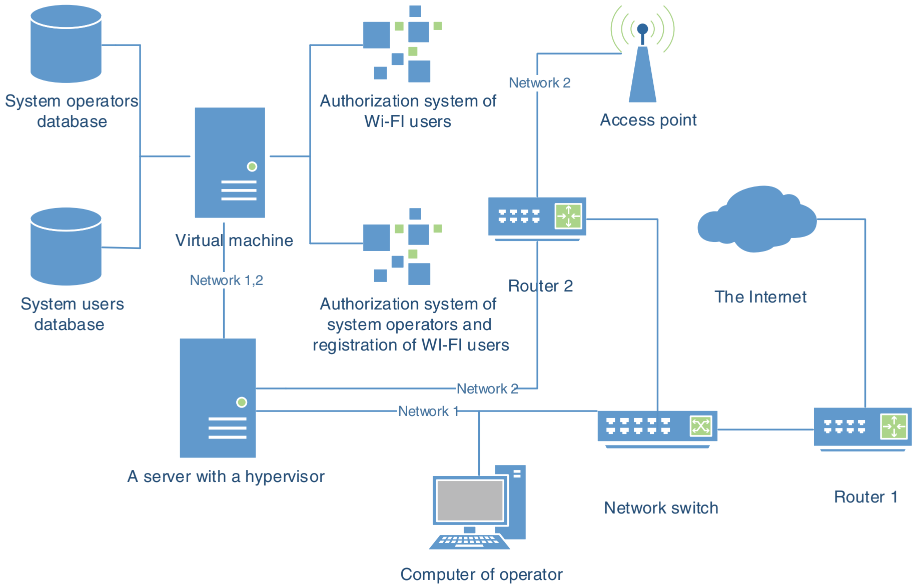

System operators database
Authorization system of Wi-FI users
Network 2
Access point
Virtual machine
System users database
Network 1,2
Router 2
The Internet
Authorization system of system operators and registration of WI-FI users
Network 2
Network 1
A server with a hypervisor
Network switch
Router 1
Computer of operator

**Fig. 3.** Network implementation of authorization system

Isolation of users and operators in this structure is carried out by means of two configured in the virtual machine LANs (Network 1 and Network 2), which are connected to the server with a hypervisor.

Users can access Network 2 through a wireless access point. Router 2 operates as the second DHCP-server that distributes IP-addresses to users and the access point. The operators of the system are located in Network 1. This separation allows the isolation of users from the registration subsystem and the database of operators, as well as from the operators' computers. Since the virtual machine is accessed by registration and authorization subsystems, it is connected to both the first and to the second networks.

However, the only possible way to connect to MySQL database installed on the virtual machine is to use phpMyAdmin. The IP-address filter is also set up on the machine, but all other resources are denied access, which eliminates the possibility for users to connect to the database.

The algorithm of authorization system operation is shown in Fig. 4.

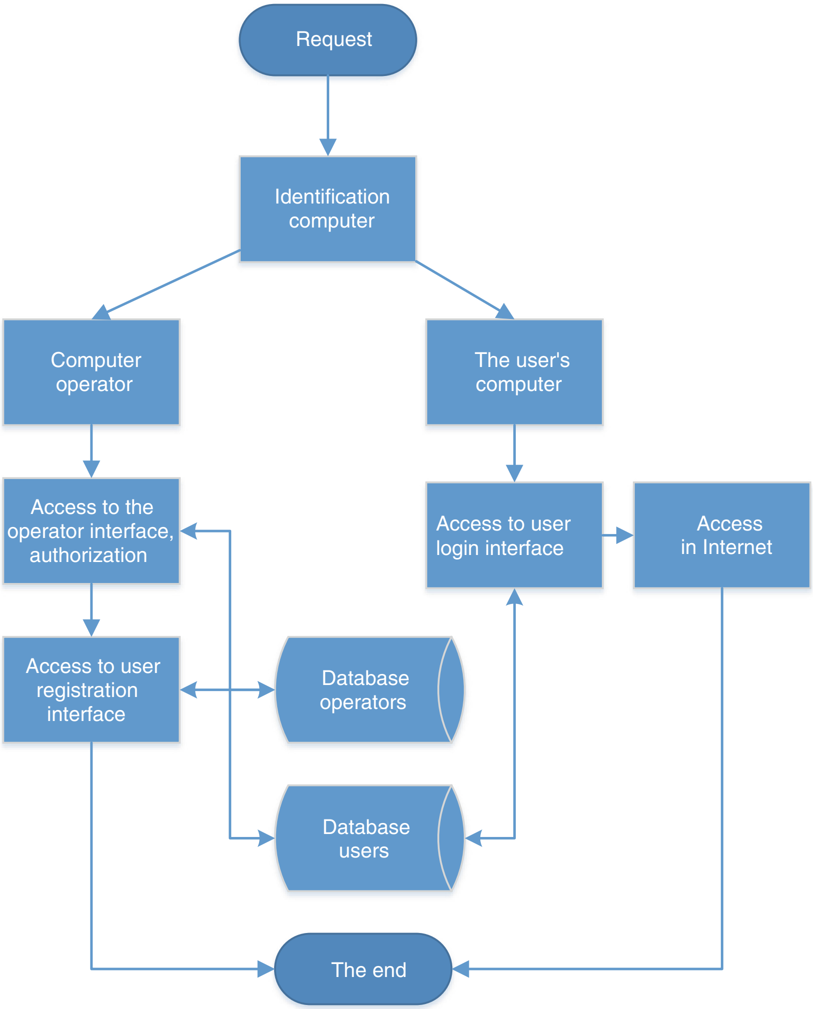


**Fig. 4.** General algorithm of the system's operation

User registration subsystem has several stages of the operator authorization. The first step to get access to the subsystem is to carry out the identification of the computer that requests access through a web-browser; if the identification fails, the access to the system interface is prohibited. If authorization is successful, the operator get an access to the interface where it can request a new User Registration Form. User Registration Form asks for identification data (username and password); in case of successful authorization, the operator has the right to add a new user to the database; in case of unsuccessful attempt of authorization, access to the entry of a new user is prohibited. The subsystem has its own database with the indicator data operator, which is independent of the user database and is monitored by the system administrator.

The table consists of the following fields:

- ID – ID, the primary key;
- user_login – username to login;
- user_pass – user password to login in encrypted form;
- user_email – e-mail for communication;
- user_registered – the date of user registration;
- user_activation_key – activation key is removed after activation;
- display_name – the name to display.

At the first attempt of the user to enter the Internet, the user authorization system queries his data for authorization; after that, the correctness of the entered data is verified by comparing it with the information existing in the database. If the entered ID is verified, the system sends a shell-script authorization with the user's data to the access point; after receiving the data the access point provides access to Internet. In case of discrepancy between the entered user's data and the data existing in the database, the system displays an error message and performs redirection of the user to the login page.

Table of user's data consists of 4 fields:

- Id – displays the user identification number in the table;
- Login – contains information for user authorization;
- Date – contains the date of user registration;
- IP – provides information about computer's IP-address from which the user was registered.

Only the system administrator has full access to the database.

After the successful connection to a wireless network and the attempt to go to any website using the browser, the user is redirected to the login page where he must enter his login. This page has a fairly simple interface; since most users connect to the Internet with various mobile devices, the login page display has been significantly simplified to help the users avoid problems.

Due to the two networks, users are isolated from the network operators and are unable to get to "add users" system as well as to the database that contains information on operators and users. Access to user authorization system is granted by identification of the operator's computer and his authorization with login and password.

For experimental verification of the proposed approach, the test system was installed on the server with Intel Xeon e5520 2.27 GHz CPU and 32 GB RAM that plays the role of multi-hypervisor with a few actively used virtual machines. This was done to determine the effect of the proposed system on the amount of used resources and the possibility to install it on a less powerful server.

A graph of hypervisor's CPU load is shown in Fig. 5.

As the figure shows, the maximum amount of CPU resources used in the experiments was a little more than 15%. Virtual machines are available for more than 80% of CPU resources.

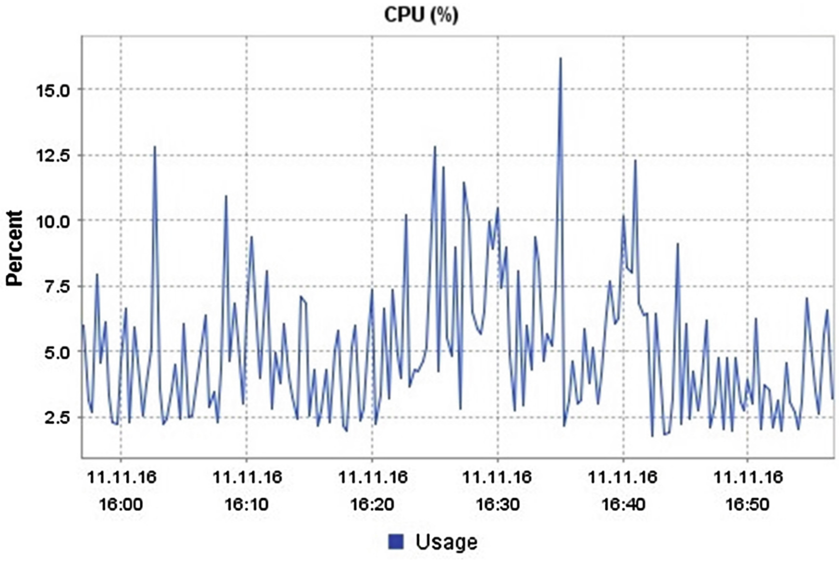


**Fig. 5.** The use of CPU resources

A graph of hypervisor's memory usage by virtual machines is shown in Fig. 6.

The graph shows that the maximum amount of memory usage does not exceed 5 GB.

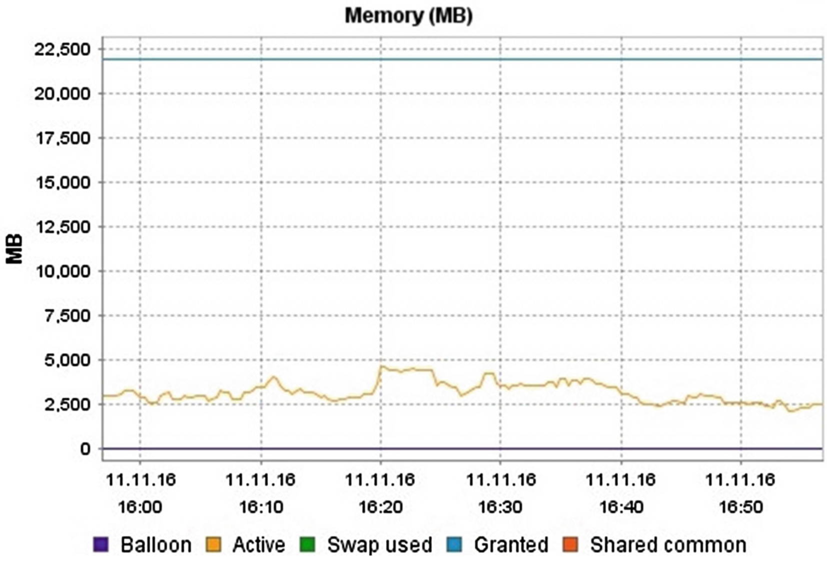


**Fig. 6.** The use of RAM

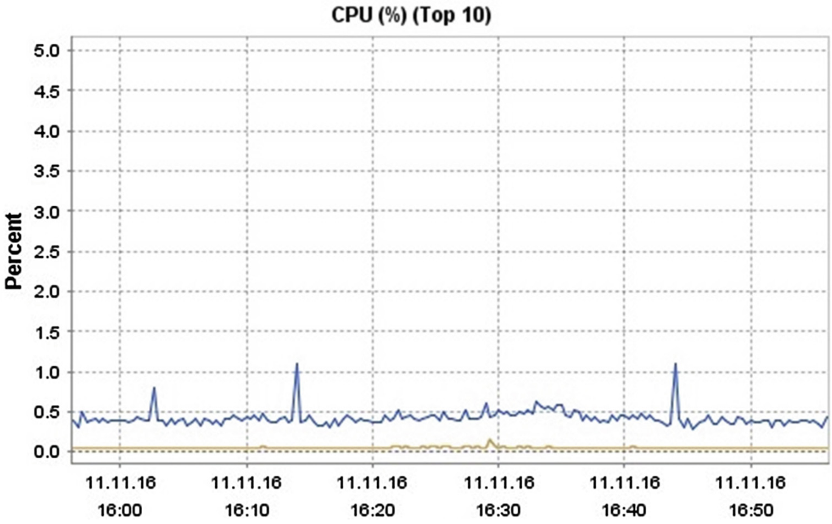


**Fig. 7.** Using the CPU resources by virtual machine with the authorization system

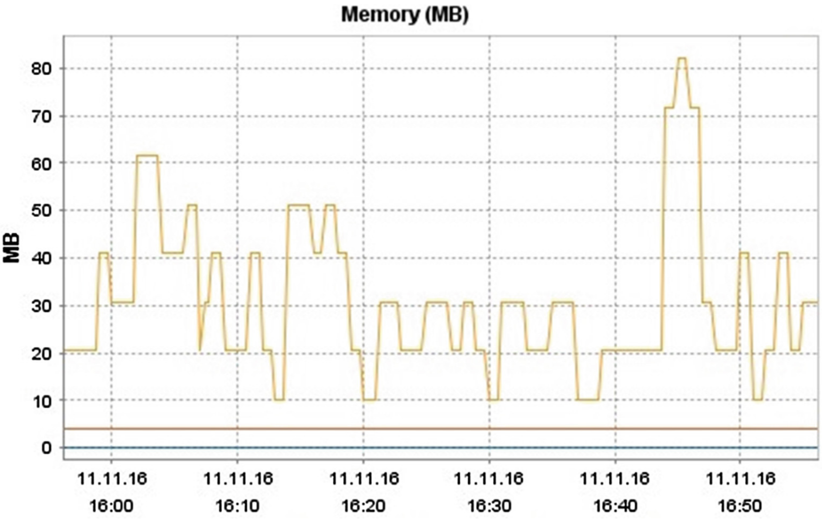


**Fig. 8.** Using RAM by a virtual machine with the authorization system

It should be noted that the presented graphs show the use of the resources by all the virtual machines installed on the hypervisor. Consider the following graphs showing the use of the server resources only by the virtual machine with an installed authorization system (Figs. 7 and 8).

## 4 Conclusion

The proposed and developed authorization system consumes a low amount of resources and can be deployed on servers with low power. The system components are isolated from each other and have independent interface. The filter of IP-addresses available in the operator's interface eliminates the possibility to access the interface of other devices. The need to enter a separate password to add new users serves as an additional security measure.

Through the use of virtualization technology and the implementation of solutions in the database, the system has the ability to create and maintain a large number of users. Virtualization also allows you to install the system on a single server and to organize the possibility of creating the backup copy of the virtual machine, which will allow a quick recovery of the system in case of cancellation or change of equipment. Using the hypervisor provides the possibility to monitor the state of the virtual machine, the amount of disk space it occupies, the memory consumption, and CPU usage; it also allows changing the configuration if necessary. Based on the published data [7, 8] and the graphs above, we can conclude that the virtual machine will consume low amount of resources. Meanwhile, users are separated from the network operators, and the authorization system is available to them on conventional computers, laptops, and mobile devices.

**Acknowledgments.** This work was financially supported by the Ministry of Education and Science of the Russian Federation under Project No. 2.3928.2017 (Project part of State task).

## References

1. Kusek, Ch., Noy, V.V., Daniel, A.: VMware vSphere 5 Administration Instant Reference. Sybex, Indianapolis (2011)
2. Wireless networking products for broadband and enterprise. Ubiquiti Networks. https://www.ubnt.com
3. Nixon, R.: Learning PHP, MySQL, JavaScript, CSS & HTML5. O'Reilly Media, Sebastopol (2014)
4. Coggeshall, J.: PHP 5 Unleashed. Sams, Indianapolis (2004)
5. Tansley, D.: Linux and UNIX Shell Programming. Addison-Wesley, Boston (2000)
6. PhpMyAdmin, https://www.phpmyadmin.net
7. Kostyuk, A., Polenov, M., Lukyanov, V., Muntyan, E., Nikolava, A.: Research of virtualization deployment possibility in smart house control systems. In: Informatization and Communication, vol. 3, pp. 72–77, Moscow (2015). (in Russian)
8. Polenov, M., Kostyuk, A., Muntyan, E., Guzik, V., Lukyanov, V.: Application of virtualization technology for implementing smart house control systems. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Silhavy, P., Prokopova, Z. (eds.) Software Engineering Perspectives and Application in Intelligent Systems. AISC, vol. 465, pp. 329–339. Springer, Cham (2016). doi:10.1007/978-3-319-33622-0_30

# Strategic Modeling of Secure Routing Decision in OLSR Protocol in Mobile Ad-Hoc Network

C.K. Vanamala[1(✉)] and G. Raghvendra Rao[2]

[1] Department of Information Science and Engineering, NIE, Mysuru, India
ckvanamala@nie.ac.in

[2] Department of Computer Science and Engineering, NIE, Mysuru, India
grrao56@gmail.com

**Abstract.** Optimized Link State Routing (OLSR) is known for its potential features for leveraging the communication requirement in mobile ad-hoc network. However, there is less number of investigations to confirm its sustainability as well as resiliency against potential threats in mobile ad-hoc network. Hence, this paper presents a modeling of misbehaved nodes to compliment intrusion detection system for mobile nodes using OLSR using the strategic construction of secure routing decision in OLSR protocol. The technique allows each communicating mobile nodes to monitor reputation, suspicion factor, and contingency of threat to compute the degree of its vulnerability. The modeling is quite capable of truly exploring the latent adversaries within the network in the form of normal nodes. The study outcome shows proposed algorithm on OLSR to possess reduced delay, reduced algorithm processing time and maximized throughput in contrast to conventional OLSR.

**Keywords:** Attacks · Mobile ad-hoc network · OLSR · Probability theory · Security · Routing · Reputation · Trust

## 1 Introduction

The concept of Mobile Ad-hoc Network is a big boon in the area of cost-effective communication system [1]. One of an effective application of it can be already seen in the Vehicular Ad-hoc Network (VANET) [2]. There is various research work already carried out towards routing [3], security [4], energy efficiency [5], traffic management [6], etc. in the mobile ad-hoc network and significant improvement has also been observed in this regards. However, studies towards proactive routing called as Optimized Link State Routing (OLSR) is recently gripping a good pace in research community. Basically, OLSR has some of the significant advantages e.g. known information of destination before routing, availability of routes, highly controlled communication over-head, easy to integrate with Internet/cloud, etc. [7, 8]. Hence, such features are quite better than frequently used Ad-hoc On-Demand Distance Vector (AODV) [9]. From the security viewpoint, there is less number of research attempts towards investigating the robustness of OLSR against various types of adversaries. It was also seen that existing research work towards secure routing is quite specific to attack types and hence reduces the applicability on other attack types. In this regards,

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_21

there is a need of study that formulates its intrusion detection system based on the malicious behavior of a node dynamically and can offer intrusion prevention system to secure the communication.

The present paper discusses the modeling technique of malicious behavior of the node using OLSR offers resists the attacks. The applicability of the technique is more of different forms of attacks in mobile ad-hoc network. The technique uses Sect. 2 discusses the existing research work followed by problem identification in Sect. 3. Section 4 discusses proposed methodology followed by the elaborated discussion of algorithm implementation in Sect. 5. Comparative analysis of accomplished result is discussed in Sect. 6 followed by conclusion in Sect. 7.

## 2 Related Work

This section discusses the most recent implementation work being carried out towards security problems in OLSR. The most recent work carried out by Kumar and Gayathri [10] have presented a security technique using trust-based communication over OLSR in mobile ad-hoc network. The study outcome was found to have a better outcome compared to conventional OLSR on packet delivery ratio, latency, and overhead. Similar problems and approach were also seen in the work of Khan et al. [11] where multiple attributes have complimented the security of OLSR resulting in 98% of accurate detection rate. Schweitzer et al. [12] have presented a technique to thwart denial of service attack. Dutta and Biswas [13] have introduced a procedure to mitigate wormhole attack in Mobile Ad-hoc Network considering OLSR protocol using both simulation and experimental approach. Investigation of selfishness in communication pattern of the node was discussed in the work of Amraoui et al. [14]. Study towards node isolation attack was carried out by Malik and Rizvi [15] towards strengthening security feature of OLSR. Usage of cryptography was adopted for securing OLSR protocol in Mobile Ad-Hoc Network as found in the work of Selvi and Kuppuswami [16]. Cryptographic-based technique was further found to be leveraged in the work of Alattar et al. [17] where the digital signature is used for intrusion detection on OLSR protocol. Marimuthu and Krishnamurthi [18] have used the trust-based technique to secure OLSR against node isolation attack. The issue of wormhole attack was investigated by Sadeghi and Yahya [19]. The authors have used both AODV and OLSR. Campos et al. [20] have addressed the anonymity problems in mobile ad-hoc network using OLSR protocol, host identity protocol, and pseudonyms. Adonis and Tavildar [21] have introduced a secure modeling by trust factor the in order to escalate the communication performance of OLSR in Mobile Ad-Hoc Network. Collusion attack was addressed in the investigation work of Suresh et al. [22] where the authors have presented an intrusion detection system using OLSR. The study was assessed for its effectiveness by control packets, detection time, and accuracy. Hence, it is observed that few works being done towards using OLSR for resisting adversaries in Mobile Ad-Hoc Network. The next section presents problem identification.

## 3 Problem Description

After reviewing the prior section, it was found that there are very less number of research journals addressing security problems in OLSR about Mobile Ad-Hoc Network. Although, there is work being carried out towards addressing wormhole attack, denial-of-service, collusion attack, still OLSR is highly vulnerable to link spoofing attack, replay attack, node capture attack, etc. Such attacks were never addressed and hence the scope of applicability of existing research work is too much limited and lacks flexibility in case the attack scenario changes. The problem statement of the study is *to design and develop a novel framework that offers security solution by the malicious behavior of a mobile node in OLSR*. The next section presents a brief of the methodology adopted to address this problem.

## 4 Proposed Methodology

The main objective of the proposed study is to optimize the performance of OLSR protocol by incorporating strategic construction of routing decision for modeling node misbehavior in Mobile Ad-Hoc Network. Figure 1 shows the schematic architecture.

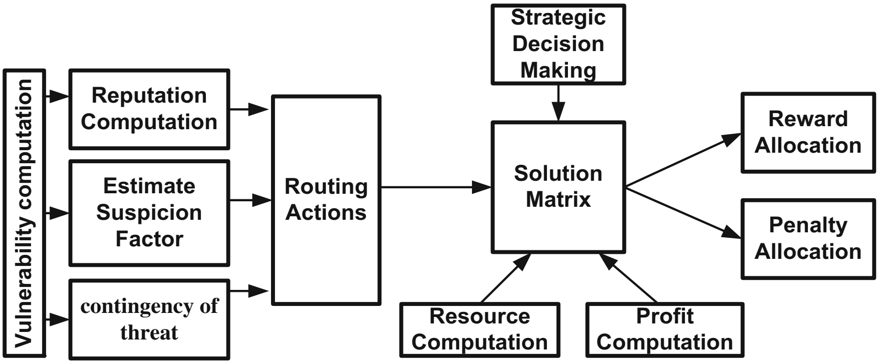


**Fig. 1.** Schematic architecture of proposed system

The proposed system computes the vulnerability assessment, reputation, suspicion factor, and contingency of threat using probability theory. All the new routing actions over OLSR are now compared with the solution matrix and is kept over routing table over OLSR. The solution matrix offer correctness in routing decision based on strategic decision making as well as computation on the ground of resource utilized and profit accomplished in case if the routing action to be implemented. The final routing action results in allocation of reward for normal nodes or penalty for malicious nodes. Hence, the proposed system does not only assists in modeling node misbehavior but also assists in resisting (or rather discouraging) the malicious node to participate in routing process in mobile ad-hoc network. The next section discusses about algorithm.

## 5 Algorithm Implementation

An algorithm is developed using strategic statistical modeling over multiple number of nodes in Mobile Ad-hoc Network when each node implements OLSR algorithm for routing. The algorithm takes the input of *i* (Number of instances of packet forwarding) and *j* (Number of instances of packet dropping), which upon processing gives the output of contingency of threat ($C_t$). The algorithm steps as follows:

```
Algorithm for Modeling Misbehaved Nodes in OLSR
Input: $i, j$
Output: $C_t$
Start
1. Estm i, j
2. For (n=act_Com)
3.     $v = k.(i.j)/(i+j)^2$
4.     $r \rightarrow i/(i+j).T_v$ & $s \rightarrow T_v - v$
5.     $C_t \rightarrow j/(i+j)$
6. If $C_t \leq f(ce(Sol\_Mat) \mid ce \in pf_{profit})$
7.     flag node as normal & update r,
8.     update $C_t \rightarrow$ lower $C_{ct}$
9. Else, flag node as malicious & update r,v.
10     update $C_t \rightarrow$ max $C_{ct}$
11. End
End
```

The logic of the algorithm design is as follows-We consider that there are various forms of attackers in disguise of normal node present in the simulation area and so the algorithm has no predefined idea of good / malicious nodes. It will mean that modeling targets to any forms of attack applicable over OLSR protocol. Initially, the algorithm computes total instances of packet transmission *i* and total instances of packet dropping *j* (Line-1). Interestingly, adoption of *i* and *j* could be applicable both for normal and malicious node and no possibility to discretize or associate *i* and *j* with vulnerability. Hence, we empirically compute vulnerability *v* using probability theory (Line-3), where k is a network constant. We also compute reputation *r* and suspicion factor *s* (Line-4). The variable $T_v$ corresponds to cumulative vulnerability i.e. (1-*v*). A closer look into the formulation will show that i/(i + j) is equal for any cases of i = j, which is the only possibility that an intruder will select to get itself avoided from getting caught. Hence, we ignore any possibility of $i \neq j$. Therefore, we use $T_v$ in computing both reputation *r*. Because, although i/(i + j) may be same for two nodes, but value of *v* will differ (line-3). In this case, chances are high for identifying any nodes where possibility of misbehaving is high and this will be the *n* considered to be best case of malicious node. Finally, we compute contingency of threat $C_t$ for communicating nodes (Line-5). A closer look into the equation in Line-5 will show that we use probability theory, which is the best way to quantify the outcome of computations. At next, considers designing a solution matrix (*Sol_Mat*) which is similar to truth table for possible allocating reward or penalty based possible combination of routing operation in OLSR. Hence, we define

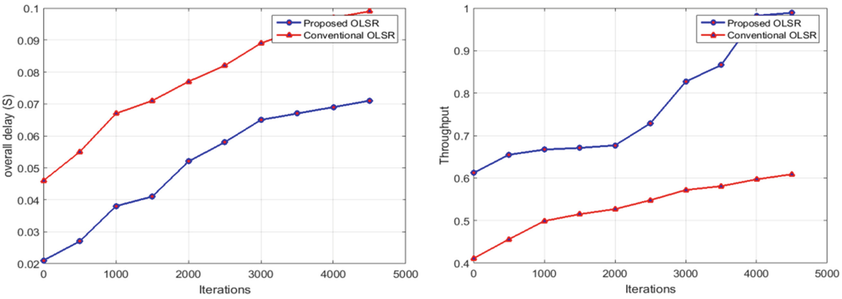
$$Sol\_Mat = \sum_{r=1}^{R} \sum_{c=1}^{C} g[\Pr of(\text{int}, upd), \text{Re}s(\text{int}, pf, alterGroup)] \tag{1}$$

Hence, all the possible solution of routing action of OLSR will be now controlled by the algorithm holding a global function *g* with respect to profit *Prof* and resource utilization *Res*. It is quite an imperative that a normal node will look for obtaining profit of updating *Prof*(*upd*) while malicious node will look for profit of intruding *Prof*(*int*). Similarly, from resource viewpoint, normal node will only compute resource to be utilized for packet forwarding *pf* as *Res*(*pf*), while malicious node will definitely compute resource utilization for intrusion *Res*(*int*). We develop much worst adversarial challenge by which an adversary after launching an attack will change its present group to new group that is recorded in *alterGroup* attribute. This phenomenon is almost similar to a vehicle when it changes from one Road side unit to another. An interesting trait in our attack modeling is that although we specifically donot highlight any attacks, but these are the generalized and mandatory operations to be carried out by any attackers initiating threats over OLSR. Hence, an adversary will only initiate an attack when its resource utilization is found to be less and its profit of launching intrusion *int* is found more. Therefore, if the contingency of threat $C_t$ is found to be less than component elements *ce* of solution matrix (Line-6) with respect to intrusion event than the algorithm declares the node as normal and updates it reputation r and records the matrix with lower value of $\text{C}_\text{t}$. Otherwise, the algorithm declares the node as malicious and re-computes contingency $\text{C}_\text{t}$ as [Res(int)-Res(*pf*)] / Prof(*int*). It finally updates both reputation *r* and vulnerability *v* along with new value of maximum contingency of threat $C_t$.

Also, the algorithm declares some normal node to be malicious nodes. Then, use a cut-off value of its reputation, which is private information not meant for sharing while exchanging control messages using OLSR. Therefore, the algorithm uses dynamic threshold in order to countermeasure the adverse effect of malicious request. All the reputation being computed will only be disseminated once the communicating node has been declared to be normal or malicious in order to resist false record dissemination. Hence, the lower values of vulnerabilities are obtained while implementing the algorithm by allocating rewards for normal nodes and penalty for malicious nodes. The penalty will be in the form of either node partition or packet forwarding to targeted node by malicious node.

# 6 Results Discussion

It is evident from the prior section that the algorithm does not use any form of cryptographic operations to rendered security on potential threats over OLSR. The algorithm mainly ensures that if any malicious node than they will directly declare as a partitioned node (penalty) or else the malicious nodes must assist in packet forwarding as long as possible (reward). It is obvious that malicious node will not be assisting in routing for a long run. At the same time, they will not be able to initiate attack directly. This directly makes the proposed OLSR much robust towards any form of attacks in Mobile Ad-Hoc Network. The study outcome was accomplished with 500 mobile nodes (90% are normal, and 10% are malicious) with random mobility model in simulation area of 1000 × 1000 $m^2$. The effect of proposed algorithm was testified using performance parameter ofthroughput and end-to-end delay along with a comparison with conventional OLSR scheme.



(a) Delay Performance (b) Throughput Performance

**Fig. 2.** Comparative analysis of proposed algorithm

The outcome shows that both end-to-end delays (Fig. 2 (a)), as well as throughput (Fig. 2 (b)), are found to be better for proposed OLSR scheme in comparison to conventional OLSR scheme. A closer look into the delayed outcome (Fig. 2 (a)) will show that delay increases till 1000th round and then its rate decreases till 1500th round. It is again found to repeat the similar trend till 3000th round. It is because of the initial computation of reputation, suspicion factor, and vulnerability in the preliminary stages of routing, whose dependencies further reduce as the matrix storing solution (Sol_Mat) is consistently updated on any communication being undertaken by any node within the simulation area. This also causes minor degradation in throughput till 2000th round as shown in Fig. 2 (b). This can be seen in the trend from 3000th to 4500th round. Hence, previously declined rate of delay increase was found between 1000th – 1500th rounds while it repeats back the similar trend from 3000th–4500th round. This outcome is also visible in throughput performance where throughput significantly increases from 2500th rounds. Hence, better delay compensation is observed along with enhanced throughput. Algorithm processing time for the conventional scheme was found to be approximately 4.5521 s while that of proposed scheme was found to be approximately 0.2774 s tested over Windows 10 machine with a core-i7 processor.

## 7 Conclusion

Although there are various techniques introduced in the past to resist the security loopholes in the mobile ad-hoc network, it is truly a challenging work owing to its decentralized architecture and dynamic topology. Usage of OLSR protocol can enhance the communication need and meet the traffic demands of the user, but it is still a questionable fact about its resiliency towards different forms of adversaries. Hence, this study implements a technique (which assumes that attacker is within the simulation area in the form of the normal or selfish node) to perform intrusion detection system. The mechanism assists the node to compute vulnerability based on a number of the data packet being exchanged by the nodes and need to profile of normal and malicious node maintained in the solution matrix that is indexed in the routing table. Hence, any decision of action to be made by proposed OLSR will now confirm with solution matrix to find if the node it is communicating with is a normal node or malicious node. A dynamic thresholding is also introduced to resist false positive of any wrong declaration by normal nodes. Although, our study does not consider any specific forms of attacks, we believe that irrespective of the type of attack the initial phase of all the attackers is to gain trust by assisting in data forwarding and this is where the adversaries will be entrapped by our model. Our future modeling will further strength the features with more number of analysis.

## References

1. Sarkar, S.K., Basavaraju, T.G., Puttamadappa, C.: Ad Hoc Mobile Wireless Networks: Principles, Protocols, and Applications. CRC Press, Boca Raton (2016)
2. Lin, X., Lu, R.: Vehicular Ad Hoc Network Security and Privacy. Wiley, Hoboken (2015)
3. Patel, D.N., Patel, S.B., Kothadiya, H.R., Jethwa, P.D., Jhaveri, R.H.: A survey of reactive routing protocols in MANET. In: IEEE-International Conference on Information Communication and Embedded Systems, Chennai, pp. 1–6 (2014)
4. Abdelaziz, A.K., Nafaa, M., Salim, G.: Survey of routing attacks and countermeasures in mobile ad hoc networks. In: UKSim 15th International Conference on Computer Modelling and Simulation, Cambridge, pp. 693–698 (2013)
5. Chawda, K., Gorana, D.: A survey of energy efficient routing protocol in MANET. In: IEEE-2nd International Conference on Electronics and Communication Systems, Coimbatore, pp. 953–957 (2015)
6. Gupta, H., Pandey, P.: Survey of routing base congestion control techniques under MANET. In: IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology, Tirunelveli, pp. 241–244 (2013)
7. Loo, J., Mauri, J.L., Ortiz, J.H.: Mobile Ad Hoc Networks: Current Status and Future Trends. CRC Press, Boca Raton (2016)
8. Prasad, B.V.V.S.: Routing Issues in MANETs. Educreation Publishing, Dwarka (2016)

9. Lakhtaria, K.I.: Technological Advancements and Applications in Mobile Ad-Hoc Networks: Research Trends: Research Trends. IGI GLobal, Hershey (2012)
10. Kumar, S.R., Gayathri, N.: Trust based data transmission mechanism in MANET using sOLSR. In: Subramanian, S., Nadarajan, R., Rao, S., Sheen, S. (eds.) CSI 2016. CCIS, vol. 679, pp. 169–180. Springer, Singapore (2016). doi:10.1007/978-981-10-3274-5_14
11. Khan, M.S., Khan, M.I., Malik, S-Ur-R., Khalid, O., Azim, M., Javaid, N.: MATF: a multi-attribute trust framework for MANETs. EURASIP J. Wireless Commun. Networking (2016)
12. Schweitzer, N., Stulman, A., Shabtai, A., Margalit, R.D.: Mitigating denial of service attacks in OLSR protocol using fictitious nodes. IEEE Trans. Mobile Comput. **15**(1), 163–172 (2016)
13. Dutta, C.B., Biswas, U.: Specification based IDS for camouflaging wormhole attack in OLSR. In: IEEE-23rd Mediterranean Conference on Control and Automation (2015)
14. Amraoui, H., Habbani, A., Hajami, A.: Effect of selfish behaviour on OLSR and AODV routing protocols in MANETs. In: IEEE Global Summit on Computer & Information Technology, Sousse, pp. 1–6 (2014)
15. Malik, D., Rizvi, M.A.: Prevention of node isolation attack on OLSR by DFOLSR. In: ACM-International Conference on Information and Communication Technology for Competitive Strategies (2014)
16. Selvi, K.T., Kuppuswami, S.: Enhancing security in optimized link state routing protocol for MANET using threshold cryptography technique. In: IEEE-International Conference on Recent Trends in Information Technology (2014)
17. Alattar, M., Sailhan, F., Bourgeois, J.: On lightweight intrusion detection: modeling and detecting intrusions dedicated to OLSR protocol. Int. J. Distrib. Sens. Netw. (2013). Hindawi Publishing Corporation
18. Marimuthu, M., Krishnamurthi, I.: Enhanced OLSR for defense against DOS attack in ad hoc networks. IEEE J. Commun. Netw. **15**(1), 31–37 (2013)
19. Sadeghi, M., Yahya, S.: Analysis of Wormhole attack on MANETs using different MANET routing protocols. In: IEEE-Fourth International Conference on Ubiquitous and Future Networks, Phuket, pp. 301–305 (2012)
20. Campos, J., Calafate, C.T., Nacher, M., Manzoni, P., Cano, J-C.: HOP: achieving efficient anonymity in MANETs by combining HIP, OLSR, and pseudonyms. EURASIP J. Wireless Commun. Netw. (2011). Hindawi Publishing Corporation
21. Adoni, K.A., Tavildar, A.S.: Trust aware routing framework for OLSR protocol to enhance performance of mobile ad-hoc networks. In: IEEE-International Conference on Pervasive Computing (2015)
22. Suresh, L., Kaur, R., Gaur, M.S., Laxmi, V.: A collusion attack detection method for OLSR-Based MANETs employing scruple packets. In: ACM-Proceedings of the 3rd International Conference on Security of Information and Networks, pp. 256–262 (2010)

# Guaging the Effectivity of Existing Security Measures for Big Data in Cloud Environment

Chhaya S. Dule[1,2,3](✉) and H.A. Girijamma[1,3]

[1] Department of CSE, RNS Institute of Technology, Bangalore, India
chhaya0671@gmail.com, girijakasal@gmail.com
[2] Jyothy Institute of Technology, Bangalore, India
[3] Department of CSE, Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India

**Abstract.** As the technology is improving, today's era is focusing towards big data. In that sense, various organizations are demanding an efficient data storage mechanism and also the efficient data analysis. The Big Data (BD) also faces some of the security issues for the important data or information which is shared or transferred over the cloud. These issues include the tampering, losing control over the data, etc. This survey work offers some of the interesting, important aspects of big data including the high security and privacy issue. In this, the survey of existing research works for the preservation of privacy and security mechanism and also the existing tools for it are stated. The discussions for upcoming tools which are needed to be focused on performance improvement are discussed. With the survey analysis, a research gap is illustrated, and a future research idea is presented.

**Keywords:** Big data · Cloud · Privacy · Security · Tools

## 1 Introduction

The expanded technology in various areas has made interest towards the big data and has become a trending research subject offering various applications in different areas like social, climate, government, etc. The research topic big data falls under the category of big data with machine learning [1, 2]. This top research scenario will not be completed without the networking, as the real-time applicability demands complex, huge data processing. Still, it is observed that the BD is mysteriously challenging research subject. In this still there is a need of addressing various problems under various scenarios and also better algorithms are needed to be developed to solve these ongoing issues. The major issues include privacy and security in data operation over the cloud under real time scenario. The collected data in BD are easily accessible when the heterogeneous data is transferred over the cloud. The confidential data which includes clinical data, research data, government data, and military data and when these data is transferred over the internet it can be accessed and malfunctioned by the intruders. The tools which are used or developed to handle the tremendous amount of data are successful in handling but fail to maintain or preserve the security & privacy [3]. The existing tools are falls below the security protection zone which come for large

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_22

scale of data [1, 2, 4]. As the study analysis says that the conventional method or security techniques are failed to handle loads of huge data having huge volume, velocity, complexity, etc. [5]. The privacy in all the prospects is the overwhelmed and unresolved concern. The method can be used when the user uses a new service [6]. Also, the differential method can also be used for privacy preservation in big data. But in real time application, the privacy is unresolved. This paper gives the survey for the privacy and security preservation in Big Data [7].

One of the biggest problems with the big data security is that ultimately the big data are stored in warehouse which still uses conventional authentication mechanism. As analytical process is an expensive process that not only requires specialized tools but also requires different forms of infrastructure. Although, there are various security challenges in cloud environment e.g. integrity, non-repudiation, confidentiality, privacy, availability, etc., the service provider is yet to ensure resiliency against potential threats over cloud environment apart from the technical challenges in processing big data. This paper reviews such problems and techniques. The sectional organization of the paper is done as follows: Sect. 2 discusses the survey method while Sect. 3 discusses about a conceptual description of big data followed by discussion of privacy issues in big data in Sect. 4. Discussion about existing security tools is elaborated in Sect. 5. The advantage points on Big Data are discussed in Sect. 6. The briefing of an elaborated research work towards security of big data is carried out in Sect. 7 followed by highlights of research gap in Sect. 8. Finally, Sect. 9 outlines the conclusion of the paper.

## 2 Survey Method

This survey paper has implemented the systematic review of literature presented by Kitchenham [8]. Followed the steps of the [8] for literature review and is described as below.

### 2.1 Research Questions

This section addresses the research questions of the study as below:

Research Question-1: How the existing authentication mechanisms emphasis over the big data analytics?

Research Question-2: What are the drawbacks of the existing researches over the privacy preservation?

Research Question-3: What is the drawback of existing mechanisms for third party privacy mechanism?

Research Question-4: Is the existing storage system in big data has any security issue?

### 2.2 Search Process

The searching for the survey is performed on top journals and conference publication i.e. IEEE, where the collected statistics of publications are tabled in below Table 1. For literature survey we have considered the latest IEEE publications from the year 2011 related to big data cloud and security in it.

**Table 1.** Published IEEE research statistic in Big data

| IEEE statistics on big data | Numbers |
|---|---|
| Conference publications | 17015 |
| Journals & Magazines | 1,789 |
| Early access articles | 271 |
| Books & eBooks | 121 |
| Standards | 14 |
| Courses | 3 |

The above shown statistics are obtained by typing “Big data” keyword in IEEE Xplore.

### 2.3 Inclusion and Exclusion Criteria’s

In this survey paper we have included only the most relevant research papers from the IEEE Xplore. The inclusion criterion includes the researches, which describes the big data cloud and security techniques. The non-relevant researches are excluded from this survey work.

### 2.4 Quality Assessment

The quality assessment is performed on the basis of the inclusion and exclusion criteria’s subjected to the research questions.

## 3 Essentials of Big Data

### 3.1 Architecture of Big Data

The term big data can be said to be a large voluminous data that is characterized by more complex form of data with sophisticated relationship existing among such data sets. The main advantage of big data is that it performs the better analysis of huge data than conventional analysis methods. Due to this reason the big data has gained very much interest in the present generation, which has advancement in the data collection, data storage as well as data interpretation. From last few decades, the use of digital media is being increased in many areas which generates the tremendous amount of data, e.g. hospital data, bank data, social networking data, etc. The data storage cost is

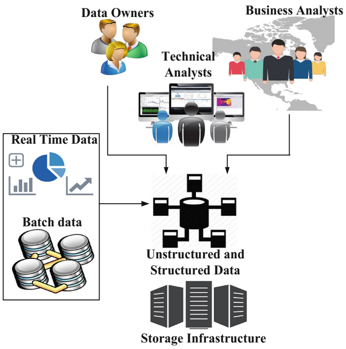


**Fig. 1.** Architecture of big data

decreasing day-by-day by which we can store the entire data rather than discarding it. In addition to this, many of the data analyzing techniques are developed but very few of them have succeeded in efficient data analysis [1, 7]. The big data in the real world is like the collection of huge resources which can be used regularly. The architecture (Fig. 1) of the big data consists of (i) data generation point (unstructured/structured), (ii) data owners, (iii) business analysts, and (iv) data batch.

### 3.2 Big Data Characteristics

There are five different big data characteristics [9], i.e., (i) Data volume, (ii) Data velocity (iii) Data Variety, (iv) Data Value and (v) Data Complexity.

- *Data Volume:* -The tremendous amount of data itself form high volume of big data. The existing data volume size is 1015 terabytes, and it has been predicted that the data volume size is predicted as 1021 zettabytes in future.
- *Data Velocity:* -The data volume is a problem that deals with the data speed for various sources. Owing to higher speed of data generation, it is quite a difficult task to capture the live data and apply analysis on it.
- *Data Variety:* -The data variety is a problem arising from different forms of the data over the distributed and large network. The data is in the form of video, audio, text, etc. measures the data representation. Owing to different forms, it quite a hard task to write a dynamic query system considering such heterogeneity.
- *Data Value:* -The data value can measure the data usefulness in decision making. The user can compile the data stored and which can offer the filtered data. However, capturing data value considering above mentioned problem in big data is really a difficult one.
- *Data Complexity:* -The data complexity measures interconnection and independence of big data structures that needs large data changed.

### 3.3 Issues in Big Data

There are few conceptual points which define the big data issues that need to be analyzed by the organization and need to adopt the technology efficiently. The issues and problems of the big data need to be handled separately.

- *Inter-Related Issues with Characteristics:* There are some issues which are inter-related with the big data characteristics. When the volume of data rises, the value of various kinds of data will fall. Today, the social sites are generating terabytes of data, and it has become more difficult to handle the data using traditional techniques [4].
- *Inter-Related Issues with Transport and Storage:* Today, data created with many social networking sites are generating the huge volume of data, i.e., the data is generating with various kinds of devices like mobiles, computers, etc. The data quantity is more than Exabyte. The current or traditional data techniques limit terabytes of data. With the existing techniques to transfer the Exabyte of data, it takes nearly 3 k hours. If the data transfer is sustained, it needs some more time.
- *Inter-Related Issues with Data Management:* The management of data is one of the biggest issues in big data. The data with different size, format, etc., and validation of these complex data is impractical in nature. The representation of present digital data in a rich manner is acceptable with collection of methods, but there is no efficient mechanism for data management.

## 4 Privacy Issues in Big Data

The big data provides the vast application advantages but the conventional data analysis methods fail to provide the proper privacy mechanism. The privacy concern of the big data includes the private data disclosure to the world. The open-end issue with the big data is privacy and security.

The big data may also include some of the personal data which is shared in the social networking sites. These data can be combined with the other data sets in the real world leading to the exposure of personal data to others. The shared personal data in the social media can be compromised or used by others for the purpose of business in illegitimate manner without any knowledge of the user. Some of the personal data can misuse for criminal activities hence it needs a better focus [10]. Some of the security and privacy challenges are stated below:

- *Protection Issue:* Some of the data which we store in the cloud will not be encrypted for efficiency purpose which will be compromised or encounter protection issue for the important data.
- *Administration Issue:* Every administrative node has an accessibility of any data which may introduce a malicious code and by which the data can be manipulated or compromised easily.
- *Communication Issue:* Many of the Hadoop-based data communication may perform over the wired or wireless connection by which anyone can tap or hack the network node and collect the important data.

- *Cloud Technologies:* The existing technologies of the cloud are not much efficient in offering the personal data security.
- *Conventional Methods Drawbacks:* In previous data management, tools are not much efficient in handing the huge volume of data, by which data gets leaked in the real world.

*Authorization Issue:* The joining of the third party service provides make security and privacy issue for users in any network activity.

## 5 Existing Security Tools

The emerging and growing Big Data need an extraordinary technology that can process the data with greater volume within a shortest time. The mechanism adopted for big data such as massively parallel processing database, distributed database, cloud computing, scalable data storage units, etc. In real time applications, big data analytics plays a major role. In recent past, there are various tools or techniques are developed or examined to store, aggregate, manipulate, visualize and analyze the data. These all the above methods have considered by computer science, economics, mathematics and also even the statistics. This gives an idea that the organizational units are adopted or interested to adopt value from BD in such a way that it can achieve flexibility, discipline in the method [11]. The big data technique is giving an idea how this data can realize. A big data technique should meet the following performance factors:

- The technique should be able to define the issues like variety, velocity, volume, veracity, etc.
- The technique should have the capability of enhancing the data performance, reliability, security, etc.
- It should have the ability to get connected with databases, warehouses, etc.
- The technology must have the scalability and extensibility.
- The technology must allow the ad-hoc queries along with minimal maintenance.

Currently, the Apache Foundation based Hadoop tool is an open source composed of numerous small sub systems of infrastructures facilitating distributed cloud computing. The subsystems in Hadoop can be given as (i) Hadoop Distributed File (HDF) system is also considered as file system and (ii) Map Reduce may also noted as programming paradigm. The other subsystems face issues while working with the huge volume of data.

The above-mentioned issues can be resolved with the help of HDF system and the data combining issues by Map Reduce. The advantage of Map Reduce is that it will minimize the computation risks dealing with reading and writing. Thus we can say that the Hadoop-based system can offer a reliable solution for data storage and present an efficient analysis mechanism. The HDF system and Map Reduce can offer storage.

### 5.1 HDF System

The HDF [12] system is a file system and is designed to store a large volume of data and streaming large volume of data access, run the clusters, etc. Normally the block size in it is 64 MB, and it helps to reduce the required disk for storage. The cluster of the HDF system consists of two nodes master or name node, workers or data nodes.

### 5.2 Map Reduce System

This is a programming paradigm allows huge scalability. This paradigm performs Map and reduces task [13]. The Map tasks are the inputs taken from the distributed FS, which generates a key value sequence pair as per written code for map function. The generated sequence pair will be collected by the master controller and are separated with the reduced task after sorting by key. In sorting a key having same value will end up with same reduce task. The function of reduce task is to merge the entire working key values with a key in same time.

Some of the recent techniques are addressed below:

- IBM Infosphere (IBMIs) Insights: This is an open source which composed of IBM big sheet along with Apache Hadoop platform offering better data analysis without imposing the schematic in its format and does the speedy analysis.
- Kognitio platform: This is an analytical format offers faster scalable database analysis.
- ParAccel: This is a parallel processing database analysis platform offers strong compilation, optimization, etc.
- SAND: This is an analytical platform that will give the linear data scalability via parallel processing.

Recently, there is some research performed to improve the Hadoop security by some of the industries [14] i.e., Apache Ranger, Apache Rhino, Apache Sentry and Apache Knox.

The brief discussion of the existing tools is as follows:

- Apache Ranger: This is a data security framework for the Hadoop platform that enables the enterprises to run the different workloads in the multi-tenant environment. The ARr function is to give the centralized administration security to control each and every security tasks. Also, it aims to provide better authorization.
- Apache Rhino: This security method has been initiated and developed by the Intel Corporation in the year 2013 and significantly it obtained Hadoop ecosystem (HE) security. This method can offer many management, logging, authorization security of Hadoop.
- Apache Knox: This is also a recent effort for security preservation for the secure and authorized access to Hadoop cluster via organizational policies. This is an enhanced version of job execution and cluster data execution. This offers easy web service integration between Kerberos authentication and existing authentication providers.
- Apache Santry: This enforces the fine-grained authorization for metadata and data of Hadoop cluster. This helps to define access control.

## 6 Big Data Advantages

Today in every sector we find the applications of BD it may be social, technology, science, etc. The prime application of the Big Data is to perform analysis of the large and massive streams of data which cannot be analyzed using conventional analytical algorithms.

The applicability depends on how the human can use it according to his necessity. In following some of the BD, advantage is addressed.

- *Customer targeting and Understanding*: The above-titled advantages are uncovered in different sectors. In this scenario, the BD can be used to target the customers by providing some of the offers and understanding them with customer opinions about their products.
- *Optimization of Business Process and Understanding:* The technology of BD is also applied in much business process analysis. With BD a retailer can be able to understand the stock rates by which it gives a route how the delivery can be optimized. For example in human resources, the BDA is adopted that includes talent acquisition and optimization.
- *In Advancement of Research:* The use of BDA in the research area is making a lot of buzzes as it is offering new ideas. The availability of huge data storage can help to analyze any research logic in-depth.
- *In Heathcare Industry*: With the help of BDA, it is possible to decode complete DNA string in less time and which impacts on disease finding. Clinically it helps to maintain the patients past and present health data.
- *In Machine Performance Optimization*: With the BDA improvement, many of the machines are become automated and smarter.
- *In Security Enforcement*: The BDA is also adopted in the prevention of cyber crimes, unauthorized banking transactions and also in identifying the terrorist attacks.

## 7 Research Performed on Big Data Security

The concept which is demonstrated in Lei et al. [15] gives privacy and security based data mining for big data information. Also, it is mentioned that some of the sensitive data may get disclosure to the unauthorized access in various ways during the data processing, data collection, etc. The study gives some of the privacy concerns and idealizes some of the interesting solutions for it. The work described in Hu et al. [16] gives the prominent energy efficient mechanisms in which the current power issues with technological issues are addressed. In the work authors have considered one of the privacy and security issue of BD i.e. architectural issue is addressed. Yan et al. [17] introduced a unique concept of Encrypted BD duplication over the cloud computing. The data which can be stored in the cloud by encrypting it but it will emit the data duplication in the cloud. The performance analysis of simulation results from a prominent solution for data duplication. A significant framework of categorization and application of privacy preservation mechanism in BD mining stated in Xu et al. [17].

The simulation results of this privacy preservation framework are matching with the game theory analysis concepts. The frequently used framework scenario [18] is presented, in which the data publishing, data mining, and data collection process are involved. In this, the data collector will collect the data from various data providers and sell it to the users, who can do some mining processes. To give some compensation for the provider's privacy losses, the data collector will offer some incentives. The data records in this consist of various attributes.

Liang et al. [19] discussed a cipher text multi-sharing data control mechanism is presented to preserve the privacy in BD storage. This mechanism combines both the anonymous method with the proxy re-encryption method, by which a cipher text can be conditionally shared multiple times without disclosure of data. Perera et al. [20] addressed the necessity of BD privacy in the Internet-of-Things (IoT) world. It is admitted that the IoT can help to collect the large set of data on a single platform, but will face privacy issues for the user's data. Nedelcu et al. [21] have addressed the BD challenges and the advantages of BD in manufacturing field. This study analysis includes some of the significant criteria's like the scope of BD and advantages of BD in companies prospective. The privacy and BD concept was also emphasized by Gaff et al. [22] where it has mentioned that only providing the security with encryption mechanism is not enough but it is a privacy solution for the data. Some future recommendations are also addressed which may give the better idea for a security solution. The study of Cardenas et al. [23] gave the consequences of big data analytics for security purpose. In this, it is addressed that there is need of certain researches to be kept exploring the various intrusion mechanisms evolving with technology advancement. Yu et al. [24] has expressed the networking concept in big data. In this work several research articles presented for the networking issues in BD, survey overview is described. The work carried out by Baek et al. [25] has discussed about a security technique towards safeguarding the enterprise applications towards smart grids. A hierarchical structure was developed that offers extensive analytical service in most secured manner using digital signature, identity-based encryption mechanism, and proxy re-encryption policies. Problem of data sharing in most secured manner was discussed in the work carried out by Dong et al. [26]. The authors have used the technique of re-encryption that explicitly uses transformation operation over the ciphertext over the virtual machine in order to secure the sensitive data. The security is achieved using identity-based cryptography mainly. There are also various studies towards data privacy.

## 8 Research Gap

This section discusses about the problem of existing research work that is still left unsolved. The survey analysis of recent works and some good study for BD privacy and security suggests than the issue is very critical and it needs a proper solution in real time applications scenario. An exhaustive study towards all the recently explored research contribution shows that they provide an efficient technique to solve certain security problems over big data approach in cloud. The above listed research questions from Sect. 2.2 are answered below:

1. *Less emphasis on authentication mechanism*: There is a less number of improvement being carried out towards ensure a robust and fault-free authentication policy for big data users. Although, there are various novel ideas towards authentication system in cloud but they were never testified for their applicability over big data analytic-based application.
2. *Few studies focusing on privacy*: At present majority of the studies towards big data security have addressed the problem of data security and access control mainly. There is a very less emphasis on the design security that is mainly due to massive data size.
3. *Less focus towards data brokers*: At present, many of the cloud service providers are using multi-tenancy. They also have a practice of sharing certain segment of the data to the third party that tremendously increases potential risk. Privacy is the first thing to get compromised. The existing solution doesn't address such problems.
4. *Storage insecurity in big data*: It is well known fact that NoSQL database is still evolving and quite problematic to retain optimal security as per the demands. Normally, the big data are stored in multiple tiers where existing system doesn't really focus on how such existing encryption strategy is compliant of tier-based storage strategy.

## 9 Conclusion

Ensuring security towards cloud is never an easy task for any service provider especially in the presence of such malicious activities over internet. With the technology of cloud modernizing, the attackers are also becoming smart. From storage viewpoint, storing the operational data takes the storage cost but storing the analyzed data takes the cost of both storage and processing carried out to transform it. Hence, big data is basically an expensive affair toward cloud storage system in order to store it. The problem might turn worst, if such expensive data is subjected to common or potential threats. Hence, this paper significantly discusses the major topics related to big data security and privacy. With this paper, we provided the recent works that are given IEEE Xplore. From the analysis of research gap analysis, it is pointed than no much studies offered efficient security. There are some recent methods can be executed properly to make better security system form BD. Still, these recent methods need continuous research to make them more efficient with increasing real-time data. Our future work will be focused towards ensuring modeling towards strengthening the privacy problems in big data security. Possibilities of using optimization theory will be quite higher as we don't want to increase the resource cost over big data infrastructure while at the same time we want to achieve optimal privacy protection and resiliency from major potential threats.

## References

1. Lohr, S.: The age of Big Data, vol. 11. New York Times (2012)
2. Swan, M.: The quantified self: fundamental disruption in big data science and biological discovery. Big Data **1**(2), 85–99 (2013)

3. Provost, F., Fawcett, T.: Data science and its relationship to big data and data-driven decision making. Big Data **1**(1), 51–59 (2013)
4. Analytics, Big Data. Big data analytics for security (2013)
5. Wang, C.: Privacy-preserving public auditing for data storage security in cloud computing. In: Infocom, 2010 Proceedings IEEE (2010)
6. Tene, O., Polonetsky, J.: Big data for all: Privacy and user control in the age of analytics. Nw. J. Tech. Intell. Prop **11**, xxvii (2012)
7. Villars, R.L., Olofson, C.W., Eastwood, M.: Big Data: What It Is and Why You Should Care. White Paper, IDC, Framingham (2011)
8. Kitchenham, B., et al.: Systematic literature reviews in software engineering–a systematic literature review. Inf. Softw. Technol. **51**(1), 7–15 (2009)
9. Rubinstein, I.: Big data: the end of privacy or a new beginning? In: International Data Privacy Law (Forthcoming), pp. 12–56 (2013)
10. James, B.D.: Security and privacy challenges in cloud computing environments. **8**, 24–31 (2010)
11. Katal, A., Wazid, M., Goudar, R.H.: Big data: issues, challenges, tools and good practices. In: Sixth International Conference on Contemporary Computing (IC3). IEEE (2013)
12. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**(05), 557–570 (2002)
13. Hashem, I., Targio, A.: The rise of “big data” on cloud computing: review and open research issues. Inf. Syst. **47**, 98–115 (2015)
14. 5 Hadoop Security Projects. https://www.xplenty.com/blog/2014/11/5-hadoop-security-projects/
15. Lei, X., Jiang, C., Wang, J., Yuan, J., Ren, Y.: Information security in big data: privacy and data mining. IEEE Access **2**, 1149–1176 (2014)
16. Hu, J., Vasilakos, A.V.: Energy big data analytics and security: challenges and opportunities. IEEE Trans. Smart Grid **7**(5), 2423–2436 (2016)
17. Yan, Z., Ding, W., Yu, X., Zhu, H., Deng, R.H.: Deduplication on encrypted big data in cloud. IEEE Trans. Big Data **2**(2), 138–150 (2016)
18. Xu, L., Jiang, C., Chen, Y., Wang, J., Ren, Y.: A framework for categorizing and applying privacy-preservation techniques in big data mining. Computer **49**(2), 54–62 (2016)
19. Liang, K., Susilo, W., Liu, J.K.: Privacy-preserving ciphertext multi-sharing control for big data storage. IEEE Trans. Inf. Forensics Secur. **10**(8), 1578–1589 (2015)
20. Perera, C., Ranjan, R., Wang, L., Khan, S.U., Zomaya, A.Y.: Big data privacy in the internet of things era. IT Professional **17**(3), 32–39 (2015)
21. Nedelcu, B.: About big data and its challenges and benefits in manufacturing. Database Syst. J. **4**(3), 10–19 (2013)
22. Gaff, B.M., Sussman, H.E., Geetter, J.: Privacy and big data. Computer **47**(6), 7–9 (2014)
23. Cárdenas, A.A., Manadhata, P.K., Rajan, S.P.: Big data analytics for security. IEEE Secur. Priv. **11**(6), 74–76 (2013)
24. Yu, S., Liu, M., Dou, W., Liu, X., Zhou, S.: Networking for big data: a survey. IEEE Commun. Surv. Tutorials **PP**(99), 1 (2016)
25. Baek, J., Vu, Q.H., Liu, J.K., Huang, X., Xiang, Y.: A secure cloud computing based framework for big data information management of smart grid. IEEE Trans. Cloud Comput. **3**(2), 233–244 (2015)
26. Dong, X., Li, R., He, H., Zhou, W., Xue, Z., Wu, H.: Secure sensitive data sharing on a big data platform. Tsinghua Sci. Technol. **20**(1), 72–80 (2015)

# Synthesis of Expert System for the Distributed Storage of Models

Maxim Polenov(✉), Sergey Gushanskiy, and Artem Kurmaleev

Department of Computer Engineering,
Southern Federal University, Taganrog, Russia
{mypolenov, smgushanskiy}@sfedu.ru,
art.kurmaleev@gmail.com

**Abstract.** This paper is about the synthesis of expert systems for the Distributed Storage of Models software package. This package is intended for storing and translating models from various languages. Application of expert system was proposed to automatize process of models' multi-language translation. All steps of synthesis and decisions were described. Pre-alpha version of expert systems based on translation module from Modelica language was developed using object-oriented approach. Approach choice was substantiated on specifics of modeling languages translation. Functionality for pre-alpha was examined and results were claimed to analyze.

**Keywords:** Models' translation · Multitranslator · Distributed Storage of Models · Expert system · Object-oriented approach · Modelica-models translation

## 1 Introduction

Nowadays with presence of huge amount of modeling tools researchers come to the problem when they have to recreate some of their models for different tools depends on their needs. To solve this problem the Multitranslator environment [1] was created to support translation from one language into another in the automatized way. After developing the standalone version of the Multitranslator [2] it was decided to split it into client-server model, using distributed architecture [3], reusable models data base [4], and called as the Distributed Storage of Models [5].

Multi-language translation is based on translation module development in Multitranslator IDE for each pair of languages. But in some cases language translation was complicated to implement [6] due to insufficient input data or when uncertainty appeared upon solving tasks with many possible solutions considering languages structures.

It brought us to proposing the expert system as a solution of this problem. Of course user-expert that obtains knowledge about the Multitranslator and input and output languages can solve this problem as well as expert system. But usually such expert is not available at any time, so development of such solution is quite necessary.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_23

## 2 Development of Expert System

To start with, the management of the requirements and plan for development are needed. As for the plan, the basic stages of expert system development [7] are used:

- Create pre-alpha version (prototype) – basic version with much fewer possibilities;
- Improve pre-alpha version – make experiments with real tasks and fill knowledge base with help of experts for chosen languages;
- Beta-testing – involving collaborators for testing;
- Release – expert system after all the tests with filled knowledge base and user documentation;
- Support and expansion – error recovery, adding more knowledge bases for different languages.

These requirements take research approach so for each stage they are not as strict as if it was commercial product. But of course they have to be strict enough to be approved during experiments. Pre-alpha version has to have most functionality of the parsing part of translation module used by Multitranslator to translate models from chosen language.

Improved version of expert system has to be same functional as if it was usage of parsing part of translation module by Multitranslator with expert help. And it has to content new expert knowledge gathered during the experiments in knowledge base. Beta-test is required to minimize all the types of faults [7] that could have been done while filling the knowledge base. Upon releasing expert system have to contain no faults in at least one knowledge base as well as full project and user documentation. Support will be made by correcting faults in knowledge bases and adding new ones to expand expert system for supporting new languages.

### 2.1 Justification for Language and Data Representation Choice

The real choice was made only between the basic CLIPS [8] language, which could be used to represent data as facts, and COOL (CLIPS Object-Oriented Language) [9] that allows us to use all the advantages of the object-oriented programming for development.

Let’s compare data representation types that can be used and take a look at the pros of COOL comparing to CLIPS:

- Inheritance. It allows creating more structural and modular definitions of data;
- Objects can keep corresponding procedural information by using message handler;
- Template comparison in object-oriented representation provides greater flexibility than fact representation;
- Abstraction. It simplifies operating with objects comparing to facts where it doesn’t exist.

For cons it will be lower performance comparing to fact method. But this won’t be an issue, since Multitranslator has already done most of the job instead of expert system and the main task of expert system is to resolve exceptional cases.

After comparing pros and cons described above it becomes obvious that COOL have to be used to implement object-oriented data representation of expert system.

## 2.2 Expert System Development Stage

Let's consider the steps of the expert system development [7]:

1. Planning;
2. Defining the knowledge – translation module is being knowledge source for alpha version of expert system and have the highest priority;
3. Synthesis of knowledge;
4. Code development and debugging;
5. Knowledge verification;
6. System evaluation.

All of the steps above have to be completed in cycle for the first knowledge base and some of them for each following base.

## 2.3 Expert System Interaction with the Distributed Storage of Models

Let's consider the Distributed Storage of Models [10] interaction with the expert system based on client-server model. This interaction is shown in Fig. 1.

After being parsed input data from the Multitranslator is transferred into the memory of the expert system. And next expert system is processing input data. Processing is based on the knowledge base of the expert system.

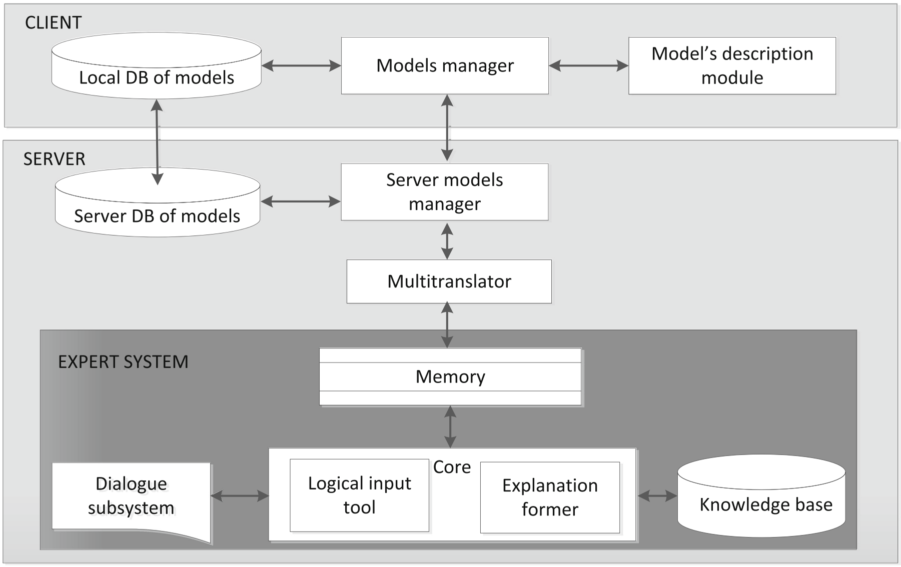


**Fig. 1.** The Distributed Storage of Models interaction with the expert system

The expert system operation becomes finished when there is no data to process and the resulting data is transferred to the Multitranslator to finish the output code generation. Otherwise processing goes on. If the expert system has already used the whole knowledge base to process and input data is still not processed it means that expert system requires more knowledge and it have to be provided.

The logical input tool allows user better understand the processing of knowledge in the expert system. It provides a possibility to watch the order of functions of processing and changes they've made.

Since an expert system is intended to replace the user-expert so it's final version has to solve the same tasks in same consequence.

Application of the expert system allows us to specify the section of model by making identifiers for processed code. This process is made by comparing input data with knowledge data base (fast search). If there is more than one equal element then the additional search parameter (deep search) can be made. It allows us to speed up the operation of the expert system.

## 3 Gathering Knowledge from Translation Module

Let's consider the structure of the Multitranslator's translation module (TM) that allows converting models from Modelica [11] language into C# [12] (Fig. 2). This structure is recursive since translation module contains instructions for input code parsing.

Let's treat some key elements, which aren't shown in Fig. 2:

- IDENT – responsible for parsing all the identifiers, it contains instructions that allow reading the name and printing it into the corresponding place in the generated output code;
- What clause – defines handlers;
- Inner or outer – just splits into "inner" and "outer" handlers;
- Import and extend clauses – it handles "import" and "extend" parameters. After import it is expected to find identifiers and names. And after "extend" it's expected to find names and class modifiers;
- Annotation – handles "annotation" parameter and expects class modifier afterwards;
- Algorithm and equation clauses – they handle "algorithm" and "equation" parameters. The algorithm clause contains instructions for algorithm parsing that expects function calls, expressions, and component references. And the equations in Modelica can contain connections, expressions, asserts.

To understand better let's detail elements in Fig. 2:

- "Modelica" element also can include all the required preload code fragments to integrate them into the project;
- "Name" handles identifiers parsing recursively referring to "IDENT";
- "Class definition", "Class specifier", "Class type" – all together they describe class parsing for prefixes, types, specifiers, class names and, of course, comment afterwards;

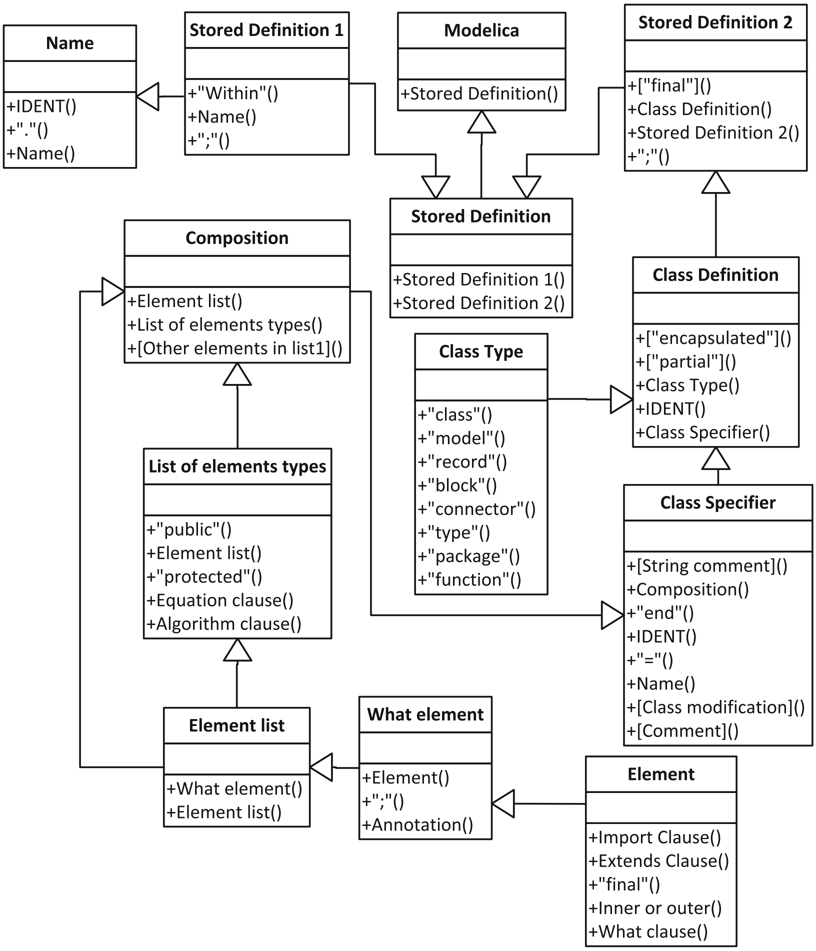


**Fig. 2.** Structure of Multitranslator's translation module from Modelica into C#

- "Composition", "List of elements types", "Element list", "What element", "Element", at first, "Composition" defines what is being parsed, elements or types and forces to be completed "Element list" rule or "List of elements types" rule correspondingly;
- "What element" doesn't only define element but also generate target language units into resulting file of model translation;
- "Element" describes rules about "import" and "extend" prefixes, as well as "inner" and "outer" ones and "final" suffix.

But we've take into consideration that class structure for knowledge base should be made in the same way except "Modelica" and "Stored Definition" elements, since it's not necessary to make them as classes in COOL.

## 4 Synthesis of Knowledge Base

First of all let's consider that most of the code conversion was already done by Multitranslator so information about the code section can be transferred to simplify code recognition process. Logically it can be only variable declaration section right after a model declaration section and only next the equation and the algorithm sections.

Since object oriented approach was chosen then our knowledge base is implemented as class instances. Our expert system is separated into couple of modules for current needs:

```
(defmodule INPUT (import MAIN ?ALL))
(defmodule DATA (import MAIN ?ALL))
```

Classes for INPUT data and knowledge base data are:

```
(defclass INPUT::DATA-SOURCE
   (is-a USER)
   (multislot namedata
    (storage shared)
  (visibility public))
   (multislot namespec
    (storage shared)
  (visibility public)))
(defclass DATA::DATA-TYPE
   (is-a USER)
  (multislot namedata
   (storage shared)
   (visibility   public))
  (multislot namespec
   (storage shared)
   (visibility   public)))
```

"Namespec" is required to define the code section. "Namedata" contains keywords that are applicable for the defined code section. In the current pre-alpha version of expert system we have next specifiers for "namespec":

1. "Type specifier" – contains all the variable types;
2. "Prefix type specifier" – contains all the prefixes for variable types;
3. "Conditional operator" – contains all the conditional operator types;
4. "Loop operator" – contains all the loop operators;
5. "Logic operator" – contains all the conditional logical operators;
6. "Mathematical function" – contains Modelica mathematical functions.

*Synthesis of rules.* In pre-alpha version there're only 3 levels of rules, the higher the level is the higher priority it has: expert; request and management rules.

*Expert rules.* Such class instances behave as expert rules for knowledge base. Here's an example of class instance for knowledge base containing data that was decided to separate depends on given code:

```
(definstances DATA::data-source-type
 ([Type-specifier] of DATA-TYPE
 (namedata Integer Real Integer Boolean)
 (namespec Type specifier)))
```

As you can see from the code above "namespec" field allows specifying its application for the exact code section. So "Type specifier" is only applicable for variable declaration code section. All six types of specifiers that were described above have similar class instances to parse input data and make results.

*Request rules.* In COOL message handlers can behave as rules. Such rules as you can see below have instructions for gathering knowledge from knowledge base. To retrieve data from INPUT (input data) and DATA (knowledge base data) classes message handler was used:

```
(defmessage-handler INPUT::DATA-SOURCE put-namedata
(?value)
(dynamic-put namedata ?value))
(defmessage-handler DATA::DATA-TYPE put-namedata (?value)
(dynamic-put namedata ?value))
```

*Management rules.* Management rules operate expert and request rules to finalize the recognition process and generate the result:

```
(do-for-instance
((?tl INPUT::DATA-SOURCE  ) (?t2 DATA::DATA-TYPE) )
(not(neq ?tl:namedata ?t2:namedata))
 (printout t ?tl:namedata "is a" ?t2:namespec crlf))
```

The construction "do-for-instance" sorts out both "INPUT::DATA-SOURCE" and "DATA::DATA-TYPE" class instances with condition that comes in the next brackets. Statement "neq" compares all the fields and returns "TRUE" only in case if all the fields have not equal values and/or different types. The "action part" of the rule prints the definition of input data (?tl:namedata) from knowledge base (?t2:namespec).

## 5 Example of Code Parsing from Modelica

To have a better explanation and share results of synthesis it was decided to make an experiment that will also showcase some functions that are already implemented.

To do it lets suppose that Multitranslator was parsing Modelica-model of petrol engine:

```
model PEngine "Petrol Engine"
  parameter Real T0=60.0 "Torque applied to the load at
zero speed";
  parameter Real J=0.5 "Moment of the inertia";
  parameter Real b=0.1 "Coefficient of friction";
  input Real T=0.0 "Torque applied to the load";
  output Real w(start=0.0) "Angular rotor speed";
equation
  der(w)*J=T0-T-b*w;
end PEngine;
```

And when it has reached "der" mathematical function that calculates derivative and was not recognized by Multitranslator's translation module correctly, so the expert system has to do it instead. To do it an input data for "INPUT" class have to be created:

```
(definstances INPUT::user-data-source
  ([data] of DATA-SOURCE
  (namedata der)
  (namespec none)))
```

As "do-for instance" statement finds the right condition it prints out:

```
der is a Mathematical function
```

Afterwards, required information will be transferred to Multitranslator to continue parsing process properly.

From this example it is certain that pre-alpha version is working as intended and development went as expected.

## 6 Conclusion

The proposed approach for automatization of translation from modeling languages has been synthesized and pre-alpha version of expert system was developed. It already allows us to use expert system as parsing part of the translation module of Multitranslator in some cases to resolve parsing uncertainty. It utilizes Multitranslator within Distributed Storage of Models as translation tool, extends its functionality and improves universality.

And also, this approach allows us implementing the automatization to make translation in other direction from programming languages into modeling languages.

## References

1. Chernukhin, Yu., Polenov, M.: Instrumental subsystem of multilanguage translation of virtual modeling systems. Izvestiya SFedU. Eng. Sci. **3**, 115–120 (2004). Taganrog (in Russian)
2. Chernukhin, Yu., Guzik, V., Polenov, M.: Multilanguage Translation for Virtual Modeling Environments. Publishing house of Southern Scientific Center of Russian Academy of Sciences, Rostov-on-Don (2009). (in Russian)
3. Tanenbaum, E., Van Sten, M.: Distributed Systems: Principles and Paradigms, 2d edn. Prentice-Hall, Upper Saddle River (2006)
4. Robinson, S., Nance, R.E., Paul, R.J., et al.: Simulation model reuse: definitions, benefits and obstacles. Simul. Model. Pract. Theory **12**, 479–494 (2004)
5. Polenov, M.: Distributed tools of models conversion and storage. Inf. Commun. **2**, 58–61 (2014). Moscow (in Russian)
6. Chernukhin, Yu., Guzik, V., Polenov, M.: Multilanguage translation usage in toolkit of modeling systems. WIT Trans. Inf. Commun. Technol. **58**, 397–404 (2014). vol. 1, WIT Press, Southampton
7. Giarratano, J.C., Riley, G.D.: Expert Systems: Principles and Programming, 4th edn. Course Technology, Boston (2004)
8. CLIPS official site. http://www.clipsrules.net/
9. CLIPS Object Oriented Language (COOL). https://www.csie.ntu.edu.tw/~sylee/courses/clips/bpg/node9.html
10. Polenov, M., Guzik, V., Gushanskiy, S., Kurmaleev, A.: Development of the translation tools for Distributed Storage of Models. In: Proceedings of 9th International Conference on Application of Information and Communication Technologies (AICT 2015), pp. 30–34. IEEE Press (2015)
11. Modelica and the Modelica Association. https://www.modelica.org/
12. C# Programming Guide. https://msdn.microsoft.com/en-us/library/67ef8sbd.aspx

# Finding Relationships in Industrial Data with the Use of Hierarchical Clustering

Martin Nemeth[1,2(✉)] and German Michalconok[1,2]

[1] Faculty of Materials Science and Technology in Trnava,
Slovak University of Technology in Bratislava, Bratislava, Slovakia
{martin.nemeth, german.michalconok}@stuba.sk
[2] Institute of Applied Informatics, Automation and Mechatronics,
Trnava, Slovakia

**Abstract.** The aim of this paper is to describe the cluster analysis of the failure data from the industrial process. The failure data used in the research were obtained from the automotive industry. The purpose of this analysis is to look at the data in broader view, and to discover various relationships in the data considering different parameters using data mining technique. The data analysis was performed by using hierarchical clustering for finding relationships between failures. We chose the hierarchical clustering analysis to find previously unknown relationships between given failure types, which is the type of task cluster analysis is mostly used for.

**Keywords:** Data mining · Hierarchical clustering · Failure · Reliability

## 1 Introduction

Safety is crucial for industrial processes. When the safe operation cannot be maintained, the risk of dangerous situations increases. These situations can be harmful for human health and life, property and can cause financial losses. In industrial production process, dangerous situations can be caused by occurring failures of various types. Developing a prediction model, which will be able to predict time and conditions under which a failure will occur is a difficult task. A precise prediction model should be based on huge amount of data with various parameters. However, data collected from the production process do not always satisfy these conditions. It is common, that the data often consist of only a few parameters, which are not ready to serve as a training set for predictive algorithms. It is needed to analyze the data to get a broader view and to be able to derive new parameters that can be helpful in the process of developing of the prediction model.

This paper deals with the process from automotive industry. The process of the assembly of the rear wheel happens on an assembly line which is separated from the rest of the process. This means that this assembly line has its own set of failures which have no relationship with failures from the rest of the process. This makes this smaller assembly line ideal for the failure analysis. There are various groups of failures, which can occur in mentioned assembly line and are described in the Table 1. In this table are described given groups of failures with their defined consequences on the process.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_24

**Table 1.** Overview of the failure groups.

| Group | Consequence | Examples |
|---|---|---|
| Immediate stop with out-of-order state | Causing immediate stop of one or more stations. These stations are put to out-of-order state | Emergency stop |
| | | Interruption of the intangible barriers. |
| Immediate stop | Causing the stop of the cycle of the station | Failure of the operation |
| | | Failure of the sensor's self-check |
| | | Failure of the breaker |
| Delayed stopping | Causing the stop of the station at the end of the cycle | Faulty component |
| | | Level error |
| Alarms | They do not cause the stop, but they need to be resolved | |
| Warnings | They do not cause the stop, and they do not need to be resolved | |
| Waitings | They inform that the station is stopped | Waiting for loading |
| | | Waiting for unloading |
| | | Conveyor saturation |

The failure dataset consists of approximately 65000 records. These records were exported from the time frame of 6 months. Each record has several parameters like failure description, start date (with time in dd/mm/yyy hh:mm:ss format), end date (with time in dd/mm/yyy hh:mm:ss format), duration of the failure (in hh:mm:ss format), localization of the failure and belongs to one of mentioned failure groups.

This article further deals with various methods to analyze the data and to find relationships in the data.

## 2 Methods

To find relationships between failures a clustering method was used. Clustering is one of various methods of data mining. According to Friedman data mining is set of methods used to discover relationships in data in large data bases. It overlaps at some degree with fields like artificial intelligence, machine learning, pattern recognition and data visualization [1]. There are however many other definitions which largely depend on the purpose of use of the data mining methods. Data mining consists of multiple steps, which are shown in the Fig. 1.

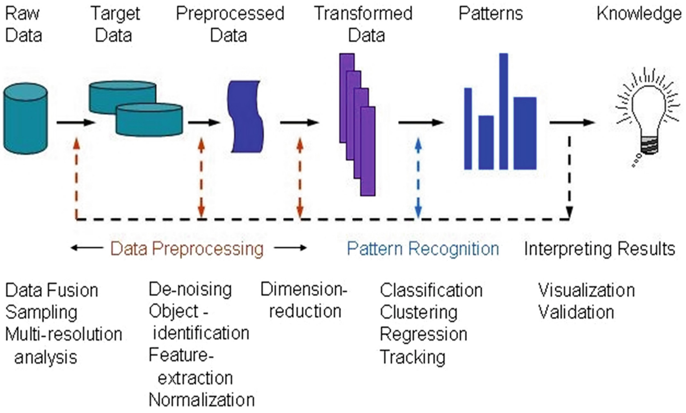


**Fig. 1.** Data mining scheme [3].

## 2.1 Clustering

Clustering algorithms are one of the methods of data mining. They are aimed to automatically partition the data into a set of regions. These regions are called clusters. In these clusters, all data are similar to each other. This similarity is most often defined by the Euclidean distances. In our case, the Euclidean distances were computed between failure types in the data set. The algorithms of the cluster analysis are most often used to find the hidden structure in the raw data, or existing patterns in these data.

For the purpose of the research, the hierarchical clustering algorithm was used. The result from hierarchical clustering is a hierarchical tree which is produced by recursive partitioning of the data in a top-down or bottom-up structure. The hierarchical clustering works with a distance matrix, or similarity matrix as an input and can be subdivided as following [4, 5]:

- Agglomerative hierarchical clustering: A bottom-up approach which each object initially represents a cluster of its own, then similar clusters are iteratively merged until the desired cluster structure is obtained. This algorithm for N samples begins with N clusters and each cluster contains a single sample.
- Divisive hierarchical clustering: A top-down approach that all objects initially belong to a single root cluster and iteratively partitions existing clusters into sub-clusters.

For the purpose of this research, the agglomerative approach was used.

## 2.2 The Distance Matrix

The values of the distance matrix, which is used as an input to hierarchical clustering algorithm, represent distances between each failure type from the data set. These values are calculated as a Euclidean distance between two random points [$x_1$, $x_2$,…, $x_d$] and [$y_1$, $y_2$,…, $y_d$] is computed as follows [2]:

$$\sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \tag{1}$$

A Euclidean distance matrix in $R_{+}^{N \times N}$ is an exhaustive table of distance-square $d_{ij}$ between points taken by pair from a list of N points $\{x_\ell, \ell = 1 \ldots N\}$ in $R_n$. Each point is labelled ordinally, hence the row or column index of a Euclidean distance matrix, i or j = 1 … N, individually addresses all the points in the list [3].

## 3 Results

### 3.1 Failure Similarity with Respect to the Start Date and Duration Parameters

The hierarchical clustering was first used to find overall relationships between failures. The initial step was to describe each failure type by chosen properties. So, each failure type can be represented as a point in the Euclidean space. Start date and duration of the failure were chosen from the set of parameters. The time distances between emerging of the same failure type were calculated with the use of the first parameter "start date". These distances were then calculated for each failure type. Another step was to calculate the average time between emerging failures of the same type also for each failure type. Subsequently, the average failure duration time was calculated for each failure type. After calculating these values, each failure type was described by two parameters. These parameters were the average time between emerging of the same failure type and average duration of the failure type. These calculated parameters could be understood as coordinates of a point in the Euclidean space. Then the failure type itself could be considered as a point in Euclidean space. The distance matrix was then calculated as Euclidean distance values between failure types.

The agglomerative hierarchical clustering approach was used for the initial data analysis. The cluster analysis was performed in Statistica 13 software. The output of this method is a set of clusters organized in a tree structure. Following table shows cluster membership of each failure type in each layer of the tree.

The Table 2 shows relationships between failure types. The strength of the relationship between failure types grows with the number of tree levels in which are these failure types in the same cluster. For example, failure type F1 has strong relationship with failure type F5, F8 and F10, because they belong to the same cluster through all levels of the hierarchical tree. As for the failure types F2, F7, F12 and F13 applies that the strongest relationship is between failure types F2 and F13 and the relationship with failure type F7 and F12 should be repeatedly assessed.

### 3.2 Failure Similarity with Respect to the Derived Parameter

In the next step the failures F1, F5, F8 and F10, which had the strongest relationship, were assesed. For each of these failures a histogram was created. These histograms capture the frequency of each failure type for the period of 6 months (Figs. 2, 3 and 4).

**Table 2.** Cluster membership of failure types after performing hierarchical clustering.

| Failure type | Cluster membership | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **10 clusters** | **9 clusters** | **8 clusters** | **7 clusters** | **6 clusters** | **5 clusters** | **4 clusters** | **3 clusters** | **2 clusters** |
| **F1** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **F2** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **F3** | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 2 |
| **F4** | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 3 | 2 |
| **F5** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **F6** | 6 | 6 | 6 | 6 | 5 | 3 | 3 | 3 | 2 |
| **F7** | 7 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **F8** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **F9** | 8 | 7 | 7 | 7 | 6 | 5 | 3 | 3 | 2 |
| **F10** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **F11** | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 |
| **F12** | 9 | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **F13** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **F14** | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 |
| **F15** | 10 | 9 | 8 | 2 | 2 | 2 | 2 | 2 | 2 |

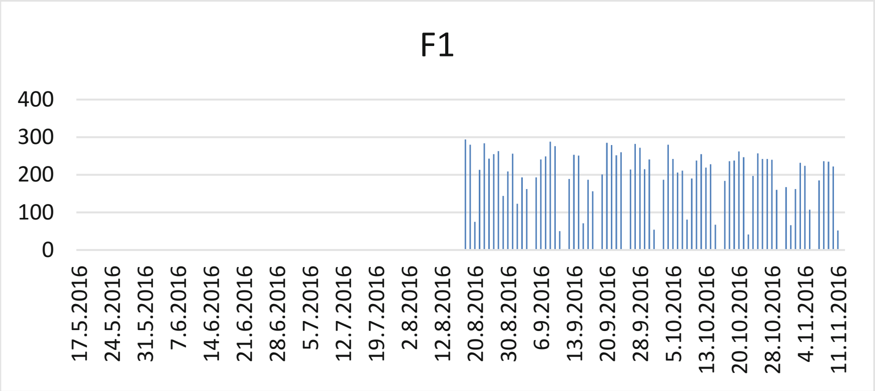


**Fig. 2.** Histogram of the failure F1

When looking closely at these histograms a similar repetitive pattern can be seen through these failure types. These anticipated patterns are in the histograms divided by blank spaces. All these blank spaces appeared at the same day of the week. According

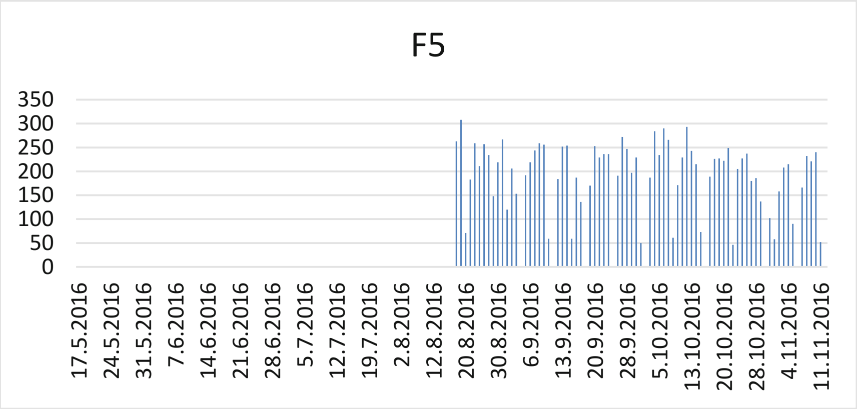


**Fig. 3.** Histogram of the failure F5

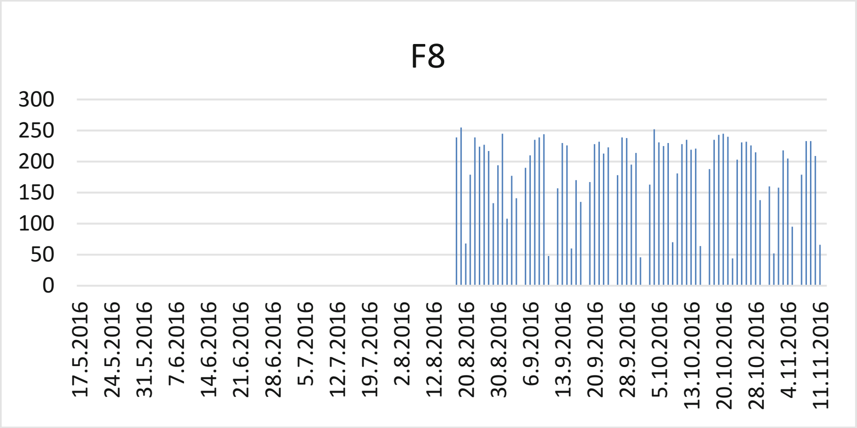


**Fig. 4.** Histogram of the failure F8

to this observation, we have decided, to assess the failure types again, but with respect to the day of the week, in which they occurred.

To assess the failure types with respect to the day of the week we also decided to use hierarchical clustering. In the first cluster analysis of the failure data we chose start date and duration as parameters to compute the distance matrix. To compute the second distance matrix, we had to create new derived parameter “day of the week” for each record in the data set. Then the matrix itself was also computed as the Euclidean distance values between failure types using parameters start date and day of the week. Subsequently, this matrix was used as an input to the hierarchical clustering algorithm also in the Statistica 13 software.

Table 3 shows the cluster membership of each failure type in each layer of the hierarchical tree. It is clear, that failure types F1, F5, F8, F10 and F2, F7, F12, F13

**Table 3.** Cluster membership of clustered failure records.

| **Failure type** | **Cluster membership** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **10 clusters** | **9 clusters** | **8 clusters** | **7 clusters** | **6 clusters** | **5 clusters** | **4 clusters** | **3 clusters** | **2 clusters** |
| **F1** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **F2** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **F3** | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 2 |
| **F4** | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 3 | 2 |
| **F5** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **F6** | 6 | 6 | 6 | 6 | 5 | 3 | 3 | 3 | 2 |
| **F7** | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **F8** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **F9** | 8 | 7 | 7 | 7 | 6 | 5 | 3 | 3 | 2 |
| **F10** | 7 | 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **F11** | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 |
| **F12** | 9 | 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| **F13** | 8 | 5 | 3 | 5 | 2 | 2 | 2 | 2 | 2 |
| **F14** | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 |
| **F15** | 10 | 9 | 8 | 2 | 2 | 2 | 2 | 2 | 2 |

remained in the same clusters as in first cluster analysis. However, the strength of the relationships between some failure types inside clusters has changed.

Tables 2 and 3 are showing that there is a relationship between mentioned failures even from different perspectives.

## 4 Conclusion

This paper was devoted to the cluster analysis of the failure data from industrial production process of automotive industry. The dataset consists of approximately 65000 records and each record represents one occurrence of a failure type. Each record has its parameters such as failure type description, start date, end date, duration and localization. To analyze the data, the hierarchical clustering analysis was used. Since the hierarchical clustering needs a distance matrix as an input, the matrix was computed by computing the Euclidean distances between failure types. In the first analysis, the parameters start date and duration were used as coordinates of a point (failure type) in Euclidean space to compute the distance matrix. Based on the results from this initial cluster analysis, the secondary analysis was performed. This analysis was aimed to consider the day of the week, when the failures emerged in the system. This analysis resulted in a cluster organization, which confirmed the cluster organization from the first cluster analysis and showed, that the day of the week has also impact on emerging failures in the production process. Future research will be aimed to find various aspects that have impact on the emerging failures and to build a prediction model for failure prediction. It is clear, that it is possible to find relationships in failure databases and from this relationship derive valuable knowledge, which can help to predict failures in the production processes and to maintain safe operation of such processes.

**Acknowledgments.** This publication is the result of implementation of the project VEGA 1/0673/15: "Knowledge discovery for hierarchical control of technological and production processes" supported by the VEGA.

This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO" (ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

## References

1. Please Friedman J.H.: Data mining and statistics: what's the connection? Stanford University, Stanford, 10 November 2016. http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf
2. Babcock, B., Datar, M., Motwani, R., O'Callaghan, L.: Maintaining variance and k-medians over data stream windows. In: Proceedings of ACM Symposium on Principles of Database Systems (2003)
3. Kamath, C.: On the role of data mining techniques in uncertainty quantification. Int. J. Uncertainty Quantification **2**(1), 73–94 (2012)
4. Nazari, Z., et al.: A new hierarchical clustering algorithm. In: ICIIBMS 2015, Track2: Artificial Intelligence, Robotics, and Human-Computer Interaction, Okinawa, Japan (2015)
5. Alpydin, E.: Introduction to Machine Learning, pp. 143–158. The MIT Press (2010)

# Assessing of the Importance of Medical Parameters on the Risk of the Myocardial Infraction Using Statistical Analysis and Neural Networks

Andrea Peterkova(✉) and German Michalconok

Faculty of Materials Science and Technology in Trnava, Institute of Applied Informatics, Automation and Mechatronics, Slovak University of Technology in Bratislava, Bratislava, Slovakia
{andrea.peterkova,german.michalconok}@stuba.sk

**Abstract.** The aim of this article is to assess or to complete the medical hypothesis on the further prepared clinical data with the use of data mining methods. In our research, we focus on cardiological datasets of patients who underwent coronary angiography and who were indicated for the ischemic heart disease. These patients are divided into four stages of clinical diagnosis. The clinical hypothesis is pointing on the clinical parameters, which have significant impact on the probability of the occurrence of the myocardial infraction. For the data analysis, we use STATISTICA 13 software.

**Keywords:** Statistical analysis · Data mining · Clinical data · Ischemic heart disease

## 1 Introduction

Cardiovascular diseases are the leading cause of death in the European countries. The most common of these diseases is the coronary atherosclerotic disease, responsible for 19% of all deaths [1]. Therefore, the focus state of the methods for finding the best solutions to allow early diagnosis of this disease or at least to deploy the most appropriate treatment to alleviate the symptoms of the disease. Contemporary medicine has a large amount of raw data about patients. Based on the data obtained from blood tests and screening diagnostics, we have the complete data of individual patients suitable for further processing.

The aim of our research is to use machine learning algorithms to design the decision support tool which will help to specify the optimal treatment for patients with coronary artery disease. The aim of this article is to present the confirmation of medical hypothesis pointing on the medical parameters important for the final prognosis of the patient. The important parameters obtained by statistical analysis were subsequently verified using neural network.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_25

# 2 Methods

To verify the medical hypothesis, we used the statistical analysis and data mining methods.

## 2.1 Formulation of the Problem

As mentioned before, cardiovascular diseases are the leading cause of death in the European countries. One of the most serious complications of ischemic disease is acute coronary syndrome, which occurs most often on the basis of a ruptured atherosclerotic plaque. These atherosclerotic plaques can be divided into stable and vulnerable. In predicting vulnerability of atherosclerotic plaque, it is necessary to optimally display its quantitative but also qualitative characteristics, as well as take into account also the different parameters. These parameters are comorbidities, laboratory indicators, therapy, cardiac imaging methods etc. and they have an important impact on the vulnerability of the plaque.

In our research, we focus on examining how the laboratory indicators, cardiac imagining methods and therapy affect the final prognosis of the patients with coronary syndromes. In the table below are presented the most important parameters impacting the myocardial infraction on the based of the medical hypothesis. Among these parameters were selected the three the most important parameters which are diabetes, the median platelet volume and the number of stenosis above 50% (Table 1).

**Table 1.** Important medical parameters chosen based on the hypothesis.

| Medical Parameters and Their Types | | | | |
|---|---|---|---|---|
| Clinical | Biochemistry hematology | ECHO | CT | Therapeutic |
| Age | Troponin TROP-T-hs | Left ventricular ejection fraction LE-EF | The number of vascular stenosis above 50% | The num. of implanted stents |
| Diabetes | Cholesterol HDL | Kinetic disorders | The number of vascular stenosis above 90% | The dose of statin |
| Chronic renal insufficiency K/DIGO | The median platelet volume - MPV | Diastolic dysfunction | The proximal stenosis RIA | Nitrate |
| | | | The presence of stenosis | |

It has to be taken into account variable parameters of each patient such as clinical characteristics, biochemistry and hematology indicators, ECHO and CT results and therapeutic parameters including deployment treatment.

### 2.2 Data Mining Methods

Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data [2]. Data Mining involves the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns in data. Than the model is used for understanding phenomena from the data, for analysis and prediction. Data mining problems are often solved by using a mosaic of different approaches drawn from computer science, including for example multi-dimensional databases, machine learning, soft computing and data visualization and statistical techniques [3]. In addressing the specific problem it is difficult to take one of the approaches, since arriving at the correct results is usually synergy of several mentioned techniques. Data mining represents a big potential for the healthcare industry and medicine to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs.

### 2.3 Classification with the Use of Neural Networks

Neural networks can be used in any situation, where the main objective is to determine an unknown variable or attribute from known observations or registered measurements. It can be use for various forms of regression problems, classification, and time series, where there is a sufficient amount of historical (previous gained) data, and where there actually exists a tractable underlying relationship or a set of relationships between these data.

Medical data must be first specifically prepared for input into neural networks. It is also important that the network output can be interpreted correctly. For the input values were used parameters of the patients. The outcome represented the output value. The neural network classifier assigns class membership to an input *x*. For example, if the input set has three categories (healthy, recurrent angina pectoris, fatal myocardial infraction), a neural network assigns each and every input to one of the three classes. The class membership information is then carried in the target variable *t*. For that reason, in a classification analysis the target variable must always be categorical.

Dataset consists of over 8000 entries, which represent clinical data about patients, also their blood indicators, screening examinations and recommended therapy. 75% were use as training set, 15% as test set and the remaining 15% as validation set. Neural network was set follows: the type of neural network was a multilayer perceptron with minimum 8 and maximum 25 hidden units. There were 20 networks to train and 5 networks to retrain. The training performance on the data shows 99,45%.

## 3 Results

The statistical analysis of the data demonstrated the important parameters, which have significant impact on the outcome. The Fig. 1 shows the overlap with the medical hypothesis. In the set of 10 most important parameters obtained with initial statistical analysis, there are the majority of the parameters from the medical hypothesis.

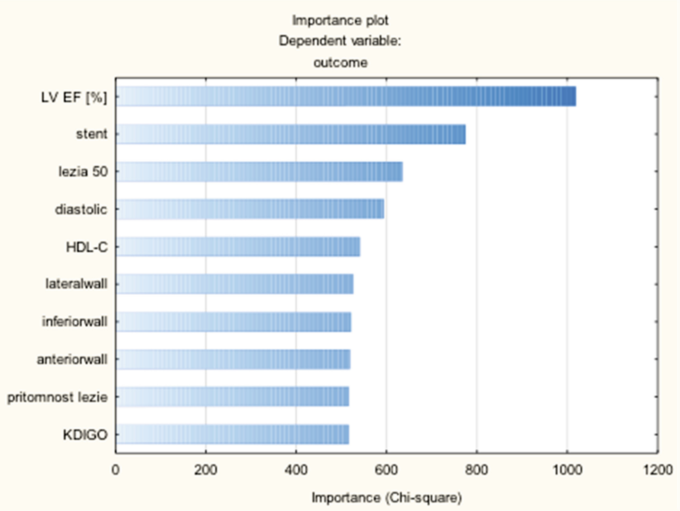


**Fig. 1.** The importance plot of clinical parameters

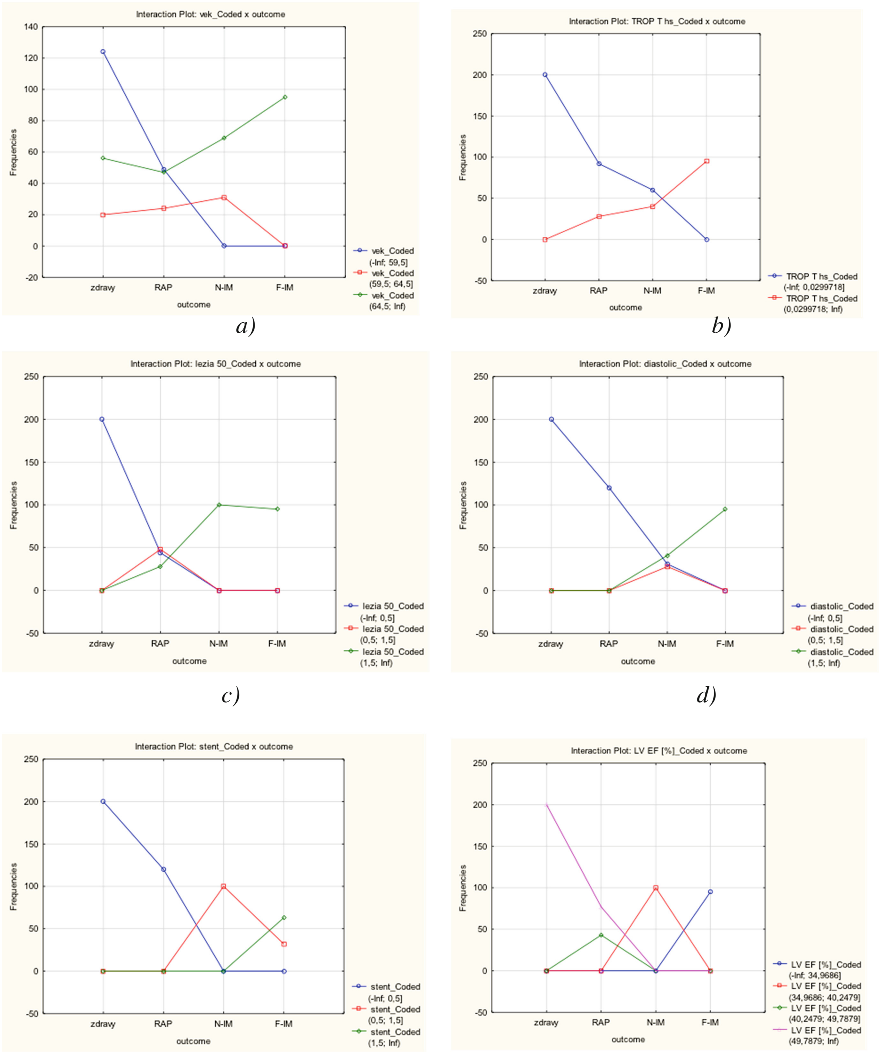
We have computed chi square values for each parameter to display the best k predictors. The predictors with the largest chi square values were chosen for the importance plot.

The Fig. 2 shows the interaction plots between chosen important parameters and the outcome. On the x-axis there are 4 predefined stages of the outcome and on the y-axis there is frequency of occurrence. On the left side of the x-axis there is stage where the patient is healthy, next stage capture patients with recurrent angina pectoris then patients with non fatal myocardial infraction and on the right side are patients with the fatal myocardial infraction. This interaction plots are also confirming the medical hypothesis. First interaction plot shows that with the increasing age increases also the risk of the fatal myocardial infraction, which means that the age has significant impact on the final prognosis of the patient. The remaining interaction plots also confirms that the computed important parameters have impact on the final prognosis.

The impact of chosen parameters on the outcome is also shown in the Fig. 3. Plots reflect the range of each parameter with respect to the outcome. As it is clear for example from the plot c, the left ventricular ejection fraction (LV EF) has a downward trend, which means that with the lower values of LV EF, the risk of the myocardial infraction increases. Another example is the impact of the UREA value, where it is clear that this parameter has not significant impact on the outcome because the range of values is the same for every stage of the outcome.

For the verification of the medical hypothesis was also used the mentioned neural network. The classification method was provided in STATISTICA 13 software. For the purpose of verification of the important parameters only the ten selected important parameters were used as a patients parameters to train the neural network.

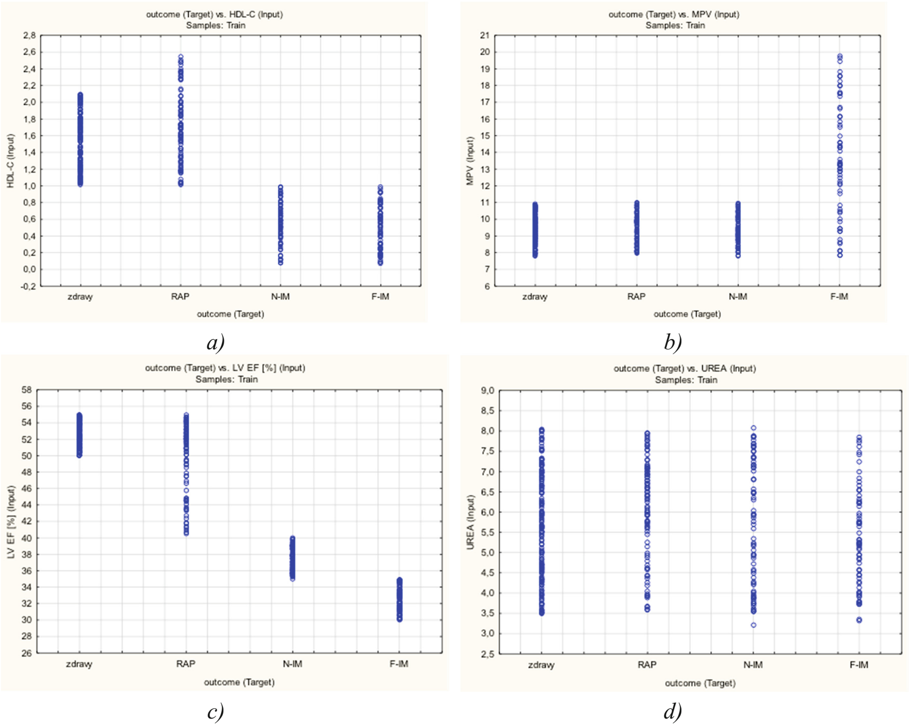
After the network training, the custom prediction was performed to test the learned network. To test the network with custom prediction, new patients, who were not



*a)* *b)* *c)* *d)* *e)* *f)*

**Fig. 2.** The interaction plot of important parameters. (a) Age (b) TROT T hs (c) The number of vascular stenosis above 50% (d) diastolic dysfunction (e) number of stents (f) Left ventricular ejection fraction

included in the training set, was chosen. It has been proved that the performance of the neural network was sufficient to classify new cases to the right classes. This demonstrates that the use of artificial neural networks will be further used for the development of the medical decision support system. This system will serve for the medical purposes to help recommend the correct treatment for patients with the coronary artery disease.



*a)* *b)*

*c)* *d)*

**Fig. 3.** The impact of chosen parameters on the outcome. (a) HDL-C (b) MPV (c) LV EF (d) UREA

## 4 Conclusion

In this paper, we are dealing with the medical parameters and monitoring their importance on the risk of the myocardial infraction. The outcome is divided into four stages and the importance of chosen parameters on the outcome of each stage is clear from Figs. 1, 2 and 3. The aim of this article was to assess the medical hypothesis, which dealt with the importance of various parameters on the final prognosis of the patients who were indicated to undergo CT angiography with the ischemic heart disease diagnosis. The hypothesis was confirmed with the methods of statistical analysis using the STATISTICA 13 software. The results broadly confirm the hypothesis, so the hypothesis was initially verified and therefore it is clear that the next development of the decision-support tool, the data can be further filtered out the parameters that has no significant influence on the outcome.

Future work is leading to the development of the medical decision support system. These systems are computerized decision support systems for problems whose solution is not clear at first sight. Such systems may be fully automated, partially supported by man, or may be a combination of both. The anticipated decision support systems will be used for determination of the optimal treatment for patients with coronary artery disease with respect to their final prognosis.

**Acknowledgments.** This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU – CAMBO" (ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

This publication is the result of implementation of the project VEGA 1/0673/15: "Knowledge discovery for hierarchical control of technological and production processes" supported by the VEGA.

## References

1. Townsend, N., Nichols, M., Scarborough, P., Rayner, M.: Cardiovascular disease in Europe–epidemiological update 2015. Eur. Heart J. **36**(40), 2696–2705 (2015). doi:10.1093/eurheartj/ehv428. Epub 25 Aug 2015. PubMed PMID: 26306399
2. Mohammed, J.Z., Wagner Jr., M.: Data mining and analysis: fundamental concepts and algorithms, First published 2014. ISBN: 978-0-521-76633-3
3. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: current issues and guidelines. Int. J. Med. Inf. **77**(2), 81–97 (2008)

# The Approach to Provide and Support the Aviation Transportation System Safety Based on Automation Models

Alexander Rezchikov[1], Vadim Kushnikov[1], Vladimir Ivaschenko[1], Aleksey Bogomolov[1], Leonid Filimonyuk[1], Olga Dolinina[2], Ekaterina Kulakova[2](✉), and Konstantin Kachur[2]

[1] Institute of Precision Mechanics and Control,
Russian Academy of Sciences, Saratov, Russia
iptmuran@san.ru

[2] Yuri Gagarin State Technical University of Saratov, Saratov, Russia
olga@sstu.ru, kulakova@petrol.chem.msu.ru

**Abstract.** We offered an approach to provide and support the safety of the complex system functioning based on the coordination of the interaction between its elements. The automation model of functioning process was created on the basis of the cause-effect approach. This model allows to present synchronous communication between elements of these systems. The model can be used in aviation training centers, which provide such programs as a periodic ground training and a professional development of aviation staff.

**Keywords:** Complex system · Aviation transportation system · Finite automation · Synchronization · Safety · Accidents

## 1 Introduction

The problem of the safety of the complex system functioning support on the basis of the elements interaction synchronization becoming increasingly important [2, 7, 8, 11]. For example for the aviation transportation systems (ATS) it is due to a number of factors, including the correct flow distribution of aerial vehicles (AV) near airports – the workload of large airports is so great that airstrips need to take or send the AV every 40–45 s. At the same time decision makers (DM) need to analyze the air situation for the short time and quickly make right decisions, which provide the full time airport work.

Today the interaction synchronization of the man-machine system separate elements can be simply organized in aggregative systems, Petri nets and automation models. Some aspects of such organization are presented in publications for example [5, 10].

This article describes the approach which allows more complete solution for such problem on a formal basis. We use the finite deterministic automation theory as the formal technique which provides the proof high level, the explanation simplicity of the decisions taken by the DMs and the time and computing operations amount effectiveness.

This approach was proved practically with the example of the ATS synchronization functioning processes.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_26

## 2 Task Definition

Suppose that the system has a number of subsystems $a_1$, …, $a_{n,}$ working capacity of these subsystems identifies the safety of the system functioning. The system safety condition is considered as a vector, coordinate of which $[s]_i$ with index $i$ is equal to 1, if the subsystem $a_i$ is working, and it is equal to 0 if the subsystem is not working, $i \in \{1, \ldots, n\}$. So the system has $2^n$ safety conditions.

The control input u of a set of admissible control inputs $U$ converts the system from the condition $s$ to some safety condition $s'$ with available resources and its required combinations. The safety condition $s'$ is identified by the control input $u$, the condition $s$ and the external impact. The physical sense of control inputs is a subsystem failures compensation, which requires to use the systems resource complex [1]. The resource costs are identified as the weight *W(u)* of each admissible control input $u$, which is equal to the weight sum of system elements.

Suppose that the experimental system has $S_{\text{st}}$ – a set of standard safety conditions. Setting the system in such conditions allows to avoid the appearance and development of critical event combinations [2].

We need to create an algorithm which identifies the minimal weight sequence of the control inputs, which converts the system from any initial condition $s_0$ to some safe condition $s^* \in S_{st}$.

## 3 The Method of Task Solution

For the task solution we use the finite deterministic Mealy automation $(S, U, Y, \delta, \lambda)$ as a formal solution technique. The conditions of this machine are the elements of S set and are identified with the system safety conditions, and signal inputs, $U$ elements are identified with control inputs applied to the system. In view of the fact that control inputs can be not enough for the adverse events compensation, each element of the $U$ set is presented as a trio ($u_{\text{con}}$, $u_{\text{mon}}$, $u_{\text{comp}}$), where $u_{\text{con}} \in U_{\text{con}}$ – a set of control signals, $u_{\text{mon}} \in U_{\text{mon}}$ – a set of special monitoring diagnostic signals, $u_{\text{comp}} \in U_{\text{comp}}$ – a set of adverse events compensation signals. Also it is suggested the existence of $U_{\text{comp}}$ in the alphabet. It's an empty symbol for the situation, when the reaction to the monitoring signal means the standard functioning and the compensation signals are not required.

For the task solution we use input sequences, which convert the automation from any initial condition to some finite condition $s \in S$, which depends on these sequences. Each sequence with such characteristic is synchronizing. The set of such $\bar{U}$ sequences and conditions of the automation (synchroconditions) is described in the following manner:

$$\bar{U}_{\text{syn}} = \{\bar{U} \in U* \mid (\exists s \in S)(\forall s \in S_0)(\delta(s, \bar{U}) = s)\}, \tag{1}$$

$$S_{\text{syn}} = \{s \in S| \, (\exists \bar{U} \in \bar{U}_{\text{syn}})(\forall s \in S_0)(\delta(s, \bar{U}) = s)\}. \tag{2}$$

If the automation can be synchronized and has standard synchroconditions, then corresponding synchronization sequences allow to convert the system to such conditions. If we search in ascending order of sequence length, then the first found sequence will be the shortest.

If the conditions of the set $S_{\mathrm{st}}$ are characterized by the subsystem working capacity, then it can be presented with a generalized condition, described in [9] for the automations with vector conditions.

The $[M]_{\mathrm{v}}$ symbols are indicated as the first $v$ rows of a matrix $M$, particularly the first v coordinates, if $M$ is a column matrix.

Suppose that automation conditions are indicated as binary vectors with $n$ length and some natural number $v \leq n$ is given. Each of such sets of $\bar{s} \subseteq S$ automation conditions is called the generalized condition, and $[s_1]_v = [s_2]_v$ fulfills for any $s_1, s_2 \in \bar{s}$. A part of elements which is general for all $s$ elements is indicated as $[\bar{s}]_v$.

The input sequence is called as generalized synchronizing sequence (GSS), if the automation is converted to the same final generalized condition regardless of the initial condition after the infeed of this sequence. If the GSS exists then the automation is called generalized synchronized, and generalized conditions, which are created by the GSS, are the synchroconditions.

So we have the following task: it is required to identify the finding algorithm of the minimal weight GSS, which converts the automation to the given generalized condition $\bar{s}$.

The task solution of the synchronizing sequences identifying for general automations comes down to the searching for the solutions with synchronizing trees, which requires serious computing resources. At the present time the best evaluation for the length of the synchronizing sequence is $O(n^3)$ [4], and according to the Černý conjecture this evaluation can be a number $(n - 1)^2$, where $n$ is a number of automation conditions. Also it is required to establish whether there are synchronizing sequences which convert the system to such generalized condition.

Therefore we consider the situation when the automation, which represents the system safety conditions, is linear or can be converted to the linear automation (LA) by the methods described in [13].

At a first approximation the LA can be used for the modeling of the controlling input sequences search, which convert ATS to safety conditions. Such description allows to get solutions for the given task, which correspond to the situations in real ATSs.

The advantage is that the results received for the LA allow to speed up the synchronizing sequences identifying and reduce the task solution to solving systems of linear equations.

The method implementation of the task solution requires to use some theorems. And it is possible that the GSS existence criterion is known [9].

**Theorem 1.** For the LA with the main characteristic matrix A GSS with length $k$ exists when and only when $[A^k]_v = [0]$.

And if this requirement fulfills for some $k$, then all sequences with length $k$ or more are the GSSs [9].

Let's indicate the smallest $k$ as $k_{\min}$, in which the requirement of the Theorem 1 is fulfilled, and

$$Q(k) = [A^{k-1}B, \ldots, B]_v, \quad \bar{U}(k) = \begin{pmatrix} u(0) \\ \cdots \\ u(k-1) \end{pmatrix} \tag{3}$$

Let's show that all generalized synchroconditions (GS) are achieved by applying the GSS with length $k_{\min}$. So the set of the GS coincides with $S_{\text{syn}}(k_{\min})$, if it is achieved by applying the GSS with length k with any $k \geq k_{\min}$.

**Theorem 2.** For the generalized synchronized LA a set of $S_{\text{syn}}(k)$ with any $k \geq k_{\min}$ coincide with a set of all linear combinations in line with independent matrix columns $Q(k_{\min})$.

**Proving.** Let's suppose that $k \geq k_{\min}$ and the GSS *u(0)*, ..., $u(k-1)$ converts the automation to the GS $\bar{s}$. According to the Theorem 1, the existence criterion of the GSS with length $k$ for the LA lies in the fact that $[A^k]_v = [0]$. And the full reaction formula for the linear automation implies that

$$[A^k]_v\, s(0) \; + \; [A^{k-1}B]_v\, u(0) + \ldots + [B]_v\, u(k-1) = [\bar{s}]_v. \tag{4}$$

Taking into account that $[A^k]_v = [0]$ we get

$$[A^{k-1}B]_v\, u(0) + \ldots + [B]_v\, u(k-1) = [\bar{s}]_v \quad \text{i.e. } Q(k)\bar{U} = [\bar{s}]_v. \tag{5}$$

Let's consider such correlation as a system of linear equations (SLE) reliable to unknown $kl - \bar{U}$ vector coordinates. As solvability criterion of the SLE is the representability of an absolute terms column as a linear combination in line with independent system matrix columns, so vector $[\bar{s}]_v$ in this situation is presented as a linear combination of matrix columns $Q(k)$. As the LA has the GSS with length $k_{\min}$, so according to Theorem 1 there is for any $k \geq k_{\min}$

$$[A^k]_v = [A^{k_{\min}} A^{k-k_{\min}}]_v = [A^k]_v A^{k_{\min}} A^{k-k_{\min}} = [0], \tag{6}$$

Where [0] is a matrix with zero elements.

So

$$Q(k) = [[0], \ldots, [0], \; Q(k_{\min})]_v. \tag{7}$$

So $[\bar{s}]_v$ is a linear combination in line with independent matrix columns [[0], …, [0], $Q(k_{\min})]_v$ or in line with independent matrix columns $Q(k_{\min})$. The arguments above also show the reverse: if $[\bar{s}]_v$ is a linear column combination $Q(k_{\min})$, which means that the GSS $\bar{U}$ with length $k_{\min}$ exists and converts the LA to the GS $\bar{s}$, then any input sequence beginning with $\bar{U}$ converts LA to the same OC $\bar{s}$. Q.E.D.

The next theorem shows that the GSS with minimal weight can be found among the GSSs with minimal length.

**Theorem 3.** If $W(u) \geq 0$ for any input signal $u$, then for any GSS $\bar{U}$ with length $k \geq k_{\min}$ there is the GSS $\bar{U}_{\min}$ with length $k_{\min}$, which converts the LA to the same GS as the GSS $\bar{U}$ , and $W(\bar{U}_{\min}) \leq W(\bar{U})$.

**Proving.** Let's suppose that the GSS $u(0), \ldots, u(k-1)$, where $k \geq k_{\min}$, converts the LA to the GS $\bar{s}$. According to the LA full reaction formula it means that

$$[A^{k-1}B]_v\, u(0) + \ldots + [B]_v u(k{-}1) = [\bar{s}]_v. \tag{8}$$

In this equation augends, which contain $A^{i}$ with $i > k_{\min} - 1$, are equal to [0] according to the Theorem 1, so

$$\begin{gathered}[0] + \cdots [0] + \left[A^{k_{min}-1}B\right]_v u(k - k_{min}) + \cdots \\ \cdots + [B]_v u(k-1) = [\bar{s}]_v\end{gathered} \tag{9}$$

It means that the GSS $u(k - k_{\min}), \ldots, u(k - 1)$ with length $k_{\min}$ converts the automation to the GS $\bar{s}$. And if the input signals weights are not negative, then the subsequence weight doesn't exceed the sequence weight which contains it. Q.E.D.

**Comment.** In the situation if $W(u) < 0$ with some $u$ the task for finding the minimal weight GSS has no solution, because according to the Theorem 1 it's always possible to specify rather long GSS with module unlimited negative weight.

Let's summarize the arguments mentioned above in the Theorem 4.

**Theorem 4.** The construction of the minimal weight GSS, which converts the given generalized synchrocondition $\bar{s}$, can be reduced to the task of the integer programming with linear restrictions and $lk_{\min} + v$ variables, where $k_{\min}$ is a minimal GSS length for the given LA, $l$ is a dimension of input vectors, $v$ is a characteristic of the generalized condition.

**Proving.** According to the Theorem 3 the minimal weight GSS can be found among the shortest GSSs. The shortest GSSs which convert the LA to the GS according to the Theorem 2 are the solution of the equations system.

$$Q(k_{\min})\bar{U}(k_{\min}) = [\bar{s}]_v. \tag{10}$$

Rewriting this equation over the Galois field with module 2 in form of comparison system we get the following

$$Q(k_{\min})\bar{U}(k_{\min}) \equiv [\bar{s}]_V \mod 2 \tag{11}$$

Let's consider the last system in the equivalent form

$$Q\bar{U} = [\bar{s}]_v + 2\,\bar{d}, \tag{12}$$

where $\bar{d} = (d_1, \ldots, d_v) = (d_1, \ldots, d_v)^{\mathrm{T}}$ is an integer vector, $Q$ is an integer matrix, $\bar{U}$ is a vector of unknown quantities. So the task of the minimal weight GSS construction, which converts the given generalized synchronized LA to the given generalized GS $\bar{s}$, is equivalent to the following task of the linear Boolean programming:

$$W(\bar{U}(k_{\min})) \to \min, \tag{13}$$

$$\begin{aligned} &Q(k_{\min})\bar{U}(k_{\min}) = [\bar{s}]_v + 2\,\bar{d}, \\ &\bar{U}(k_{\min})_i \in \{0, 1\}, \quad 1 \le i \le lk_{\min}. \end{aligned}$$

The evaluation $0 \le di \le lk_{\min}$, $1 \le i \le v$ is correct for coordinates $d$.

According to the arguments mentioned above the original task comes down to the task of the integer programming with linear restrictions and $lk_{\min} + v$ variables. Q.E.D.

The next theorem allows to evaluate the length of the minimal GSSs.

**Theorem 5.** If the GSS exist for the LA over the Galois field with module 2, then its minimal length doesn't exceed the quantity

$$(n-1)^2 + 2^{2-(n-1)\,\mathrm{mod}3} 3^{n-2-2[(n-1)/3]} + 1, \tag{14}$$

where $n$ is a dimension of the main characteristic matrix A.

**Proving.** Suppose that in the theorem hypotheses the minimal length of the GSS is equal to $k$, where $k > (n-1)^2 + 2^{2-(n-1)\ \mathrm{mod}3}\ 3^{n-2-2[(n-1)/3]} + 1$. According to the Theorem 1 it means that the requirement $[A^k]_v = [0]$ fulfills for this k. According to the Schwartz theorem the degrees of any Boolean matrix with a dimension n are periodical, starting from the degree $(n-1)^2 + 1$. The period length of this sequence is identified as the lowest common denominator of the greatest common denominators of the minimal cycle lengths in the strongly connected components of the graph corresponding to this matrix. As the greatest common denominators of the minimal cycle lengths are not exceeding the number of the strongly connected component vertexes, and the lowest common denominator of natural numbers doesn't exceed its multiplication, then the period length doesn't exceed the numbers multiplication maximum, the sum of which is equal to $n$. According to [3] this maximum is equal to $2^{2-(n-1)\ \mathrm{mod}3}\ 3^{n-2-2[(n-1)/3]}$, so the degrees sequence period $A^{\mathrm{k}}$, starting from the degree $k = (n-1)^2 + 1$, doesn't exceed the quantity $2^{2-(n-1)\ \mathrm{mod}3}\ 3^{n-2-2[(n-1)/3]}$. It is also correct for the submatrixes sequence $[A^k]_v$, so for each $k > (n-1)^2 + 2^{2-(\mathrm{n}-1)\ \mathrm{mod}3}\ 3^{\mathrm{n}-2-2[(\mathrm{n}-1)/3]} + 1$ there is such $k' \le (n-1)^2 + 2^{2-(\mathrm{n}-1)\ \mathrm{mod}3}\ 3^{\mathrm{n}-2-2[(\mathrm{n}-1)/3]} + 1$ that $[A^k]_v = [A^{k'}]_v$. Therefore if the requirement $[A^k]_v = [0]$ fulfills for some $\mathrm{k} > (\mathrm{n}-1)^2 + 2^{2-(\mathrm{n}-1)\ \mathrm{mod}3}\ 3^{\mathrm{n}-2-2[(\mathrm{n}-1)/3]} + 1$ then it also fulfills for some $k' \le (n-1)^2 + 2^{2-(\mathrm{n}-1)\mathrm{mod}3}\ 3^{\mathrm{n}-2-2[(\mathrm{n}-1)/3]} + 1$. It means that according to the Theorem 1 the existence of the GSS with length $k'$ contradicts the basic premise about the minimal length $k$. Q.E.D.

**Comment.** Let's consider the situation when the GS identifying relies on $v = n$, it means that GS coincides with a condition in the usual sense and the classical problem of synchronization is solving. In this situation the evaluation from the Theorem 5 can be essentially improved: the minimal GSS length doesn't exceed n, because the right part is replaced by $n$, because the Theorem 1 fulfillment with some $k$ means a nilpotency of the main characteristic matrix, whence it follows [12] that $A^{\mathrm{k}} = [0]$ with some $k \leq n$.

According to the Theorem 5, the requirement check $[A^k]_v = [0]$ makes a sense for the values $k \leq (n-1)^2 + 2^{2-(\mathrm{n}-1)\ \mathrm{mod}3}\ 3^{\mathrm{n}-2-2[(\mathrm{n}-1)/3]} + 1$. The first degree value which fulfill the requirement $[A^k]_v = [0]$ is a number $k_{\min}$, mentioned in the Theorem 2. All sequences with length $k_{\min}$ and more are generalized synchronizing, but according to the Theorem 2 for the calculation of all synchroconditions it's enough to find $S_{\mathrm{syn}}\ (k_{\min})$.

In terms of proved theorems let's sum up an algorithm with steps 1–3 to find the minimal weight GSS which converts the system to the safe GS $\bar{s}$:

1. Take $k = 1$.
2. Check the requirement of the Theorem 1. If the requirement of the Theorem 1 fulfills, then take $k_{\min} = k$ and go on to the step 3. If the requirement of the Theorem 1 doesn't fulfill and $k < (n-1)^2 + 2^{2-(\mathrm{n}-1)\ \mathrm{mod}3}\ 3^{\mathrm{n}-2-2[(\mathrm{n}-1)/3]} + 1$ (in the situation when $v = n$ take the right part equal to $n$), then increase $k$ by 1 and repeat the step 2. If the requirement of the Theorem 1 doesn't fulfill and $k \geq (n-1)^2 + 2^{2-(\mathrm{n}-1)\ \mathrm{mod}3}\ 3^{\mathrm{n}-2-2[(\mathrm{n}-1)/3]} + 1$, then finish the algorithm with report that the GSS doesn't exist.
3. Solve the task of the linear Boolean programming with $lk_{\min} + v$ variables:

$$\begin{aligned} &W(\bar{U}(k_{\min})) \to \min, \\ &Q(k_{\min})\bar{U}(k_{\min}) = [\bar{s}]_v + 2\,\bar{d} \\ &\bar{U}(k_{\min})_i \in \{0,1\}, \quad 1 \leq i \leq lk_{\min}, \\ &0 \leq d_i \leq lk_{\min}. \end{aligned}$$

This task has the solution when and only when $\bar{s}$ is one of the generalized synchroconditions.

**Comment 1.** If the input signals are balanced, then task of the integer linear programming is replaced by the linear equations system in integers with restrictions given above.

So the algorithm described above can be used for the system synchronization to the standard condition.

**Comment 2.** For the question about the complexity of finding of the generalized synchronizing sequences let's mention the following. If we consider n subsystems each of it can be in a functioning condition or in a failure condition then the general amount of safety conditions is $N = 2^{\mathrm{n}}$. If the identifying of the synchronizability and finding of the synchronizing sequences of such system uses methods applying for the general automations, then in a worst case scenario we need to find sequences with length $O(N^3) = O(2^{3\mathrm{n}}) = O(8^{\mathrm{n}})$. At the same time the length of the shortest synchronizing sequence for the linear automation is limited with the quantity

$(n-1)^2 + 2^{2-(n-1)\bmod 3}\,3^{n-2-2[(n-1)/3]} + 1 = O(2^{n/3})$ according to the Theorem 5. This quantity limits the search area of the smallest number $k_{\min}$, for which the requirement of the Theorem 1 fulfills. If this requirement fulfills then generalized synchronizing sequences exist and we can identify them from the system contained of $v \leq n$ linear algebraic equations with $lk_{\min} + v$ unknown quantities, where l is a dimension of the input vector, the number compared with $n$. If we talk about the task of Boolean programming, then these equations turn to task linear restrictions. From the mentioned above it follows that the introduction of the linear automation model highly reduces the procedures complexity for the identifying of the synchronization and synchronizing sequences.

## 4 The Task Solution Example

For the task solution, we use the following objects classes characterizing system conditions of the air traffic control [6]: "assigned flight level" – "FL$_1$", "FL$_2$", "FL$_3$", …, "FL$_m$"; AV – "AV$_1$", "AV$_2$", "AV$_3$", …, "AV$_k$"; "flight operations officer" - {"AV$_i$ command to take FL$_j$"}, $i = \{1, 2, \ldots, k\}$, $j = \{1, 2, \ldots, m\}$.

Let's suppose that safety requirements are fulfilled, if each flight level has no more than one AV. Let's consider the situation when $m = k = 10$. Then $X = \{x_1, x_2, \ldots, x_{100}\}$, where $x_1$ – AV$_1$ to take the flight level FL$_1$, $x_2$ – AV$_1$ to take the flight level FL$_2$, …, $x_{10}$ is AV$_1$ – to take the FL$_{10}$, $x_{11}$ – AV$_2$ to take the flight level FL$_1$, $x_{12}$ – AV$_2$ to take the flight level FL$_2$, $x_{20}$ – AV$_2$ to take the flight level FL$_{10}$, …, $x_{99}$ – AV$_{10}$ to take the flight level FL$_9$ и $x_{100}$ – AV$_{10}$ to take the flight level FL$_{10}$.

Figure 1a displays disturbed (non-standard) processes synchronization in the ATS corresponding to the result of the synchronizing sequence applying $p_1 = x_1\, x_{12}\, x_{24}\, x_{34}\, x_{44}\, x_{56}\, x_{68}\, x_{78}\, x_{88}\, x_{100}$ (wrong command to take AV$_3$ and AV$_5$ of the 4th flight level, AV$_7$ and AV$_9$ of the 8th flight level), which converts the system to the critical condition $s_{14}$, which means the near-midair collision of AV$_3$, AV$_4$ and AV$_5$, and also AV$_7$, AV$_8$ and AV$_9$.

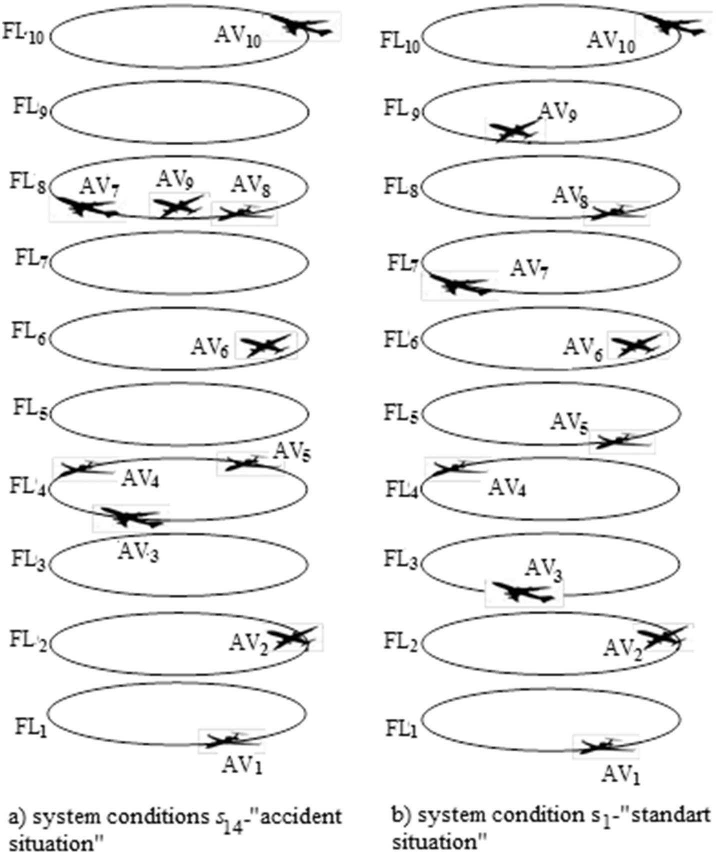
For this situation according to the Theorem 1 the synchronizing sequence $p \in X^*$ exists. So if all aerial vehicles are monotypic and have the same priority, then the weight of the sought sequence is equal to the symbol amount from $X$ to $p$.

According to the Theorem 4 let's formulate the task of the integer linear programming for this example.

It is required to minimize the performance function, which has a sense of the synchronizing sequence length $p$, which depends on the elements $x_i \in X$:

$$c_1x_1 + c_2x_2 + \ldots c_jx_j + \ldots c_nx_n \rightarrow \min \tag{15}$$

with restrictions



a) system conditions $s_{14}$-"accident situation"

b) system condition $s_1$-"standart situation"

**Fig. 1.** Disturbed (non-standard) and standard processes synchronization in the ATS

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + \ldots a_{1j}x_j + \ldots + a_{1n}x_n = a_1,\\
&a_{21}x_1 + a_{22}x_2 + \ldots a_{2j}x_j + \ldots + a_{2n}x_n = a_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,\\
&a_{i1}x_1 + a_{i2}x_2 + \ldots a_{ij}x_j + \ldots + a_{in}x_n = a_i,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,\\
&a_{m1}x_1 + a_{m2}x_2 + \ldots a_{mj}x_j + \ldots + a_{mn}x_n = a_m,\\
&x_j \geq 0,\ (j = 1, \ldots, n).
\end{aligned} \tag{16}$$

The applying of precise methods for the task solution (Branch and bound method, Gomory method) isn't always possible because of the timing. So for the solution of this task we use approximate approaches, which allow to get the solution in a time satisfying the requirements of the operating control.

Figure 1b displays the variant of conversion from the $s_{14}$ condition to the $s_1$ condition, which corresponds to the standard arrangement of all AVs by the holding area flight levels, which is provided with the synchronizing sequence applying $p = x_1\ x_{12}\ x_{23}\ x_{34}\ x_{45}\ x_{56}\ x_{67}\ x_{78}\ x_{89}\ x_{100}$.

## 5 Conclusion

We offered an approach providing the time and amount effective task solution of the providing and supporting the safety of the complex system functioning with forming the minimal length sequence of the control inputs applying to the system and converting it to the safety condition. Such sequence forming is based on the usage of the mathematics model in place of the finite automation.

This model can be used for the ATS staff training, and then it can be used for the operational control of aerial vehicles streams in real-life environment. The results of this work are used at JSC "IL" (Open Joint Stock Company "Ilyushin Aviation Complex") and can be used as the integrated part of the federal system of the flight safety providing and supporting.

## References

1. Kluev, V.V., Rezchikov, A.F., Bogomolov, A.S., Filimonyuk, LYu.: The conception of complex resource for research of «man – object – environment» systems' safety. Test. Diagn. **8**, 44–55 (2013)
2. Kluev, V.V., Rezchikov, A.F., Kushnikov, V.A., et al.: An analysis of critical situations caused by unfavorable concurrence of circumstances. Test. Diagn. **7**, 12–16 (2014)
3. Sloane, N.J.A.: An encyclopedia of integer sequences. SIAM Rev. **38**, 333–337 (1996)
4. Trahtman, A.N.: Modifying the upper bound on the length of minimal synchronizing word. In: Owe, O., Steffen, M., Telle, J.A. (eds.) FCT 2011. LNCS, vol. 6914, pp. 173–180. Springer, Heidelberg (2011). doi:10.1007/978-3-642-22953-4_15
5. Lapkovsky, R., Ivanov, A., Ivaschenko, V.: Cause-effect approach to traffic modeling on problem intervals of road network. Large Scale Syst. Control **35**, 283–303 (2011)
6. Neymark, M.S., Tsesarskiy, L.G., Filimonyuk, L.Y.: The decision-making model for an entrance of airplanes into an airport's responsibility zone. All-Russ. Sci.-Tech. J. "Polyot" ("Flight") **3**, 31–37 (2013)
7. Novozhilov, G.V., Rezchikov, A.F., Neymark, M.S., et al.: Human factor in aviation transport systems. All-Russ. Sci.-Tech. J. "Polyot" ("Flight") **5**, 3–10 (2013)
8. Novozhilov, G.V., Rezchikov, A.F., Neymark, M.S., et al.: Cause-and-effect approach to the analysis of aviation transport systems. All-Russ. Sci.-Tech. J. "Polyot" ("Flight") **7**, 3–8 (2011)
9. Speransky, D.V.: Generalized synchronization of linear sequential machines. Cybernetics **3**, 17–25 (1998)
10. Tsukanov, M.A.: Coordination of units process as difficult structured of manufacture as an economic measure. Large Scale Syst. Control **3**, 318–321 (2013)

11. Sharov, V.: Application of bayesian approach to update events' probabilities in automated system of aviation accidents forecasting and prevention. Large Scale Syst. Control **43**, 240–253 (2013)
12. Rosenblatt, D.: On the graphs and asymptotic forms of finite boolean relation matrices. Naval Res. Logistics Quart. **4**, 151 (1957)
13. Gill, A.: Linear Sequential Circuits: Analysis, Synthesis, and Applications. McGraw-Hill, New York (1967)

# Models of Minimizing the Damage from Atmospheric Emissions of Industrial Enterprises

Alexander Rezchikov[1], Vadim Kushnikov[1], Vladimir Ivashchenko[1], Aleksey Bogomolov[1], Tatyana Shulga[2], Nataliya Gulevich[2], Nataliya Frolova[2], Elena Pchelintseva[2], Elena Kushnikova[2], Konstantin Kachur[2], and Ekaterina Kulakova[2](✉)

[1] Institute of Precision Mechanics and Control of Russian Academy of Sciences,
Saratov, Russia
iptmuran@san.ru

[2] Yuri Gagarin State Technical University of Saratov, Saratov, Russia
olga@sstu.ru

**Abstract.** The article suggests mathematical models of reducing the damage from atmospheric emission of industrial enterprises when the characteristics of the state of the environment are undefined, which is caused by the character, the size and the time of appearance of the external perturbation actions.

**Keywords:** Industrial enterprises · Atmospheric emission · Models of minimizing the damage · Undefined characteristics of the state of the environment

## 1 Introduction

At the moment there are practically no works dedicated to the changing conditions of the functioning of industrial enterprises and aimed at minimizing various kinds of damage inflicted by their atmospheric emission on the adjacent region, its population, agriculture, environment and even on the enterprises themselves [1–4]. It is explained by the complexity of the task of evaluating the economical damage from the pollutants of the atmosphere when the characteristics of the state of the environment are undefined because of:

- Frequent changes in the structure and performance of the technological equipment of industrial enterprises and the difficulty of affecting the rhythm of its work
- Environmental perturbations continuously appearing in the process of functioning of industrial enterprises, the character, the size and the time of appearance of which are, as a rule, unknown.

In the article the statement of problem is offered that covers minimizing the damage from atmospheric emissions of industrial enterprises when the characteristics of the state of the environment are undefined. Also, mathematical models and algorithms of solving this problem are offered.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_27

## 2 Method of Solving the Problem

We have the following:

- $Cf_1$ is damage related to the growth of the population's sickness rate;
- $Cf_2$ is agricultural losses resulting from atmospheric emission;
- $Cf_3$ is losses related to environmental changes;
- $Cf_4$ is losses related to the degradation of the population's quality of life;
- $Cf_5$ is losses of an industrial enterprise caused by the turning off of technological equipment, reduction of its performance while tuning the size of the atmospheric emission as well as by paying fines for violating the legal limits on acceptable concentrations of poisonous substances;
- $S$ is fines imposed on an industrial enterprise for violating hygienic standards;
- $L$ is losses of an industrial enterprise caused by the turning off of the technological equipment or reducing its performance.

This problem is a problem of vector optimization, which is solved by means of compression of the vector criterion using weight factors but not assigning the absolute priority to any of them.

This is achieved by projecting the vector $Cf$ onto five chosen directions matching the vector of weight factors.

$$\mu = (\mu_1, \ldots, \mu_i, \ldots, \mu_5), \mu_i \geq 0, \sum\nolimits_{i=1}^{5} \mu_i = 1 \quad (1)$$

As a result, the target function after the conversion of the criteria will look as follows:

$$Cf_S = \sum\nolimits_{i=1}^{5} \mu_i Cf_i \quad (2)$$

The stated problem belongs to the class of variational problems, the solution of which can involve certain difficulties:

- The controlling actions may not belong to the class ofcontinuous functions
- The period of time, for which is problem is being solved, doesn't always allow determining the character of change in the vector of the environment state for the moment of the problem solution
- The solving of a system of non-linear differential equations of high number of dimensions requires significant consumption of time and computer's resources.

The above makes it feasible to use the method of piecewise-linear approximation. A quantization step equal to 3 to 4 months is used for this.

To solve the problem related to minimizing the damage from atmospheric emissions of industrial enterprises we use compression of the vector criterion using weight factors but not assigning the absolute priority to any of them. The choice of the scheme of component compression of the vector criterion can be affected by the external conditions existing at the moment of solving the problem and influencing the amount of the damage. For example, if we observe significant damage caused by the population's sick rate because of the influence of the atmospheric pollutants, the absolute priority should be assigned to the task of optimizing criterion $Cf_1$.

It is known that for many atmospheric pollutants the amount of damage from them can be represented as the following dependency:

$$Cf_i = Cf_i(C, t), i \in \{1, 2, \ldots, 5\} \tag{3}$$

where

$C$ is concentration of pollutants;

$t$ is the duration of their influence upon the objects and territories in checkpoints.

To calculate the concentration of pollutants $C$ at checkpoints we need a complex set of mathematical models to determine:

- The composition and the amount of the polluting substances contained in the atmospheric emissions of an industrial enterprise;
- The altitude at which the pollutants are located;
- The patterns of distribution of the polluting substances in the atmosphere when weather conditions are different.

In order to calculate the data describing various kinds of damage resulting from atmospheric pollution and to control the performance of the technological equipment of an enterprise a solution matrix is used. Let us examine the use of this matrix and take as an example the calculation of the amount of damage related to the growth of the sickness rate in the population caused by the atmospheric pollutants $Cf_1$.

This kind of damage consists of:

- damage $Y_1$, caused by short-timed negative influence of the atmospheric emission upon public health
- из ущерба $Y_2$, caused by the accumulation over a period of time of negative influences of the atmospheric emission upon public health.

To determine the amount of damage, we are using the matrix of solutions $||e_{ij}||^{Y_1}$ (Table 1).

**Table 1.** Matrix of solutions $||e_{ij}||^{Y_1}$ to determine the $Y_1$ part of the damage

| $F_j$ <br> $E_i$ | $F_1$ | … | $F_j$ | … | $F_m$ |
|---|---|---|---|---|---|
| $E_1$ | $e_{11}$ | … | $e_{13}$ | … | $e_{1m}$ |
| … | … | … | … | … | … |
| $E_i$ | $e_{i1}$ | … | $e_{ij}$ | … | $e_{im}$ |
| … | … | … | … | … | … |
| $E_n$ | $e_{n1}$ | … | $e_{nj}$ | … | $e_{nm}$ |

In the cells of Table 1 there are values if damage $e_{ij}$, $i \in \{1, 2, \ldots, n\}$, j $\in \{1, 2, \ldots, m\}$, which is caused by short-timed negative influence of the atmospheric emission upon public health and corresponds to different versions of external conditions and the decisions that are made. The calculation of the amount of damage is based on the evaluation made by a group of experts.

In order to minimize the total damage from atmospheric emissions $Cf_S = \sum_{i=1}^{5} \mu_i Cf_i$, minimax criterion (or Savage's criterion) is used. Using minimax criterion we exclude the possibility of a situation, in which the damage caused by atmospheric pollutants exceeds the pre-set value determined by the following expression:

$$Z_{MM} = \max e_{ir}, i = \overline{1, n}; \tag{4}$$

$$e_{ir} = min\, e_{ij}, j = \overline{1, m}$$

where $Z_{MM}$ is the value of the maximum damage inflicted by the atmospheric pollutants. This problem can have more than one solution.

The set of the problem's solutions looks like:

$$E_0 = \left\{E_{i0} | E_{i0} \in E : e_{i0} = \max_i \min_j e_{ij}\right\}, e^* = e_{i0} \tag{5}$$

When it is not necessary to set the limit to the damage caused by the atmospheric emission, other criteria can be used, for example, the Bayes-Laplace criterion.

Evaluation in points of the efficiency of one or another method of damage calculation (1–8 points) is given in Table 2.

**Table 2.** Area of applying methods of calculating damage from pollutants' influence

| Kind of damage | Method of calculation | Expert evaluations | Metric functions | Piecwise functions | S-shaped functions | Recipient methods | Combined methods |
|---|---|---|---|---|---|---|---|
| Damage caused by public sickness rate | | 1, 2 | 3–5 | 3, 4 | 1, 2 | 1, 2 | 6–8 |
| Agricultural losses | | 1, 2 | 3, 4 | 3, 4 | 3, 4 | 1, 2 | 6–8 |
| Damage caused by environmental changes | | 1, 2 | 2, 3 | 2, 3 | 3, 4 | 2, 3 | 6, 7 |
| Damage caused by the degradation of life quality | | 1, 2 | 3, 4 | 3, 4 | 3, 4 | 1, 2 | 6, 7 |
| Enterprise-related damage | | 1, 2 | 2, 3 | 3, 4 | 3, 4 | 3, 4 | 6–8 |

- 1 means effective application of the method.
- 2 means the method is used frequently.
- 3 means it is possible to use the method.

- 4 means the method is used as an auxiliary one.
- 5 means the method is used seldom.

In the last column of Table 2 the combined methods of damage calculation are given, for which the following marks are used:

- 6 means a method of controlled districts
- 7 means statistical methods
- 8 means a method combining controlled districts, statistical methods and artificial neural networks.

It is necessary to note that the above mentioned approach to the choice of the method of calculating the damage from the pollutants' influence doesn't allow taking into account of the full effect of the degree of change of the external conditions, for which the problem is being solved. In order to eliminate this drawback it is possible to use the corresponding indicator of the change of the external conditions, which is a continuous function, monotone on the [0; 1] depending on the changes of $F_1$, $F_2$, …, $F_\mathrm{m}$.

## 3 Software

The structure of the complex system of applications implementing the models of minimizing the damage from atmospheric emissions of an industrial enterprise is shown on Fig. 1.

Brief description of all subsystems is given below:

*The subsystem of information collection* provides for collecting, verification and storing in memory of the software system of the information necessary for the solution of the problem:

- The list of the hazardous substances emitted into the atmosphere while an industrial enterprise functions
- The list of the items of technological equipment whose work results in forming of hazardous substances
- The number of working shifts (days) the equipment has been functioning
- The laws of distribution of random values describing the times of turning the equipment on and off
- The height and the diameter of the industrial enterprise's chimney
- Weather conditions
- Etc…

*The subsystem responsible for transferring information* receives this information from the subsystem of collection. In order to transfer information from remote sources Internet technologies and web applications are used. Informational interaction between gauges and the database (DB) and knowledge base (KB) server is carried out over TCP/IP protocols and open connection channels of the Internet. The information transfer boils down to message exchange within the limits of the HTTP protocol using SSL. Such informational interaction simplifies data exchange and allows using wide-spread types of Internet connection and the means of interaction. The role of communication channels is assigned to *GPRS*, *WiMAX* and *ADSL* connections.

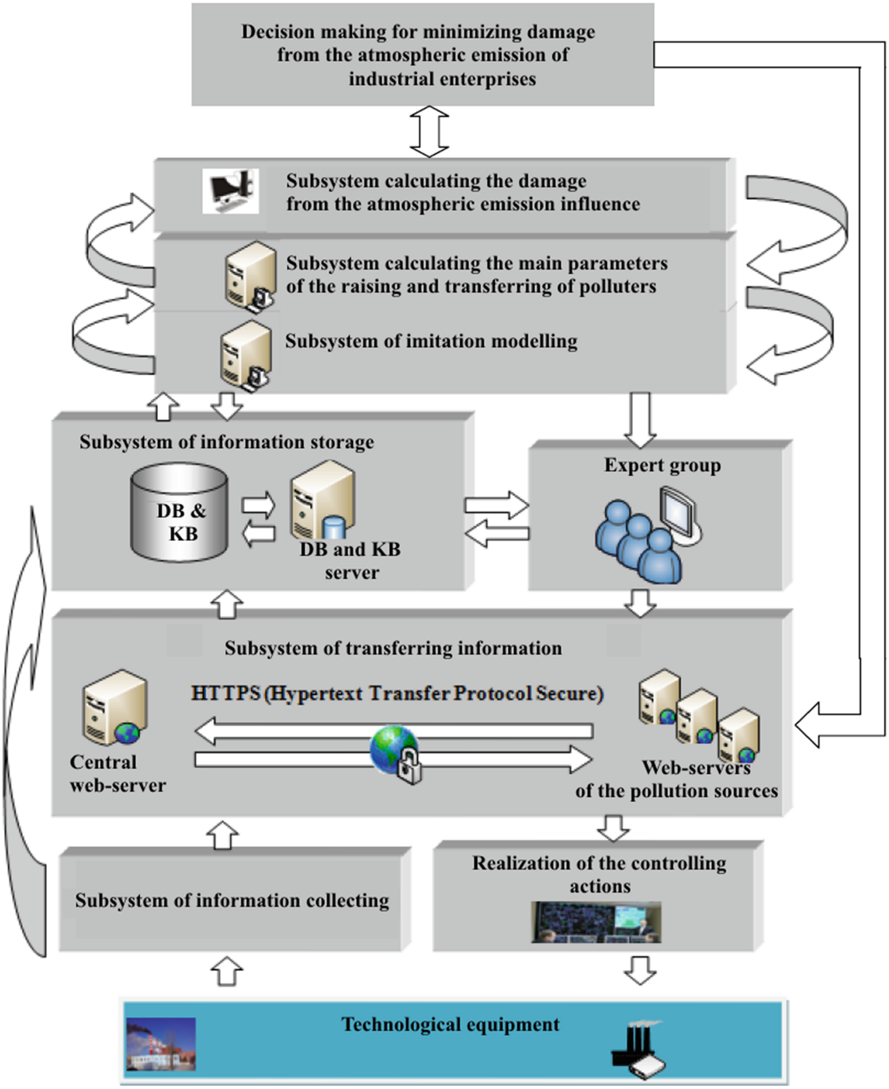

Decision making for minimizing damage from the atmospheric emission of industrial enterprises
Subsystem calculating the damage from the atmospheric emission influence
Subsystem calculating the main parameters of the raising and transferring of polluters
Subsystem of imitation modelling
Subsystem of information storage
DB & KB
DB and KB server
Expert group
Subsystem of transferring information
HTTPS (Hypertext Transfer Protocol Secure)
Central web-server
Web-servers of the pollution sources
Subsystem of information collecting
Realization of the controlling actions
Technological equipment

**Fig. 1.** Structure *of the automated system of control which realizes the software for* of minimizing the damage from atmospheric emissions of an industrial enterprise

*Subsystem of information storage*, which is used to solve the problem, has to provide access to the information for all problems being solved by the department of the chief ecologist of an industrial enterprise. As a database control system, *Microsoft SQL SERVER is used*. It guarantees safety and integrity of the data and provides for correct execution of inbound and outbound operations when a client accesses the data. Also, integration of the database system with a web server is achieved. The web server in turn organizes access of all servers of the pollution sources to the central server.

The results of the problem solution, which contain recommendations regarding changes of structure and performance of the technological equipment, are sent to the chief technologist of the enterprise for approval.

*Subsystem of imitation modelling* of the process of forming the pollutants during the work of technological equipment of an enterprise is based on the imitation modelling technique *AnyLogic 7*, which is implemented in the Java programming language using *Eclipse IDE.*

Calculation of the total amount of the emission of atmospheric pollutants by an industrial enterprise is done using well known mathematical models, which are also used to calculate the amount of emission of pollutants by separate groups of technological equipment of the enterprise.

*Subsystem of calculating the parameters of the process of raising and transferring the pollutants* allows calculation of the concentration of polluting substances on the monitored territories. To calculate the parameters of the process of raising and transferring the pollutants the built-in *MATLAB* programming language is used.

*Subsystem of calculating the damage from the atmospheric emission* carries out not only the calculation of different kinds of damage, but their minimization according to the minimax criterion or Savage's criterion depending on the preferences of the person who makes decisions (PMD).

Minimizing the damage from the atmospheric emission of industrial enterprises is done by the following algorithm:

- The damage is determined from the sickness rates of the population $Cf_1$, from the agricultural losses $Cf_2$, from the changes in the state of the environment $Cf_3$, from the degradation of the life quality $Cf_4$, as well as from the paying of ecological fines by the enterprise $Cf_5$;
- Selection of the compression coefficients

$$Cf_S = \sum\nolimits_{i=1}^{5} \mu_i Cf_i, \mu_i \geq 0, i \in \{1, 2, \ldots, 5\}, \sum\nolimits_{i=1}^{5} \mu_i = 1 \tag{6}$$

of the target function is done, as well as forming the scalar target function.

- Options for the decision making are formed $E_i$, $i \in \{1, 2, \ldots, n\}$ to reduce the performance of the technological equipment.
- External conditions $F_j$ are set, $j \in \{1, 2, \ldots, m\}$, which affect the amount of damage in the zone influenced by the atmospheric pollutants of an enterprise.
- The values of the amount of damage from the influence of pollutants are calculated: $e_{ij}$, $i \in \{1, 2, \ldots, n\}$, $j \in \{1, 2, \ldots, m\}$, based on the data from Table 2 and the solution matrix $\| e_{ij} \|$ is formed.
- Minimizing of the target function is carried out using the minimax criterion.
- The choice of the option is made for the solution of the problem of the reduction of the technological equipment performance aimed at controlling the concentration of pollutants at checkpoints.
- Messages are sent to the PMD about the best option for solving the problem, which contain the information on the achieved concentration of the pollutants at

checkpoints, the length of the time period and the required performance of the technological equipment.

*Subsystem of realization of controlling actions* realizes the recommendations on changing the structure and performance of the technological equipment of the enterprise, in the process of the exploitation of which atmospheric pollutants are formed. The dispatcher service of the enterprise is used for this.

## 4 Conclusion

The article offers a problem statement to minimize the damage from the atmospheric emission of industrial enterprises, which takes into account the undefined status of the parameters of the environment state. It also offers mathematical models and algorithms for its solution.

Minimizing the atmospheric emission of industrial enterprises allows us to reduce the damage inflicted upon the adjacent region, the population, the agriculture, the environment, as well as to increase the effectiveness of the functioning of the enterprises themselves by means of reducing ecological fines, preventing ecological crimes and offences.

At the moment the work is carried out to include the developed software that minimizes the damage of atmospheric emission of industrial enterprises into the automated management system of “SEPO-ZEM”, Ltd. Also, it is planned to use this software in “Liga”, JSC while developing new forward-looking systems of ecological monitoring.

## References

1. Kushnikova, E., Rezchikov, A., Ivaschenko, V., Filimonyuk, L.: Models of minimization of harm from the enterprises’ atmospheric pollution in a case of environment uncertainty. Ecol. Ind. Prod. **4**, 60–65 (2015)
2. Kushnikova, E., Rezchikov, A., Ivaschenko, V., Filimonyuk, L.: Models and algorithms of damage minimization from industrial pollution. Large-Scale Syst. Control **57**, 158–190 (2015)
3. Rezchikov, A., Dolinina, O., Kushnikov, V., Ivaschenko, et al.: The problem of a human factor in aviation transport systems. In: Proceedings of the 3rd International Conference on Computing, Technology and Engineering (ICCTE 2016), Singapore, pp. 16–20 (2016)
4. Rezchikov, A., Kushnikov, V., Ivaschenko, V., Bogomolov, A., Filimonyuk, L., Kachur, K.: Control of the air transportation system with flight safety as the criterion. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Silhavy, P., Prokopova, Z. (eds.) Automation Control Theory Perspectives in Intelligent Systems. AISC, vol. 466, pp. 423–432. Springer, Cham (2016). doi:10.1007/978-3-319-33389-2_40

# Fairness and Load Balancing Optimization via Association Control in Multi-rate WLANs

Jianjun Lei(✉), Shanshan Yang, and Chang Su

School of Computer Science and Technology,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, China
{leijj, changsu}@cqupt.edu.cn, yangss@yahoo.com

**Abstract.** To resolve the problems of load imbalance and performance anomaly which lead to the increase of packet transmission delay, the waste of limited bandwidth resources and the decline of network performance, a categorized AP association algorithm based on the demands of users is proposed in this paper, which uses AP (Access Point) association control to achieve fairness and load balancing and can support various traffic businesses. Furthermore, a dynamic AP handoff based on load-balancing algorithm is presented to cope with the clients of gusty traffic, which can balance the network load with minimal AP handoff. Analysis and simulation results show that the proposed scheme obtains significant performance in terms of AP utilization, and it improves system throughput by 29.8% compared with the SSF algorithm in terms of throughput under different user densities.

**Keywords:** WLANs · Load balancing · Demands of users · AP association · Handoff

## 1 Introduction

Recently, the density of IEEE 802.11 based wireless local area networks (WLANs) deployments has increased dramatically. In the IEEE 802.11 standards, the conventional AP association scheme based on Received Signal Strength Indicator (RSSI) may lead the stations (STAs) to make associations with congested APs, while leaving adjacent APs to carry very light load or even to be idle. Load imbalancing could result in the decline of network performance or even network congestion. So load balancing has been the major objective of AP association strategy. However, it is well-known that the MAC protocol of 802.11 provides equal long-term transmission opportunities to all users associated with the same AP [1]. In multi-rate WLANs, throughput-based fairness requires that clients with lower rates occupy the channel for a longer time than those do with higher rates, which leads to the unfairness of accessing the channel among clients and reduces the aggregated throughout. To solve these problems, the AP association strategy should be improved according to the transmission rate of STA, so as to increase the system throughout and realize the proportional fairness in the STA initial

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_28

access of the network. In addition, with the rapid development of WLANs and the increase of the demands of users for communication, the resulting WLAN Quality of Service (QoS) [2] problem is increasingly outstanding. QoS is regarded as the integrated embodiment of business performance, and it decides the users' satisfaction of different businesses. QoS management must ensure that each client can get an acceptable end-to-end delay and an affordable maximal jitter delay.

Therefore, this paper focuses on the problems of AP association in multi-rate 802.11 WLANs. To achieve proportional fairness and load balance, a categorized AP association based on the demands of users algorithm is proposed, which brings a significant throughput increase, especially in hotspot distribution. And a dynamic AP handoff based on load-balancing algorithm is presented to cope with the clients of gusty traffic. It can balance the network load with minimal AP handoff.

The remainder of the paper is organized as follows. The related work is presented in Sect. 2. In Sect. 3, the network model is described and the two algorithms are detailed. Section 4 gives the performance evaluations. Section 5 is the conclusion of this paper.

## 2 Related Work

Improving the fairness and load balancing via AP association control has been extensively studied in many literatures. The conventional AP association scheme, by default RSSI-based method, may lead to unfairness among clients and imbalanced load among APs. The load balancing strategies for WLANs are conceived in many literatures, and various metrics also are proposed to determine the load of AP, rather than the default Strong Signal First (SSF) scheme.

Bejerano et al. [3] have proved that strong correlation exists between load balancing and fairness in AP association, which indicates the other one can be easily achieved if one of them has been achieved. Then, by utilizing load balancing techniques, it can realize the approximate optimal fair bandwidth allocation. Wendong et al. [4] proposes an optimal AP association scheme, in which throughput and energy consumption are both considered to indicate the AP load. The estimated file download time for web browsing is considered to indicate the AP load in [5]. In [6], the differentiated access service is selected according to the users' running applications. The effect of hidden terminals and the traffic intensity of clients are both considered as the main factors to impact the AP load in AP association scheme In [7]. In [8], according to some innovative metrics to estimate AP load, such as RSSI, transmission rate and throughput, a distributed adaptive load balancing algorithm is presented for multi-rate WLANs, in which the throughput of newly arriving users and the negative influence of the throughput of users associating with APs are both considered. In addition, load balancing AP association strategy is investigated in the light of game theory with local information in [9]. However, it could be challenging to apply these strategies in realistic environment, since most of them require some changes to clients, which is not viable for different operators. Thus, a multi-constraint load balancing based on cell breathing which requires no change to clients has been proposed in [10]. It solves the tradeoff between data power loss of users and load balancing among APs, which has been not cared about by most of existing load balancing schemes based on cell breathing. In [11], a time-based fair AP algorithm has been proposed to jointly consider AP association and power control for proportional fairness and aggregated throughput in multi-rate WLANs. But this scheme formulates AP association as a NP-hard problem, which needs multiple iterations to find the appropriate optimal solution, and each iteration process would trigger the power adjustment. It is easy to lead the clients to switch among APs frequently. An on-line AP association algorithm is presented for 802.11n WLANs in [12], in which the impact of legacy clients is considered. It first gives a bi-dimensional Markov model to estimate the throughput of clients from uplink and downlink, and then formulates AP selection as an optimization problem, aiming at achieving proportional fairness among clients. But in this algorithm, each client must obtain timely information from all clients associated with the nearby AP, which results in high time complexity. For high density WLANs, another on-line AP association algorithm, namely Categorized algorithm (Categorized) with lower complexity is proposed in [13], which classifies APs by types of clients associating with them. A newly arriving client will give priority to associating with the AP with the same type to ensure achieving proportional fairness. It can solve the performance anomaly caused by heterogeneous clients, but load imbalance may emerge.

Most of the above AP association schemes concentrate on load balance among APs to a certain extent, but the bandwidth demand of users is usually ignored, which is a key factor that influences the load balance. In this paper, the categorized AP association algorithm based on the demands of users is proposed, which introduces a novel metrics of AP load such as the aggregated transmission time demands of the clients associating with it, and classifies APs based on the types of different associating clients. The clients will give preference to the AP with the same type and the least allocated transmission time. This algorithm effectively minimizes the impact of performance anomaly by associating different types of clients with different APs, and achieves load balance among APs. In addition, in order to cope with AP congestions caused by the clients with gusty traffic, the dynamic AP handoff based on load-balancing algorithm is presented, which improves the utilization of channel effectively and achieves load balancing according to clients' application priority and quantity demanded.

# 3 Algorithm Description

## 3.1 Network Model

We consider an IEEE 802.11-based multi-rate WLAN consisting of N mobile clients and M access points. Let set A and set S respectively denote the set of APs and the set of STAs. Each AP serves only STAs that reside in its coverage area. At any given time, a STA can be allowed to choose one and only one AP to associate with whereas each AP can serve multiple STAs simultaneously. The coefficient $x_{ij}$ is intended to indicate the association relationship between STA $i$ and AP j, which is defined as follows.

$$x_{ij} = \begin{cases} 1, & \textit{if STA i is associated to AP j}, \\ 0, & \textit{otherwise}. \end{cases} \tag{1}$$

Where, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$.

**Definition 1 (Demands of Users).** The demands of users specify the transmission time demanded. The transmission time demanded by user $i$ is the length of time to transmit the data that the user requires from AP j. It is formulated as follows.

$$T_i = \frac{\mathrm{S}_i}{r_i} = \frac{B_i}{r_i} T. \tag{2}$$

Where, T is the total allocated transmission time of AP j, which satisfies $\mathrm{T} = \sum_{i=1}^{\mathrm{N}} x_{ij} T_i$, $T_i$ is the transmission time demanded by user $i$ after the association is made; $s_i$ is the size of data that user $i$ requires; $r_i$ is the transmission rate of user $i$ after the association is made; $B_i$ is the size of data required by user $i$ during per unit of time.

**Definition 2 (Priorities of Businesses).** Since different application businesses have different demands of service, users are divided into four classes according to their application businesses. Each class determines priority in accordance with its demand for bandwidth and delay. The priorities for voice stream (class-1), video stream (class-2), best-effort stream (class-3), and background stream (class-4) are from high to low respectively.

### 3.2 Problem Description

**Definition 3 (Performance Anomaly).** It refers to the effect that, when links operating at different rates coexist with an AP, the throughput of high rate link will all degrade to the level of the lowest rate link [7]. Thus the aggregated throughput also will decline.

**Definition 4 (Proportional Fairness).** It is designed to provide fair service to users in multi-rate WLANs and solve performance anomaly. It adopts the method of Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) to communicate among clients in WLANs, in which each user has the same probability to access the channel. Thus as the density of users increases, the conflicts caused by contending for the channel among users will also increase, which makes the most of bandwidth be wasted at packet retransmissions, and then degrades system throughput. So to improve the network performance in multi-rate WLANs, it is necessary that each user be allocated with bandwidth in a proportional-fairness manner so that the allocated bandwidth could be proportional to its transmission rate.

**Definition 5 (Load Balance).** In the IEEE 802.11 standard, the conventional AP association scheme based on the received signal strength indicator (RSSI) may lead STAs to make associations with congested APs, while leaving adjacent APs to carry very light load or even to be idle, which leads to two problems: for the network, it may degrade the aggregate system throughput; from the users' perspective, the packet transmission time may be prolonged [3]. These two problems can cause significant bandwidth waste. The load of AP j is denoted by $\rho_j$, which is equal to the allocated transmission time of AP j. The load variance is expressed by the following equation.

$$\delta^2 = \frac{\sum_{j=1}^{M}\left(\rho_j - \bar{\rho}\right)^2}{M} \tag{3}$$

Where, M is the number of APs, and $\bar{\rho}$ is the average load, where $\bar{\rho} = \frac{\sum_{j=1}^{M}\rho_j}{M}$. When the load variance is minimized, the network can achieve load balancing.

$$\begin{aligned} min\ \delta^2 &= \frac{\sum_{j=1}^{M}\left(\rho_j - \bar{\rho}\right)^2}{M}, \\ s.t.\ \bar{\rho} &= \sum_{j=1}^{M}\rho_j/M. \end{aligned} \tag{4}$$

In the actual environment, the performance of each AP can be different, thus the maximum load sustained by an AP is different. We can give an approximate minimum load variance algorithm to achieve load balancing, which is the proposed load-balancing algorithm based on dynamic AP handoff.

### 3.3 AP Association Algorithm

In this section, the categorized AP association algorithm based on the demands of users is proposed, which can achieve proportional fairness and load balance among APs. It can be implemented by appending a category field and a popular transmission rate field in the Beacon frame from APs. The category field is decided by the popular transmission rate of the current AP. The popular transmission rate specifies the data rate used by the most of associated users of the same AP.

The pseudo code of AP association algorithm is given as follows.

**Algorithm 1 The Categorized AP Association Based on the Demands of Users Algorithm**

```
Input: Set of APs A, Set of STAs N, STA Type Vector $Type = \{t_i \mid \forall i \in N\}$.
Output: AP association matrix $X = \{x_{i,a} \mid \forall i \in N, a \in A\}$
Initialize $x_{i,a} = 0, \forall i \in N, a \in A$
for each AP $a \in A$
  set the category of AP a $c_a$ to zero;
end for
for each STA $i \in N$
  for each AP $a \in A$
    if STA$i$ receives beacon frames from AP$a$
      Get the category of AP a $c_a$ and the popular data rate among STA of AP$a$;
      Add AP$a$ into subset $A_i$;
    end if
  end for
  if AP$j \in$subset $A_i$ and$c_j = t_i$
    $T_j = arg \min_j \left( \sum_{k=1}^{i-1} x_{kj} T_{kj} + T_{ij} \right)$
    Set $x_{ij}$ to 1
  else if AP$j \in$subset $A_i$and $c_j = 0$
    $T_j = arg \min_j \left( \sum_{k=1}^{i-1} x_{kj} T_{kj} + T_{ij} \right)$
    Set $x_{ij}$ to 1,, and set $c_j$ to$t_i$
  else if$\nexists$subset $A_i$and $c_j = 0$
    $T_j = arg \min_j \left( \sum_{k=1}^{i-1} x_{kj} T_{kj} + T_{ij} \right)$
    Set $x_{ij}$ to 1
  end if
end for
```

Initially, the categories of all APs are set to be zero. When a new user joins the network and before making an association decision, it firstly acquires a list of candidate APs according to beacon frames of nearby APs. From the list of candidate APs, the user will choose the APs with the same category of its type. And then from these APs, the user associates with the AP whose allocated transmission time is the least. If there is no such an AP, the user will associate with the AP whose category is zero and allocated transmission time is the least. If there is no AP whose category is zero, the user will give preference to choosing the AP with the least allocated transmission time to associate with.

Note that we classify APs by the popular transmission rate among all the associated users of the same AP, which makes the transmission rate of all the associated users close to each other. The load of AP is the allocated transmission time, which is the sum of transmission time of all users associated with the AP, including the newly arriving user. The allocated transmission time is regarded as the metrics of AP load. Thus it can achieve proportional fairness among the clients of the same AP, and effectively solve the problem of performance anomaly.

### 3.4 Dynamic Handoff Algorithm

To ensure the QoS of user businesses, a dynamic AP handoff based on load-balancing algorithm is presented to maximize the AP utilization and the entire system capacity of the network, which also can achieve the global load balance of the network.

The pseudo code of dynamic handoff algorithm is given as follows.

**Algorithm 2 . The Dynamic AP Handoff based on Load-Balancing Algorithm**

```
Input: Set of APs  A, Set of associated STAs with AP$j$ $A_j$
, Set of AP load $\rho_j$, Set of the AP load threshold $\rho_{jmax}$.
Output: AP association matrix $X_a = \{x_{i,a} \mid \forall i \in N, a \in A\}$,
$X_j = \{x_{i,j} \mid \forall i \in N, j \in A\}$.
if $APj \in A$ and $\rho_j > \rho_{jmax}$ do
    Sort the users by their business priority in the
    descending order;
    With the same priority, sort the users by their
    candidate APs numbers n in the descending order;
    Calculate $\bar{\rho} = \frac{\sum_{j=1}^{M} \rho_j}{M}$;
```

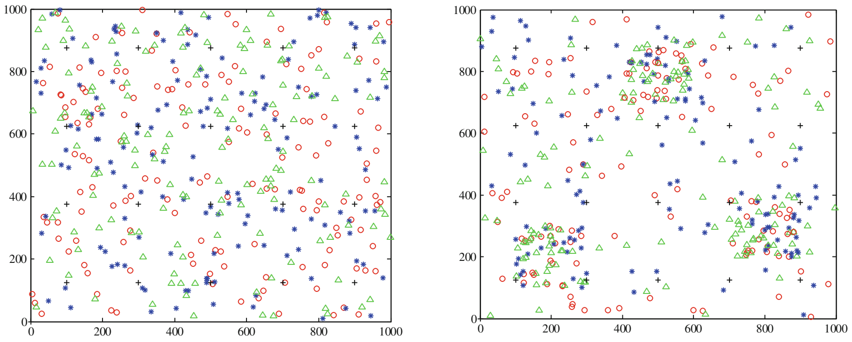
```
  for STA $k$ in the sorted users(STA $k \in A_j$)and $n_k \geq 2$
    Using algorithm 1 to choose a better AP $a$;
    if $a \neq j$ and $\rho_a < \rho_{amax}$ and $\rho_a < \bar{\rho}$ do
      Disassociate STA $k$ from AP $j$ and reassociate to
       AP $a$;
      Update $\rho_j$ for AP $j$;
      Update $\rho_a$ for AP $a$;
    else k+=1;
    end if
    if $\rho_j < \rho_{jmax}$ do
      break;
    else
      k+=1;
    end if
  end for
end if
```

According to the different performance of the APs, a load threshold is set for each AP. The real-time AP load condition is monitored by Access Controller (AC). Once the load of an AP exceeds its load threshold, the load balancing control algorithm will be triggered. The load migration objective ought to be the STAs with the lower priority in the overlapping area. Then, the STAs find a better AP with light load for them to associate with according to Algorithm 1. And then the algorithm enters into the next loop until it achieves load balancing. This algorithm can easily decrease the network load jittering caused by AP handoff, reduce the times of handoff and avoid the ping-pong effect.

## 4 Performance Evaluation

### 4.1 Evaluation Methodologies

In this section, we compare our algorithm with the Strong Signal First (SSF) that is the default AP association scheme in the 802.11 standards and Categorized Algorithm in [12] via OPNET 14.5 simulations. The performance is evaluated in terms of system throughput, delay and fairness among clients. To get closer to the real network environment, two kinds of network simulation scenarios are adopted, namely (a) uniform distribution in which all clients are randomly distributed and (b) hotspot distribution in which most of the clients are deployed in a few circle-shaped hotspot areas, which are shown in Fig. 1.



(a) Uniform distribution

(b) Hotspot distribution

**Fig. 1.** Network simulation scenarios (the different shapes represent 802.11b and 802.11g clients with different rates and the black plus signs denote APs).

The network topology contains a server, a switch, 20 APs and multiple mobile users. To simulate the real network environment, the simulation scenario is designed in the practical situation that the density of users is dynamically changing. To compare the performance of the three algorithms in WLAN in scenarios of different scales, the number of clients ranges from 40 to 200 at the interval of 20. Under different network scales, each algorithm simulation runs 30 times, and then obtains the average performance evaluation standard for comparison. The 20 APs are deployed uniformly in $1000 \times 1000\,\mathrm{m}^2$. The overlapping area exists between two adjacent APs. The transmission rate of the clients is obtained by utilizing the relationship between the transmission rate $r_{ij}$ and the Signal to Interference plus Noise Ratio (SINR) $\gamma_{ij}$, which is shown in Table 1 [13].

**Table 1.** The relationship between the effective bit rates and the SINRs in IEEE 802.11 standard

| $\gamma_{ij}(dB)$ | 6–7.8 | 7.8–9 | 9–10.8 | 10.8–17 | 17–18.8 | 18.8–24 | 24–24.6 | 24.6 |
|---|---|---|---|---|---|---|---|---|
| $r_{ij}(Mbps)$ | 6 | 9 | 12 | 18 | 24 | 36 | 48 | 54 |

The performance evaluation metrics includes system throughput, delay and the fairness index. The fairness index that ranges from 0 to 1 and reflects the user fairness in time is defined as follows:

$$\beta = \frac{\left(\sum_{i=1}^{M} y_i\right)^2}{M\left(\sum_{i=1}^{M} y_i^2\right)}, \beta \varepsilon [0, 1]. \tag{5}$$

Where, $y_i$ is the allocated transmission time.

## 4.2 Evaluation Results

### 4.2.1 Throughput

As Figs. 2 and 3 show, with the growth of user scale, the tendencies that denote the aggregated throughput of all algorithms are plotted under two network scenarios respectively. With the growth of user scale, we can notice that the aggregated throughput of SSF algorithm is much less than those of the other two algorithms. When the number of clients grows beyond 120 in uniform distribution and beyond 80 in hotspot distribution, our algorithm can achieve higher throughput than SSF and Categorized algorithms can do, which can be due to the impact of load imbalance among APs in SSF and Categorized algorithm. However, it is alleviated in our algorithm. As SSF and Categorized algorithms only make the user associate with the AP with the highest signal strength, a number of users may gather on the same AP. When the number of users is 200, the aggregated throughputs of our algorithm in uniform and hotspot distributions are respectively 10% and 21.6% higher than that of Categorized algorithm. As Fig. 3 shows, when the number of clients is beyond 120, the aggregated throughput of the Categorized algorithm starts to decrease, for the AP load is not considered when making association decisions in Categorized algorithm, thus resulting in APs overload in hotspots. But our algorithm cooperating with the dynamic handoff algorithm can make an appropriate handoff decision when the AP is overload.

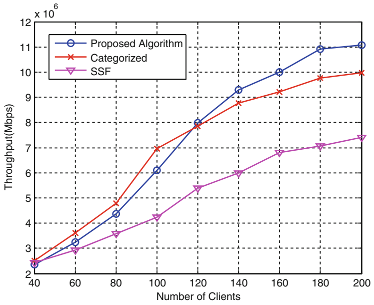


**Fig. 2.** Aggregated throughput in uniform distribution

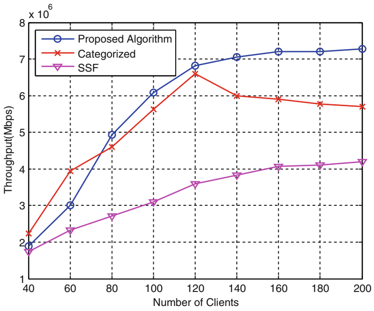


**Fig. 3.** Aggregated throughput in hotspot distribution

### 4.2.2 Delay

Furthermore, we examine the average delay of our algorithm in uniform and hotspot distributions, as Figs. 4 and 5 show. It can be observed that the delay increases along with the increase of the number of users. The reason is that as the user density increases, the contention will get more intense. Under uniform distribution, the delay of SSF is the shortest. This is mainly because each user gives preference to associating with the AP that has the strongest signal and thus utilities a high transmission rate in uniform distribution. In addition, the delay of our algorithm is the shortest under hotspot distribution. The reason is that our algorithm avoids the clients gathering, and makes full use of APs, thus alleviating the conflict among users. For SSF and Categorized algorithms, some APs are overloaded and the others are idle or light-loaded.

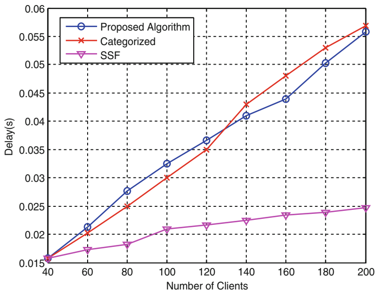


**Fig. 4.** Delay in uniform distribution

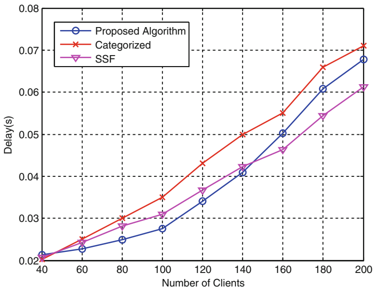


**Fig. 5.** Delay in hotspot distribution

The other reason is that our algorithm can redistribute the transmission time of the overloaded APs to the idle or light-loaded APs when gusty traffic happen to occur in some users or some APs happen to be overloaded.

### 4.2.3 Fairness Index

Finally, we evaluate the fairness index of each user for all algorithms when the user density increases under the two distributions. The fairness index of our algorithm is higher than those of the other algorithms regardless of the client distribution and density. The fairness index of our algorithm is above 0.7, the fairness indexes of SSF and Categorized algorithms are below 0.6 and 0.7 respectively, under both uniform and hotspot distributions. Under uniform distribution, it can be noted that the three algorithms change with the similar tendency when the number of users grows. And in the hotspot distribution case, the Categorized algorithm happens to decline when the number of clients grows beyond 140. It is reasonable since load balancing among APs is ignored by SSF and Categorized algorithm. In other words, when the user density is high, the most of users gather on a few of APs, which leads the rest of APs to be underutilized. It illustrates that our algorithm obviously outperforms SSF and Categorized algorithms in terms of fairness index (Figs. 6 and 7).

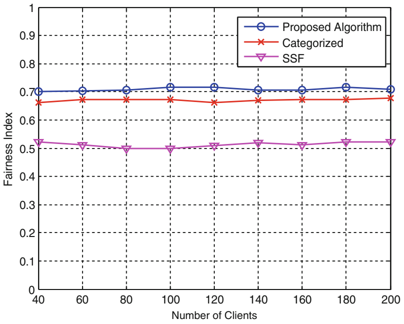


**Fig. 6.** Fairness index in uniform distribution

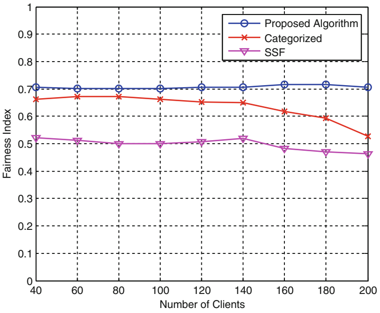


**Fig. 7.** Fairness index in hotspot distribution

## 5 Conclusion

In this paper, the AP association schemes for multi-rate IEEE 802.11 based WLANs have been studied. We first formulate the AP association problem as an optimization problem. To achieve proportional fairness and load balance, we propose the AP association algorithm which brings about a significant throughput increase, especially in hotspot distribution. And then we present the dynamic handoff algorithm to cope with AP congestions caused by the clients with gusty traffic. Finally, we conduct simulations to confirm that the proposed AP association algorithm cooperating with the dynamic handoff algorithm can achieve fairness among clients and load balancing among APs.

**Acknowledgments.** This research is supported by Program for Innovation Team Building at Institutions of Higher Education in Chongqing (CXTDX201601021), Fundamental and Advanced Research Project of Chongqing Municipal (cstc2016jcyjA2213) and Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJ1500439).

## References

1. Tan, G., Guttag, J.V.: Time-based Fairness improves performance in multi-rate WLANs. In: General Track, 2004 Usenix Technical Conference, pp. 187–195(2004)
2. Chen, Z., Xiong, Q., Liu, Y., Huang, C.: A strategy for differentiated access service selection based on application in WLANs. In: INFOCOM-IEEE Conference on Computer Communications Workshops, pp. 317–322 (2014)
3. Bejerano, Y., Han, S.J., Li, L.: Fairness and load balancing in wireless LANs using association control. In: IEEE/ACM Transactions on Networking, pp. 560–573 (2007)
4. Ge, W., Ji, H., Leung, V.C.M.: Access point selection for WLANs with cognitive radio: a restless bandit approach. In: IEEE International Conference on Communications, pp. 1–5 (2011)
5. Pradeepa, B.K., Kuri, J.: An estimated delay based association policy for web browsing in a multirate WLAN. IEEE Trans. Netw. Serv. Manage. **9**(3), 346–358 (2012)
6. Keranidis, S., Korakis, T., Koutsopoulos, I.: Contention and traffic load-aware association in IEEE 802.11 WLANs: algorithms and implementation. In: International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, pp. 334–341 (2011)
7. Gong, H., Nahm, K., Kim, J.W.: Distributed fair access point selection for multi-rate IEEE 802.11 WLANs. In: Proceedings of the 5th IEEE Consumer Communications and Networking Conference, pp. 528–532 (2008)
8. Chen, X., Yuan, W., Cheng, W.: Access point selection under QoS requirements in variable channel-width WLANs. IEEE Wirel. Commun. Lett. **2**(1), 114–117 (2013)
9. Wang, S., Cui, Y., Xu, K.: Multi-constraint load balancing based on cell breathing in WLAN. J. Chin. J. Comput. **32**(10), 1947–1956 (2009)
10. Li, W., Cui, Y., Cheng, X.: Achieving proportional fairness via AP power control in multi-rate WLANs. IEEE Trans. Wirel. Commun. **10**(11), 3784–3792 (2011)
11. Gong, D., Yang, Y.: AP association in 802.11n WLANs with heterogeneous clients. In: 2012 Proceedings of IEEE INFOCOM, pp. 1440–1448, 25–30 March 2012

12. Gong, D., Yang, Y.: On-line AP association algorithms for 802.11n WLANs with heterogeneous clients. J IEEE Trans. Comput. **63**(11), 772–786 (2014)
13. Cui, Y., Li, W., Cheng, X.: Partially overlapping channel assignment based on node orthogonality for 802.11 wireless networks. In: IEEE INFOCOM, pp. 361–365, 10–15 April 2011

# Multi-stage Optimization Over Extracted Feature for Detection and Classification of Breast Cancer

S.J. Sushma[1(✉)] and S.C. Prasanna Kumar[2]

[1] Instrumentation Technology,
Visvesvaraya Technological University, Belagavi, India
sushjgowda@gmail.com
[2] Department of Instrumentation Technology, RVCE, Bengaluru, India

**Abstract.** Although, there are various forms of medical image processing techniques for identification of breast cancer, but majority of the technique are either expensive, or time dependent, or doesn't produce accurate outcomes. We reviewed the recent techniques of breast cancer detection to find that the sole conclusion of presence or absence of disease depends on the skills of a radiologist. The present manuscript introduces a very simple modeling of breast cancer detection followed by multiple level of optimization carried out towards its extracted feature. A transform-based technique is used for feature extraction which is further optimized using particle swarm optimization for precise detection of cancerous tissues within a mammogram. Finally, we use a set of simplified fuzzy rules in order to identify the type of the cancer. The presented system offers faster response time with negligible computational complexity.

**Keywords:** Breast cancer detection · Mammogram · Classification · Optimization · Benign · Malignant

## 1 Introduction

The usage and adaption of radiology in the form of medical image processing has significantly contributed in identifying various forms of complicated diseases from more than a decade [1, 2]. In this regards, this paper pivots around the breast cancer. There are various techniques in order to diagnose breast cancer e.g. mammogram, molecular breast imaging, ultrasound, magnetic resonance imaging, etc. [3]. One of the biggest problems is to perform detection of cancer in its early state [4]. Another significant problem is that existing system doesn't confirm the presence of cancer to 100% accuracy [5]. It is only the skill of a professional doctor to provide inference of the correct condition of the disease [6]. At present, there has been various review being carried out towards techniques of breast cancer [7–9]. Review of existing techniques towards breast cancer concludes that (i) there is a less specificity for mammogram, (ii) diversified ability to interpret mammograms by doctors, (iii) similar density of tissue, and (iv) incorrect positioning of the subject, etc. Apart from all these, there is still a problem of possible exposure to radioactive rays in all these techniques.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_29

Various forms of nuclear imaging techniques e.g. Positron Emission Tomography and Sestamibi imaging make use of administered contrast agents as well as radioactive tracers. Such techniques are quite expensive in nature and are out of reach of common man. Apart from this, the most recent techniques are found to less prioritize the detection techniques and more on image capturing techniques using antenna or radar based techniques.

Hence, this paper presents a novel optimization technique towards leveraging the detection and classification process involved in breast cancer detection. Section 2 discusses about the existing research work followed by problem identification in Sect. 3. Section 4 discusses about proposed methodology followed by elaborated discussion of algorithm implementation in Sect. 5. Comparative analysis of accomplished result is discussed under Sect. 6 followed by conclusion in Sect. 7.

## 2 Related Work

This section discusses about the significant research work being carried out towards breast cancer detection. Our prior review has already discussed certain research contribution and their effectiveness [10]. Kwon et al. [11] have presented a technique about image enhancement pertaining to temporal domain using Gaussian filtering process. Singh et al. [12] introduces wavelet as well as neural network based technique in order to enhance the classification performance by eliminating the human-based errors. Santorelli et al. [13] have presented a technique that uses microwave systems for identifying the presence of breast cancer. The authors have used experimental-based approach with wideband antenna. Li et al. [14] have used decomposition mechanism using imaging modalities of ultra-wide band for realizing the reconstructed image pertaining to breast cancer. Reference waveforms were used for identifying the tumor over the breast image. Yin et al. [15] have formulated a correlation-based mechanism for mitigating the artifact effect. Song et al. [16] have used experimental approach by considering integrated circuits in order to design radar for effective breast cancer detection. Usage of thermo acoustic imaging was used in the study of Wang et al. [17]. The detection mechanism was carried out using contrast agents. Bahrami et al. [18] have presented a similar technique of microwave antenna design for detection of breast cancer. Usage of ultrasound for similar purpose was seen in the work carried out by Denis et al. [19]. The technique uses region of interest over the shear elastography for quantizing the stiffness characteristics of breast tissue. It also uses Young's modulus for discretizing the forms of cancer (i.e. classification). George et al. [20] have used a set of optimization technique for performing an effective classification system. Experimental analysis based study was adopted by Bolado et al. [21] for investigating the effect of compression in identifying the breast cancer. Similar form of research work was also carried out by Guardiola et al. [21] considering microwave imaging. The technique is known for its generation of images with higher extent of information. The next section discusses about the problems being identified by the existing research work.

## 3 Problem Description

From the previous section, it can be seen that recent research work towards detection of breast cancer is mainly inclined towards microwave imaging. However, microwave imaging system also incorporates a significant trade-off between penetration of wave and the resolution of the signal being generated. A closer look into the recent techniques will show that less work is carried out towards detection mechanism and very few for classification techniques. Lesser computational modeling and more experimental modeling with breast phantoms also reduces the scope to understand the implication in real-time environment. Therefore, there is a need of an efficient computational model to perform detection and classification of breast cancer cost effectively. The next section discusses about the proposed methodology to overcome such issues.

## 4 Proposed Methodology

The present work is a continuation of our prior work [22] where the emphasis was on suitably enhancing mammograms for cancer detection. The prime aim of the proposed system is to develop a simple model that has multiple-levels of optimization for assisting in breast cancer detection and classification. The schematic diagram of the proposed system is shown in Fig. 1.

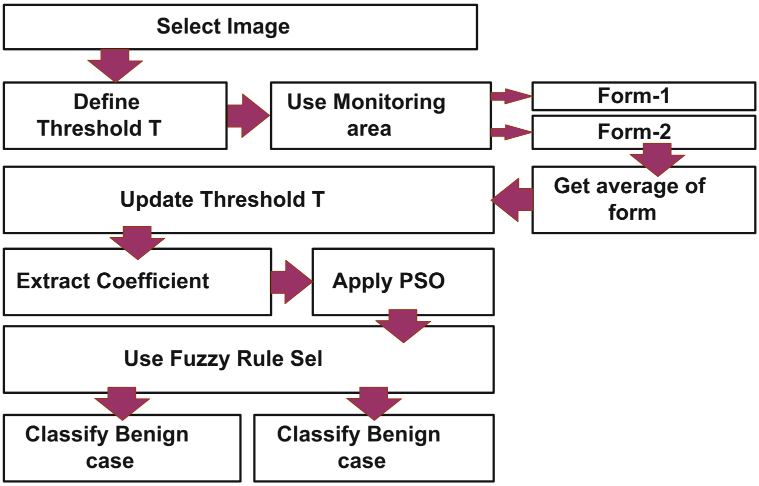


**Fig. 1.** Schematic architecture of proposed system

The proposed system mainly emphasizes on precise feature extraction and this process if further leveraged by discrete wavelet transform and particle swarm optimization. The technique also makes use of fuzzy rule set for performing classification of the mammogram. Algorithm responsible for this is discussed in next section.

## 5 Algorithm Implementation

This section discusses about an algorithm that is responsible for identifying the suspected mass within a breast image and then further performs classification of the type of cancer. The input of the algorithm is basically a grey scale image (*I*) which after processing will lead to outcome image (*out*) with an inference of type of the cancer. The steps of the algorithm are discussed below:

```
Algorithm for Identifying and Classifying Breast Cancer
Input:I (Input Image)
Output:out (image detected with benign/malignancy)
Start
1. init I(x,y)
2. $\phi_{x,y} = \frac{\phi^2(x,y).255}{\arg_{\max}[\phi^2(x,y)]}$
3. $M_a(T) \rightarrow [f_1\ f_2]$
4. Update T as $(f_1+f_2)/2$
5. Apply 2D DWT
6. $g(t) \Rightarrow \frac{|\sum_t \alpha(t).\varphi_m^*.\chi|}{C}$
7. if ($\theta_1$=L, ..&&..$\theta_n$=M)
8.    flag out → benign case
9. if ($\theta_1$=H, ..&&..$\theta_n$=L)
10.   flag out → malignant case
End
```

The input of gray scale image is transformed in matrix where its size is evaluated (Line-1). An interesting fact about the input is auto-correction of orientation of the breast image in a specific direction that takes place while the system takes it as an input. The next part of the algorithm deals in elimination of the lighter levels of gray scale using the formulation defined in Line-2. It assists in maximizing the processing speed and minimizing the artifacts. The variable $\phi^2$(x, y) represents the gray scale size. One of the difficult parts of the algorithm implementation is to remove the muscular area with maximum intensity of brightness. This is because the presence of such area doesn't allow differentiating between the actual tissues of breast with other areas e.g. pectoral regions. Hence, it is not feasible for using normal thresholding in such cases.

This problem is overcome by using a matrix of monitoring area ($M_a$) with specific width and height is considered. This will keep a trace of all the essential gray scale information that will assist in extracting original breast tissue from background. The monitoring area ($M_a$) is defined using two forms $f_1$ and $f_2$ that is generalized to (Line-3)

$$f \to i.G(i)/G(i) \tag{1}$$

The above generalized expression with mean that for $f_1$, the computation is carried out for *i* value greater than *t* while $f_2$ is computed considering *i* value residing with *t* and (*t*-max-level of gray scale). Initially, the threshold *T* is computed considering the similar empirical expression of $f_1$, but for overall updating of local threshold is carried out considering mean of $f_1$ and $f_2$ (Line-4), which is now new threshold T. If the new threshold is not found changed compared to the old threshold, it will mean that final converging point of threshold is met or else the process iterates to find new value of $f_1$ and $f_2$. The next step is to apply two dimensional Discrete Wavelet Transform (DWT) in order to extract the coefficients that are further considered as significant feature (Line-5). The next part of the implementation is to apply optimization technique in order to further confirm the extracted feature. We apply particle swarm optimization for this purpose considering adaptive weight $\alpha(t)$ for the purpose of updating. An empirical expression is formed in Line-6 that extracts the elite feature considering mean of gray scale level. The proposed algorithm implements PSO for optimizing the performance. Here the dimension of the particle swarm is evaluated as per the dimension of the given input image of breast cancer. It is done prior to initializing every particle depending on their velocity and position. The next step will be to apply for fitness function. We apply wavelets to the image to find out the malignant section of image as represented by maximized grayscale value in that image section and minimal values for background and benign masses of breast. We apply this strategy PSO in order to explore the regions with maximized grayscale values in order to identify the malignancy. The variable $\alpha(t)$ represents weight, $\varphi^*_m$ represents global function, while the variable $\chi$ will represents the form (t-$2^m$n), where m and n represents rows and columns, and *C* represents cardinality of gray scale levels (Line-6). After the elite feature is extracted, the next step is to develop a set of fuzzy rule with membership variables $\theta_{1\ldots}\theta_{n,.}$ Examples of fuzzy If and Then Ruleset are as shown below.

$$\textbf{If}\ ((\theta_1 = \text{LOW}))\ \textbf{And}\ \theta_n = \text{MEDIUM}\ \textbf{Then}\ \text{outcome} = \text{BENIGH}$$

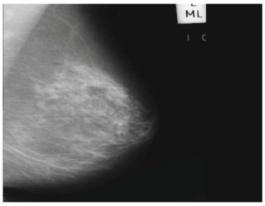
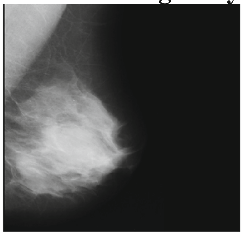
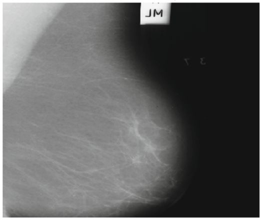
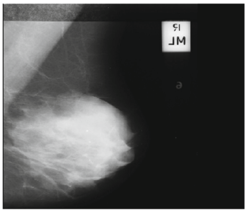
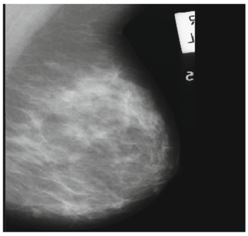
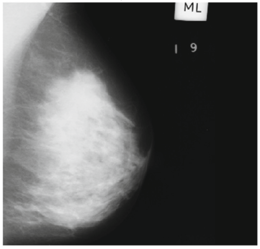
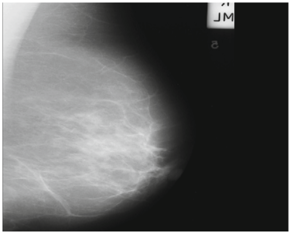
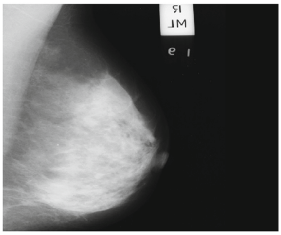
$$\textbf{If}((\theta_1 = \text{HIGH}))\ \textbf{And}\ \theta_n = \text{LOW}\ \textbf{Then}\ \text{outcome} = \text{MALIGNANT}$$

A set of logic with lower to medium values of membership function is created for benign case while different set of logic with higher to lower values of membership function is considered for malignant cases of cancer.

## 6 Results Discussion

The implementation of the proposed study was carried out using Matlab and a dataset of MIAS [23] and DDSM has been used for this purpose. The study computes statistical parameters e.g. mean standard deviation, skewness, and kurtosis. A close observation of

**Table 1.** Visual outcomes of the proposed study

| **Cases of Benign** | **Cases of Malignancy** |
|---|---|
|  |  |
| Mean: 0.00773, SD: 0.3908,<br>Skewness: 0.82844, Kurtosis:1 | Mean: 0.93873, SD: 0.94667<br>Skewness: 0.55677, Kurtosis:0 |
|  |  |
| Mean: 0.11432, SD: 0.62124,<br>Skewness: 0.28216, Kurtosis: 0.2895 | Mean: 0.90217, SD: 0.83498,<br>Skewness: 0.33094, Kurtosis:0.02338 |
|  |  |
| Mean: 0.1211, SD: 0.5771,<br>Skewness: 0.2755, Kurtosis: 0.4118 | Mean: 0.1729, SD: 0.62124,<br>Skewness: 0.3187, Kurtosis: 0.2111 |
|  |  |
| Mean: 0.1109, SD: 0.6611,<br>Skewness: 0.4221, Kurtosis: 0.3115 | Mean: 0.11432, SD: 0.62124,<br>Skewness: 0.1181, Kurtosis: 0.1164 |

the numerical outcome shows a peculiar trend of the statistical parameters, especially standard deviation and kurtosis. It was explored that standard deviation of mammograms with benign cases is lesser as compared to that of malignancy cases. Similarly, the kurtoses of the mammograms with benign cases are much more than that of malignancy cases. Some of the sample of visual outcome of proposed study is shown in Table 1.

However, such trend is not uniform for all the images; however, for majority of images the trend persists. Apart from this, we also estimated the computational time of the cumulative processing is found to be approximately 0.655 s carried out in normal windows machine with 4 GB RAM and core i7 processor. From computational complexity, the proposed system uses only run time memory to perform its execution and hence, it doesn't store any significant information in the system memory. Therefore, the proposed system offer cost effective detection and classification of breast cancer. We also measure the accuracy of the entire dataset image as 97%.

## 7 Conclusion

Initially, the input image is subjected for elimination of unwanted regions followed by normalization of the image. A new monitoring area is constructed followed by elimination of unwanted non-breast tissue. A set of coefficients were extracted from the finalized area using transform-based technique that is considered as extracted feature. Such coefficients were again subjected to cluster-based optimization technique for further fine tuning the extracted feature from accuracy viewpoint. The system than applies simple fuzzy logic to further confirm about the classification of the type of cancer possibilities from the given mammogram image. The final output of the proposed system results in giving precise inference of benign case or malignant case from the given dataset.

## References

1. Velusamy, P.D., Karandharaj, P.: Medical image processing schemes for cancer detection: a survey. In: IEEE-International Conference on Green Computing Communication and Electrical Engineering, Coimbatore, pp. 1–6 (2014)
2. Hollingsworth, A.B.: Mammography and Early Breast Cancer Detection: How Screening Saves Lives. McFarland-Health & Fitness (2016)
3. Dabbs, D.J.: Breast Pathology. Elsevier Health Sciences (2016)
4. Islam, M.S., Kaabouch, N., Hu, W.C.: A survey of medical imaging techniques used for breast cancer detection. In: IEEE International Conference on Electro-Information Technology, Rapid City, SD, pp. 1–5 (2013)
5. National Cancer Institute. https://www.cancer.gov/types/breast/hp/breast-screening-pdq. Accessed 13 Jan 2017
6. Gouze, A., Kieffer, S., Van Brussel, C., Moncarey, R., Grivegnee, A., Macq, B.: Interactive breast cancer segmentation based on relevance feedback: from user-centered design to evaluation. In: Proceedings of SPIE-Medical Imaging (2009)

7. Hassan, A.M., El-Shenawee, M.: Review of electromagnetic techniques for breast cancer detection. IEEE Rev. Biomed. Eng. **4**, 103–118 (2011)
8. Li, Y., Chen, H., Cao, L., Ma, J.: A survey of computer-aided detection of breast cancer with mammography. J. Health Med. Inf. **7**(4) (2016)
9. Arya, C., Tiwari, R.: Expert system for breast cancer diagnosis: a survey. In: IEEE-International Conference on Computer Communication and Informatics, Coimbatore, pp. 1–9 (2016)
10. Sushma, S.J., Prasanna Kumar, S.C.: Advancement in research techniques on medical imaging processing for breast cancer detection. Int. J. Electr. Comput. Eng. **6**(2), 717–724 (2016)
11. Kwon, S., Lee, H., Lee, S.: Image enhancement with Gaussian filtering in time-domain microwave imaging system for breast cancer detection. IEEE Electron. Lett. **52**(5), 342–344 (2016)
12. Singh, S.P., Urooj, S., Lay-Ekuakille, A.: Breast cancer detection using PCPCET and ADEWNN: a geometric invariant approach to medical x-ray image sensors. IEEE Sens. J. **16** (12), 4847–4855 (2016)
13. Santorelli, A., Porter, E., Kang, E., Piske, T., Popović, M., Schwartz, J.D.: A time-domain microwave system for breast cancer detection using a flexible circuit board. IEEE Trans. Instrum. Meas. **64**(11), 2986–2994 (2015)
14. Li, Q., et al.: Direct extraction of tumor response based on ensemble empirical mode decomposition for image reconstruction of early breast cancer detection by UWB. IEEE Trans. Biomed. Circ. Syst. **9**(5), 710–724 (2015)
15. Yin, T., Ali, F.H., Reyes-Aldasoro, C.C.: A robust and artifact resistant algorithm of ultrawideband imaging system for breast cancer detection. IEEE Trans. Biomed. Eng. **62**(6), 1514–1525 (2015)
16. Song, H., et al.: A radar-based breast cancer detection system using CMOS integrated circuits. IEEE Access **3**, 2111–2121 (2015)
17. Wang, X., Qin, T., Witte, R.S., Xin, H.: Computational feasibility study of contrast-enhanced thermoacoustic imaging for breast cancer detection using realistic numerical breast phantoms. IEEE Trans. Microw. Theory Tech. **63**(5), 1489–1501 (2015)
18. Bahrami, H., Porter, E., Santorelli, A., Gosselin, B., Popovic, M., Rusch, L.A.: Flexible sixteen monopole antenna array for microwave breast cancer detection. In: IEEE-Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, pp. 3775–3778 (2014)
19. Denis, M., et al.: Update on breast cancer detection using comb-push ultrasound shear elastography. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **62**(9), 1644–1650 (2015)
20. George, Y.M., Zayed, H.H., Roushdy, M.I., Elbagoury, B.M.: Remote computer-aided breast cancer detection and diagnosis system based on cytological images. IEEE Syst. J. **8** (3), 949–964 (2014)
21. Diaz-Bolado, A., Laurin, J.J.: Experimental validation of the effect of compression on simplified phantoms in microwave tomography applied to breast cancer detection. IEEE Antennas Wirel. Propag. Lett. **11**, 1602–1605 (2012)
22. Sushma, S.J., Kumar, S.C.P.: Image enhancement using bio-inspired algorithms on mammogram for cancer detection. In: IEEE-International Conference on Emerging Research in Electronics, Computer Science and Technology, Mandya, pp. 11–16 (2015)
23. Mamographic Image Analysis Homepage. http://www.mammoimage.org/databases/. Accessed 13 Jan 2017

# Secure Framework of Authentication Mechanism Over Cloud Environment

Ramesh Shahabadkar[1](✉), S. Sai Satyanarayana Reddy[1],
Chinthakunta Manjunath[2], Ugranada Channabasava[3],
and Krutika Ramesh Shahabadkar[4]

[1] Vardhaman College of Engineering, Kacharam, Shamshabad,
Hyderabad, Telangana 500018, India
ramesh.shahabadkar@gmail.com
[2] Faculty of Engineering, Christ University, Bangalore, India
[3] K S School of Engineering and Management, Bangalore, India
[4] RNS Institute of Technology Channasandra, Bangalore, India

**Abstract.** Cloud computing offers a cost effective virtual infrastructure management along with storage and application-oriented services to its customers. This innovation quickly turns into a generally very widely accepted worldview for conveying administrations through web. In this way, this administration expert provider must be offer the trust and information security, on the grounds that there is a most vital and profitable and most delicate information in extremely secure using cryptographic techniques to secure the data in cloud. So for ensure the privacy of essential information, it must be secured utilizing encryptions algorithms and afterward transferring to cloud. This paper presents a novel technique for electronic distributed computing administrations utilizing two-variable validation (2FA) access control framework. The prime target of the projected framework is to guarantee a optimal security for all the actors involved in the component design of proposed authentication system. Furthermore, property based control in the framework likewise authorize cloud servers to maximum the access to those clients with the same arrangement of properties while saving client privacy. At long last, we additionally do a reproduction to show the practicability of our proposed framework. The assessment work is done by utilizing expense of communication, data transfer capacity and proficiency of the framework as an execution metric.

**Keywords:** Access control · Attributed-based control system · Cloud computing · Two-Factor authentication (2FA) · Web services

## 1 Introduction

Cloud computing has turns into a generally utilized worldview for dispersing services through the internet. Along these providers, this server must be giving the trust and the information security, on the grounds that significant and extremely delicate information are put away in substantial sum in clouds. To ensure the imperative data present in cloud, it must be encoded before transferring to the clouds utilizing cryptographic strategies. We have predominantly three distinctive trademarks in cloud administration,

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_30

which are unique in relation to routine facilitating. Basically, sold on interest, actually by minutes or 60 min; Elasticity, a client could have as much as of administrations they need at various circumstances by provider [1]. Cloud registering gives a critical enhancement in virtualization and scattered processing, and it enhances access to rapid of web alongside weak economy. There are numerous uses of distributed computing, for example, information. Sharing, information stockpiling, enormous information administration, medicinal in-arrangement framework and so forth. End clients entrance cloud-based purposes during a web plan, delicate customer or transportable submission whereas the commerce programming and client's in sequence are set away on servers at a distant district. The advantages of electronic distributed computing administrations are huge, which incorporate the simplicity of openness, decreased expenses and capital consumptions, expanded operational efficiencies, adaptability, adaptability and prompt time to advertise. Although, the superior features of cloud computing offers a new arena of distributed clock, but it also suffers from security loopholes. There are in the interim additionally worries about security and protection particularly for electronic cloud administrations. As delicate information might be put away in the cloud for sharing reason and qualified clients might likewise get to the cloud framework for different applications and administrations, client validation has turned into one of the most important factor of safety over cloud interface [2]. In order to utilize cloud services, should access their privilege account using standard authentication mechanism of user ID and password. Unfortunately, such conventional mechanism of authentication is no more secure in cloud that uses internet protocol shrouded with massive number of Trojans. To begin with, the conventional record/secret word based authentication is not security saving. Nonetheless, it is all around recognized that protection considered in distributed computing frameworks. Second, it is regular to share a PC among various individuals. It perhaps simple for programmers to introduce some spyware to take in the login secret word from the web-program. An as of late proposed access control model called characteristic based access control is a decent candidate to handle the primary issue. It gives unknown validation as well as further characterizes access control strategies in view of various properties of information object. In a quality based access control framework, every client has a client master key issued by the power. Practically speaking, the client master key is put away inside the PC. When we consider the aforementioned second issue on online administrations, it is normal that PCs might be shared by numerous clients particularly in some extensive endeavors or associations. The point of this dad per is to outline a novel procedure for electronic distributed computing administrations utilizing two-variable verification (2FA) access control framework. Accurately, in our plan 2FA air conditioning access control framework, a characteristic based access control component is executed client mystery key and a lightweight security gadget. Lastly, we additionally complete to show the practicability of our proposed framework. The evaluation work is carried out by using cost of communication, bandwidth and efficiency of the system as a performance metric. This manuscript has been prearranged as follow. Segment 2 explains the related works done by different authors. Segment 3 explains proposed framework as well as implementation part. Segment 4 provides consequences and discussion then, finally Sect. 5 concludes this paper along with future research direction.

## 2 Related Work

This segment studies is mostly cantered around looking into the current systems and contributory considers talked about by earlier literary works, it is vital for examination that what the current status in the same area is. There are different specialists who have utilized this system on different issues spaces of cloud computing. This paper demonstrates existing condition of research paper, its year of publications, and the name of the distributers. Along these lines, we audit the current number of exploration papers and investigated the viability in them (Table 1).

**Table 1.** Existing survey on data mining classification methods

| Authors | Problem focused | Techniques used | Performance parameters |
|---|---|---|---|
| Fotiou et al. [3] | Security problems and to offers data owners flexibility | Lightweight access control | Lifetime, Number of messages exchanged, Average number of Token |
| K. Punithasurya et al. [4] | To enhance the Security on cloud | Analysing various access control mechanism | User's convenience, Reusability, Node overhead, Authentication failure |
| R. Wu et al. [5] | To achieve highly configurable security requirements of cloud | Role-based access control | Activation & Deactivation time, Network traffic |
| V. Harika et al. [6] | Security and privacy challenges in cloud | Hierarchical attribute based encryption | Execution & Decryption time |
| A. Sirohi et al. [7] | Data security at cloud | Hash based message authentication, Dual substantiation & access management | Confidentiality, Overhead, Authorization, Encryption, Cost effective |
| Pandey et al. [8] | Improve cloud security | Trust dependent ciphering process, policy of key management, encryption | Throughput, Time to generate secret generation time, High end reliability |
| Rashmi et al. [9] | Security challenges in Software | Software as a Service model | Data confidentiality, Authentication |
| F. I. Oyeyinka et al. [10] | Security challenge and to reduce cost of services | Modified things Role Based Access Control model (T-RBAC) | Network traffic, Cost effective, Confidentiality |

(*continued*)

**Table 1.** (*continued*)

| Authors | Problem focused | Techniques used | Performance parameters |
|---|---|---|---|
| S. Kshatriya et al. [11] | Data sharing and security challenges in cloud computing | Survey on data sharing utilizing different encryption technique | Confidentiality |
| P. Kuppuswamy et al. [12] | Securing cloud storage systems | Partitioning and Role based access control | Client security, Identity and Access management, Authentication |
| Talib [13] | Distributed data access control, security in cloud computing | Formula-Based Cloud Data Access Control (FCDAC) | Round trip time, security, confidentiality |
| V. Echeverría et al. [14] | Security in cloud and privacy preserving | Permission as a Service (PaaS), attribute based encryption (ABE) | Confidentiality, Client security |
| Z. Iqbal et al. [15] | Service access policies representation | Enhance attribute based access manager and rule based representation method | Throughput, high end reliability and confidentiality |
| S. Yu et al. [16] | Data security & to reduce heavy computation overhead | Ciphering based on attributes | Confidentiality, highly efficient for security |
| D. A. Gondkar et al. [17] | Protected health record sharing process over cloud | Ciphering based on attributes | Confidentiality, highly efficient for security |
| Z. Liu et al. [18] | Data security and data for sharing on cloud | Identity-based access control | Overhead, Authorization, Confidentiality |

## 3 Proposed System with Implementation

The projected scheme develops an apparatus of the secret key management over cloud. Owing to insecure cloud environment, the proposed systems divide the secret key. The mechanism that perform localization of each of these two secure splits of key, where one part of the secure key resides over the client's machine while second part of the split key is stored over the secured device. The system performs further security incorporations by using two factor authentication processes which lets attacks know that there are multiple dependencies to perform cryptanalysis. Hence, attackers find it near to impossible to locate another split of the secured key even if compromised the first key split. Hence, the proposed data over the security device where work for further encrypting the client's secret key. There is additionally a connecting relationship

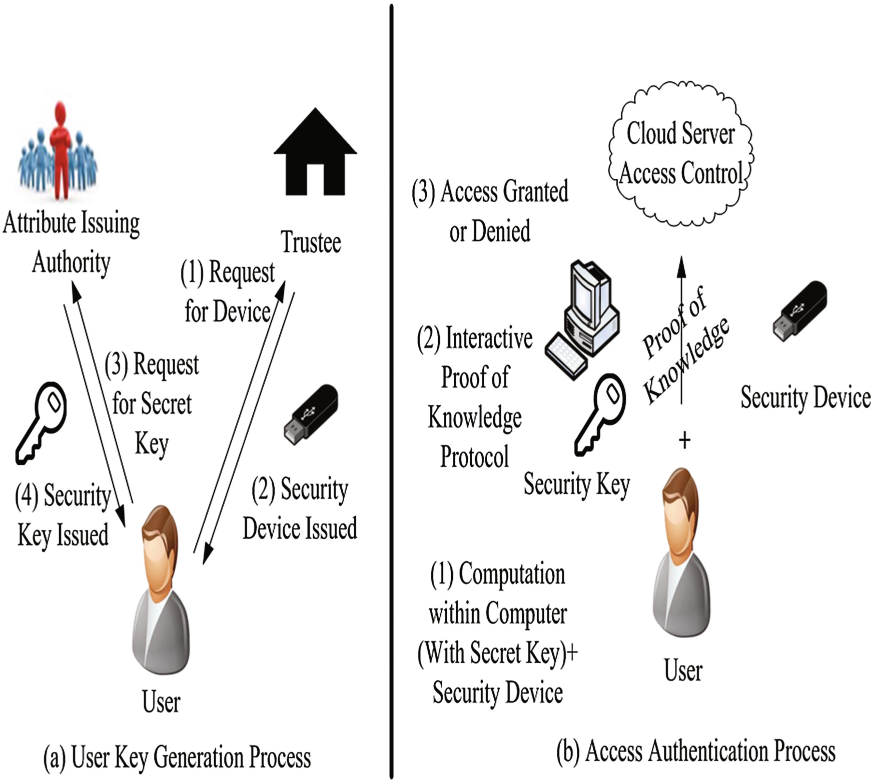

Attribute Issuing Authority
Trustee
(1) Request for Device
(3) Request for Secret Key
(4) Security Key Issued
(2) Security Device Issued
User
(a) User Key Generation Process
Cloud Server Access Control
(3) Access Granted or Denied
(2) Interactive Proof of Knowledge Protocol
Proof of Knowledge
Security Device
Security Key
+
(1) Computation within Computer (With Secret Key)+ Security Device
User
(b) Access Authentication Process

**Fig. 1.** Overview idea of proposed method

between the client's gadget and the mystery key so that the client can't utilize another client's gadget for the verification. The correspondence overhead is negligible and the calculation required in the gadget is only some light-weight calculations, for example, hashing or exponentiation over gathering. All the substantial computations, for example, matching is done on the PC. The thought of our framework is illustrated in Fig. 1.

## 4 Results and Discussion

This section, gives the assessment of the proposed strategy is being assessed and authorised. Assume the aggregated number of features in the framework is 100. At the day end, the features universe A $= \{1...100\}$. The analysis processing of the services in order to validate the client is highlighted in Fig. 2. In case of normal strategy, say, comprising of 2 conditions with 2 properties for every statement for a sum of 4 qualities, the time is under 0.3 s. For an approach of 10 conditions with 10 traits for every statement, the processing time is found to be approximately 3 s. Similar trends of the outcomes related to processing time can be seen in the server side too. The outcome

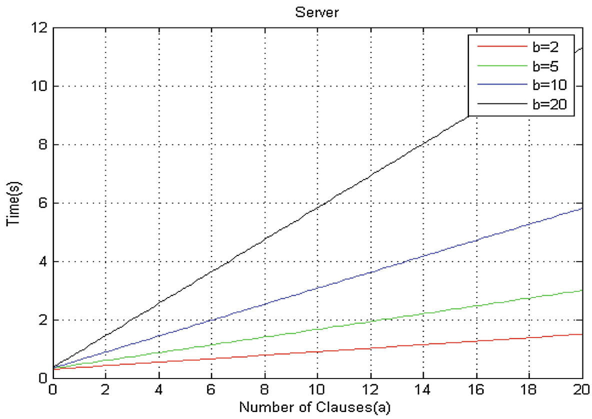


**Fig. 2.** Time consumption during service-side authentication (sec)

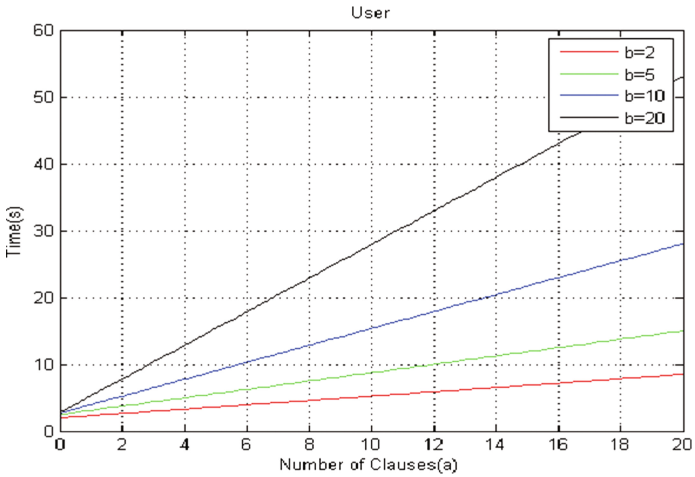


**Fig. 3.** Time consumption during client-side authentication (sec)

shows that time consumed for operating the client-side application is approximately five times slower owing to the usage of poor security devices for registration.

The outcome shown in Fig. 3 highlights the interesting trends of the processing time for authentication over client side application. Considering more than 100 charecteristics, the cumulative validation time is found to be approximately 18 s. Similar trend is also observed in Fig. 2 where the aggregate data transfer capacity prerequisite is around 45 KB, which is satisfactory throughout today's system. One could accomplish that our protocol is conceivable for extremely straightforward arrangement is still not functional yet for strategy of medium size.

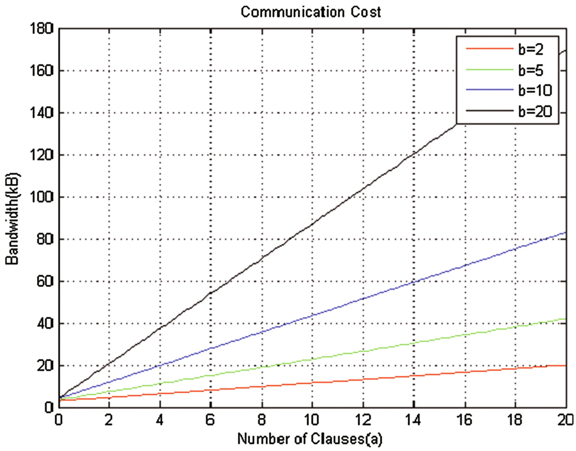


**Fig. 4.** Communication expense of the Auth protocol (KB)

The correspondence expense of our convention is portrayed in Fig. 4. Specifically, for a policy of 100 qualities, the aggregate data transmission prerequisite is originate to be in the order of 45 KB which is found to be within acceptable limit.

## 5 Conclusion and Future Research Direction

This article displayed a new 2FA access control framework for online distributed computing administrations. The presented technique not only enhances the mechanism of secure authentication but also leverages the communication system over cloud environment. Point by point security examination demonstrates that the proposed 2FA access control framework accomplishes the coveted security necessities. Through execution assessment, we exhibited that the development is "feasible". The future work to facilitate enhances the effectiveness while keeping every decent element of the framework.

## References

1. Nelson, M.R.: Building an open cloud. Science **34**(5935), 1656–1657 (2009)
2. Zhiguo, W., Liu, J., Deng, R.H.: HASBE: a hierarchical attribute-base solution for flexible and scalable access control in cloud computing. IEEE Trans. Inf. Forensics Secur. **7**(2), 743–754 (2012)
3. Fotiou, N., Machas, A., Polyzos, G.C., Xylomenos, G.: Access control as a service for the Cloud. J. Internet Serv. Appl., 6–11 (2015). Springer
4. Punithasurya, K., Jeba, P.S.: Analysis of different access control mechanism in cloud. Int. J. Appl. Inf. Syst. (IJAIS) **4**(2), 34–39 (2012)
5. Wu, R., Zhang, X., Ahn, G.-J., Sharifi, H., Xie, H.: Design and Implementation of access control as a service for IaaS cloud. Automot. Serv. Excell., 1–16 (2013)

6. Harika, A.V., Haleema, P.K., Subalakshmi, R.J., Iyengar, N.Ch.S.N.: Quality based solution for adaptable and scalable access control in cloud computing. Int. J. Grid Distrib. Comput. **7**(6), 137–148 (2014)
7. Sirohi, A., Shrivastava, V.: Implementing data storage in cloud computing with HMAC encryption algorithm to improve data security. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **5**(8), 678–684 (2015)
8. Pandey, V.K., Patel, S.K., Bedre, S.: A novel trust dependent attribute based encryption (TD-ABE) for improving the cloud security. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **4**(12), 875–881 (2014)
9. Rashmi, Sahoo, G., Mehfuz, S.: Securing software as a service model of cloud computing: issues and solutions. Int. J. Cloud Comput. Serv. Archit. (IJCCSA) **3**(4), 1–11 (2013)
10. Oyeyinka, F.I., Omotosho, O.J.: A modified things role based access control model for securing utilities in cloud computing. Int. J. Innov. Res. Inf. Secur. (IJIRIS) **5**(2), 21–25 (2015)
11. Kshatriya, S., Chaware, S.M.: A survey on data sharing using encryption technique in cloud computing. Int. J. Comput. Sci. Inf. Technol. **5**(4), 5351–5354 (2014)
12. Kuppuswamy, P., A-Khalidi, S.Q.Y.: Analysis of security threats and prevention in cloud storage: review report. Int. J. Adv. Res. Eng. Appl. Sci. **3**(1), 1–10 (2014)
13. Talib, A.M.: Ensuring security, confidentiality and fine-grained data access control of cloud data storage implementation environment. J. Inf. Secur. **6**, 118–130 (2015)
14. Echeverría, V., Liebrock, L.M., Shin, D.: Permission management system: permission as a service in cloud computing. In: IEEE 34th Annual IEEE Conference on Computer Software and Applications Conference Workshops (COMPSACW), pp. 371–375 (2010)
15. Iqbal, Z., Noll, J.: Towards semantic-enhanced attribute-based access control for cloud services. In: IEEE Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 1223–1230 (2012)
16. Yu, S., Wang, C., Ren, K., Lou, W.: Achieving secure, scalable, and fine-grained data access control in cloud computing. In: IEEE Proceedings on INFOCOM, pp. 1–9 (2010)
17. Gondkar, D.A., Kadam, V.S.: Attribute based encryption for securing personal health record on cloud. In: 2nd International Conference on Devices, Circuits and Systems (ICDCS), pp. 1–5 (2014)
18. Liu, Z.: A secure anonymous identity-based access control over cloud data. In: Fourth International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), pp. 292–295 (2013)

# Scheduling of Parabolic-Type Tasks Arrays in GRID Systems

A.E. Saak(✉), V.V. Kureichik, and A.A. Lezhebokov

Southern Federal University, Rostov-on-Don, Russia
saak@tgn.sfedu.ru, vkur@sfedu.ru, legebokov@gmail.com

**Abstract.** The paper is devoted to the scheduling problem in Grid systems with centralized structure of scheduling system and resources co-allocation that are modeled by resource quadrants. The user's task is represented as a resource rectangle. Heuristic scheduling algorithms' quality is estimated by non-Euclidian heuristic measure accounting occupied resource area and its form. The article also considers the issue of level polynomial algorithms adaptation for arrays of parabolic-type tasks.

**Keywords:** Grid system · Centralized scheduling system structure · Resource rectangle · Parabolic-type task · Non-Euclidean heuristic measure · Level polynomial scheduling algorithm

## 1 Introduction

Grid systems of centralized architecture [1, 2], composed of sites containing parallel systems are modeled by resource quadrant [3, 4] under condition of multiprocessor task processing possibility at several sites simultaneously [2, 5].

Authors assume that the user determines and inputs in the GRID system a number of requested processors in accordance with a task. This number stays the same during task processing (rigid job) [6]. The processing time is taken as an integer [7–10] and determined before task processing begins (clairvoyant rigid jobs) [7]. The task is processed without interruptions from the beginning to the end (parallel rigid non-preemptive jobs) [8, 10]. To process a task the dispatcher allocates consecutively numbered processors (contiguous parallel tasks) [9, 10].

Assumptions mentioned above allows us to represent the user's task by a resource rectangle with horizontal and vertical dimensions equal to number of time unites and processor unites requested for task processing respectively [10]. Symbol $a(j) \times b(j)$ denotes job $j$ which requires $a(j)$ time unites and $b(j)$ processor unites.

## 2 Problem Statement

The scheduling problem in Grid systems with centralized architecture and multisite mode, which is characterized by the possibility of multiprocessor task processing at several sites simultaneously, can be reduced to the problem of Packing resource

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_31

Rectangles into a Square in the Oriented case (PRSO) [11]. The authors of [11] show that the PRSO problem is NP-difficult.

Exponential complexity of optimum resource allocation problem requires using polynomial heuristic algorithms for its solving. In [3, 12–15] authors suggested a resource rectangle environment and defined a resource rectangle, operations over resource rectangles, quadratic arrays of at least two tasks and quadratic single task. Polynomial algorithms based on proposed operations are developed in the resource rectangle environment [3, 12–15]. The scheduling quality is estimated by non-Euclidean heuristic measure accounting occupied resource area and its form. In [14] authors described V-level polynomial algorithms considering the rectangle vertical dimension: with the lack, with the exceeding, with minimal deviation and H-level polynomial algorithms considering the rectangle horizontal dimension: with the lack, with the exceeding, with minimal deviation.

This paper also considers the issue of level polynomial algorithm adaptation for arrays of parabolic-type tasks.

## 3 Scheduling of Parabolic-Type Tasks Arrays with the Use of Level Algorithms

In [15] authors assumed that parabolic-type resource rectangle is considered as quadratic resource rectangle of aspect ratio meeting the condition

$$\frac{a(j)}{b(j)} + \frac{b(j)}{a(j)} > 3.$$

Thus, resource rectangles $j \times 3j, j = 1, 2, \ldots, k$ are classified as parabolic. Arrays of such resource rectangles are ranked by height and defined as follows: the array **I** for k = 19, the array **II** for k = 20, the array **III** for k = 21, the array **IV** for k = 22, and the array **V** for k = 23.

Figure 1 shows results of V-level algorithm for the array **V** with the lack. Figure 2 shows results of V-level algorithm for the array **V** with the exceeding. Figure 3 shows results of V-level algorithm for the array **V** with minimal deviation. Horizontal dimension of each rectangle is shown at its center.

Heuristic measures of resource packing with the use of vertical-level algorithm with the lack, with the exceeding, and with minimal deviation for the array composed of parabolic- and quadratic-type tasks are shown in the Table 1.

As it's shown in the Table 1, heuristic measures of resource packing with the use of level algorithm by vertical dimension with the lack are less than $\frac{1}{2} + 0,42$, by vertical dimension with the exceeding are less than $\frac{1}{2} + 0,3$, and by vertical dimension with minimal deviation are less than $\frac{1}{2} + 0,19$.

Figures 4, 5, 6 show results of H-level algorithm for the array **V** by horizontal dimension with the lack, with the exceeding and with minimal deviation respectively.

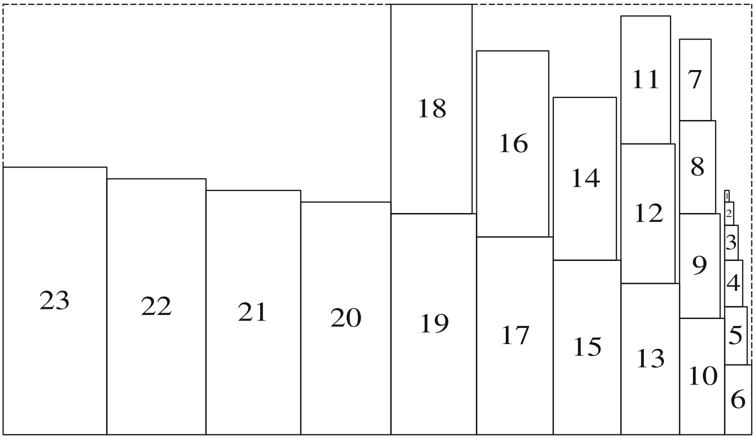


**Fig. 1.** V-level algorithm packing by vertical dimension with the lack for parabolic-type resource rectangle array

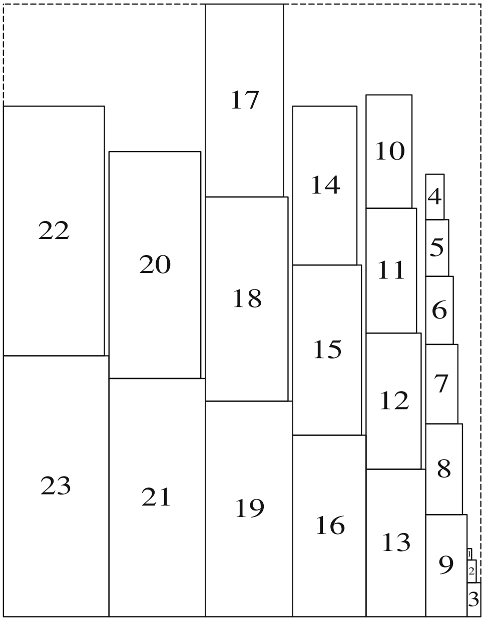


**Fig. 2.** V-level algorithm packing by vertical dimension with the exceeding for parabolic-type resource rectangle array

Heuristic measures of resource packing with the use of H-level algorithm by horizontal dimension with the lack, with the exceeding, and with minimal deviation for array of parabolic- and quadratic-type tasks are shown in the Table 2.

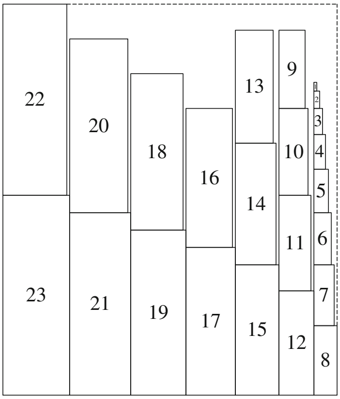


**Fig. 3.** V-level algorithm packing by vertical dimension with minimal deviation for parabolic-type resource rectangle array

**Table 1.** Heuristic measures of V-level algorithm resource packing

| Array number | By vertical dimension with the lack | By vertical dimension with the exceeding | By vertical dimension with minimal deviation |
|---|---|---|---|
| I | 0,86 | 0,66 | 0,69 |
| II | 0,81 | 0,64 | 0,67 |
| III | 0,82 | 0,80 | 0,65 |
| IV | 0,92 | 0,79 | 0,63 |
| V | 0,83 | 0,78 | 0,61 |

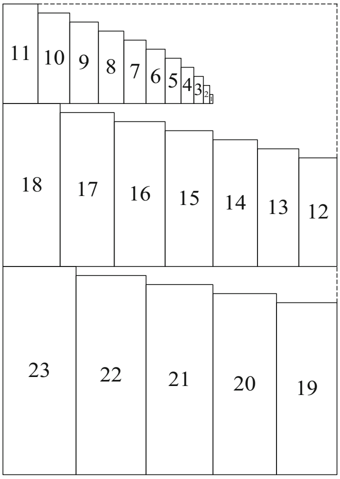


**Fig. 4.** H-level algorithm packing by horizontal dimension with the lack for parabolic-type resource rectangle array

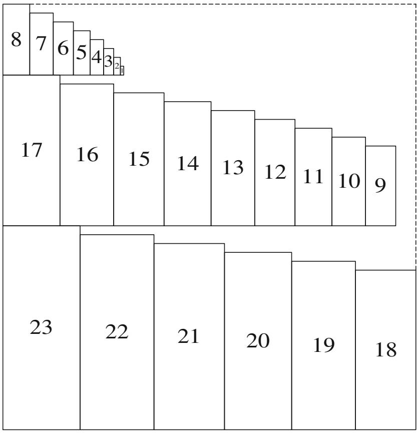


**Fig. 5.** H-level algorithm packing by horizontal dimension with the exceeding for parabolic-type resource rectangle array

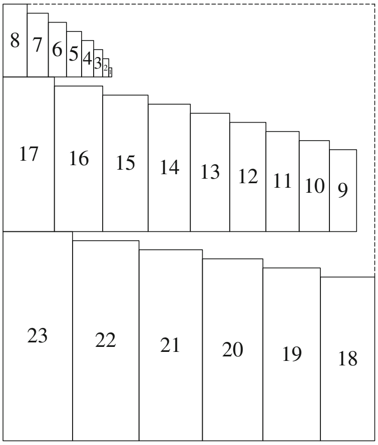


**Fig. 6.** H-level algorithm packing by horizontal dimension with minimal deviation for parabolic-type resource rectangle array

**Table 2.** Heuristic measures of H-level algorithm resource packing

| Array number | By horizontal dimension with the lack | By horizontal dimension with the exceeding | By horizontal dimension with minimal deviation |
|---|---|---|---|
| I | 0,74 | 0,68 | 0,74 |
| II | 0,74 | 0,70 | 0,74 |
| III | 0,74 | 0,69 | 0,74 |
| IV | 0,74 | 0,71 | 0,74 |
| V | 0,73 | 0,70 | 0,70 |

As it’s shown in the Table 2, heuristic measures of resource packing with the use of H-level algorithm by horizontal dimension with the lack are less than $\frac{1}{2} + 0,24$, by horizontal dimension with the exceeding are less than $\frac{1}{2} + 0,21$, and by horizontal dimension with minimal deviation are less than $\frac{1}{2} + 0,24$.

The diagram of heuristic measures of resource packing with the use of level algorithms by vertical dimension and by horizontal dimension in the context of arrays **I-V** scheduling is shown in Fig. 7.

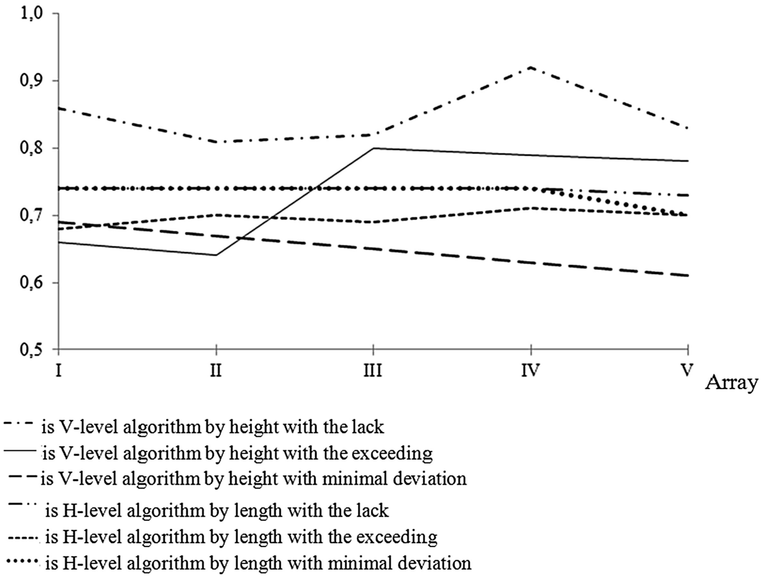


**Fig. 7.** Heuristic measure of resource packing with the use of level algorithms

The diagram shows that V-level algorithm with minimal deviation has the minimum value of heuristic measure maximum $\frac{1}{2} + 0,19$ in the context of considered arrays of parabolic- and quadratic-type resource rectangles. The research allows us to recommend the usage of polynomial algorithms in GRID system of centralized structure and resource co-allocation while servicing arrays of parabolic- and quadratic-type tasks.

## 4 Conclusion

To schedule arrays of parabolic- and quadratic-type resource rectangles in resource rectangle environment vertical-level and horizontal-level algorithms are proposed to use. Authors estimated heuristic measures for benchmark arrays of parabolic- and quadratic-type tasks. The adaptation of developed polynomial algorithms to the mentioned class of GRID system users’ tasks is described in the paper.

**Acknowledgements.** The study was performed by the grant from the Russian Science Foundation (project #14-11-00242) in the Southern Federal University.

## References

1. Rahman, M., Ranjan, R., Buyya, R., Benatallah, B.: A taxonomy and survey on autonomic management of applications in grid computing environments. Concurr. Computat. Pract. Exper. **23**(16), 1990–2019 (2011)
2. Hamscher, V., Schwiegelshohn, U., Streit, A., Yahyapour, R.: Evaluation of job-scheduling strategies for grid computing. In: Buyya, R., Baker, M. (eds.) GRID 2000. LNCS, vol. 1971, pp. 191–202. Springer, Heidelberg (2000). doi:10.1007/3-540-44444-0_18
3. Saak, A.E.: Polynomial algorithms for resource allocation in grid-based systems for quadratic typing, arrays applications. Inf. Technol. 7, 32 p (2013)
4. Saak, A.E.: Resource and multiprocessor task management in grid system of centralized architecture. In: Proceedings of XII All-Russian Conference "Control problems" RCCP'2014, Moscow, 16–19 June 2014, pp. 7489–7498 (2014)
5. Sonmez, O., Mohamed, H., Epema, D.: On the benefit of processor coallocation in multicluster grid systems. IEEE Trans. Parallel Distrib. Syst. **21**(6), 778–789 (2010)
6. Feitelson, Dror G., Rudolph, L., Schwiegelshohn, U., Sevcik, Kenneth C., Wong, P.: Theory and practice in parallel job scheduling. In: Feitelson, Dror G., Rudolph, L. (eds.) JSSPP 1997. LNCS, vol. 1291, pp. 1–34. Springer, Heidelberg (1997). doi:10.1007/3-540-63574-2_14
7. Bougeret, M., Dutot, P.-F., Jansen, K., Otte, C., Trystram, D.: A fast 5/2-approximation algorithm for hierarchical scheduling. In: D'Ambra, P., Guarracino, M., Talia, D. (eds.) Euro-Par 2010. LNCS, vol. 6271, pp. 157–167. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15277-1_16
8. Bougeret, M., Dutot, P., Trystram, D., Jansen, K., Robenek, C.: Improved approximation algorithms for scheduling parallel jobs on identical clusters. Theor. Comput. Sci. **600**, 70–85 (2015)
9. Błądek, I., Drozdowski, M., Guinand, F., Schepler, X.: On contiguous and non-contiguous parallel task scheduling. J. Sched. **18**(5), 487–495 (2015)
10. Caramia, M., Giordani, S., Iovanella, A.: Grid scheduling by on-line rectangle packing. Networks **44**(2), 106–119 (2004)
11. Martello, S., Monaci, M.: Models and algorithms for packing rectangles into the smallest square. Comput. Oper. Res. **63**, 161–171 (2015)
12. Saak, A.E.: Algorithms scheduling in grid-based systems for quadratic typing, arrays applications. Inf. Technol. 11, 9–13 (2011)
13. Saak, A., Kureichik, V., Kuliev, E.: Ring algorithms for scheduling in grid systems. In: Silhavy, R., Senkerik, R., Oplatkova, Z., Prokopova, Z., Silhavy, P. (eds.) Proceedings of the 4th Computer Science On-line Conference, CSOC2015. Advances in Intelligent Systems and Computing, vol. 349, pp. 201–209. Springer, Cham (2015)
14. Saak, A.E.: Scheduling of sets of circular-type and hyperbolic-type tasks in grid systems. Inf. Technol. 5, 323–332 (2016)
15. Saak, A.E.: Circular-typed multiprocessor tasks scheduling in grid systems. Inf. Technol. 1, 37–41 (2016)

# The Lukov Castle – A Historical 3D Visualization in Different Time Periods

Pavel Pokorný(✉) and Zuzana Jarošová

Department of Computer and Communication Systems,
Faculty of Applied Informatics, Tomas Bata University in Zlín,
Nad Stráněmi 4511, 760 05 Zlín, Czech Republic
pokorny@fai.utb.cz, zuzana.jarosova93@seznam.cz

**Abstract.** This paper briefly describes a visualization method for the Lukov Castle. This castle was probably founded at the beginning of the 13th Century and grew rapidly over the following centuries, so it became an important residence of the local aristocracy. All available historical materials of the Lukov Castle and its surrounding area were collected. The focus was mainly directed on historical sketches, historical paints and photos. All the collected information was chronologically sorted and, on this basis, seven 3D visualizations of this castle were created that demonstrate its development and decay. To begin with, the grounds terrain model was created - based on the Daftlogic database [14]. All buildings and accessories were separately modeled by period (using standard polygonal representation) and textured using UV mapping techniques. In this way, seven complex 3D scenes of the Lukov Castle in these periods were created: the first half of the 13th Century, the second half of the 13th Century; then, in the 14th, 15th, 17th and 18th Centuries. The last model corresponds to its current appearance. The visualization output is performed by rendered images and animations in these time periods. The Blender software suite was used for visualization purposes.

**Keywords:** Computer graphics · 3D visualization · Historical visualization · Modeling · Texturing · Animation

## 1 Introduction

Data visualization is a hot topic. A simple definition of data visualization states: “It is the study of how to represent data by using a visual or artistic approach - rather than a traditional reporting method [1]. The represented data is usually displayed through texts, diagrams, images, or animations to communicate a message.

Today, visualization has found ever-expanding applications in education, science, engineering (e.g. product visualization), medicine, interactive multimedia, etc. A typical example of a visualization application is the Computer Graphics field. The invention of computer graphics may be the most important development in visualization since the invention of central perspective in the Renaissance period. The development of animation has also helped the advance of visualization [17].

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_32

Along with the development and computer technology improved performance, the possibilities and limits of computer graphics continue to increase. The consequences of this trend are 3D visualizations, which are being used ever more frequently, image outputs are of better quality [5], and the area of the visualized environment that can be show at the same moment is rising [6]. As mentioned above, these visualizations are used in many scientific and other areas of human interest [2].

One of the fields is historical visualizations. Based on historical documents, drawings, plans, maps and photographs, it is possible to create 3D models of objects that no longer exist - things, products, extinct animals or buildings [4]. Using common 3D software modeling tools to create 3D models and to assign suitable materials and the corresponding textures to these models, we can derive a very credible appearance of these historical objects.

This paper describes a 3D visualization method of Lukov Castle in history - and today.

## 2 A Short History of Lukov Castle

Lukov Castle dates back to the early 13th Century. Historical stone fragments illustrate the significant effect of stonemasons from this period. There is no doubt that, at that time, the castle was in the ownership of a royal crown – although, in the early part of the 14th Century, it was to be found in the possession of the powerful Sternberg family. The castle remained under their administration for almost two hundred years [8].

During the Czech-Hungarian wars, the castle was captured and burned by the troops of King Matthias Corvinus. It is probable that the extensive damage to the fortifications led to the following major reconstruction and expansion of the castle. In 1548, the castle was bought by the Neksove family from the Landek family - under whose ownership the castle was reconstructed in the Renaissance style and adapted to meet the need to live more comfortably. The Neksove Family, were good landlords and so their estates grew significantly and became rich.

In the 17th Century, during the Thirty Years War, Lukov Castle became a center of the rebellion of the local population – so-called “Wallachians”, against the Habsburgs on several occasions. Later, the castle was captured by the Swedish army, and when they left, the castle was burned. In later times, Lukov Castle came into the possession of the Seiler family. This family belonged to Viennese courtly aristocracy and they did not take care of it. The castle lost its importance and, at the end of the 18th Century, it was completely abandoned and became a source of cheap building materials.

Attempts to rescue the castle complex started in the last quarter of the 20th Century. In 1983, the process of continuous archaeological research began in the castle area. Since 1983, gradual restoration works of the castle monuments have been undertaken. Thanks to many years of volunteer work, Lukov Castle was transformed from a neglected ruin to a popular tourist destination, where ongoing educational programs are aimed at children and young people [7].

## 3 Resources and Software

The first phases needing to be done were to collect all available historical materials of Lukov Castle and to select suitable programs for visualization creation.

### 3.1 Acquiring Resources

The overall progress of this work was initiated by the collation of available historic materials and information about the castle. The main resources were found in the National Heritage Institute in Kroměříž [12], the State District Archive in Zlín [11], the castle webpage [8], and the book on this subject [7]. Attention was mainly focused on sketches, drawings and photos.

Most of the resources described above included historical documents and photos of this castle. The most important information was obtained from Radim Vrla, an employee of National Heritage Institute in Kroměříž. This historian participated in archaeological excavations of Lukov Castle and created nice sketches of the castle´s appearance in different periods. Some of them are shown in Figs. 1, 2 and 3. Apart from these, Radim Vrla also provided consultations during the construction of 3D models in order to achieve maximum historical authenticity.

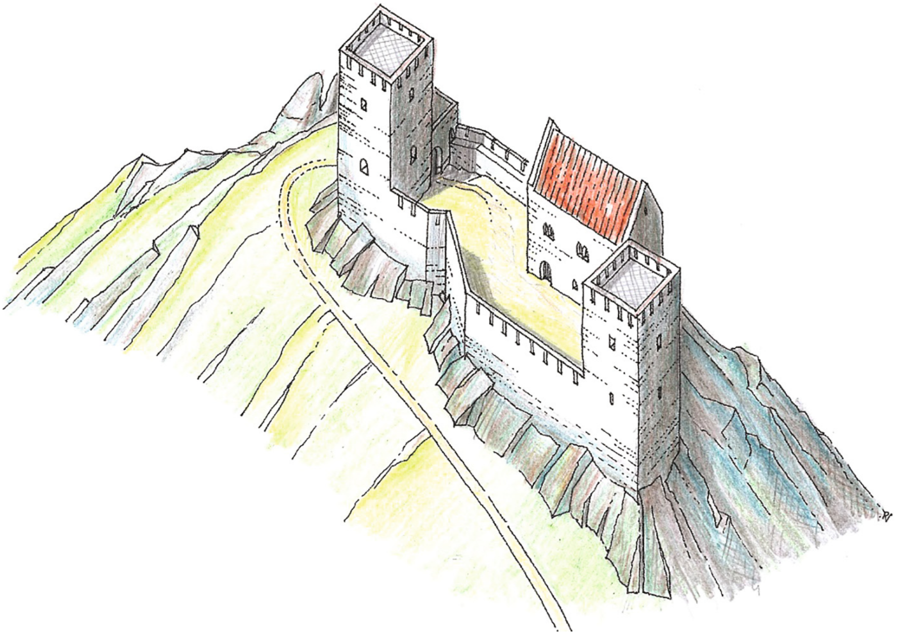


**Fig. 1.** Lukov Castle artwork from the first half of the 13th Century; Author: Radim Vrla [12]

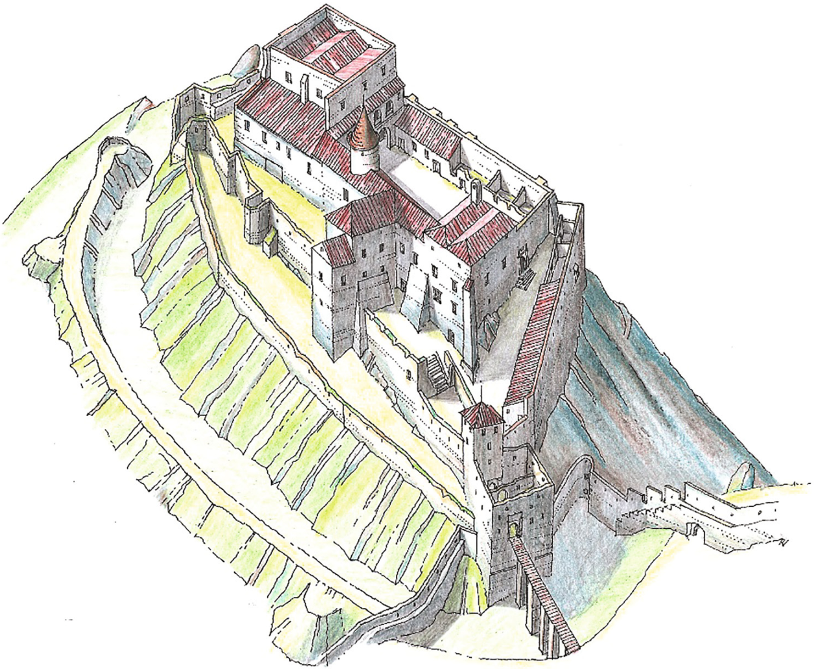


**Fig. 2.** Lukov Castle artwork from the 18th Century; Author: Radim Vrla [12]

Finally, seven historical sketches were derived covering these periods: the first half of the 13$^{th}$ Century, the second half of the 13$^{th}$ Century; then, in the 14$^{th}$, 15$^{th}$, 17$^{th}$ and 18th Centuries. The last model corresponds to the present-day castle.

### 3.2 Programs Used

Preference was given to the use of open-source software; therefore, the Blender software suite was used for 3D modeling, texturing and rendering purposes [13]; textures were drawn in GIMP [15]; and Microdem [18] was the last software program used.

Blender is a fully integrated creation suite that offers a broad range of essential tools for the creation of 3D content – including modeling, skinning, texturing, UV mapping, animation, rigging, particles and other simulations, scripting, rendering, compositing, post-production, and game creation [3, 9]. Blender is based on cross-platform OpenGL technology, and is available under GNU GPL license.

GIMP is an acronym for GNU Image Manipulation Program, a free distributed program under the GNU General Public License. It is mainly a drawing tool and a digital image editor. It allows one to retouch photos by fixing problems affecting the whole image or parts of the image, image composition, and color adjustment in the photos to bring back a natural look or for image authoring [10].

Microdem is a freeware microcomputer mapping program designed for displaying and merging digital elevation models, scanned maps, vector-mapping data, satellite

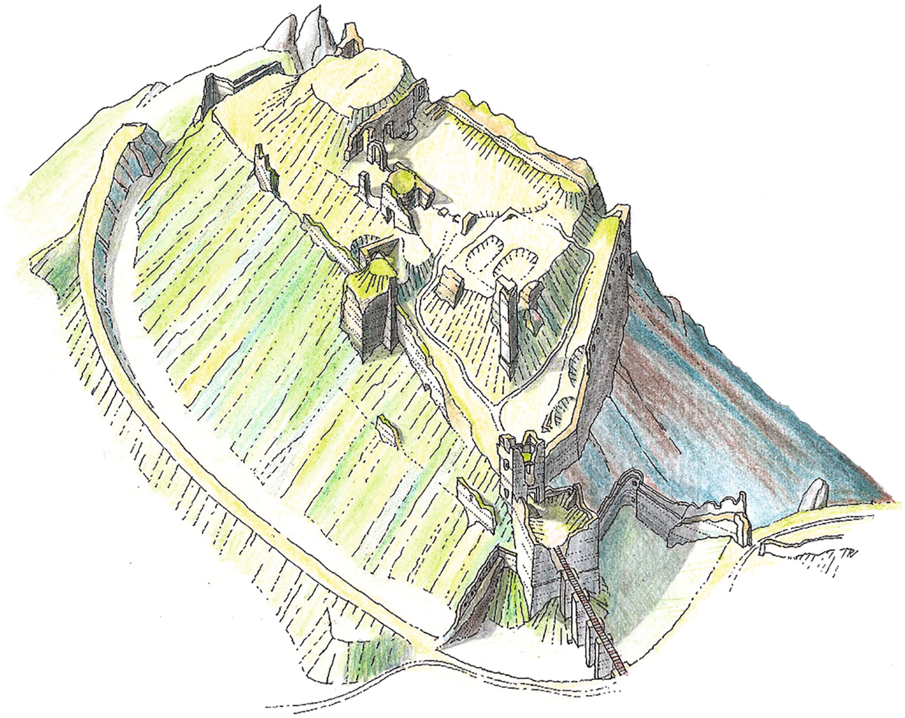


**Fig. 3.** Lukov Castle artwork from the present; Author: Radim Vrla [12]

imagery, or GIS databases [19]. This software was utilized to convert the landscape elevation map data into a bitmap image (i.e. a height-map).

## 4 A Landscape Model

Data files which contained text information about the earth elevations were used to create the landscape model of Lukov Castle and its vicinity.

Digital Elevation Model data (i.e. DEM) which was provided by the NASA Shuttle Radar Topographic Mission (SRTM) in 2007 was also used. Data for over 80% of the globe is stored on this site [18] and can be freely downloaded for noncommercial use.

The data relating to the region around the castle was downloaded, and then opened using the Microdem freeware program - described above. This software is able to convert the obtained data into a bitmap image. Microdem can clip and convert these images to grayscale; this was then applied to the Lukov Castle's surroundings. Then, a bitmap was obtained in which the brightness of each pixel represents the altitude of the corresponding location.

Blender was used to insert a square (i.e. plane object) in the new scene, which was then divided several times by the Subdivide tool to derive a grid with a density of several hundred vertices. Then, the Displace modifier was used to assign textures (for

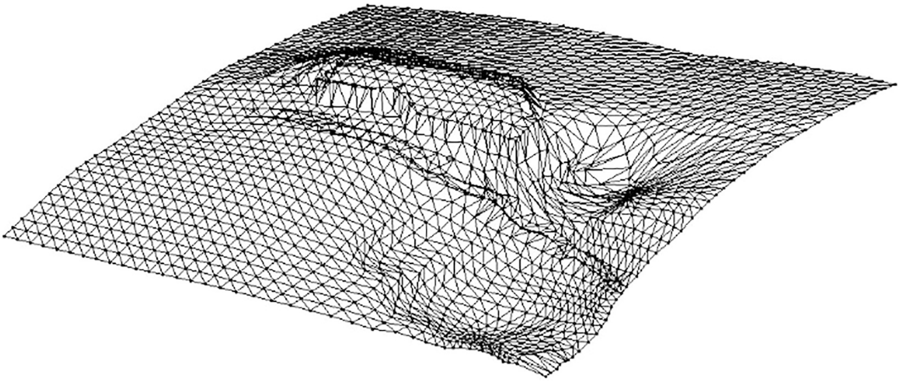


**Fig. 4.** The final mesh model of Lukov Castle's surroundings

the obtained bitmap). The Displace modifier deforms an object - based on its texture and setting parameters. The model of Lukov Castle's landscape was derived using this method (Fig. 4). Although this model is not entirely accurate given the scale, the whole scene and the quality of the castle-area model buildings, is sufficient for our purposes.

## 5 Modeling and Texturing the Castle and Accessories

As mentioned above, seven historical sketches of the Lukov Castle were obtained. Because these sketches are the most important resources, it was decided that seven different 3D complex castle models based on these sketches be made, supplemented with corresponding textures. Each model was created in the same way - but formed a separate scene.

### 5.1 Modeling

The first important resource of the castle is the top-down view of the present-day castle. This was acquired from the castle´s homepage [10]. This image was loaded into the Blender environment and placed on the background screen. In the following step, it was compared to the historical sketch from the period we wanted the model to create, as well as drawings from other resources. This enabled the deep insight of the appearance of the model planes. This was then followed by the modeling phase.

A standard polygonal representation for the models of the castle and accessories was used. Blender supports a large number of modeling tools for these so-called "mesh objects" - basic editing commands, transformation tools, modifiers, Extrude, Knife, Bevel, Loop Cut and Slide, etc. [9]

The advantage was that the buildings usually have a box shape - so modeling was not difficult for this reason. To begin with, the shape of the top view of the castle was traced and then extruded into the third dimension. Subsequently, more specific shapes like holes for the windows and doors, stairs, roofs or crenellations on the walls were

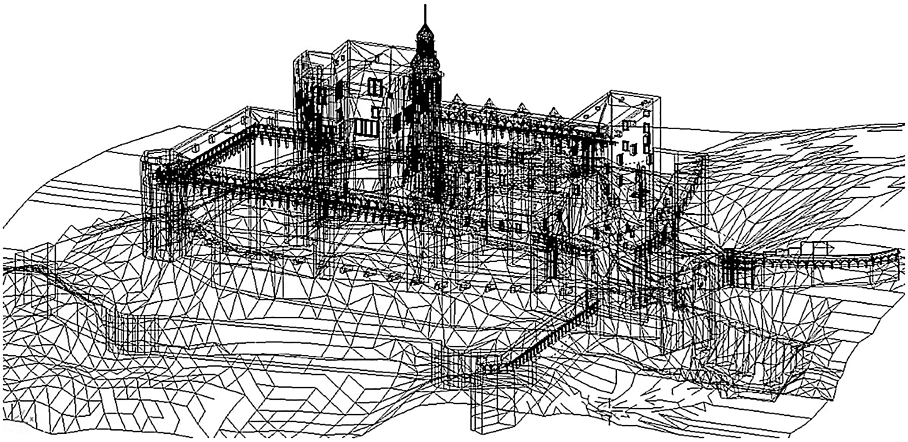


**Fig. 5.** The mesh model of Lukov Castle from the 17th Century (wireframe shading)

modeled. An example of the finished castle model from the 17th Century in wireframe shading is shown in Fig. 5.

It is important to make “pure” models, i.e. to ensure that the models do not have unnecessary vertices, edges and surfaces. The model in Fig. 5 contains approx. 30 000 vertices for the entire scene.

## 5.2 Texturing

UV mapping techniques were used for texturing the objects. This process starts by the decomposition of each object into 2D sub-surfaces (e.g. a UV map). Blender supports the Unwrap tool for decomposing purposes. The UV map created in this way is saved into the .png raster graphic format (this requires the use of a lossless compression algorithm). The resolutions: 512 × 512 or 1024 × 1024 pixels were used for the UV maps in order to increase the speed of the rendering process used at the end of this work.

All textures were drawn in the GIMP software environment and all of the UV maps created in GIMP were also opened. In these pictures, the location of each part of the 3D object is visible. With this information, one can fill each individual sub-surface as necessary. Suitable textures were downloaded from the CGTextures website [16] – these textures were edited and modified in order to use them on the models. For texture creation and editing purposes, standard GIMP selection, transforming and coloring tools [10] were used. Once this process was completed, all of the created textures were saved back into the same files and opened and then, using the Blender environment mapped on the appropriate 3D models. The mesh model of Lukov Castle from the 17th Century - including mapped textures, is shown in Fig. 6.

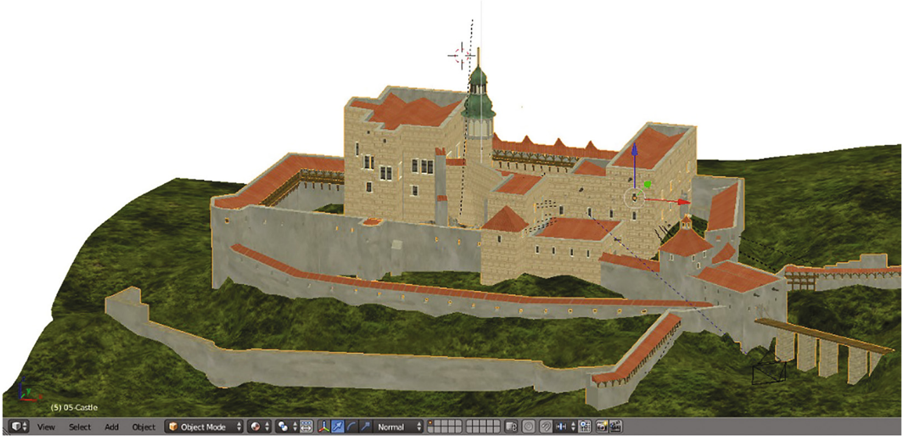


**Fig. 6.** The model of Lukov Castle from the 17th Century, with textures in Blender

## 6 Rendering and Animations

After the completion of the modeling phase, the whole scene was completed – the castle model was joined with the landscape model – and, in addition, other suitable parameters like the surroundings, camera and lighting were set. The surrounding area setting parameters were performed by using the Blender World window. It is possible to set simple colors for horizon and zenith, and to blend them, or to use the internal (procedural) or external texture (any bitmap file). Simple light blue colors were used for the background.

The lighting of the scenes can be realized in several ways in Blender. It is possible to use light objects, (called Lamps in Blender), for local lighting or global influences (i.e. Environment Light, Ambient Light, Ambient Occlusion and Indirect Light). The Environment Ligh technique was combined with small Emit material value settings of selected objects.

The last step before the rendering process was to select a suitable position for the camera with the aim of creating more rendered images from different positions. Due to this, more different camera objects were added and oriented correctly in order to capture the best graphic images of the whole scene.

Other render output can be realized via animation. In the Blender environment, this can be implemented with the help of animation curves that can be freely shaped and transformed. The camera can then follow this in a defined time. The closed curve was then inserted into the Blender scene and shaped in order for the camera to be able to capture the 3D model created in this way from all directions.

The Render command performs the rendering calculation process in the Blender environment. Additionally, one can set many of the accompanying parameters. The basic parameters include - the choice of a rendering algorithm, image or animation resolution, type of output file format, anti-aliasing, motion blur, enable/disable ray-tracing and shadows. The decision was made to use Blender's internal renderer

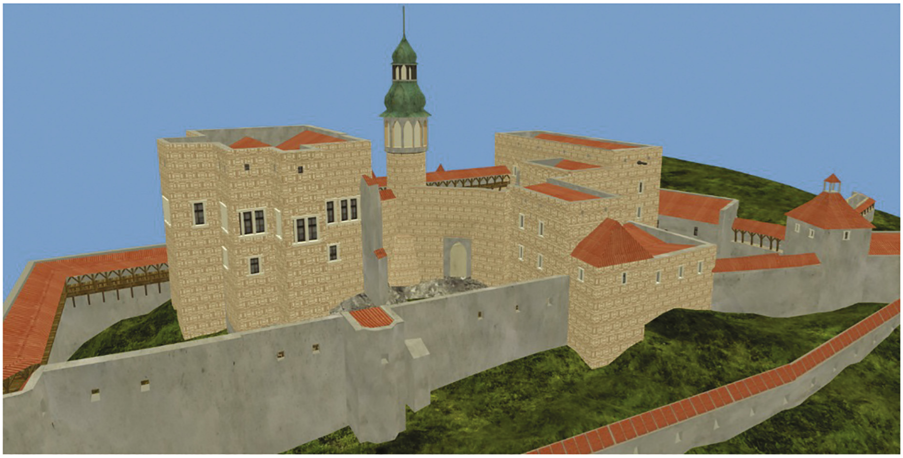


**Fig. 7.** A rendered image of Lukov Castle from the 17th Century

with an image resolution of 1920 × 1080 pixels, 25 frames per second, and the MPEG-2 output format to render animations. Figure 7 shows a sample rendered image of Lukov Castle from the 17th Century.

## 7 Conclusion

This paper briefly describes the visualization method used for depicting the Lukov Castle in different time-periods. Based on historical materials (mainly sketches and drawings), seven more complex 3D models in these periods were created covering: the first half of the 13th Century, the second half of the 13th Century; and then in the 14th, 15th, 17th and 18th Centuries; where the last model corresponds to its current appearance.

All models were modeled and textured in the Blender software suite and the final complex scenes were created with rendered images and animation. These outputs demonstrate the probable appearances of the castle in the above-mentioned centuries.

The future goal is to expand and improve these models. This process will include creating more detailed parts of the castle - including its courtyard, mainly with the help of high-quality textures. Further improvements would be to create virtual web application outputs that can be performed with the help of a Blend4web plug-in, which Blender can then use to export the whole 3D scene into a single html file that can be opened and controlled in any web browser.

**Acknowledgments.** This work arose from collaboration with the National Heritage Institute and the Kroměříž Regional Council, which provides both professional and methodological assistance to heritage site owners and managers in the Zlín Region.

## References

1. Brand, W.: Data Visualization for Dummies. Wiley, New Jersey (2014)
2. Qiu, H., Chen, L., Qiu, G., Yang, H.: An effective visualization method for large-scale terrain dataset. WSEAS Trans. Inf. Sci. Appl. **10**(5), 149–158 (2013)
3. Kent, B.R.: 3D Scientific Visualization with Blender. Morgan & Claypool, San Rafael (2015)
4. Centofanti, M., Brusaporci, S.: Architectural 3D modeling in historical buildings knowledge and restoration processes, Unione Italiana Deisegno Repository (2013), http://95.110.184.161:8080/handle/123456789/27
5. Qiu, H., Chen, L., Qiu, G., Yang, H.: Realistic simultaion of 3D cloud. WSEAS Trans. Comput. **12**(8), 331–340 (2013)
6. Wettel, R., Lanza, M.: Codecity: 3D visualization of large-scale software. In: Companion of the 30th International Conference on Software Engineering, Leipzig, Germany, pp. 921–922. ACM, New York (2008)
7. Holík, J.: Lukov – střípky z historie záchrany hradu. Spolek přátel hradu Lukova, Lukov (2013)
8. The Lukov Castle, http://www.hradlukov.cz/
9. Hess, R.: Blender Foundations–The essential Guide to Learning Blender 2.6. Focal Press, Amsterdam (2010)
10. Peck, A.: Beginning GIMP: From Novice to Professional. Apress, New York (2008)
11. State District Archive in Zlín, Moravian Provincial Archives in Brno, http://www.mza.cz/zlin/
12. National Heritage Institute in Kroměříž, https://www.npu.cz/cs/uop-kromeriz
13. Blender.org - Home, http://www.blender.org
14. Google Maps Find Altitude, http://www.daftlogic.com/sandbox-google-maps-find-altitude.htm
15. GIMP - The GNU Image Manipulation Program, http://www.gimp.org
16. CG Textures – Textures for 3D, graphic design and Photoshop!, http://www.cgtextures.com/
17. Visualization (computer graphics), http://en.wikipedia.org/w/index.php?title=Visualization_(computer_graphics)
18. Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E.: Hole-filled seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT) (2008), http://srtm.csi.cgiar.org
19. Microdem Homepage, http://www.usna.edu/Users/oceano/pguth/website/microdem/microdem.htm

# RaESS: Reliable-and-Efficient Statistical Spreading Data Fusion Mechanism in Wireless Sensor Network

B.S. Jayasri(✉) and G. Raghavendra Rao

Department of Computer Science and Engineering,
National Institute of Engineering, Mysore, India
jayasriphdl4ni@gmail.com, grrao56@gmail.com

**Abstract.** Ensuring continued sustainable communication characteristics are still questionable fact to be obtained by existing data fusion techniques. We have reviewed existing studies to find more scope towards reliability. This paper has presented a novel model called as RaESS or Reliable-and-Efficient Statistical Spreading Data Fusion Mechanism which mainly aims to achieve higher number of unique transmission and lower utilization of resources. We introduced Degree of Information that compliments to increase reliable transmission while minimizing packet drops. Compared to existing technique, proposed technique shows reduced energy consumption and enhanced communication performance (data delivery ratio, delay, algorithm processing time).

**Keywords:** Data fusion · Resource utilization · Wireless sensor network · Fuser node · Energy efficiency · Optimization · Communication

## 1 Introduction

The area of wireless sensor network has witnessed a significant revolution in its applicability from conventional monitoring to Internet-of-Things [1, 2]. With the modernization of technology and demands for data acquisition (or processing), the multi-sensor data fusion is gaining a faster momentum in wireless sensor network [3, 4]. In data fusion, it is necessary that all the sensors must capture, only the unique data and thereby forward it to its parent fuser node [5]. Although this phenomenon can significantly increase the data quality, but it is less likely to occur in real-time. The prime reason behind this is that a similar event can occur at multiple places at same time. As the sensor node works on principle of Time Division Multiple Access, so it is quite feasible that two different cluster could possess nearly similar forms of acquired data. This process of data fusion leads to transmittance of similar forms of data that finally leads to communication overhead in base station. Although, there are various studied towards data fusion techniques [6, 7], till date, there is no standard or benchmarked data fusion technique to ascertain sustainable communication performance. Apart from this, the limited computation capabilities with restricted resources don't let a sensor to work on other communication protocols or information diffusion principle [8]. Usage of sensor is now seen in almost every commercial place but very

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_33

few of them make use of real concept of wireless sensor network. The real upcoming applications are also related to large scale implementation. Confirming the deterministic and predictable communication performance is yet to be seen in future and such research work is highly required to be carried out. Hence, this paper presents a novel technique towards leveraging the reliability in transmission process in wireless sensor network. Section 2 discusses about the existing research work followed by problem identification in Sect. 3. Section 4 discusses about proposed methodology followed by elaborated discussion of algorithm implementation in Sect. 5. Comparative analysis of accomplished result is discussed under Sect. 6 followed by conclusion in Sect. 7.

## 2 Related Work

This section discusses about the recent work being carried out towards data fusion in wireless sensor network. Our prior work has already discussed certain techniques of data fusion and their effectiveness [9, 10]. The most recent work carried out by Liu et al. [11] have formulated a regression fusion rule of logistic type in order to enhance the operation of data fusion. Chen et al. [12] have presented an experimental model of multi-modal data fusion where varied significant data from the road surface are being captured for application design. Multimodal data fusion has been also investigated by Farias et al. [13] where an integrated maximum a posteriori was applied for enhancing the fusion capability to process larger dataset. Habib et al. [14] have investigated fusion operation of bio-sensors and presented a technique to aggregate data where the decision of fusion is carried out by fuzzy rule set. Baccarelli et al. [15] have discussed about energy efficient and environmental friendly data fusion technique and have also discussed about the optimization requirement associated with resource management. Yassine et al. [16] have discussed the technique where location is prioritized for performing data fusion considering three metrics e.g. fingerprint of received signal strength, angle of arrival, and time of arrival. The technique also uses Kalman filter as well as genetic algorithm for further enhancing fusion performance. Zhang et al. [17] have discussed about achieving energy efficient data fusion techniques considering multimedia sensor networks. Larios et al. [18] have presented a technique for obtaining lower energy consumption and higher accuracy using self-organizing map. Neves et al. [19] have presented a study that integrates services from actuators and sensors. Tan et al. [20] have investigated effect on coverage due to data multi-sensor fusion of stochastic nature. Nemati et al. [21] have studied data fusion for enhancing the estimation of rate of respiration. Yue et al. [22] have used neural network for enhancing communication performance of sensor network with mobility attributes. Therefore, there has been various research work already carried out towards improving the data fusion performance in wireless sensor network. The next section discusses about the problems associated with the existing work.

## 3 Problem Description

This section discusses about the problems being explored from existing research work. Although, there are various techniques to enhance communication performance, very less focus is laid on incorporating deterministic measures towards data fusion in wireless sensor network. Majority of the techniques have focused on energy efficiency but in that case there are various aspects of data fusion which is being left ignored e.g. filtration of redundant packets, increasing number of unique transmission, packet dropping due to channel problems, etc. Hence, the problem statement is "to design a deterministic form of data fusion that ensures better equilibrium between continued sustainable transmissions with minimal resource consumption." The next section discusses about adopted research methodology to overcome this problem.

## 4 Proposed Methodology

The proposed work is a continuation of our prior model of data fusion [23]. The present work uses analytical research methodology and emphasizes on incorporating reliable transmission in wireless sensor network. The prime target is to enhance the performance of data fusion. The schematic architecture of proposed system is shown in Fig. 1.

The proposed system uses graph theory to model data fusion structure and implements a principle of statistical spreading for maintaining a good balance between a new variable called as Degree of Information and Transmission reliability. The model also focuses on data quality by calculating packet error rate and reduces the probability of dropping a packet during the transmission state, using the process of distributed

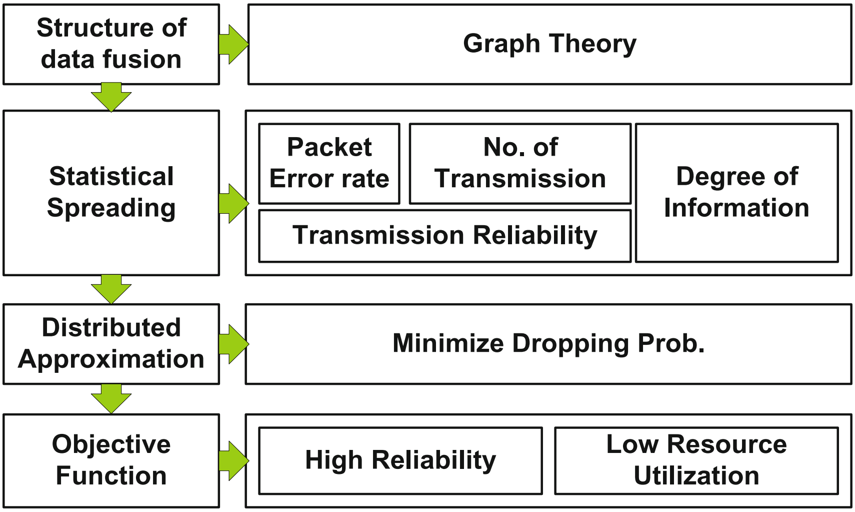


**Fig. 1.** Schematic Architecture of RaESS

approximation. Finally, an objective function is designed to ensure higher reliability and lower resource usage.

## 5 Algorithm Implementation

This section discusses about the algorithm that is responsible for mainly incorporating transmission reliability as well as statistical spreading mechanism while performing data fusion in wireless sensor network. The algorithm takes the input of the communication request from the node that after processing leads to generation of optimal transmission reliability. The steps of the proposed system are as shown below:

```
Algorithm for RaESS
Input: $\theta$ (Reliability), $\phi$ (Degree-of-Information), $\eta$ (No. of sensors), BS (Base Station)
Output: Optimized Reliability
Start
1. $BS \xrightarrow{\mathrm{Re}\,q} \eta, \eta_i \xrightarrow{\phi_i} BS$
2. $BS \xrightarrow{query} n[n \subseteq \eta]$
3. If ($\phi_p \geq \phi(1- \theta_q)/\eta$
4.     $T_i \rightarrow \log_\theta(1-\frac{\phi(1-\theta_q)}{\eta\phi_i})$
5.     $EO_{sol} \rightarrow \frac{[(\theta_q)^{-1}-1]}{[(\theta_p)^{-1}-1]} \approx \frac{\phi_p^*}{\phi_q^*}$
6.     $D = \sum_{i=1}^{\eta} \phi_i(1-\theta_i),[D \leq \eta\phi_o(1-\theta_q)]$
7.     $\theta_i = \alpha.s_i.\theta_q$
8. Else
9.     $T_i \rightarrow 0$
End
```

The proposed algorithm implements graph theory where the routes (G) are formed using nodes (V) and communication links (E) i.e. G = {V, E}. The study considers $\eta$ as number of homogenous sensors. We introduce a new parameter called as $\phi$ (Degree-of-Information) that is defined as cumulative amount of significant and accurate information residing in one data packet. An empirical formulations of the reliable data transmission is proposed to be $\phi_{node} = \sum\phi.\ \Theta$

$$\phi_{node} = \sum_{i=1}^{m} \phi_i \theta_i, m \leq \eta \tag{1}$$

Not only this, the transmission of the replicates forwarded by a specific sensor can be empirically expressed as $t_p \rightarrow \log_\theta(1 - \phi_p)$. On the other side, the reliability of the transmission $\theta$ can be computed by Degree-of-Information value received by base station divided by that of total value of Degree-of-Information. The complete formulation of the study considers transmission reliability of $\theta$ is found to be much more than specific $\theta$ value of a specific node during data fusion. The study maintains the optimal converging point when (i) data fused at sink is found to be statistically enough and (ii) at reduced transmission. In this algorithm, the base station broadcasts a request to the sensors where the sensors upon receiving the request forwards their reliability data to the base station (Line-1). This mechanism is called statistical spreading that is mainly introduced to (i) ensure minimal resource drainage (leading to resource efficiency), and (ii) enhancing the quality of the fused data. The bases station than forwards a query beacon to those nodes interested in communication (Line-2). Using information from channels, all the sensors computes rate of packet error *err* and then computes its degree of information. The sensors than uses a logical condition (Line-3) to check if it has good amount of information to be transmitted, than it computes cardinality of transmissions i.e. T (Line-4). After computing transmission T, the sensor doesn't wait for any form of beacons for acknowledgement and forward its reporting beacon corresponding to T times. Otherwise, there is no transmission (Line-9). The algorithm than establishes a relationship with reliability and degree of information (Line-5) in order to evolve up with an Elite Outcome ($EO_{sol}$).

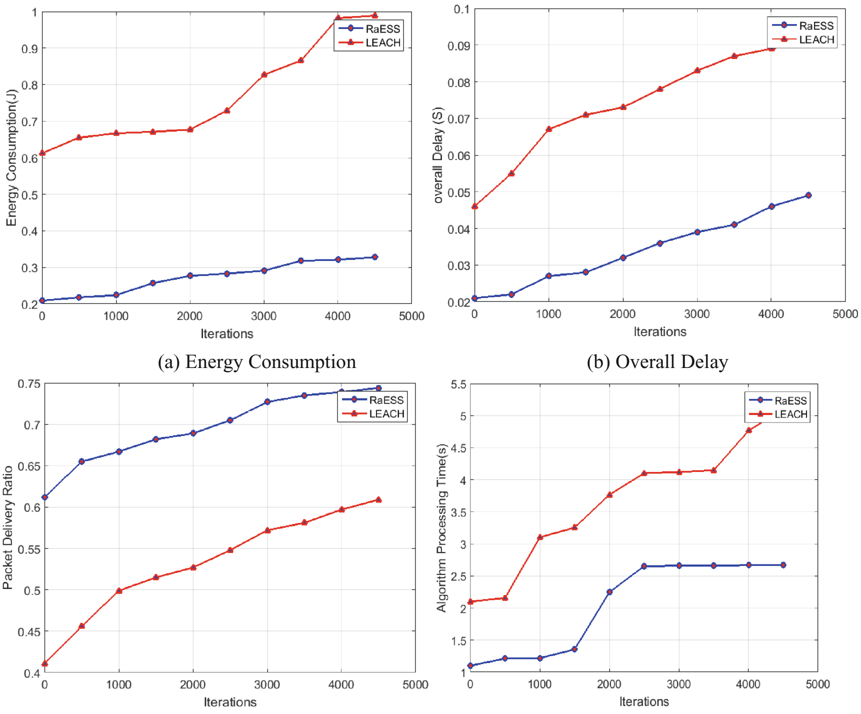
As the algorithm targets its implementation over large scale wireless sensor network, and the sensor nodes are highly distributed, hence it also assumes the possibilities of packet drop *D* where the computation of *D* is shown in Line-6. We consider a condition that the rate of drop of data is equivalent for all the candidate nodes. Therefore, cumulative degree of information is equally classified corresponding to data drop in order to calculate $\theta$ (Reliability). It is essential that value of the dropped packet *D* should satisfy the logical condition of $\eta\phi_o(1-\theta_q)$. It will mean that when the drop in cumulative degree of information could be equivalent among all sensors, than the probability of degree of information that could be dropped at specific node can be empirically expressed as $D_i = \phi_i(1-\theta_i) = \phi_o(1-\theta_q)$. We consider *s* as size of the graph of data fusion used by a specific node and compute the final reliability as shown in Line-7. The variable $\alpha$ will corresponds to $1/[1 + (s_p-1). \phi_q]$.

The process of data fusion continues in similar fashion where the member nodes transmit their data to fuser node while one fuser node forwards their fused data to another fuser node using multihop network. This increases data quality as well as reduces chances of re-transmission owing to elimination of, forwarding redundant data to the base station. Therefore, the proposed algorithm of RaESS is capable of (i) maintaining reliability of transmission and (ii) perform statistical spreading of fused data using graph theory. As the complete study is based on probability theory, hence approximation logic assists in speeding up the whole computation process. This process methodology will have direct advantage in the energy conservation process too as well as to minimize wastage of other valuable resources within a node. The next section discusses about the outcome accomplished from the proposed study.

## 6 Results Discussion

The evaluation of the study was carried out with 600 sensor nodes compliant with MEMSIC node configuration dispersed randomly under simulation area of 1000 × 1000 $m^2$. A hypothetical data of 3500 bits has been used for communication purpose. As the proposed RaESS focuses towards achieving reliable transmission, it was essential for evaluating its effectiveness with respect to energy consumption, delay, packet delivery ratio, as well as algorithm processing time. The study outcome was compared with standard LEACH model to briefing its effectiveness. Figure 2 showcases the graphical outcome of comparative performance analysis.

The proposed RaESS incorporates higher degree of reliability while ensuring reliable selection of path for transmission on the basis of low energy consumption, high data flow, more filtration of redundant data, and frequent update of routing. A closer look into Fig. 2(d) shows proposed algorithm ensure nearly similar processing time after 2500 rounds proving its sustainability and longevity as a direct outcome of



(a) Energy Consumption

(b) Overall Delay

(c) Packet Delivery Ratio

(d) Algorithm Processing Time

**Fig. 2.** Comparative Performance Analysis

reliability. RaESS algorithm has also significant benefit, with respect to highly reduced energy consumption (Fig. 2(a)) as well as reduced delay (Fig. 2(b)) and increased packet delivery ratio (Fig. 2(c)) with increasing rounds of simulation. Hence, proposed RaESS algorithm ensures better communication performance in wireless sensor network.

## 7 Conclusion

For a sensor node to work effectively, it is required that its transmission should be maintained at highest degree of reliability, while also ensuring energy efficiency. However, just as the occurrences of an event is nearly impossible, it is also challenging to ensure the reliable work performance of a sensor node irrespective of its placement within the monitored area. After reviewing the existing and recent literatures on data fusion, we find that there is little work towards reliability incorporation in data fusion. The existing system is meant for ensuring fault tolerant data transmission from one transmitting node to a base station, but it never ensures the same for the entire network. Therefore, the present paper discusses and formulates the problem pertaining to reliability of the transmission when the sensor node carries out data fusion. The proposed system offers a mechanism of transmission where the information is forwarded with maximized reliability and lesser resources. The study outcome is found to offer its positive effect in minimizing energy consumption and maximizing communication performance in large scale wireless sensor network.

## References

1. Rehmani, M.H., Pathan, A.-S.K.: Emerging Communication Technologies Based on Wireless Sensor Networks: Current Research and Future Applications. CRC Press-Computers, Boca Raton (2016)
2. Behmann, F., Wu, K.: Collaborative Internet of Things (C-IoT): for Future Smart Connected Life and Business. Wiley, Chichester (2015)
3. Kamila, N.K.: Handbook of Research on Wireless Sensor Network Trends, Technologies, and Applications. IGI Global (2016)
4. Fahmy, H.M.A.: Wireless Sensor Networks: Concepts, Applications, Experimentation and Analysis. Springer, Singapore (2016)
5. Mahmoud, M.S., Xia, Y.: Networked Filtering and Fusion in Wireless Sensor Networks. CRC Press, New York (2014)
6. Castanedo, F.: A review of data fusion techniques. Sci. World J. (2013). Hindawi Publishing Corporation
7. Sidek, O., Quadri, S.A.: A review of data fusion models and systems. Int. J. Image Data Fusion **3**(1), 3–21 (2012). Taylor & Francis
8. Braca, P., Goldhahn, R., Ferri, G., LePage, K.D.: Distributed information fusion in multistatic sensor networks for underwater surveillance. IEEE Sens. J. **16**(11), 4003–4014 (2016)
9. Jayasri, B.S., Raghavendra Rao, G.: Need For Energy Efficient Data Fusion in Wireless Sensor Networks. Int. J. Eng. Res. Technol. (IJERT) **3**(1), January 2014

10. Jayasri, B.S., Rao, G.R.: Reviewing the research paradigm of techniques used in data fusion in WSN. In: IEEE International Conference in Computing and Communications Technologies (ICCCT), pp. 83–88, 26–27 February 2015
11. Liu, L., Luo, G., Qin, K., Zhang, X.: An algorithm based on logistic regression with data fusion in wireless sensor networks. EURASIP J. Wirel. Commun. Networking (2017). Springer
12. Chen, Y.L., et al.: Inexpensive multimodal sensor fusion system for autonomous data acquisition of road surface conditions. IEEE Sens. J. **16**(21), 7731–7743 (2016)
13. Farias, R.C., Cohen, J.E., Comon, P.: Exploring multimodal data fusion through joint decompositions with flexible couplings. IEEE Trans. Sig. Process. **64**(18), 4830–4844 (2016)
14. Habib, C., Makhoul, A., Darazi, R., Salim, C.: Self-adaptive data collection and fusion for health monitoring based on body sensor networks. IEEE Trans. Ind. Inf. **12**(6), 2342–2352 (2016)
15. Baccarelli, E., et al.: Green multimedia wireless sensor networks: distributed intelligent data fusion, in-network processing, and optimized resource management. IEEE Wirel. Commun. **21**(4), 20–26 (2014)
16. Yassine, A., Nasser, Y., Awad, M., Uguen, B.: Hybrid positioning data fusion in heterogeneous networks with critical hearability. EURASIP J. Wirel. Commun. Networking (2015)
17. Zhang, Z.J., Lai, C.F., Chao, H.C.: A green data transmission mechanism for wireless multimedia sensor networks using information fusion. IEEE Wirel. Commun. **21**(4), 14–19 (2014)
18. Larios, D.F., Barbancho, J., Rodríguez, G., Sevillano, J.L., Molina, F.J., León, C.: Energy efficient wireless sensor network communications based on computational intelligent data fusion for environmental monitoring. IET Commun. **6**(14), 2189–2197 (2012)
19. Neves, P.A.C.S., Rodrigues, J.J.P.C., Lin, K.: Data fusion on wireless sensor and actuator networks powered by the zensens system. IET Commun. **5**(12), 1661–1668 (2011)
20. Tan, R., Xing, G., Liu, B., Wang, J., Jia, X.: Exploiting data fusion to improve the coverage of wireless sensor networks. IEEE/ACM Trans. Networking **20**(2), 450–462 (2012)
21. Nemati, S., Malhotra, A., Clifford, G.: Data fusion for improved respiration rate estimation. EURASIP J. Adv. Sig. Process. **2010**, 926305 (2010)
22. Yue, Y., Fan, H., Li, J., Qin, Q.: Large-scale mobile wireless sensor network data fusion algorithm. In: 2016 IEEE International Conference on Big Data Analysis (ICBDA), Hangzhou, pp. 1–5 (2016)
23. Jayasri, B.S., Raghavendra Rao, G.: EEDF: energy efficient data fusion supportive of virtual multipath propagation in WSN. Int. J. Appl. Eng. Research (IJAER) **10(**86) (2015)

# Analysis of Temperature Impact on Production Process with Focus on Data Integration and Transformation

Michal Kebisek[✉], Lukas Spendla, and Pavol Tanuska

Faculty of Materials Science and Technology, Slovak University of Technology, Trnava, Slovakia
{michal.kebisek,lukas.spendla,pavol.tanuska}@stuba.sk

**Abstract.** The proposal is focused on initial steps of data mining process, specifically on the data integrations and transformation stages. In the proposed paper, we have described integration process of production and weather data for analysis and knowledge discovery process that is based on the CRISP-DM methodology. The data integration process was designed and performed using RapidMiner software platform. From the integrated data we have presented use case that is suitable for further detailed data analysis and utilisation in knowledge discovery process.

**Keywords:** Data integration · CRISP-DM · Temperature impact · Production data · RapidMiner

## 1 Introduction

The increase in competition and globalisation of businesses have brought great changes in the structure of companies, and made them to pay attention to total quality management, technological changes, security and environmental questions. These modifications have been carried over to the production and manufacturing area, because this is the one most directly involved in the efficiency and sustainability of the industrial processes [1].

The current trend in this area is introduction of Industry 4.0 concept into the companies that is directly focused on the production and manufacturing area, including Internet of Things, Big Data concept, Cloud computing, etc. [2, 3]. Implementing this concept in production companies with large amounts of data from the production processes, can be further used in failure prediction, quality improvement, predictive maintenance, production optimisation, etc.

Large amounts of data collected in the manufacturing companies opens up various opportunities for data analysis, since it provides different view on the production process as tradition business reports. [4] This however, also raises issue, that the collected data are stored in various systems, applications and databases. Therefore, integration is required to obtain comprehensive view on the production processes.

Consequently, it is necessary to perform data integration of collected and stored data that can be used for analysis and predictions of production processes. [5]

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_34

This paper focuses on the integration and transformation process of production and weather data that will serve as a basis for further knowledge discovery process.

## 2 CRISP-DM

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts. [6]

As a methodology, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks. As a process model, CRISP-DM provides an overview of the data mining life cycle.

The life cycle model on Fig. 1. consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary.

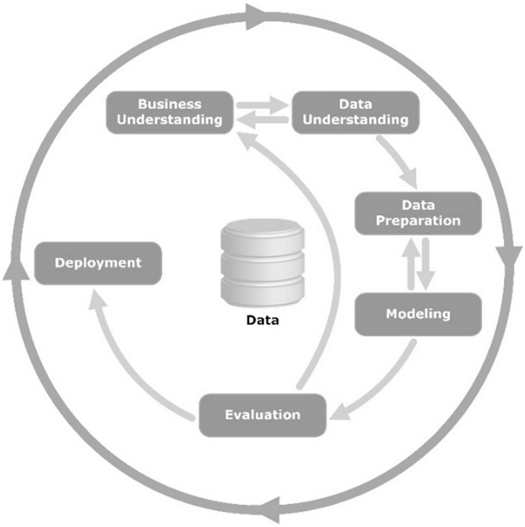


**Fig. 1.** CRISP-DM - Cross-Industry Standard Process for Data Mining

The CRISP-DM model is flexible and can be customized easily. For example, if your organization aims to detect money laundering, it is likely that you will sift through large amounts of data without a specific modelling goal. Instead of modelling, your work will focus on data exploration and visualization to uncover suspicious patterns in financial data. CRISP-DM allows you to create a data mining model that fits your particular needs. In such a situation, the modelling, evaluation, and deployment phases might be less relevant than the data understanding and preparation phases. However, it is still important to consider some of the questions raised during these later phases for long-term planning and future data mining goals.

## 3 Analysis of Temperature Impact on Production Process

Production process as a whole is usually monitored and analysed based on various key performance indicators (KPI), e.g. production time, batch size, production volume, etc. [7] All these parameters are based on production data that are gathered and processed from various heterogeneous systems. However these data do not take into account environmental influences on the production process. Therefore, adding data from the environment can provide additional, but very important, point of view.

Significant impact on the production process can have various environmental parameters, e.g. temperature, humidity, dust, noise, etc. In our initial analysis, we have focused on integration of available weather data, due to the fact that to obtain other environmental parameters it would be necessary to install additional sensors for their collection.

Production company that serves as a basis for our proof of concept is situated in the south east of Slovakia. The production processes are performed in the production hall that is exposed to various weather conditions. The outside temperature can be as low as –21°C during the winter, to maximum up to 35°C in the summer. Although the production hall is air conditioned, it is not possible to maintain constant temperature during various temperature fluctuations across the whole day. Therefore one of our main objectives was to analyse the influence of outside temperature on the overall production.

Since there are no weather data directly from the production hall (the temperature in the building is measured by standard digital thermometers, however the data are not stored), therefore we are using the data collected from three weather stations that are situated in the vicinity of the production hall, i.e. in the industrial area. Available weather data are from the period between 1.6.2016 and 30.9.2016.

To perform the future analysis the weather data must be integrated together with production data in a way that will be suitable for further analysis and data mining stages. This process was captured as process diagram on Fig. 2. CRISP-DM methodology served as a basis for steps performed in this process.

The initial phase, Problem definition, focuses on understanding the main objectives and requirements from the use case point of view, to propose a data mining problem definition and design plan to achieve the main objective.

The data collection and identification, as part of the data understanding phase in CRISP-DM, focuses on obtaining suitable weather data that can be used to analyse the influence of outside temperature on the overall production.

Three weather stations are located outside the production hall that serves as a basis for our use case. Due to the fact that these stations are from different manufacturers, data obtained from them are also different and inconsistent. Two of the weather stations measure temperature every hour, the third one, is measuring only four times a day–at 00:00, 06:00, 12:00 and 18:00. Data from this station are shown on Fig. 3. In addition to temperature the data contains also additional data like average daily temperature, atmospheric pressure, humidity etc.

The cause of this is that data format is not consistent across individual weather stations. Therefore it is necessary to design the process to gather, store and integrate

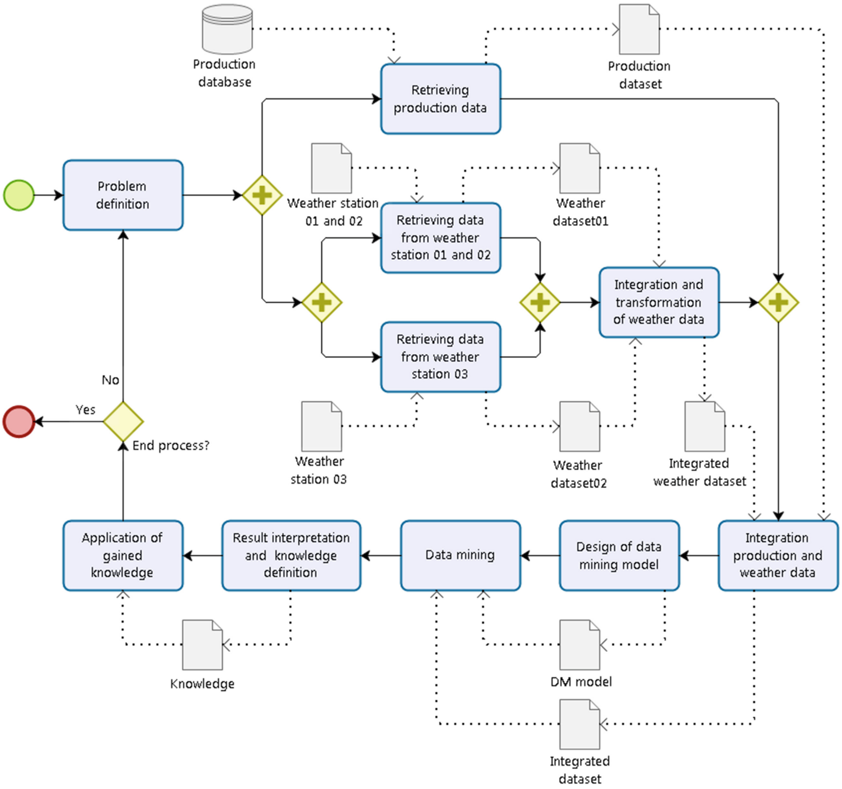

Production database
Production dataset
Retrieving production data
Problem definition
Weather station 01 and 02
Retrieving data from weather station 01 and 02
Weather dataset01
Integration and transformation of weather data
Retrieving data from weather station 03
No
Yes
End process?
Weather station 03
Weather dataset02
Integrated weather dataset
Application of gained knowledge
Result interpretation and knowledge definition
Data mining
Design of data mining model
Integration production and weather data
Knowledge
DM model
Integrated dataset

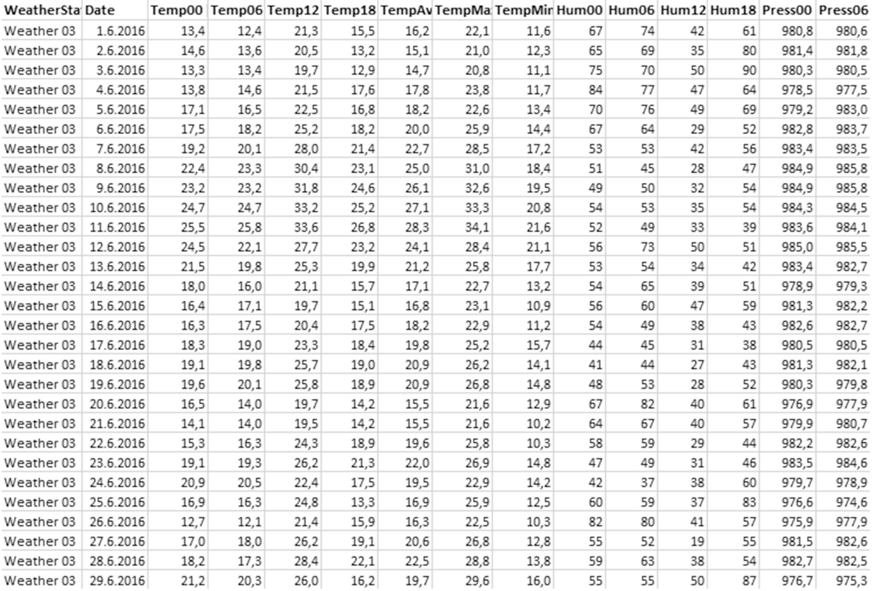
**Fig. 2.** Data integration process diagram

these data together. The following Data preparation phases covers all activities to unify and integrate the weather data do the form usable in the final data set.

Subsequently these three different integrated datasets were supplemented with the data obtained from the production process. The final dataset from this process can serve as a basis for subsequent Modelling phase in our future work that utilises and applies various modelling techniques and parameter combinations to achieve optimal results, i.e. performs data mining. Detailed description of this process can be find in the next chapter.

The results will be evaluated in the subsequent Evaluation phase. The data mining models will be thoroughly evaluated and reviewed to ensure that the results properly achieved the main objective.

Application of gained knowledge represents the final stage of this iteration. The achieved results and gained knowledge will be represented in a way that the customer can further use.



| WeatherSta | Date | Temp00 | Temp06 | Temp12 | Temp18 | TempAv | TempMa | TempMir | Hum00 | Hum06 | Hum12 | Hum18 | Press00 | Press06 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Weather 03 | 1.6.2016 | 13,4 | 12,4 | 21,3 | 15,5 | 16,2 | 22,1 | 11,6 | 67 | 74 | 42 | 61 | 980,8 | 980,6 |
| Weather 03 | 2.6.2016 | 14,6 | 13,6 | 20,5 | 13,2 | 15,1 | 21,0 | 12,3 | 65 | 69 | 35 | 80 | 981,4 | 981,8 |
| Weather 03 | 3.6.2016 | 13,3 | 13,4 | 19,7 | 12,9 | 14,7 | 20,8 | 11,1 | 75 | 70 | 50 | 90 | 980,3 | 980,5 |
| Weather 03 | 4.6.2016 | 13,8 | 14,6 | 21,5 | 17,6 | 17,8 | 23,8 | 11,7 | 84 | 77 | 47 | 64 | 978,5 | 977,5 |
| Weather 03 | 5.6.2016 | 17,1 | 16,5 | 22,5 | 16,8 | 18,2 | 22,6 | 13,4 | 70 | 76 | 49 | 69 | 979,2 | 983,0 |
| Weather 03 | 6.6.2016 | 17,5 | 18,2 | 25,2 | 18,2 | 20,0 | 25,9 | 14,4 | 67 | 64 | 29 | 52 | 982,8 | 983,7 |
| Weather 03 | 7.6.2016 | 19,2 | 20,1 | 28,0 | 21,4 | 22,7 | 28,5 | 17,2 | 53 | 53 | 42 | 56 | 983,4 | 983,5 |
| Weather 03 | 8.6.2016 | 22,4 | 23,3 | 30,4 | 23,1 | 25,0 | 31,0 | 18,4 | 51 | 45 | 28 | 47 | 984,9 | 985,8 |
| Weather 03 | 9.6.2016 | 23,2 | 23,2 | 31,8 | 24,6 | 26,1 | 32,6 | 19,5 | 49 | 50 | 32 | 54 | 984,9 | 985,8 |
| Weather 03 | 10.6.2016 | 24,7 | 24,7 | 33,2 | 25,2 | 27,1 | 33,3 | 20,8 | 54 | 53 | 35 | 54 | 984,3 | 984,5 |
| Weather 03 | 11.6.2016 | 25,5 | 25,8 | 33,6 | 26,8 | 28,3 | 34,1 | 21,6 | 52 | 49 | 33 | 39 | 983,6 | 984,1 |
| Weather 03 | 12.6.2016 | 24,5 | 22,1 | 27,7 | 23,2 | 24,1 | 28,4 | 21,1 | 56 | 73 | 50 | 51 | 985,0 | 985,5 |
| Weather 03 | 13.6.2016 | 21,5 | 19,8 | 25,3 | 19,9 | 21,2 | 25,8 | 17,7 | 53 | 54 | 34 | 42 | 983,4 | 982,7 |
| Weather 03 | 14.6.2016 | 18,0 | 16,0 | 21,1 | 15,7 | 17,1 | 22,7 | 13,2 | 54 | 65 | 39 | 51 | 978,9 | 979,3 |
| Weather 03 | 15.6.2016 | 16,4 | 17,1 | 19,7 | 15,1 | 16,8 | 23,1 | 10,9 | 56 | 60 | 47 | 59 | 981,3 | 982,2 |
| Weather 03 | 16.6.2016 | 16,3 | 17,5 | 20,4 | 17,5 | 18,2 | 22,9 | 11,2 | 54 | 49 | 38 | 43 | 982,6 | 982,7 |
| Weather 03 | 17.6.2016 | 18,3 | 19,0 | 23,3 | 18,4 | 19,8 | 25,2 | 15,7 | 44 | 45 | 31 | 38 | 980,5 | 980,5 |
| Weather 03 | 18.6.2016 | 19,1 | 19,8 | 25,7 | 19,0 | 20,9 | 26,2 | 14,1 | 41 | 44 | 27 | 43 | 981,3 | 982,1 |
| Weather 03 | 19.6.2016 | 19,6 | 20,1 | 25,8 | 18,9 | 20,9 | 26,8 | 14,8 | 48 | 53 | 28 | 52 | 980,3 | 979,8 |
| Weather 03 | 20.6.2016 | 16,5 | 14,0 | 19,7 | 14,2 | 15,5 | 21,6 | 12,9 | 67 | 82 | 40 | 61 | 976,9 | 977,9 |
| Weather 03 | 21.6.2016 | 14,1 | 14,0 | 19,5 | 14,2 | 15,5 | 21,6 | 10,2 | 64 | 67 | 40 | 57 | 979,9 | 980,7 |
| Weather 03 | 22.6.2016 | 15,3 | 16,3 | 24,3 | 18,9 | 19,6 | 25,8 | 10,3 | 58 | 59 | 29 | 44 | 982,2 | 982,6 |
| Weather 03 | 23.6.2016 | 19,1 | 19,3 | 26,2 | 21,3 | 22,0 | 26,9 | 14,8 | 47 | 49 | 31 | 46 | 983,5 | 984,6 |
| Weather 03 | 24.6.2016 | 20,9 | 20,5 | 22,4 | 17,5 | 19,5 | 22,9 | 14,2 | 42 | 37 | 38 | 60 | 979,7 | 978,9 |
| Weather 03 | 25.6.2016 | 16,9 | 16,3 | 24,8 | 13,3 | 16,9 | 25,9 | 12,5 | 60 | 59 | 37 | 83 | 976,6 | 974,6 |
| Weather 03 | 26.6.2016 | 12,7 | 12,1 | 21,4 | 15,9 | 16,3 | 22,5 | 10,3 | 82 | 80 | 41 | 57 | 975,9 | 977,9 |
| Weather 03 | 27.6.2016 | 17,0 | 18,0 | 26,2 | 19,1 | 20,6 | 26,8 | 12,8 | 55 | 52 | 19 | 55 | 981,5 | 982,6 |
| Weather 03 | 28.6.2016 | 18,2 | 17,3 | 28,4 | 22,1 | 22,5 | 28,8 | 13,8 | 59 | 63 | 38 | 54 | 982,7 | 982,5 |
| Weather 03 | 29.6.2016 | 21,2 | 20,3 | 26,0 | 16,2 | 19,7 | 29,6 | 16,0 | 55 | 55 | 50 | 87 | 976,7 | 975,3 |

**Fig. 3.** Example of data set from weather station 03

## 4 Data Integration and Transformation Process

This process was performed using the RapidMiner [8] software platform that provides various tools and operators suitable for datasets modifications. The model for this transformations was captured on Fig. 4.

The transformation process can be divided into two distinct parts–data from the production process and weather data. Data collected from the production process need to be cleaner from unused values. The source data were incomplete due to the outages of sensor and measuring points, therefore we had to replace the missing data with the values approximated from the existing records. After these steps the production data were prepared for integration with the weather stations data. It should be noted that these simple steps are enough, mainly because the appropriate initial production data quality, since they served as a basis for various previous analysis.

Due to this fact, the processing of weather data was more challenging. Data obtained from the weather stations 01 and 02 contains weather values (parameters) that are measured, collected and stored every hour. Every record represents weather values from the specific hour of the day. However, weather station 03 uses different data collection approach. The data at this station are measured, collected and stored four times a day, at specific intervals–at 00:00, 6:00, 12:00 and 18:00. These values represent one record in the collected data, as shown on Fig. 3. Due to this two different approaches, we have to define algorithms that convert the data from weather station 03 and weather stations 01 and 02 to a common format that can be used in the proposed to supplement the production data. These algorithms were implemented using the Python

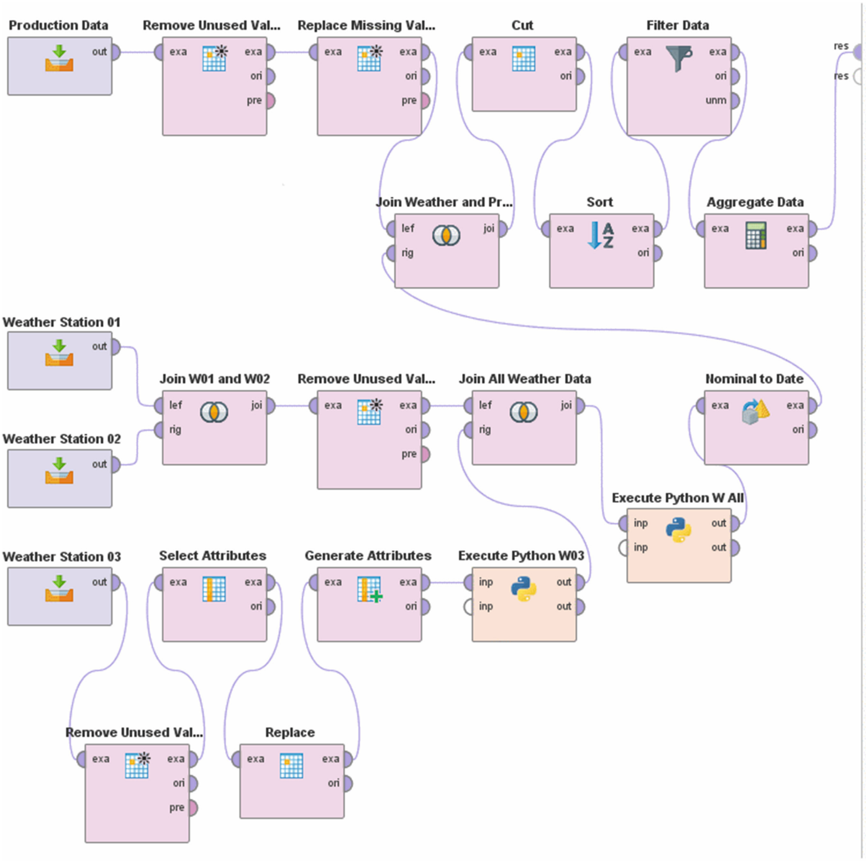


**Fig. 4.** Data integration model

programing language [9], utilising the Pandas library [10], and integrated into the transformation process.

## 5 Results

For the initial data understanding it was necessary to define use cases for future production process analysis. The production company did not provided specific use cases that should be improved or analysed. However there were defined few unconfirmed hypotheses about certain correlations.

In this paper we have described first use case that could serve as a basis for future knowledge discovery process. It should be noted that some details presented in this cases had to be anonymised, due to the company data policy.

Graph on Fig. 5, shows average daily temperature and deviation of number of manufactured products from the scheduled production plan for specific day.

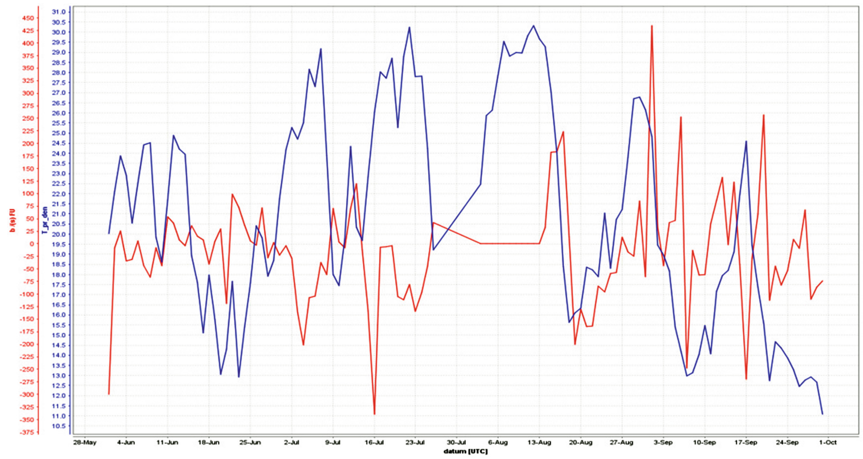


**Fig. 5.** Average daily temperature and deviation of number of manufactured products from the scheduled production plan

The graph clearly shows that there is a correlation between average daily temperature and deviation from the scheduled production plan through time in some cases.

The horizontal axis shows individual days from interval for which we have collected and obtained production and weather data, i.e. period between 1.6.2016 and 30.9.2016. It should be noted that the data on the horizontal axis are approximated to fit on the graph.

Numbers on the vertical axis show average daily temperature, represented by using blue colour, calculated from the weather data collected in 24 h. These values are supplemented by the production data, namely deviation between number of manufactured products and scheduled production plan for specific day. These data are collected from the real production process. It should be noted that the data need to be anonymised, due to the company data policy. Both of these values use linear scale, therefore the values are directly comparable.

The correlation between the average daily temperature and deviation from the scheduled production plan is clearly visible not only from this chart, but also from more detailed views.

We can assume, that if temperature rises above 30°C for 4 consecutive days (26°C in average), the production of the company will be reduced. The impact is different for each produced product. Production of products in first category is reduced by 8.97%, in the second category by 2.34% and in the third category by 7.28%. These statement is valid in 18 days in the year 2016, what caused that the production has produced 1,374 products less, than planned. These correlations could serve as a basis for further data analysis and subsequent knowledge discovery process.

# 6 Conclusion

The paper presents initials steps performed in the data mining process on real production data from the production company. It focuses on the data collection and integration phases that must be carried out before the modelling phase, according to CRISP-DM methodology.

The emphasis in the proposed data processing and transformation process is given on the data integration steps for the production and weather data. The data integration and transformation process, was captured as model utilising RapidMiner software platform. This model represents the creation of dataset that could serve as a basis for subsequent data mining activities.

Presented use case will serve as a basis for further detailed analysis and application of knowledge discovery process.

Considering the obtained results from the analysis, we can conclude, that there is significant impact of temperature on the overall production, and related production quality. Given this results, we are considering further expanding our model for the collection and integration of other weather data from inside and outside of the production hall, as well as further analysis of potential impact of other environmental parameters, e.g. humidity, atmospheric pressure, dust, etc. It should also be possible to analyse impact on overall production quality and improvement of quality KPI in the production company.

**Acknowledgments.** This publication is the result of implementation of the project VEGA 1/0673/15: "Knowledge discovery for hierarchical control of technological and production processes" supported by the VEGA.

This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO"(ITMS: 26220220179) supported by the Research & Development Operational Program funded by the EFRR.

# References

1. Trnka, A.: Control of production processes with selected data mining algorithms. Infokommunikacionnye technologii v nauke, proizvodstve i obrazovanii: četvertaja meždunarodnaja naučno-techničeskaja konferencija **1**, 183–188 (2010)
2. Bauernhansl, T., Hompel, M., Vogelheuser, B.: Industrie 4.0 in Produktion, Automatisierung und Logistik. Anwendung, Technologien, Migration. Springer Fachmedien, Wiesbaden (2014)
3. Hermann, M., Pentek, T., Otto, B.: Design Principles for Industrie 4.0 Scenarios: A Literature Review. Technische Universitat Dortmund (2015)
4. Da Cunha, C., Agard, B., Kusiak, A.: Data mining for improvement of product quality. Int. J. Prod. Res. **44**, 4027–4041 (2006)
5. Hazen, B.T., Boone, Ch.A., Ezell, J.D., Jones-Farmer, L.A.: Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. Int. J. Prod. Econ. **154**, 72–80 (2014)
6. IBM: IBM SPSS Modeler CRISP-DM Guide. IBM Software Group, Chicago (2011)

7. Parmenter, D.: Key Performance Indicators (KPI): Developing, Implementing, and Using Winning KPIs. Wiley, New Jersey (2007)
8. RapidMiner: Data Science Platform, https://rapidminer.com
9. Python Software Foundation, https://www.python.org
10. Pandas - Python Data Analysis Library, http://pandas.pydata.org

# Software-Defined Data Formats in Telecommunication Systems

Sergey V. Kuleshov, Alexey J. Aksenov,
and Alexandra A. Zaytseva(✉)

St.-Petersburg Institution for Informatics and Automation of RAS,
St.-Petersburg, Russia
{kuleshov, a_aksenov, cher}@iias.spb.su

**Abstract.** The paper considers an approach to solving the problem of formats succession and agreement in software reconfigurable telecommunication systems. This problem is relevant for many modern telecommunication systems characterized by intensive growth. Several approaches are proposed: format metadescriptions, hierarchical format analysis, active data concept.

**Keywords:** Telecommunication systems · Software-defined systems · Active data · Softwarization

## 1 Introduction

Expansion of new information technologies in society causes the need for their inclusion in the research paradigm of almost all branches of science especially in telecommunication systems.

An important trend in the telecommunication systems market – is improving access to interactive communications by expansion of relatively inexpensive software services that comes in several directions. One of them – is distribution of a variety of video conferencing software clients to smartphones and tablets. Another trend is the «softwarization» which means the conversion to software form of all that may be devoid of its physical embodiment [1–4].

In the telecommunications sector softwarization can be achieved by virtualization of the content delivery channels. Virtualization – is the way to organize a set of physical resources or their logical configuration which gives any advent ages over the original configuration.

The traditional communication channels are targeted to a specific type of content which demands a specialized terminal communication equipment (telephone for voice, telegraph to send text messages, etc.).

Using softwarization principle in content delivery allows to convert the transmitted data into a universal form that can be transmitted through the universal infocommunication environment (Fig. 1). A combination of converters between content-oriented form (images, sound) and transport-oriented (universal) form and the physical components of the communication environment can be considered as a universal virtual channel [5]. The functions of terminal units (content visualization) in this case are

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_35

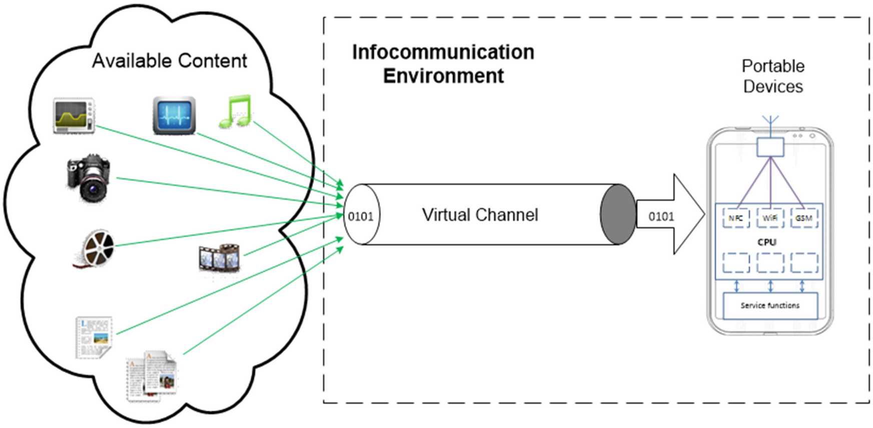


**Fig. 1.** Universal infocommunication environment

carried out by universal mobile devices (computers, smartphones) by implemented software.

Currently there are a set of software-defined technologies: Software-Defined Network (SDN), Software-Defined Data Center (Software-Defined Data Center - SDDC), Software-Defined Storage (SDS), Software-defined radio (SDR). It makes sense to consider all of these technologies as a single development direction of the software-reconfigurable environments.

## 2 The Proposed Approach

In some cases it is needed to provide access to telecommunications services in those geographic locations without public wireless network access and where the use of special radio or satellite systems is not possible. Often such access is required only for a limited amount of time without deployment of network base stations (e.g., for rescue operations).

This dynamic deployment of specialized data networks is possible by temporary reconfiguration of standard telecommunications devices (especially mobile) (Fig. 2).

The next approaches to implement dynamic temporary deployment are offered:

1. Using a regular data relay mode between mobile devices without reconfiguring.
2. The reconfiguration by switching the data relaying mode of consumer devices supporting such functionality by design.
3. Software reconfiguration of consumer units using the approach of active data for relaying data in the network of mobile devices. It has the most flexible options for configuring devices.

Approach 1 is only possible in networks built on the mesh-network technology which allows third-party traffic through the device network.

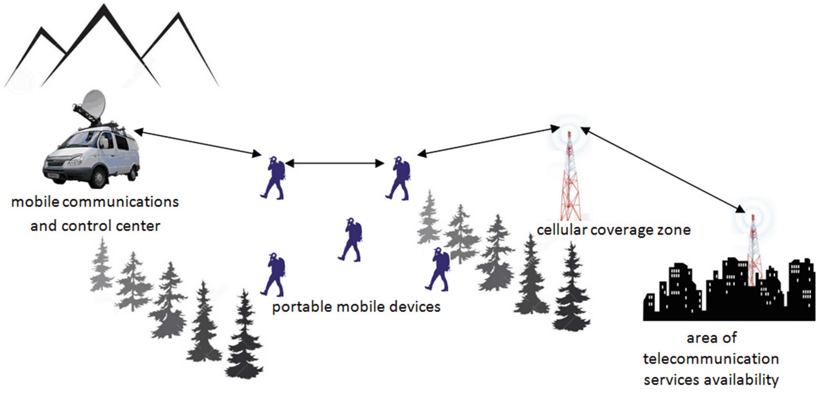


**Fig. 2.** The principle of virtual data channels deployment by reconfiguring of available consumer mobile devices

The main problem of approaches 2 and 3 is the need for hardware support of such reconfiguration (special relay mode or support for active data) by communication equipment manufacturers.

Another problem in the implementation of this approach is that the existing system of commercial cellular communication when working over long distances are focused on terminal↔base station interaction model, and the terminal↔terminal interaction is allowed only within the framework of nano- and pico-cells [6].

Software-defined system can remove these restrictions overriding terminal communication protocol and adjusting the characteristics of the equipment (given the availability of broadband receiver), which will take an “uplink” mobile channel and process it by receivers of the consumer devices.

## 3 Reconfiguration of Formats and Protocols

Reconfiguration of formats and protocols, on the one hand, is determined by the software parser functions which easily can be replaced. On the other hand, if the format is based on a data structure with a variable set of fields that are not fixed in size then the partial format correction in design-time operation is difficult, and in the run-time operation is often not possible.

Interpretation of the data streams can be carried out by the following software functions:

- Built on the basis of a formal description of a grammar or a finite state machine;
- Without the use of a formal stream description by direct reading and interpretation of individual values (the most common option).

To simplify the solution to the problem of format reconfiguration the following approaches based on software-defined systems are proposed:

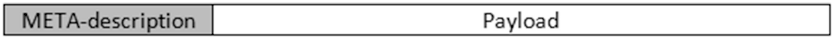


**Fig. 3.** Method of the meta-description injection

1. Introduction of a meta-description format placed in the format itself (Fig. 3). The main drawback - the increase in the volume of the data stream. Thus the meta-description format in such an approach requires a specification and so on ad infinitum. This solution will allow producing software reconfiguration of the formats both in design-time and run-time.
2. For design-time software reconfiguration we can use the principle of hierarchical containers i.e. the construction of the hierarchy format syntax elements in case of failure of the linear format wherever it is possible (Fig. 4). In this case the parser function is required to maintain the call hierarchy within the hierarchical structure analysis.

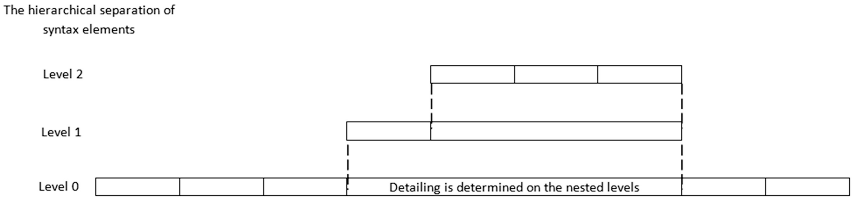

The hierarchical separation of syntax elements
Level 2
Level 1
Level 0
Detailing is determined on the nested levels

**Fig. 4.** Hierarchical separation of syntax elements

3. For run-time software reconfiguration the approach of active data (AD) can be used provided the principle of hierarchical containers in a configurable format is satisfied [7]. In this case executable AD blocks may be implemented as sub-functions each of which is responsible for parsing its syntactic level of format and can be replaced independently by software reconfiguration procedure (Fig. 5).

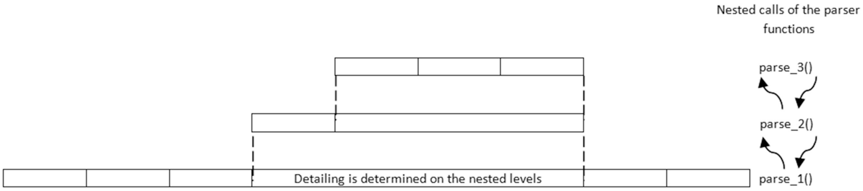

Nested calls of the parser functions
parse_3()
parse_2()
parse_1()
Detailing is determined on the nested levels

**Fig. 5.** Application of active data concept to format reconfiguration

## 4 Conclusion

The paper describes approaches to development of software-defined communication systems including active data concept. It is shown that the application of the concept of active data while respecting the principle of hierarchical containers in a configurable format makes possible complete run-time reconfiguration. These approaches are of particular importance in addressing format inheritance problems in the areas of modern telecommunications systems which are characterized by the most intensive development and rapid obsolescence of communications equipment.

The proposed approach allows providing access to telecommunication services in geographic locations without public wireless network access and where the use of special radio or satellite systems is not possible. Often such access is required only for a limited amount of time without deployment of network base stations (e.g., for rescue operations).

**Acknowledgments.** This work is supported by the Programme of RAS III.3 “The Department of Nanotechnologies and Information Technologies”, project N 0073-2015-0007.

## References

1. Alexandrov, V.V., Kuleshov, S.V., Yusupov, R.M.: Software-defined environments technology and import substitution. J. Informatization Commun. **3**, 154–157 (2016). (In Russian)
2. Saariketo, M.: Imagining alternative agency in techno-society: outlining the basis of critical technology education (EN). In: Media Practice and Everyday Agency in Europe, pp. 129–138 (2014)
3. Kuleshov, S.V., Yusupov, R.M.: Is softwarization the way to import substitution? J. SPIIRAS Proc. **46**(3), 5–13 (2016). doi:10.15622/sp.46.1. (In Russian)
4. Pretz, K.: The “Softwarization” of telecommunications systems. Drivers include 5G technology and open-source software. http://theinstitute.ieee.org/ieee-roundup/blogs/blog/the-softwarization-of-telecommunications-systems
5. Kuleshov, S.V.: Hybrid codecs and their use in programmable digital data transmission channels. J. Inf. Measur. Control Syst. **5**(10), 41–45 (2012). (In Russian)
6. Bakin, E., Borisovskaya, A., Pastushok, I.: Analysis of capacity of Picocell with dominating video streaming traffic. In: Proceedings of the 15th Conference of Open Innovations Association, FRUCT 2014, pp. 3–8 (2014)
7. Alexandrov, V.V., Kuleshov, S.V., Zaytseva, A.A.: Active data in digital software defined systems based on SEMS structures. In: Gorodetskiy, A.E. (ed.) Smart Electromechanical Systems. SSDC, vol. 49, pp. 61–69. Springer, Cham (2016). doi:10.1007/978-3-319-27547-5_6

# SADI: Stochastic Approach to Compute Degree of Importance in Web-Based Information Propagation

Selva Kumar Shekar[1](✉), Kayarvizhy Nagappan[1], and Balaji Rajendran[2]

[1] Department of Computer Science and Engineering, BMS College of Engineering, Bengaluru, India
Selva.cse@bmsce.ac.in
[2] Center for Development of Advanced Computing, Bengaluru, India

**Abstract.** The problem of information propagation (IP) is being studied theoretically but its practical implementation is quite limited as there are many underlying challenges to be resolved. One core problem found in the analysis of IP in dynamic web-based networks (DWBN) such as in social networks is the lack of light weight mechanism to compute the effective node identity. This paper presents a framework using *Stochastic Approach to compute the Degree of Importance* (DoI) to explore the most influential nodes residing in the dynamic network. The approach explores the influential nodes in any form of operational states of the nodes using probability theory. The model is evaluated with a massive set of open large data of DWBN to validate its effectiveness with the execution time to compute DoI.

**Keywords:** Information propagation · Social network · Influential node · Web-based network · Dynamic network · Graph theory

## 1 Introduction

Even after more than two decades of usage of communication protocols and technologies, it is quite a challenging part to understand the underlying concept of Information Propagation (IP). Basically, IP not only deals with the flow of data but also it deals with various underlying data processing complexities on multiple nodes in any form of the network [1–3]. With the evolution of social network analysis [4], cloud computing [5], and big data analytics [6], there is a tremendous demand for knowing the actual picture of IP. In different forms of the networking system, there is different communication protocol applicable that majorly governs the way the information has to be routed [7]. At present, massive amounts of data is being generated almost every second that poses serious challenges to data miners [8] as there are exponentially larger amount of data redundancies [9]. This problem is more on any dynamic network, and the only way to solve this problem is to extract some latent patterns or compute certain significant co-relational factor [10]. Hence, it is essential to understand the node's significance in a dynamic network which is a troublesome task to do. However, by doing so, it provides a larger scope of visualizing the data formation, significance

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_36

factor, correlation, density, etc. that could potentially assist in applying certain analytics in future. There are no benchmark models associated with IP that makes the investigation even more challenging.

Hence, this paper presents a novel modeling of the IP in the context of dynamic web-based network (DWBN) that could explore certain significance factor to further assists in information dissemination process. Section 2 describes the related work towards web-based information propagation (WBIP), whereas Sects. 3 and 4 discuss the problem formulation and research methodology respectively. Section 5 deals with in-depth discussion of algorithm implementation. Model validation by response time evaluation as result analysis is discussed in Sect. 6 followed by conclusion in Sect. 7.

## 2 Related Work

This section discusses the existing research work being carried out toward information propagation. Most recently, Liu and Xu [11] have presented a framework for representing heterogeneous users and their information propagation mechanism. The prime objective is to discretize credible message from gossip-based messages. A case study of mobile ad-hoc network was considered to investigate information propagation by Liu and Kato [12] where a Markov chain approach was applied to perform communication after any form of disaster. Mahdizadehaghdam et al. [13] have carried out a predictive-study to investigate information diffusion using Kalman filter. An interesting approach of information propagation is discussed by Park et al. [14] where a machine learning approach was used for modeling the communication system among the drones. A statistical approach with supervised learning technique is implemented to further minimize errors in the prediction of information propagation modeling. A boosting technique of multiple views has been presented by Peng et al. [15] to formulate classification-based well-defined decision making in information propagation. Zhang et al. [16] has discussed the similar concept about vehicular communication with multiple-input multiple-output in order to obtained faster response time. Zhuang and Yagan [17] modeled a technique on multi-layered clustered network Lejun et al. [18] have investigated about the patterns of data flow in the email network along with an emphasis on user's data flow. Wang et al. [19] developed a framework to compute the mean velocity of the information passing between the nodes. The study of Zhang et al. [20] conceptualized a stochastic process which is further incorporated to compute the speed of data flow of the message passing between the vehicles in a vehicular network. A similar pattern of research goal was also formulated by Han and Yang [21] where they have presented a study considering cognitive network and a model of knowledge propagation among contextually cooperating cognitive agents was presented by Balaji et al. [22]. Teodorescu [23] have presented a framework for computing the density of a specific group e.g. friends from the social network. Information Propagation Analysis in Academic networks is carried out to find current trend in the research topics by SS Kumar et al. [24]. Subhankar et al. [25] Modeled the viewing and sharing structure of the popularity videos propagation in OSN. The study mainly aims to understand how the information flows among each actor in social network and carry with a then certain feature that is contextually identical to a certain attribute. Therefore in the recent times existing review studies witness different segments of information propagation models designed and integrated to the diverse area of networks.

## 3 Problem Description

After reviewing some of the recent related research work towards information propagation, it was found that majority of the research attempts are towards wireless communication system, especially the vehicular and ad-hoc network. There is less number of works being carried out towards dynamic web-based networking system, which may pose a significant challenge towards understanding the IP in the upcoming ubiquitous network. It was also observed that IP was less investigated on social network considering larger data sets. Hence, exploring an influential node in a social network by IP is challenge to find in the existing system. The next section discuses about the proposed model to overcome this problem in IP.

## 4 Proposed Methodology

The prime purpose of the proposed study is to evolve up with a novel design of information propagation in the web-based network. The prototype design model of the proposed system is shown in Fig. 1.

The proposed system uses stochastic approach considering different forms of operational states of the dynamic network. We apply probability to compute three different attributes of stochastic state i.e. transmittance probability, receiving probability, and preserving probability of any forms of information propagation within dynamic networks. We use graph theory to model information matrix where neighboring matrix and time instance plays a crucial role. Finally, we model and compute Degree of

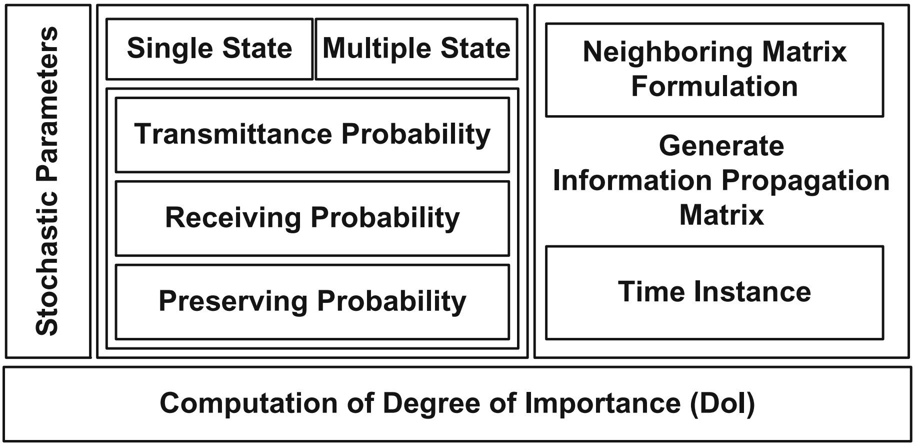

Stochastic Parameters
Single State
Multiple State
Transmittance Probability
Receiving Probability
Preserving Probability
Neighboring Matrix Formulation
Generate Information Propagation Matrix
Time Instance
Computation of Degree of Importance (DoI)

**Fig. 1.** Schematic architecture of proposed system

Importance (DoI) that signifies the extent of the significance of particular vertices for a given network. The next paragraph discusses the algorithm implementation.

## 5 Algorithm Implementation

The algorithm aims to implement a stochastic methodology over the web-based network integrated with information propagation. The prime contribution in this algorithm is its simple technique to compute a novel term called as Degree of Importance. The algorithm takes the input of $t_i$(time), $\eta_t$(Web-Based Network), $N$ (Neighboring matrix), $IP_{mat}$(Information Propagation Matrix), $\theta$ (Preserving Matrix), and $d$ (duration), which upon processing leads to generation of cumulative Degree of Importance (DoI). The steps involved in the algorithm are as follows:

```
Algorithm for Computing Degree of Importance in Information Propagation
Input:$t_i$, $\eta_t$, $N$, $IP_{mat}$, $\theta$, $d$
Output:DoI
Start
1. Init$t_i$, $\eta_t$, $N(t_i)$
2. For single state
3.    $IP_{mat} \rightarrow [p, q]$
4.    $DoI^1(IP_{mat}) \rightarrow \sum p.q^{n-1}.N(t_1)\ldots\ldots N(t_n)$.
5. End
6. For multiple state
7.    $IP_{mat} \rightarrow [p, q, r]$
8.    $DoI^2(IP_{Mat}) \rightarrow \sum p.q^{n-1}.\ \theta(t_{1,\,r})\ldots\ldots\ \theta(t_{n,}\, r)$
9. End
10. $DoI \rightarrow [\ DoI^1(IP_{mat}), DoI^2(IP_{mat})]$
End
```

The algorithm is to compute the DOI which could be incorporated into applications of IP in DWBN. A *graph-based* technique is used to model the network $\eta_t$ with specific vertices and edges existing among them for a given time instance $t_i$. A neighboring matrix $N(t_i)$ is computed that is considered to be equivalent to network graph $\eta_t$ (Line-1). The algorithm applies stochastic approach where various networks $\eta_{t1}, \eta_{t2}, \eta_{t3}, \ldots\eta_{tn}$ is considered to be evolved at a time $t$ (= $t_1, \ldots.t_n$).

The initial computation of the degree of importance will be carried out towards a single state of the network (Line-2). Here, we consider only the state with recent time stamp to understand the existing need of information propagation. So, we perform computation of two stochastic state $p$ and $q$ that corresponds with transmission and receiving of information to adjacent nodes at a particular time $t_i$ and $\mathrm{t_{i+1}}$ respectively. This computation will lead to formation of the matrix for information propagation (Line-3). It is also required to understand that the computation of $p$ and $q$ can be easily carried out using probability theory towards transmission and receiving information. However, both the factors (i.e. $p$ and $q$) are highly influenced by temporal and spatial parameters from the source node during dissemination the information in the network. The metric DoI, computes the anticipated quantity of information being transmitted by a sender node $s$ at a particular instance of time $t$ that traverses multiple hops ($s$, $d$) while finally received by destination node $d$. Therefore, formulation of DoI is carried out by developing a set of stochastic series consisting of states $p$ and $q$ along with neighboring nodes $N$ for a consecutive set of time $t_1$, $t_2 \ldots t_n$. Hence, generalization of this series gives the empirical representation of DoI considering single state (Line-4).

The next part of the algorithm will focus on computing DoI considering multiple stochastic states. The prime reason behind this consideration is that there are certain operational requirements in IP in web-based network that doesn't only depend on its single state, but may need to have an access to other heuristic state information. Such heuristics may be formed after completing a whole cycle of information propagation in one sub-set of network. It will also mean that there do exist lots of operation that will not only require its present state but will also need to have an access in its prior operational states. Hence, the matrix for information propagation (i.e. $\mathrm{IP_{mat}}$) is quite important in this regards as it keeps a record of complete state information during each cycle of computation. Therefore, the matrix will now consider a new stochastic state $r$ along with $p$ and $q$ (Line-7). The new stochastic state can be computed by estimating probability that a vertices conserve all the received information during specific instance of time $t_i$ till it ends its cycle of information propagation in time $t_{i+1}$. A new preserving matrix $\theta(\mathrm{t_n}, \mathrm{r})$ has a specific method of computation. $\theta(\mathrm{t_n}, \mathrm{r})$ is equivalent to generalized series form of $\mathrm{r^{n-1}N(t_i)}$. We use this in computing DoI as empirically shown in Line-8. Hence, the DoI in this part of algorithm depends mainly on transmitting probability $p$, receiving probability $q$, preserving probability $r$, and cumulative time duration $d$. Therefore, the anticipated extent of information propagation is calculated on the above mentioned approach.

A closer look into the algorithm structure will show that the algorithm is designed considering all the possible operational state of a web-based network in order to visualize the proposed modeling of IP. Analysis of the proposed algorithm can be carried out easily as DoI depends on multiple attributes that can be altered in order to ensure the best fit of the proposed algorithm in identifying the influential node. Hence, DoI assists in direct identification of the influential node and it is applicable on any types of large network that can be easily modeled using the proposed graph theory. Configuring the vertices and edges will be easier and so is the computation of two different forms of stochastic states in the web-based network. The next section discusses about the outcomes accomplished after implementing the proposed algorithm.

## 6 Results and Analysis

The system model is simulated using numerically controlled tool on both synthetic data as well as open source data available on Stanford Large Network Dataset Collection [26]. The runtime computational complexity is assessed by the response time on a fixed system simulation environment. The core data set of the social network of approximately of 4039 nodes and 88,234 edges. For simplicity of understanding the sample test data, is inference with different numbers of rows/edges (which also corresponds to the size of data) and its possible impact on response time. The response time is the time required to compute for DoI in all the stochastic states. Figure 2 showcase the response time of various sets of data on a fixed system environment of 64-bit machine with Core i7 processor. The overall processing of the whole 88,234 edges takes an average of 396 s using our approach - SADI. The effectiveness of SADI is computed with response time evolution with incremental rows and corresponding edges. Table 1 illustrates the observed values and the Fig. 2 a graph of No. of edges versus response time.

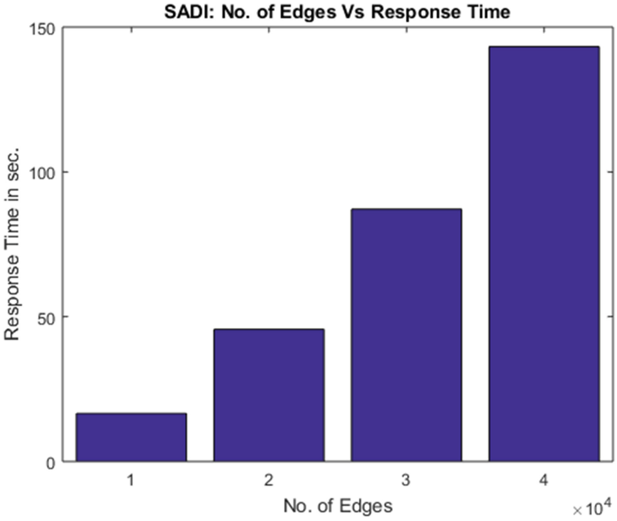


**Fig. 2.** A graph of no of edges versus response time in (sec)

The numerical outcome in Table 1 shows that consistency of the response time with increment of the number of edges or rows in consideration. The difference of the time between 10,000 edges to 20,000 edges is 29.06 s, between 20,000 to 30,000 is 41.47 and between 30,000 to 40,000 is 56.14. The trend exhibits that the pattern is almost consistent in variation, where initially it is more and further it is less. Hence it proves the stability of the system. The proposed approach, SADI, for computation of DoI can be benchmarked and can be established as sate of art work.

**Table 1.** An instance of observations of response time

| No. of edges | 10,000 | 20,000 | 30,000 | 40,000 |
|---|---|---|---|---|
| SADI | 16.59 | 45.65 | 87.12 | 143.26 |

## 7 Conclusion

With the increasing complexity of the data, it has become quite a challenging task for the researchers modeling towards information propagation. The proposed system introduces a new parameter called as DoI which will play a crucial role in analyzing any complex network. It will, therefore, have a bigger contribution towards computing significance factor of certain influential nodes in the dynamic network. The model uses stochastic approach which synchronizes with both time-series analysis and probability theory for IP problem. The unique contribution of the model is that it can perform in any operational state of the nodes and therefore it has fair chances to explore better routing path at the certain point of time from source to destination which is important in the formulation of application of IP. The proposed system is evaluated over massive data from the social network, and the study outcome shows a consistent behavior of the model with respect to the response time, which can be benchmarked for state of art method for DOI computation. In future, our stochastic approach to compute DOI can be enhanced with more complex network structures and can lead to better investigations on information propagation.

**Acknowledgment.** The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

## References

1. Geyer, R., Cairney, P.: Handbook on Complexity and Public Policy. Edward Elgar Publishing, Cheltenham (2015)
2. Wu, J., Wang, Y.: Opportunistic Mobile Social Networks. CRC Press, New York (2014)
3. Karyotis, V., Stai, E., Papavassiliou, S.: Evolutionary Dynamics of Complex Communications Networks. CRC Press, New York (2013)
4. Tayebi, M.A., Glässer, U.: Social Network Analysis in Predictive Policing: Concepts, Models and Methods. Springer, New York (2016)
5. Murugesan, S., Bojanova, I.: Encyclopedia of Cloud Computing. Wiley, New York (2016)
6. Marr, B.: Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Wiley, New York (2016)
7. Parsons, J.J.: New Perspectives on Computer Concepts 2016, Introductory. Cengage Learning-Computer, Boston (2015)
8. Chen, X., Vorvoreanu, M., Madhavan, K.: Mining social media data for understanding students' learning experiences. IEEE Trans. Learn. Technol. **7**(3), 246–259 (2014)
9. Gu, X., Yang, H., Tang, J., Zhang, J.: Web user profiling using data redundancy. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA (2016)

10. Jaradat, S., Dokoohaki, N., Matskin, M., Ferrari, E.: Trust and privacy correlations in social networks: a deep learning framework. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, pp. 203–206 (2016)
11. Liu, Y., Xu, S.: Detecting rumors through modeling information propagation networks in a social media environment. IEEE Trans. Comput. Soc. Syst. **3**(2), 46–62 (2016)
12. Liu, J., Kato, N.: A Markovian analysis for explicit probabilistic stopping-based information propagation in postdisaster ad hoc mobile networks. IEEE Trans. Wirel. Commun. **15**(1), 81–90 (2016)
13. Mahdizadehaghdam, S., Wang, H., Krim, H., Dai, L.: Information diffusion of topic propagation in social media. IEEE Trans. Sig. Inf. Process. Over Netw. **2**(4), 569–581 (2016)
14. Park, J., Kim, Y., Seok, J.: Prediction of information propagation in a drone network by using machine learning. In: 2016 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, South Korea, pp. 147–149 (2016)
15. Peng, J., Aved, A.J., Seetharaman, G., Palaniappan, K.: Multiview boosting with information propagation for classification. IEEE Trans. Neural Netw. Learn. Syst. **PP**(99), 1–13 (2017)
16. Zhang, Z., Wu, H., Zhang, H., Dai, H., Kato, N.: Virtual-MIMO-Boosted information propagation on highways. IEEE Trans. Wirel. Commun. **15**(2), 1420–1431 (2016)
17. Zhuang, Y., Yağan, O.: Information propagation in clustered multilayer networks. IEEE Trans. Netw. Sci. Eng. **3**(4), 211–224 (2016)
18. Zhang, L., Guo, L., Xu, L.: Research on e-mail communication network evolution model based on user information propagation. China Commun. **12**(7), 108–118 (2015)
19. Wang, W., Liao, S.S., Li, X., Ren, J.S.: The process of information propagation along a traffic stream through intervehicle communication. IEEE Trans. Intell. Transp. Syst. **15**(1), 345–354 (2014)
20. Zhang, Z., Mao, G., Anderson, B.D.O.: Stochastic characterization of information propagation process in vehicular ad hoc networks. IEEE Trans. Intell. Transp. Syst. **15**(1), 122–135 (2014)
21. Han, C., Yang, Y.: Understanding the information propagation speed in multihop cognitive radio networks. IEEE Trans. Mob. Comput. **12**(6), 1242–1255 (2013)
22. Rajendran, B., Iyakutti, K.: Contextually cooperating agents for user assistance in web-based knowledge gathering tasks. Int. J. Comput. Appl. (IJCA) **1**(23), 12–18 (2010)
23. Teodorescu, H.N.: On models of 'having friends' and SN friends distribution: information propagation on social networks and disaster modeling. In: 2016 International Conference on Control, Decision and Information Technologies (CoDIT), St. Julian's, pp. 659–664 (2016)
24. Kumar, S.S., Kumar, K.S., Kayarvizhy, N.: Analysis of information propagation in academic social networks. In: 2016 International Conference on Recent Trends in Information Technology (ICRTIT), pp. 1–4, Chennai (2016)
25. Ghosh, S., Kumar, S.S.: Video popularity distribution and propagation in social networks. Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTCS) **6**(1), 001–005 (2017). ISSN 2278-6856
26. Leskovec, J., Mcauley, J.J.: Learning to discover social circles in ego networks. In: Advances in Neural Information Processing Systems, pp. 539–547 (2012). http://snap.stanford.edu/data

# Actiontracking for Multi-platform Mobile Applications

Zdzisław Sroczyński(✉)

Institute of Mathematics, Silesian University of Technology,
23 Kaszubska Street, 44-100 Gliwice, Poland
zdzislaw.sroczynski@polsl.pl

**Abstract.** The article contains the proposal of the standard log file format for activity tracking purposes. The retrieval process for Activity Markup Language (ActML) in different software environments was discussed, alongside with the test suite of multi-platform mobile applications. The next part of the article deals with automatic comparison of acquired actiontracking logs in order to compute the coefficient of the difference between experimental and reference action paths. The proposed approach provides new insights in the field of the usability evaluation, especially for mobile and multi-platform applications.

## 1 Introduction

The research works in the field of the human-computer interaction (HCI) are conducted since the creation of the very first computerized man-machine interfaces. Alongside with the new focus of the development of mobile, touch and gesture controlled applications, the issues of HCI are of interest for majority of software vendors, as well as scientists working in the interdisciplinary fields of knowledge [10].

The main topics of the HCI discipline are the usability, which equates to overall quality, somewhat similar user experience (UX) and the design of the user interface (UI), which is graphical one in the basic version, but may become a very sophisticated brain-machine one in the close future. The usability of the computer system has some attributes, identified slightly different in the literature. The best general reference here is [7], which distinguishes:

1. *efficiency* – how easy is to archive the goal,
2. *satisfaction* – freedom from discomfort and positive attitude to the product,
3. *learnability* – the easiness to learn and start using the product,
4. *memorability* – the easiness to return and operate the system after some break,
5. *faultlessness* – low error rate and the ability to recover.

These attributes can be estimated by the use of particular performance and issue-based metrics [1]. The main sources of the data for further processing

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_37

are: expert analysis, user surveys and manual or automated observation of the user. The user's activities can be logged through click-tracking, eyetracking or actiontracking. The first two of the methods give the information about places of interest inside the UI, but the data is not connected with internal logic of the application. Therefore, an expert is needed to interpret the results of these trackings. More extensive discussion of the issues with UX evaluation from the perspective of the user centred design is presented in [11], while UI prototyping techniques are elaborated in [18].

There are also several well elaborated coefficients describing particular aspects of the user experience, as Fitt's law, referring the trade-off between speed and accuracy when pointing the target in the UI [3], gestures monitoring requiring special hardware environment [17] and some specific coefficients referring for example the "lostness" level while web page navigation [12]. However, these solutions are not applicable or hard to adapt in the general case.

Having the data about user's actions combined with the internal knowledge about the structure and event handling in the application, evaluation of the quality of human-computer interaction for that particular application would be much more accurate. This approach, called actiontracking, is most often used for web solutions [2,13], but it is relevant to every kind of software, especially for popular mobile and multi-platform apps. The Studies in this area are conducted from the very beginning of the mobile terminals popularity [4,9], when the researches were answering the questions about preferred interaction modes. Nowadays the investigations focus at an unsupervised monitoring and logging of user activities [5] and novel models of usage [8].

The evaluation of the UX for a given application should take into account the main goal which is a better usability of the app. It should help to determine:

- what UX is more clear and readable for the user?
- which UX ensures that goals will be reached faster and with better precision?
- which UX causes less positioning errors?
- how should be placed GUI elements and how far from each other?
- what should be changed for mobile apps, web one and desktop ones in multi-platform projects?

In the easiest approach, there is a need for time-taking statistical surveys of the software users to answer above questions. Moreover, the surveys can be subject to errors due to a subjective attitude of queried people. So, the possibility to collect the data of user interactions automatically would be a step forward. The next level in this process is automatic comparison of collected data. The complete system for independent and objective quality assessment would be created this way.

To achieve the best possible results, tracking of user's activities should be carried out in the transparent way, not disturbing the operator. As the amount of the data about actions can be huge, the second important feature should be the ability to compare the sequences of activities in the automatic way, without the assist of the expert or author of the software. Thus obtained information about the correspondence of the task performance with the previously saved pattern of the proper action path can provide a convenient solution to evaluate the quality of the HCI.

Let us introduce a new XML-based universal notation to provide a convenient method of activity data processing. Thanks to that notation it will be possible to acquire, analyse and compare the data about user's activities at different levels, getting the correct general conclusions. We will call this notation Activity Markup Language (ActML) and describe it in details in the next section.

Next, we propose to extend one of the popular multi-platform software frameworks, i.e. FMX library used in RAD Studio development environment [14], to provide automatic logging of users' actions in the ActML format. This extension utilizes dynamic code analysis and RTTI, so there is no need to provide extra commands or configuration setting for the application. Moreover, the logging process automatically includes all the activities, regardless their source or author intentions, and that is one more possible advantage of our solution.

Having the users' activities collected, we can compare them with the patterns recorded before by experts or experienced users. The difference between these "proper" action paths for given tasks can be the measure of UX consistency and quality. This methodology compares activities of experts and common users assuming that high quality user interaction design should lead to proper activities regardless how reach is the experience with the particular system.

The introduced actiontracking method was tested on three applications by respondents of different age, some of them more familiar with mobile technologies than others. The overall results are promising, as it was easy to point the flaws in the UX design, as well to distinguish the beginner users.

## 2 Activity Markup Language

Activity Markup Language (ActML) is a language to describe all the user's activities during a session with the particular software system. It is based on XML and thanks to that documents encoded with it can be processed easily. Furthermore it is universal, i.e. can be expanded with new tags or attributes not breaking down the compatibility with existing analysis algorithms and applications. On the other hand, there is a possibility to reduce a structure of saved ActML documents discarding some attributes, when needed.

The following source code contains an example of the actiontracking for a simple and short session saved in ActML format:

```
<?xml version=”1.0” encoding=”UTF−8” standalone=”no” ?>
<actions>
 <actionset time=”5676359”>
  <action control=”ButtonAddValue.OnClick” time=”5681439”/>
  <action control=”TrackBarItemHeight.OnChange”
                                   value=”30” time=”5685727”/>
  <action control=”ButtonFinish.OnClick” time=”5687619”/>
 </actionset>
</actions>
```

There are many actionsets allowed in one "actions" collection. Each actionset describes a session of the user's interaction, which begins in the particular timestamp, denoted as the attribute of the tag. Actions describe every activity of the user, all the interactions with the visual components of the GUI. The main attribute of the action tag is control, which contains the name of the control and the name of the method called. This names are concatenated with a dot sign similarly to the common object notation. In addiction there is a timestamp in the end of all action tags. Additionally, there are attributes called "value" containing values of data entered or key positions of the controls. This way an information about what happened to what controls and when is stored in ActML log file. The timestamps are especially important, as they provide information about the level of confidence and memorization of the user.

The time attribute helps to estimate the value of the time interval between subsequent actions. The result can be used to measure the level of the operational reliability of the system operator. In a common scenario, the longer intervals mean the less certain user's actions, caused by a lack of knowledge, not enough experience or system misdesign.

Summarizing Activity Markup Language (ActML) specification there are some factors to underscore:

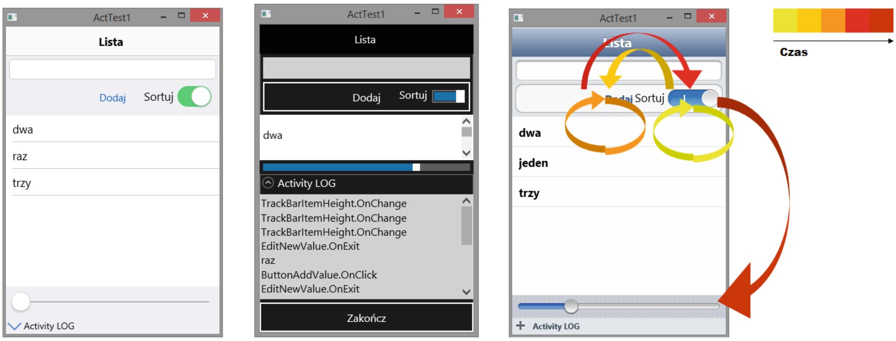
- ActML file contains the information about all activities initiated by the user, having corresponding event handlers in the application,
- data should be collected in automatic way, as it is intended for machine evaluation afterwards (that is a significant difference between usage of ActML and the other techniques of investigation user's actions, as for example Test Driven Development – TDD),
- the usage data in ActML can be generated by the evaluated application itself or through additional external software, for example a specially modified profiler,
- ActML structure is simple and does not cause any significant overhead,
- actionset is a list of events with timestamp,
- it is possible to acquire complete flow chart of actions taken by the user during work session by processing subsequent control attribute values.

### 2.1 Examples of ActML Usage

There is a simple example of multiplatform application shown in the Fig. 1. The app view contains text input control with an "add" button below it. Next there is a list of items, which can be sorted when a switch is turned on. The switch is improperly placed on the right side of the button, but this design is planned – it allows to test possible errors of positioning user's fingertips. In the bottom of the view there is a slider (TrackBar) determining the sizes of the list elements[1].

On the very bottom of the form view there is an additional control signed "Activity LOG". This control expands to the list of entries equivalent to user's

[1] all the example applications shown in the article are in Polish, as this was the actual language of the conducted experiments.



```
TrackBarItemHeight.OnChange
TrackBarItemHeight.OnChange
TrackBarItemHeight.OnChange
EditNewValue.OnExit
raz
ButtonAddValue.OnClick
EditNewValue.OnExit
```

**Fig. 1.** Sample GUI of the mobile application themed according to new iOS flat guidelines (on the left); the same application themed like Windows Metro, with activity log open (in the middle); activity tracking transition graph for the example session for the application with the iOS 6 skeumorphic look-and-feel (on the right).

actions and action tags in ActML file. This helps in debugging the application, but of course is unnecessary to build the proper XML activity log file. More complex application, or that ones which cannot allocate the space in the view project (as games), do not need visible activity log at all.

The following listing contains the transcript from the short session with the application described above. The user turned the switch twice, then added item to the list, turned the switch one again and moved the slider slightly to the right.

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<actionset time="14631046">
 <action control="SwitchSort.OnSwitch" time="14632648"/>
 <action control="SwitchSort.OnSwitch" time="14633460"/>
 <action control="EditNewValue.OnExit"
                     value="raz" time="14636152"/>
 <action control="ButtonAddValue.OnClick" time="14637172"/>
 <action control="SwitchSort.OnSwitch" time="14642186"/>
 <action control="TrackBarItemHeight.OnChange"
                     value="45" time="14647734"/>
</actionset>
```

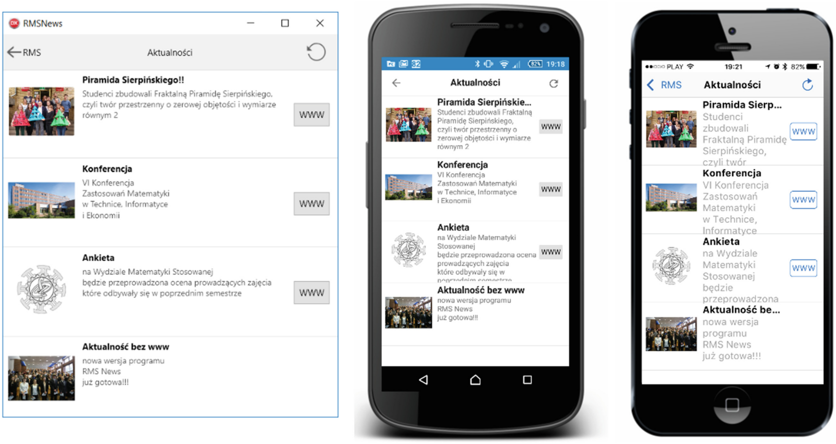
The structure of ActML file allows to reconstruct the transition graph of the user activities. Activity tracking acquired this way can highlight the aspects not visible for the other methods, based on external monitoring. This concerns for example gestures, which are much more harder to identify without very close interaction with the operating system. The ActML approach assures that all the gestures detected by the gesture manager will by logged right after they were dispatched into the monitored application.

The exemplary transition graph for activity tracking is shown in the Fig. 1 above.

## 2.2 Collecting ActML Data

ActML data can be logged in any working application with minor effort from its author, extending event handlers with the log code. Moreover, the developer can use a special library assuring automatic generation of the xml log code for all the events in the application. This way ActML log is created completely unattended without changing the source code of the application being analysed. Technically it can be done with the use of RTTI (Real-Time Type Information) technology – for example in projects written in C++ or Object Pascal, and similar solutions called Reflection in the other languages, as for example Java or C#.

This approach gives better results for activity tracking than unit tests and TDD (Test Driven Development), because every single activity is logged, regardless of the developer's focus at particular problems, which is determined by his personal previous experience. Automatic addition of activity entries to the log file does not require any changes in the application source code, nor a cooperation between the developer and GUI designer, and leaves the overall design of the user interface intact (although there is mentioned above the dedicated extra bottom section of the view called "Activity LOG" in our examples, it is only for sorely special debugging purposes).



**Piramida Sierpińskiego!!**
Studenci zbudowali Fraktalną Piramidę Sierpińskiego, czyli twór przestrzenny o zerowej objętości i wymiarze równym 2

**Konferencja**
VI Konferencja
Zastosowań Matematyki
w Technice, Informatyce
i Ekonomii

**Ankieta**
na Wydziale Matematyki Stosowanej
będzie przeprowadzona ocena prowadzących zajęcia
które odbywały się w poprzednim semestrze

**Aktualność bez www**
nowa wersja programu
RMS News
już gotowa!!!

**Fig. 2.** The test application themed according to Windows Metro Design (left), Android standard look-and-feel (middle) and iOS guidelines (right).

The example application shown in the Fig. 2 was compiled with the newest Embarcadero RAD Studio 10.1 Berlin cross-compiler set for MS Windows, with themes enabled for different operating systems (iOS, Android, Windows Metro) [15]. It is possible thanks to the special modification of style files by unlocking target systems restriction. There is a possibility to test and track activity for the application designed for mobile platforms running it in the Windows environment, preserving the proper look-and-feel of the destination operating system in the same time this way. Thanks to multiplatform capabilities of Rad Studio the application with one shared source codebase can be compiled into the machine code of the destination mobile operating system, but we found it much easier to collect the actiontracking data using the fast deployment process at Windows developer machine, especially in case of business, listview-based applications.

On the other hand, there are no contraindications to install, run and test binary versions of the application directly at mobile devices running the desired operating system and collect the ActML data with the use of appropriate input touch screen and other sensors specific for a mobile device. Only slight organizational problems with the transfer of the resulting xml log files may appear, caused by closed logical architecture of some operating systems (for example application sand-boxing in iOS).

## 3 ActML Action Tracks Analysis

The main aim of collecting ActML data is to compare user's activities during different sessions with the software. It would allow to detect sequences of events similar one to another. The recognition of such patterns of activities would be helpful to identify moments when the user is performing well, as well as these ones, when the user is confused or even makes mistakes. This way there should be a chance to distinguish experienced users from novices, elder from young ones, or people with disabilities. Based on these conclusions, the system can eventually provide a different, adjusted interaction especially designed for particular needs.

The procedure for the comparison and recognition of the activities requires pre-recording of the reference sequence by the trained, expert operator. After that the comparison process can be done without personal effort from the expert. The comparison is also possible for different configurations of the analysed application and for many different users at the same time.

The reference actiontracking log contains the minimal set of single operations needed to accomplish given task. Of course, the real user can perform slightly different, fulfilling some activities in different order or even making less important errors, resulting in unnecessary actions saved in the activity log. The exact action-to-action comparison in these circumstances would lead to incorrect conclusions. To determine if the goal was actually reached, we need a method to compare the pattern with experimental ActML logs automatically in the approximate way.

That is possible to compute the comparison factor for action logs, focusing at the sequence of events, from the ActML data after the removal of timestamps. Time-stamps are very useful for examining the efficiency of the user, but they do not matter when the general goal of the task is taken into account. The ActML data from different actionsets should be compared in the flexible way, and the solution here can be a string metric called Levenshtein distance, widely used for approximate comparison of complex and graph data structures [16].

Levenshtein edit distance $D_{Le}(x, y)$ is a measure of the difference between two finite sequences of characters. Levenshtein distance between two character strings is the minimum number of single-char edit operations, required to change one string ($x$) into another ($y$) [6]. The operation categories are: insertions, deletions and replacements, in our case all having the same weights equal 1.

The distance $D_{Le}(a_e, a_i)$ (where $a_e$ is the ordered, concatenated set of actions recorded by the expert user, and $a_i$ is the examined sequence from the $i$th ActML log) is the measure of difference between the expert pattern activity and particular user's actions.

Because actionset list of entries, describing particular events, can have different lengths, corresponding to the different numbers of events, the following normalization of the $D_{Le}$ coefficient puts its values in the range 0–100%:

$$D_{LeN} = 100 \times \frac{D_{Le}}{\max\left(|a_e|, |a_i|\right)}$$

where $|s|$ is number of characters in string $s$.

The result is the measure of difference of two sets of actions obtained from ActML log. The smaller $D_{LeN}$, the more similar is given $a_i$ sequence to the pattern $a_e$, i.e. the respondent behaves more like an expert.

The assumption $D_{LeN} < 10[\%]$ is sufficient for the identification of very similar activities in the majority of test cases. Lower values from the range $\langle 10; 30\rangle[\%]$ may correspond to somewhat important analogies in the sequences of actions.

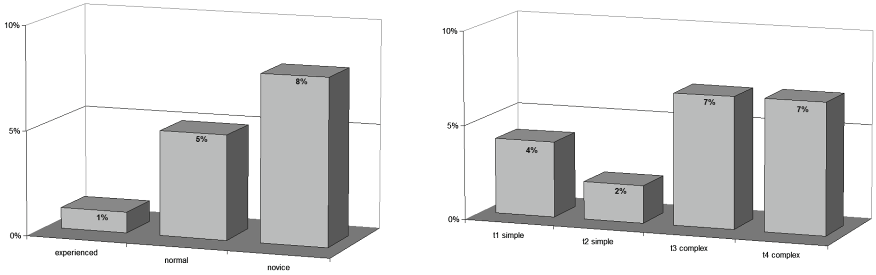


**Fig. 3.** Average $D_{LeN}$: for users with different skills (on the left), for tasks of different complexity (on the right); experiment involved 10 persons divided into 3 classes (experts, normal, novices) and 4 tasks – 2 easy, 2 more complex.

We have conducted two-phase experiment with applications of different complexity: very simple mock-up multi-platform app, more complex multi-platform app with several separated views, and mobile game. The results of tests proved that normalized Levenshtein edit distance measure $D_{LeN}$ is an efficient tool to recognize similar sequences of actions made by users of the software.

There were 10 people involved into the experiment, including three experts, three experienced mobile users and four novice users, completely not familiar with the surveyed applications. Users had to perform two kinds of tasks:

simple (consisting of single actions) and moderately complex (several actions). These moderately complex tasks were for example “check if there are any news”, “make the call to the contact number in the app”, “show the third photograph in the gallery” or “turn all the sounds in the game off”. For every task the reference sequence was previously recorded by a skilled operator. While all the actionsets in the experiments were recognized as similar to the reference ones, the average $D_{LeN}$ coefficient varied depending on user’s experience and the difficulty of tasks (Fig. 3).

Having such a tool, able to automatically identify an approximate match for a given sequence of actions, is possible to detect the level of maturity for the user and scale the complexity of the task. This – at the very end – can help to answer the key questions raised in the Introduction properly rating the overall user experience and building adaptable user interfaces, protecting users against common and predictable errors. Thanks to unified ActML model, automatic acquisition and normalized edit distance measure this objectives can be achieved easily even for production applications, without complicated monitoring environment or extensive refactoring of the application source code.

## 4 Conclusions

There is a concept of actiontracking for multi-platform applications presented in the paper. ActML – the novel format for logging the data about user’s activities was introduced. In the next part the method of measure the similarity of actions taken by the application operator was proposed, with the use of normalized Levenshtein edit distance coefficient.

The application test suite was developed using multi-platform software tools, which gives solid base for thorough investigations of user experience at different operating systems and different user interface designs. The following considerations proved, that modern software engineering methods can help in a fully automatic acquisition of activity datasets, making usability analysis possible even for complex, production level applications. Thus, the usability testing will not need a complicated hardware or software environment and moreover – the results of action comparison can be utilized on-line for adaptation of the user interface or supporting the user to avoid severe errors.

There was only a part of the proposed ActML format used in the experiments, and still unanswered questions concern memorability and learning issues for the given application. They can be answered by analysing the timestamps recorded in the actionsets. The implementation and evaluation of adaptive interfaces with the use of actiontracking is also an open issue, worth further investigations in more complex software projects and widened test cases.

## References

1. Assila, A., Oliveira, K.M., Ezzedine, H.: Towards indicators for HCI quality evaluation support. In: Indulska, M., Purao, S. (eds.) ER 2014. LNCS, vol. 8823, pp. 178–187. Springer, Cham (2014). doi:10.1007/978-3-319-12256-4_19

2. Atterer, R., Wnuk, M., Schmidt, A.: Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In: Proceedings of the 15th International Conference on World Wide Web, pp. 203–212. ACM (2006)
3. Goldberg, K., Faridani, S., Alterovitz, R.: A new derivation and dataset for Fitts' law of human motion. J. LATEX Class Files **6**(1), 1–14 (2007)
4. Kjedskov, J., Skov, M.B.: Interaction design for handheld computers. In: Proceedings of the 5th Asian Pacific Conference on Human-Computer Interaction, APCHI 2002 (2002)
5. Lettner, F., Holzmann, C.: Automated and unsupervised user interaction logging as basis for usability evaluation of mobile applications. In: Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia, MoMM 2012, pp. 118–127, New York, NY, USA. ACM (2012)
6. Navarro, G.: A guided tour to approximate string matching. ACM Comput. Surv. **33**(1), 31–88 (2001)
7. Nielsen, J., Budiu, R.: Mobile Usability. New Riders Press, Berkeley (2012)
8. Raptis, D., Kjeldskov, J., Skov, M.B.: Continuity in multi-device interaction: an online study. In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction, NordiCHI 2016, pp. 29:1–29:10, New York, NY, USA. ACM (2016)
9. Rukzio, E., Broll, G., Leichtenstern, K., Schmidt, A.: Mobile interaction with the real world: an evaluation and comparison of physical mobile interaction techniques. In: Schiele, B., Dey, A.K., Gellersen, H., Ruyter, B., Tscheligi, M., Wichert, R., Aarts, E., Buchmann, A. (eds.) AmI 2007. LNCS, vol. 4794, pp. 1–18. Springer, Heidelberg (2007). doi:10.1007/978-3-540-76652-0_1
10. Sikorski, M.: Human-Computer Interaction (in Polish). Polish-Japanese Academy of Information Technology Press, Warsaw (2010)
11. Sikorski, M.: A cross-disciplinary UX evaluation of a CRM system. In: I-UxSED, pp. 31–36 (2012)
12. Smith, P.A.: Towards a practical measure of hypertext usability. Interact. Comput. **8**(4), 365–381 (1996)
13. Sobecki, J., Żatuchin, D.: Knowledge and data processing in a process of website quality evaluation. In: Nguyen, N.T., Katarzyniak, R.P., Janiak, A. (eds.) New Challenges in Computational Collective Intelligence, pp. 51–61. Springer, Heidelberg (2009)
14. Sroczyński, Z.: Designing human-computer interaction for mobile devices with the FMX application platform. Theor. Appl. Inform. **26**(1–2), 87–104 (2014)
15. Sroczyński, Z.: Human-computer interaction on mobile devices with the FM application platform. In: Rostański, M., Pikiewicz, P. (eds.), Internet in the Information Society. Insights on the Information Systems, Structures and Applications. Academy of Business in Dabrowa Gornicza Press (2014)
16. Stapor, K.: Integration of structural pattern recognition methods into knowledge-based framework: application to geographic map image analysis. In: Studia Informatica. Silesian University of Technology Press (2000)
17. Szymanski, J.M., Sobecki, J., Chynal, P.: Actiontracking in gesture based information systems. In: Proceedings of the 2014 Mulitmedia, Interaction, Design and Innovation International Conference MIDI 2014, Warsaw, Poland, pp. 1–7, 24–25 June 2014
18. Weichbroth, P., Sikorski, M.: User interface prototyping, techniques, methods and tools. Studia Ekonomiczne **234**, 184–198 (2015)

# Towards Utilization of a Lean Canvas in the Testing Extra-Functional Properties

Padmaraj Nidagundi(✉) and Leonīds Novickis

Faculty of Computer Science and Information Technology,
Institute of Applied Computer Systems, Riga, Latvia
padmmaraj.nidagundi@gmail.com, lnovickis@gmail.com

**Abstract.** Everyday software usage is increasing with technological advancements and the deep integration of technology and human life. Moreover, software dependency also increases from day-to-day task to do it in an easy way. In a recent year software development has seen high growth in a number of tools and technology for building better software for end users. Software development adopted different approaches such as waterfall, prototyping, incremental development, iterative, spiral development, rapid application development, lightweight methodologies and other to develop software and at the same time software testing also needs to be done in the scope of time and budget. Software needs to be well tested in main functional and extra-functional properties requirements point of view. In software life cycle at some point it is important to consider the testing extra functional properties before delivering the software application to end user. It is important to create a basic level of checklist for the testing extra-functional properties to make sure the software is delivered error free. In testing, it is noticed that it is difficult to make test strategy, design tests cases in all possible way for the testing extra-functional properties. In such a scenario, the utilization of lean canvas to identify and use all testing extra-functional properties can make test planning and test design easier. It also helps to track and manage an individual tester's or test team's all possible testing extra-functional properties on one page. In business many years the lean canvas used only in business planning and strategy building, but well optimized lean principle adopted lean canvas board can bring many benefits for the testing extra-functional properties.

**Keywords:** Software quality assurance · Software testing · Testing extra-functional properties

## 1 Motivation and Introduction

It is well proven that error free software increases the user satisfaction and brings a greater business value, in this process software validation and verification play a key role. With the growing complexity of software, it becomes more challenging to all software development firms to provide the error free software in time and budget. These can direct impact on the test plan and process and tools used in the testing.

Software needs to be tested with well-defined functional requirements, but most of the time software development firms do not give much more value for extra-functional

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_38

testing. In another case, extra-functional testing contains various types such as performance, security usability and compatibility, etc. And most of the time quality control team needs to think and decide which testing type needs to follow the project.

## 2 The Problem Statement

When quality assurance team is testing extra-functional properties [1, 6], it faces many challenges while developing the test strategy for the software testing.

The most common well-known software extra-functional properties testing challenges are: Requirements are very generic with software application.

Lack of a more appropriate test environment and infrastructure. Application of performance criteria and the limited volume of sample data for testing. Tractability in between different types of testing. Understanding priority of the types of testing. Not having the proper test plan, test case, strategy, testability and it is hard to manage checklist to follow up testing. Limited test tools support & scope for the testing. Knowledge transfers challenge to the team member about collaboration and tests the software application in short time.

## 3 The Problem Statement

Since the introduction of business model canvas in the year 2008, it has still been used only for the strategic management and lean startup for business planning, making strategy and documenting them. The business model canvas provides building blocks for the business activities to bring them on one page.

This paper's main goal is to find the possibilities to adopt the lean canvas for identifying and visualizing on one page for testing extra-functional properties. Specifically:

- Possibilities of adopting the lean canvas design for the testing extra-functional properties.
- Improving test planning and test design with help lean canvas.
- Test strategy improvements.
- Finding possibilities for mapping all testing components.
- Adopting lean principles for the lean canvas design to find the test metrics.
- To measure the capacity, speed, security, availability, scalability, etc.
- Finding the most appropriate blocks from types of testing for the lean canvas design.

### 3.1 Related Research

Alex Osterwalder [7] with his co-authors mainly used business canvas to developing and testing a business idea by visualizing it on one page. This will outline the key points related to business such as opportunities, activities, financial projections and strategies.

The main significant advantages of the lean canvas are focus, speed & agility and common language [3].

The life cycle business model canvas contains six phases and iterating itself - idea, build, product, measure, data and learn (Fig. 2).

The main focused goal of the business model canvas is to identify the core objects of any business. It starts with a business idea and loop ends with learning. At each phase it removes the unwanted things from the loop. Each phase provides the constant feedback.

### 3.2 Extra-Functional Properties and Lean Canvas Life Cycle

Figure 1 shows the extra-functional testing possibilities with any system, considering these as an input and adding to the business model canvas, we are able to identify the similar terminologies. Now imagine we are testing a web application developed for the school management using cloud solutions.

- Ideas - Current software application needs testing of load/performance, compatibility, localization, security, etc.
- Build - Cloud architectures and its components.
- Product - School management application features and properties.
- Measure - Number of a concurrent request from users, transactions per XSS and SQL.
- injection testing, etc.

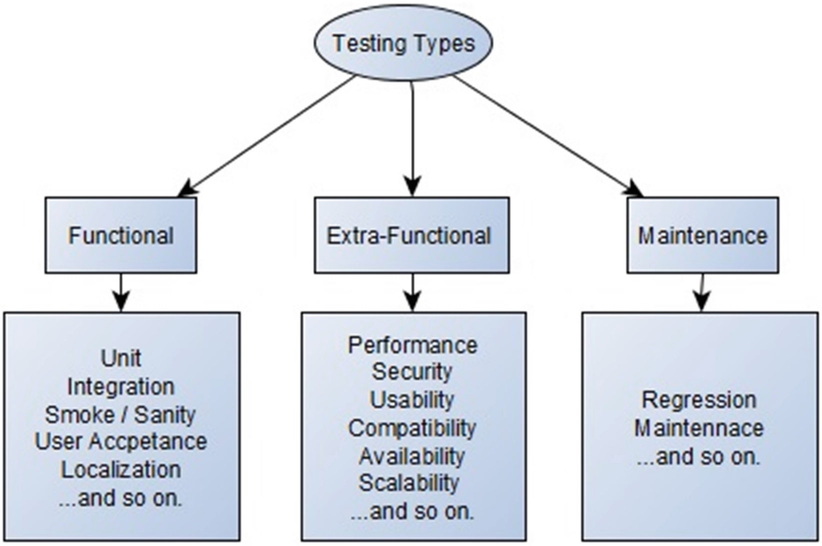


**Fig. 1.** Types of software testing in general.

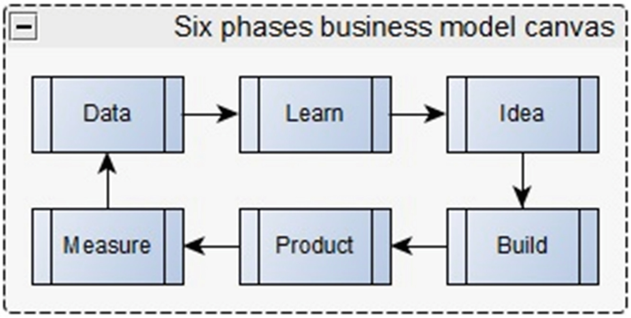


**Fig. 2.** Six phases business model canvas.

- Data - Use sample database that contains the 50 users.
- Learn - Graphs, results and logs of each test.

## 4 Extra-Functional Properties of Testing Process and Possible Waste Identification Test Metrics Using Lean Principles

It needs to rethink now about how we design test cases for any application testing because the quality assurance team needs to plan well for the testing and collect the data from number source, making sure the software is well- tested for extra-functional properties and is error free. Normally it is very difficult to manage the tractability, software features, priority for the types of testing and generic requirements make the testing process complex as we discussed in Sect. 2.

While testing software extra-functions, it is important to identify the possible waste in a process to get the desired result. This will also impact the software testing time and save budget allotted to test the application [4]. Now applying the seven lean principles for the extra-functional properties of testing we can identify the possible waste.

- Transport – Changing test environment, can give different results.
- Inventory – Not able to test all software components until they fulfill clear requirements.
- Motion – Not have a scope type of testing with software under test.
- Waiting – Quality assurance team needs to wait for approval from the development team before considering testing.
- Overproduction – Testing is finished and the business owner wants to add a new feature or change its functionality, then the application needs to re-test.
- Over processing – Not have sufficient tools to test or quality assurance team is not able to conclude the results due to varying results.
- Defects – Found new bugs need to be re-fixed and retested again.

The aim of utilization of seven principles is to mitigate the possible waste in test cycle.

### 4.1 Transformation Model to Find Relevant Test Metrics for Extra-Functional Properties

From the last section, we used lean seven principals and identified several possibilities for finding basic test metrics, later we can get more relevant test metrics using the transformation model [5]. In test report generation and making conclusions normally test metrics play a key role to define the software quality (Fig. 3).

Once we collect the possible lean metrics, we can add them on to the lean board, but on the board, these names can change after each life cycle of testing if software requirements changes.

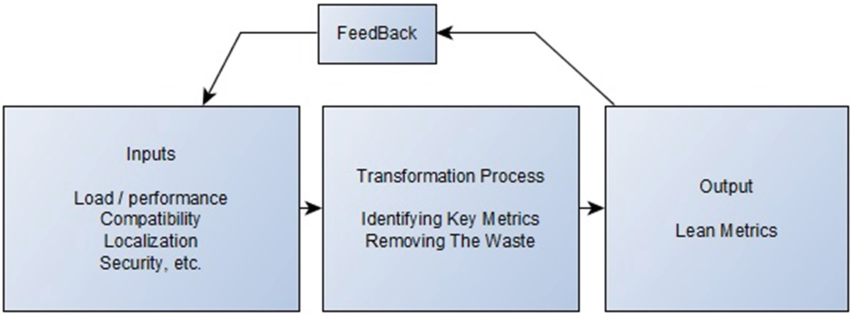


**Fig. 3.** Prototype of the transformation model, to find test metrics for lean canvas board.

### 4.2 Design Lean Canvas Board Using the Visualizing of the Blocks

From the above section, we have got to know there are a lot of metrics we get, and it is possible to design a lean canvas board with blocks on it, having an appropriate title. The main idea from collected metrics is to draw and visualize a lean canvas board for further utilizations.

## 5 Conclusion and Future Research

This paper explains the introduction and possibilities of how we are able to adopt the lean canvas board with optimized metrics. It is important now to rethink a new way of adoption of the lean canvas board design as a base design for those extra-functional properties testing and it will help to solve the problem related to testing design, test planning, team collaboration and improve the traceability.

- The next step to carry out this research, still needs to do several activities.
- Identification of all possible test metrics for the testing.
- Need to carry out an experiment for the different types of testing.
- Finding a common test metric among all types of testing.
- Need to develop the prototype that visualizes the test metrics on the single page.

- Need to investigate and develop the possible algorithm to generate optimal test metrics.
- Need to study about how we can integrate developed lean canvas boards into the possible existed collaboration tools.
- Improving the lean canvas board with continuous feedback.

The paper author wishes this paper to generate more new ideas, and many ideas implemented in the solution. Also, wishes it to bring a light on possible lean canvas adoption for the extra-functional test process to make it more simple and productive.

## References

1. Afzal, W., Torkar, R., Feldt, R.: A systematic review of search-based testing for non-functional system properties. Inf. Softw. Technol. **51**(6), 957–976 (2009)
2. Carvalho, J., Melo de Sousa, S., Paulo Fernandes, J., Pereira, N., Filipe Mendes, L., Figueired, C., Raquel Oliveira, C.: Automated analysis of non-functional requirements for web applications. In: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6 (2016)
3. Ide, M., Amagai, Y., Aoyama, M., Kikushima, Y.: A lean design methodology for business models and its application to IoT business model development. In: 2015 Agile Conference, pp. 107–111 (2015)
4. Manjunath, K., Jagadeesh, J., Yogeesh, M.: Achieving quality product in a long term software product development in healthcare application using lean and agile principles: software engineering and software development. In: Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), pp. 26–34 (2013)
5. Nidagundi, P., Novickis, L.: Introduction to lean canvas transformation models and metrics in software testing. Publ. Sci. J. Ser. Sci. J. RTU. **19**, 30–36 (2016)
6. Singh, P., Tripathi, A.K.: Issues in testing of software with NFR. Int. J. Softw. Eng. Appl. (IJSEA) **3**(4), 61 (2012)
7. Nidagundi, P., Novickis, L.: Introducing lean canvas model adaptation in the scrum software testing. Procedia Comput. Sci. **104**, 97–103 (2017)

# Hybrid SMOTE-Ensemble Approach for Software Defect Prediction

Hamad Alsawalqah[1], Hossam Faris[1], Ibrahim Aljarah[1](✉), Loai Alnemer[1], and Nouh Alhindawi[2]

[1] King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan
{h.sawalqah,hossam.faris,i.aljarah,l.nemer}@ju.edu.jo
[2] Department of Software Engineering, Faculty of Sciences and Information Technology, Jadara University, Irbid, Jordan
hindawi@jadara.edu.jo

**Abstract.** Software defect prediction is the process of identifying new defects/bugs in software modules. Software defect presents an error in a computer program, which is caused by incorrect code or incorrect programming logic. As a result, undiscovered defects lead to a poor quality software products. In recent years, software defect prediction has received a considerable amount of attention from researchers. Most of the previous defect detection algorithms are marred by low defect detection ratios. Furthermore, software defect prediction is very challenging problem due to the high imbalanced distribution, where the bug-free codes are much higher than defective ones. In this paper, the software defect prediction problem is formulated as a classification task, and then it examines the impact of several ensembles methods on the classification effectiveness. In addition, the best ensemble classifier will be selected to be trained again on an over-sampled datasets using the Synthetic Minority Over-sampling Technique (SMOTE) algorithm to tackle imbalanced distribution problem. The proposed hybrid method is evaluated using four software defects datasets. Experimental results demonstrate that the proposed method can effectively enhance the defect prediction accuracy.

**Keywords:** Software defect prediction · SMOTE · Ensemble approaches · Data mining · Software engineering

## 1 Introduction

Today, the task of developing and delivering a bug-free and high quality software product to the end users is becoming more and more challenging. Delivering high quality software product to the end users is ensured by software reliability and software quality assurance [1,2]. With the rapid growth in size and complexity

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*, Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_39

of today's software, the prediction of software reliability plays a crucial rule in software development process [3]. According to [4,5], the quality of a software product is highly correlated to the presence or absence of the faults. Software fault presents in a computer program as an error situation that is caused by wrong specification, incorrect programming logic, lack of programming and testing skills, and so forth [6–8]. Defective software module hinders the software from working in the desired manner which further increases the development and maintenance costs and is responsible for customer dissatisfaction [9,10].

Software defect prediction, which predicts defective software modules, can help the project manager and the software developers in assessing project progress, planning defect detection activities, evaluating product quality and assessing process performance for software management [11]. Software defect prediction is very useful for finding bugs and prioritizing testing efforts especially when any company does not have sufficient resources for testing the entire product [7,12]. In this context, researchers and practitioners have applied several statistical and machine learning techniques to predict the fault proneness models and reduce software development and maintenance costs. Among them, the machine learning technique is the most popular [1]. The majority of software fault prediction techniques builds models using metrics and faulty data from an earlier deployment or identical objects and then uses the models to predict whether modules presently under development contain defects, which is called supervised learning approaches [7].

During the last two decades, many supervised software defect prediction techniques have been developed and applied such as Neural Network [13], Support Vector Machine [14], Naive Bayes [15], Genetic Programming [16], Random Forest [17], Logistic Regression [18], Decision Trees [19], Fuzzy Logic [20], Association Rule Mining [21], and the Artificial Immune Systems (AISs) [22,23].

From a machine learning point of view the problem is considered very complex and very challenging problem due to the high imbalanced distribution of the classes in the datasets. Where the ratio of normal examples are much higher than the defect ones. In such case, most of conventional and basic classification algorithms tend to classify correctly the major class and ignore the smaller class, which in most cases like in software defect prediction problem, the defect-prone is the important class. Consequently, this will lead the classifier to poor performance. To tackle this problem, in this paper we proposed a hybrid approach based on the Synthetic Minority Over-Sampling Technique (SMOTE) and ensemble classifiers for detecting software defects in different imbalanced datasets.

This paper is structured as follows: the ensemble classifiers applied in this work (i.e. RF, AdaBoost, Bagging) are described in the following section. A brief overview of the SMOTE technique is given in Sect. 3. The proposed hybrid SMOTE-Ensemble approach is described in Sect. 4. The utilized datasets in this work are described in Sect. 5. Evaluation metrics are presented in Sect. 6. The experiments and results are discussed in Sect. 7. Finally, the findings of this work are given in Sect. 8.

## 2 Ensemble Classifiers

The main objective of the classification problem is to find the best hypothesis that produced the best prediction results. In many problem domains, even the well-suited problems, it is very difficult to find a good hypothesis that makes good predictions. In general, keeping many weak hypothesis and combine their output is better than selecting the best one. In this section, we will describe three of the well-known ensemble techniques.

### 2.1 Bagging (Bootstrap Aggregation)

Bagging is an ensemble technique that is used to improve the classification results by combining the prediction of multiple classification methods [24]. The idea is to generate random training sets with replacement then train these random sets using any classification technique many times. And then, we use voting to predict the class label. Bagging works better for unstable learning algorithm where a small changes in the training set result in large changes in predictions $(i.e., DecisionTrees, NeuralNetworks)$.

### 2.2 Random Forests (RF)

Random Forest [25] is a special case of Bagging where it selects random features on order to create bootstrap models using Decision Trees. The main idea of Random Forest algorithm is that it develops a large number of decision trees by selecting data and variables randomly. Random Forest relies on aggregating the output from many "shallow" trees (called stumps) that are tuned and pruned without much analysis. The idea is that the errors from many "shallow" trees will disappear when aggregated and lead to a more accurate prediction.

### 2.3 AdaBoost

AdaBoost [26], Also known as Adaptive Boosting, is the most well known ensemble technique. Boosting trains model by sequentially training a new simple model based on the errors of the previous models. In other words, we start by discovering the examples that are hard to predict using simple classifiers and focus next classifiers to predict these examples better. AdaBoost uses the predictions of $N$ weak classifiers. For a given pattern $x_i$ each classifier $c_j$ can predict the class label of $x_i \in -1, 1$ and the final prediction is the sign of weighted sum of classifier predictions.

AdaBoost seems to enhance the performance accuracy for two reasons:1. The misclassification rate of the final classifier is reduced by combining multiple classifiers that have higher misclassification rate. 2. The variance of the final combined classifier is less than the variance of the weak classifiers.

However, increasing the number of iterations will cause an overfitting. The best way to avoid overfitting is to limit the number of iterations.

## 3 Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE is an oversampling technique was first proposed in [27]. This oversampling technique modifies the class distribution in the dataset by oversampling the minority class by creating synthetic samples rather than oversampling with replacement. Synthetic samples are generated by operating in feature space. The minority class is oversampled by taking out each sample and creating synthetic samples along the line segments that join any/all of the $k$ minority class nearest neighbours.

The algorithm starts by choosing $k$ nearest neighbors then synthetic samples are generated by taking the difference between the feature vector of the sample under consideration and its nearest neighbor. Then it multiplies the difference by a random number between 0 and 1, and add it to the feature vector under consideration. Therefore, a random point is selected along the line segment between two specific features. Consequently, SMOTE broadens the data region of the minority class examples and forces the decision region of the class to become more general.

## 4 Hybrid SMOTE-Ensemble Approach

As mentioned earlier, most of the collected and available datasets used to predict software defects are highly imbalanced. This makes it more difficult for common classification algorithms to detect the rare and small classes. In machine learning literature, there are different approaches proposed for handling the challenge of imbalance class distribution. Two major approaches are by using ensemble classifiers and by using a data-level approach. The first handles the problem by applying different base classifiers on variations of the training datasets then combines their votes using a predefined mechanism. The second approach is called also an external approach because it preprocesses the training dataset using oversampling or undersampling algorithms before applying the classifier.

In this work, we investigate the application of combining both approaches sequentially in order to boost the performance of the detection rate of software defects. At first, three ensembles (i.e. RF, Adaboost and Bagging) are applied and evaluated. The best performing ensemble classifier is selected to be trained again on an oversampled datasets using the SMOTE algorithms. The hybrid approach is illustrated in Fig. 1. This hybrid approach will be refered to as SMOTE+classifier. The task of SMOTE is to create a percentage of new artificial instances from the minority class which is the defects class in our case. This process will form a more balanced distribution of the number of instances in each class which consequently could help the classifier taking into account the minority class. After that, the new oversampled dataset will be used to train the ensemble classifier. It is expected that by creating a percentage of artificial examples similar to the examples that represent the defect cases in the dataset,

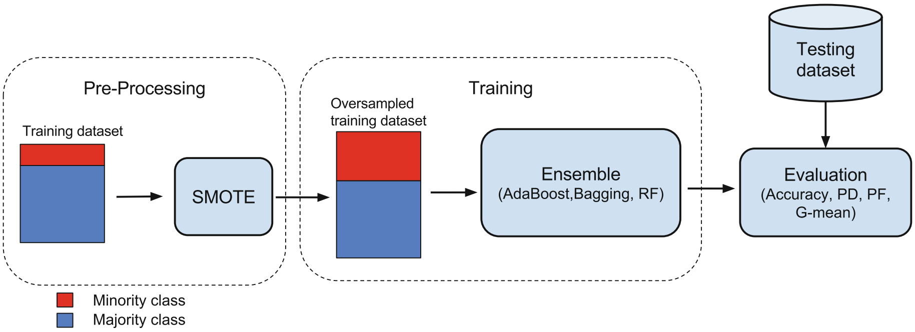

Pre-Processing
Training dataset
SMOTE
Training
Oversampled training dataset
Ensemble (AdaBoost,Bagging, RF)
Testing dataset
Evaluation (Accuracy, PD, PF, G-mean)
Minority class
Majority class

**Fig. 1.** Hybrid SMOTE+Ensemble approach for bug prediction

the defects detection rates of the ensemble classifiers will increase. An important question rises here: how much instances should we create by SMOTE before training our classifiers? To answer this question, an extensive experiment will be conducted as part of the experiments section to study the effect of this ratio.

## 5 Datasets Description

To facilitate the replication and verification of our experiments, four public benchmark datasets [28] from PROMISE Repository[1] were used. These datasets

**Table 1.** Datasets description

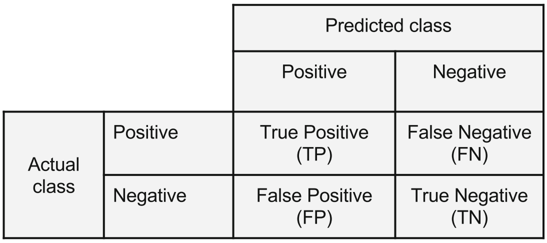
| Dataset | Language | Description | #Attributes | *LOC* | #Modules | #Non-defects | #Defects | % Non-defects | % Defects |
|---|---|---|---|---|---|---|---|---|---|
| CM1 | C | NASA spacecraft instrument | 22 | 20 *K* | 498 | 449 | 49 | 90.16 | 9.83 |
| KC1 | C++ | System implementing storage management for receiving and processing ground data | 22 | 43 *K* | 2109 | 1783 | 326 | 84.54 | 15.45 |
| JM1 | C | Real-time predictive ground system: Uses simulations to generate predictions | 22 | 315 *K* | 10885 | 8779 | 2106 | 80.65 | 19.35 |
| PC3 | C | Flight software for earth orbiting satellite metadata | 38 | 40 *K* | 1563 | 1403 | 160 | 89.7 | 10.23 |

[1] http://openscience.us/repo/.

were collected from real software projects by NASA which were developed in C/C++ for spacecraft instrumentation (CM1 dataset), storage management of ground data (KC1 dataset), scientific data processing (JM1 dataset), and satellite flight control (PC3 dataset). A detailed description of of the datasets is given in Table 1.

## 6 Model Evaluation

In order to evaluate the performance of the software faults prediction model, we used the confusion matrix that shown in Fig. 2, which is a table that is often used to describe the classification model results.



| | | Predicted class | |
|---|---|---|---|
| | | Positive | Negative |
| Actual class | Positive | True Positive (TP) | False Negative (FN) |
| | Negative | False Positive (FP) | True Negative (TN) |

**Fig. 2.** Confusion matrix

The defect detection effectiveness is evaluated using four measurements based on the previous confusion matrix; namely, Model Accuracy (Acc), Number of Predicted Defects (PD), Number of Incorrectly Predicted cases with No Defects (PF) and G-mean. Acc is the ratio of the correctly predicted faults to the total number of faults. PD is the ratio of correctly predicted faults to the total number of faults. PF is the number of not fault-prone modules that are predicted incorrectly as defects. In addition, we used the G-mean to combine the PD and PF which is a good indicator of the relationship between the two measures. The Acc, PD, PF, G-mean are calculated in Eqs. 1, 2, 3, and 4, respectively.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$PD = \frac{TP}{TP + FN} \tag{2}$$

$$PF = \frac{FP}{FP + TN} \tag{3}$$

$$G - mean = \sqrt{PD \times (1 - PF)} \tag{4}$$

## 7 Experiments and Results

The experiments are conducted on three stages; at first, three basic classifiers commonly used in literature are applied and evaluated on the investigated datasets. The classifiers are Naive Bayes (NB), Multilayer Perceptron (MLP) and the C4.5 decision trees. We used the J48 implementation of C4.5 which is available in Weka. In the second stage the ensemble classifiers including RF, Bagging and Adaboosts are applied. Finally, in third stage the hybrid approach of SMOTE combined with best ensemble method in the previous stage is applied and evaluated.

All algorithms in our experiments are applied based on a 10 folds cross-validation as a training and testing methodology. The algorithms are evaluated using the evaluation criteria described in the previous section. As can be seen in Fig. 3, evaluation results of basic classifiers show that J48 obtained better results compared to MLP in terms of PD and G-mean for all datasets. While NB tends to give higher FP rates which makes it worst.

In the second stage of the experiments, the ensemble classifiers RF, Bagging and Adaboost are applied in attempt to improve the defect detection rates. For Bagging and Adaboost, a base classifier must be selected. In our experiments, J48 is selected to be a base classifier for these two ensembles for three reasons: first due to their good performance compared to NB and MLP networks in the previous stage, second they are much faster to build than the training MLP networks, third it is preferable that the base classifiers are learning algorithms that are highly affected by any change in the training data.

Number of iterations in ensemble learners could highly affect their performance. Therefore, all ensembles are experimented using different number of iterations starting from 10 up to 100 iterations with a step of 10 iterations increased in each experiment. The best performance of Bagging is obtained with 10 iterations for all datasets. For Adaboost, it was obtained with 10 iterations for

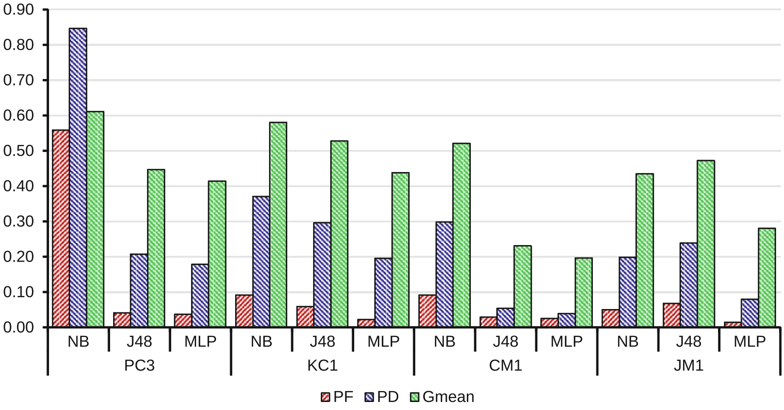


**Fig. 3.** Evaluation results of basic classifiers.

PC3 and JM1, and 20 iterations for CM1 and KC1. For RF, the best results are obtained with 30, 50, 60 and 80 for CM1, PC3, JM1 and KC1, respectively. Evaluation results of best ensembles models compared to J48 which was the best in the previous stage are shown in the Fig. 4. It can be seen that Adaboost obtained best results compared to the other classifiers with a noticeable improvement in G-mean and PD rates and very competitive low FP rates.

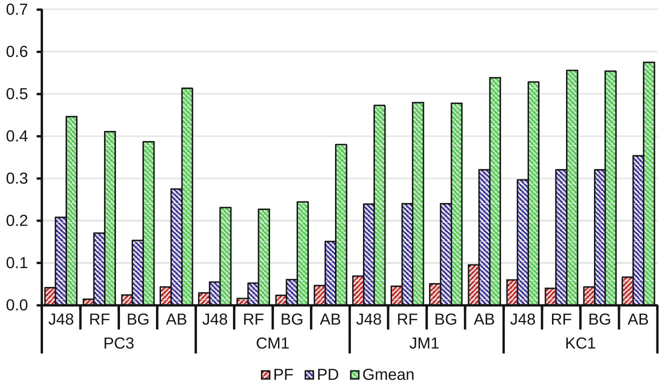
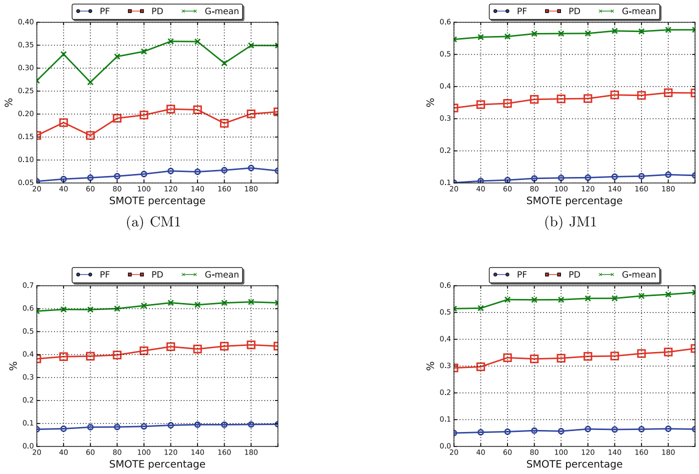


**Fig. 4.** Evaluation results of ensemble classifiers and J48 decision tree model.

In the last stage of the experiments, we experiment the effect of combining SMOTE with the best ensemble classifier from the previous stage which is Adaboost. In SMOTE, there are two different parameters that influence its performance: number of nearest neighbors ($k$) to use and the percentage of SMOTE instances ($P$) to create. For the first parameter, the default number of 5 neighbors is used which was recommended in [27]. While for the second parameter, different percentages are experimented starting from 20% up to 200%. Figure 5 shows the PD, PF and G-mean evaluation values of Adaboost when trained on datasets oversampled by SMOTE with different $P$ values. According to this figure, the best performance obtained for CM1 and KC1 datasets is at $P$ of 120% and for JM1 and PC3 datasets is at 180% and 200%, respectively.

Finally, the best obtained results of SMOTE combined Adaboost with all other classifiers are shown in Tables 2, 3, 4 and 5, for PC3, KC1, CM1, JM1 datasets respectively. In JM1 and KC1 datasets, the SMOTE+Adaboost approach managed to achieve best results in terms of PD and G-mean with very competitive low PF rate. For PC3 dataset, we can see that although NB has the best results in terms of PD and G-mean, it has the worst FP rate which is very high at value of 55.89%. On the other hand, SMOTE+Adaboost comes second with 36.56% for PD and 57.48% for G-mean with low FP rate of 6.45% which makes it more favorable. For CM1, it can be seen that SMOTE enhanced the performance of Adaboost in all criteria but it comes second after NB.



(a) CM1

(b) JM1

(c) KC1

(d) PC3

**Fig. 5.** Evaluation results of SMOTE+AdaBoost by changing number of iterations.

**Table 2.** Evaluation results for PC3 dataset

| | Accuracy | PF | PD | G-mean |
|---|---|---|---|---|
| NB | 0.4826 | 0.5589 | 0.8469 | 0.6112 |
| J48 | 0.8814 | 0.0418 | 0.2081 | 0.4466 |
| MLP | 0.8820 | 0.0378 | 0.1788 | 0.4147 |
| RF | 0.9018 | 0.0149 | 0.1713 | 0.4107 |
| Bagging | 0.8912 | 0.0247 | 0.1531 | 0.3865 |
| AdaBoost | 0.8869 | 0.0433 | 0.2750 | 0.5129 |
| SMOTE+AdaBoost | 0.8772 | 0.0645 | 0.3656 | 0.5748 |

**Table 3.** Evaluation results for KC1 dataset

| | Accuracy | PF | PD | G-mean |
|---|---|---|---|---|
| NB | 0.8246 | 0.0925 | 0.3715 | 0.5806 |
| J48 | 0.8404 | 0.0601 | 0.2964 | 0.5278 |
| MLP | 0.8565 | 0.0229 | 0.1963 | 0.4380 |
| RF | 0.8610 | 0.0402 | 0.3208 | 0.5548 |
| Bagging | 0.8582 | 0.0434 | 0.3200 | 0.5533 |
| AdaBoost | 0.8438 | 0.0665 | 0.3534 | 0.5743 |
| SMOTE+AdaBoost | 0.8329 | 0.0957 | 0.4424 | 0.6294 |

**Table 4.** Evaluation results for CM1 dataset

| | Accuracy | PF | PD | G-mean |
|---|---|---|---|---|
| NB | 0.8484 | 0.0918 | 0.2985 | 0.5207 |
| J48 | 0.8805 | 0.0294 | 0.0550 | 0.2310 |
| MLP | 0.8825 | 0.0256 | 0.0400 | 0.1974 |
| RF | 0.8920 | 0.0165 | 0.0525 | 0.2272 |
| Bagging | 0.8865 | 0.0234 | 0.0610 | 0.2441 |
| AdaBoost | 0.8743 | 0.0468 | 0.1510 | 0.3794 |
| SMOTE+AdaBoost | 0.8536 | 0.0762 | 0.2110 | 0.3585 |

**Table 5.** Evaluation results for JM1 dataset

| | Accuracy | PF | PD | G-mean |
|---|---|---|---|---|
| NB | 0.8042 | 0.0507 | 0.1993 | 0.4350 |
| J48 | 0.7973 | 0.0688 | 0.2394 | 0.4722 |
| MLP | 0.8100 | 0.0148 | 0.0799 | 0.2806 |
| RF | 0.8167 | 0.0450 | 0.2402 | 0.4789 |
| Bagging | 0.8118 | 0.0511 | 0.2406 | 0.4778 |
| AdaBoost | 0.7912 | 0.0958 | 0.3199 | 0.5378 |
| SMOTE+AdaBoost | 0.7787 | 0.1259 | 0.3808 | 0.5764 |

## 8 Conclusions

Delivering high quality software products on time within budgetary cost is crucial issue for the success of software companies. In order to help software developers in undertaking this issue, an attempt has been made in this paper for software defect prediction.

This paper formulates the software defect prediction problem as a classification task. It further investigates the impact of several ensemble methods to solve the defect prediction problem. Specifically, we have developed a hybrid ensemble classification based on over-sampled approach for detecting software defects in different imbalanced datasets. High imbalanced distribution of classes in datasets degrades the performance of classification approaches. The proposed approach is developed based on the Synthetic Minority Over-sampling Technique (SMOTE). The experiments have shown that the proposed SMOTE+Ensembles approach has better quality results than the other classification algorithms. Our future research will include the verification of the proposed method on other datasets.

## References

1. Rawat, M.S., Dubey, S.K.: Software defect prediction models for quality improvement: a literature study. IJCSI Int. J. Comput. Sci. Issues **9**, 288–296 (2012)
2. Aljarah, I., Banitaan, S., Abufardeh, S., Jin, W., Salem, S.: Selecting discriminating terms for bug assignment: a formal analysis. In: Proceedings of the 7th International Conference on Predictive Models in Software Engineering, no. 12. ACM (2011)
3. Zheng, J.: Predicting software reliability with neural network ensembles. Expert Syst. Appl. **36**, 2116–2122 (2009)
4. Hall, T., Beecham, S., Bowes, D., Gray, D., Counsell, S.: A systematic literature review on fault prediction performance in software engineering. IEEE Trans. Softw. Eng. **38**, 1276–1304 (2012)
5. Arisholm, E., Briand, L.C., Johannessen, E.B.: A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. J. Syst. Softw. **83**, 2–17 (2010)
6. Dowd, M., McDonald, J., Schuh, J.: The Art of Software Security Assessment: Identifying and Preventing Software Vulnerabilities. Pearson Education, Upper Saddle River (2006)
7. Abaei, G., Selamat, A.: A survey on software fault detection based on different prediction approaches. Vietnam J. Comput. Sci. **1**, 79–95 (2014)
8. Tomar, D., Agarwal, S.: Prediction of defective software modules using class imbalance learning. Appl. Comput. Intell. Soft Comput. **2016** (2016). Article no. 6
9. Fenton, N.E., Neil, M.: Software metrics: roadmap. In: Proceedings of the Conference on the Future of Software Engineering, pp. 357–370. ACM (2000)
10. Fenton, N., Bieman, J.: Software Metrics: A Rigorous and Practical Approach. CRC Press, Boca Raton (2014)
11. Clark, B., Zubrow, D.: How good is the software: a review of defect prediction techniques. Sponsored by the US Department of Defense (2001)
12. Wang, S., Liu, T., Tan, L.: Automatically learning semantic features for defect prediction. In: Proceedings of the 38th International Conference on Software Engineering, pp. 297–308. ACM (2016)
13. Quah, T.S., Thwin, M.M.T.: Application of neural networks for software quality prediction using object-oriented metrics. In: Proceedings on International Conference on Software Maintenance, ICSM 2003, pp. 116–125. IEEE (2003)
14. Elish, K.O., Elish, M.O.: Predicting defect-prone software modules using support vector machines. J. Syst. Softw. **81**, 649–660 (2008)
15. Menzies, T., Greenwald, J., Frank, A.: Data mining static code attributes to learn defect predictors. IEEE Trans. Softw. Eng. **33**, 2–13 (2007)
16. Evett, M., Khoshgoftar, T., Chien, P.D., Allen, E.: Gp-based software quality prediction. In: Proceedings of the Third Annual Conference Genetic Programming, pp. 60–65 (1998)
17. Koru, A.G., Liu, H.: Building effective defect-prediction models in practice. IEEE Softw. **22**, 23–29 (2005)
18. Suffian, M.D.M., Ibrahim, S.: A prediction model for system testing defects using regression analysis. arXiv preprint arXiv:1401.5830 (2014)
19. Koprinska, I., Poon, J., Clark, J., Chan, J.: Learning to classify e-mail. Inf. Sci. **177**, 2167–2187 (2007)

20. Yuan, X., Khoshgoftaar, T.M., Allen, E.B., Ganesan, K.: An application of fuzzy clustering to software quality prediction. In: Proceedings of 3rd IEEE Symposium on Application-Specific Systems and Software Engineering Technology, pp. 85–90. IEEE (2000)
21. Czibula, G., Marian, Z., Czibula, I.G.: Software defect prediction using relational association rule mining. Inf. Sci. **264**, 260–278 (2014)
22. Catal, C., Diri, B.: Software fault prediction with object-oriented metrics based artificial immune recognition system. In: Münch, J., Abrahamsson, P. (eds.) PROFES 2007. LNCS, vol. 4589, pp. 300–314. Springer, Heidelberg (2007). doi:10.1007/978-3-540-73460-4_27
23. Catal, C., Diri, B.: A fault prediction model with limited fault data to improve test process. In: Jedlitschka, A., Salo, O. (eds.) PROFES 2008. LNCS, vol. 5089, pp. 244–257. Springer, Heidelberg (2008). doi:10.1007/978-3-540-69566-0_21
24. Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996)
25. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
26. Schapire, R.E.: Explaining AdaBoost. In: Schölkopf, B., Luo, Z., Vovk, V. (eds.) Empirical Inference, pp. 37–52. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41136-6_5
27. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
28. Shepperd, M., Song, Q., Sun, Z., Mair, C.: Data quality: some comments on the nasa software defect datasets. IEEE Trans. Softw. Eng. **39**, 1208–1215 (2013)

# A Novel Design in Formal Verification Corresponding to Mixed Signals by Differential Learning

D.S. Vidhya[1](✉) and Manjunath Ramachandra[1,2]

[1] Assam Don Bosco University Guwahati, Guwahati, India
dsvidhya0770@gmail.com
[2] RV College of Engineering, Bangalore, India
manju_r_99@yahoo.com

**Abstract.** The mixed signals exhibit a characteristics called loops composed of multiple feedbacks and thus it is not feasible for apply traditional testing methods for conduction sophisticated formal verification with higher accuracy accompanied by speedy response. We surveyed the current research work in this regards to find that there are open-end issues pertaining to mixed signal formal verification. This paper has displayed a novel formal verification procedure of the mixed signal utilizing differentially-placed neural network. An analytical demonstrating is given (i) an algorithm for generating multiple mixed signal comparing to feasible operational states of mixed signal circuits, (ii) algorithm for formal verification and (iii) algorithm for training. The accomplished result of comparative analysis demonstrates 98.7% of accuracy with speed response as compared to the current learning algorithms.

**Keywords:** Analog/Digital · Differential learning approach · Formal verification · Mixed signal · Neural network

## 1 Introduction

With the increasing awareness of electrical and electronics products among the consumers, it has become essential to ensure that the entire upcoming product should adhere to industry standards. The standard technique to ensure the compliance of product design as per the industry standard is called as formal verification [1, 2]. Basically, there are two forms of formal procedures e.g. property checking and equivalence checking [3, 4]. The mechanism of formal property checking basically intends to demonstrate the preciseness of the prototype design of the product or highlight the core factors of faults with an aid of sophisticated mathematical approach [5]. This mechanism is completely free from using any form of test-bench and hence response time is faster. Various forms of languages e.g. Assertion language or Interval Temporal Logic (ITL) that are somewhat equivalent to Verilog [6, 7]. All these mechanism are normally used for assessing the design behavior under various test-environments using various types of checker tools [8]. Similarly, the formal equivalence checking is the mechanism for exploring the operation match of a specific design in contrast with certain reference design [9]. Usage

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_40

of such techniques can be frequently seen in pre and post routed Netlist, RTL, etc. [9]. In present era, the frequently used technique of this type is sequential equivalence checking and combinatorial equivalence checking [10, 11]. From theoretical viewpoint, there are three types of formal methods: model checking, deductive, and static analysis [12, 13]. Model checking techniques are more inclined towards searching the conclusive model of design whereas deductive methods use theorem provers [14] for generating precise mathematical evidence of the design. However, they are also highly dependent on human in order to obtain inductive inputs. Although the existing test-bench permits the scalable and robust modeling, it cannot ensure complete functional verification process. Such problems of simulation-based techniques are mitigated using formal method. Usage of formal verification process doesn't only enhance the assessment process but also minimize the tedious efforts consumed in the verification of the design in VLSI.

Hence, it can be seen that research work towards formal verification is still in nascent stage especially with respect to computational modeling. This research work presents a novel model of formal verification of the mixed signal where the differentially fed neural network is used as the learning algorithm. Section 2 discusses about the existing research work carried out in formal verification followed by brief highlights of problem identification in Sect. 3. Section 4 briefs about proposed system followed by elaborated discussion of research methodology in Sect. 5. Section 6 illustrates the algorithm being implemented in the proposed study while the obtained results are discussed in Sect. 7. Finally, Sect. 8 concludes the paper by summarizing the contribution.

## 2 Related Work

This section discusses about the existing research work carried out towards formal verification. The most recent work carried out by Saha et al. [15] have discussed about various architectures towards formal verification based on temporal factor. Alam et al. [16] have presented a formal verification modeling using Petri Nets focusing on the retention of the security requirements of a network model. The study has used SAT modulo theory (SMT) libraries and solver for this purpose. Calinescu et al. [17] have introduced a formal verification of quantitative nature using Markov Modeling. Considering multiple case studies, the authors have shown that their technique supports better reliability with better performance with less cost and reduced energy consumption. The presented technique was assessed using case studies. Campos et al. [18] have presented a study where safety factors of the aerospace design were subjected to formal verification. The complete verification was carried out using model checkers. The recent studies carried out by Cifuentes et al. [19] have presented a formal verification model over distributed system with an aid of queues and threads. Another recent work carried out by Webster et al. [20] has presented a case study of robotics using formal verification process. The authors have developed a scheduler system for the house assisted robots.

Our prior work has reviewed some of the effective techniques of formal verification of analog and mixed signal along with highlights of research gap [21]. Work towards designing the system with analog and mixed signal was carried out by Ain et al. [22]. The authors have deployed syntactic fabric for asserting such signals. Using Verilog and simulation-based study, the authors have used formal verification in order to check the correctness of the design factor. Validation of the mixed signal was also carried out by Lim et al. [23] using model equivalence checking. Yin et al. [24] have presented a formal verification technique for assessing analog and mixed signal using SAT modulo theory. The study outcome was found to have better applicability towards nonlinear circuits. Similar direction of work was also carried out by Little et al. [25]. The authors have presented a formal verification method using state-space search techniques. Hence, it can be seen that there are various work that has been carried out toward formal verification method. The next section discusses problem identification.

## 3 Problem Identification

From the prior section, it has been observed that a good amount of work has been carried out towards formal verification topic. However, a closer look into the existing research techniques will show that there are certain trade-offs as well as open research issues which are yet to be addressed. Hence, we highlight some of the research gaps captured from the existing studies:

- ***Less Work Towards Mixed Signal***: A closer look into the numbers of research work shows that 95% of the research work is focused on industrial applications and futuristic appliances. Hence, there is an obvious research gap towards modeling a formal verification model dedicated for verifying mixed signal circuits.
- ***Little Emphasis on Computational Modeling***: There is only little analytical work that has spoken about a computational modeling towards novel direction of formal verification. With regards to mixed signal, studies are yet to be carried out in computational modeling.
- ***Less Benchmarked Modeling***: Although there has been various numbers of research work towards complex applications like aerospace, healthcare, etc., but still none of the work is found to be compared with any existing and robust standard model.
- ***Lack of Optimization Techniques***: Optimization can increase the performance of the formal verification modeling to a large extent. The techniques e.g. neural network can significantly increase the accuracy rate and can positively affect the performance. However, still such problems are not investigated in existing cases.
- ***Cost-Effective Modeling***: It is still questionable point if such models will really be sustainable upon changing its environmental parameters. Limiting the studies towards case study also limits its applicability and hence narrow downs the formal verification process.

Hence, there is truly a need to evolve up with a novel modeling of formal verification that can overcome the above mentioned problems. The next section briefs about the proposed system that addresses such problems.

# 4 Proposed System

The prime purpose of the proposed system is to propose a computational modeling of formal verification for mixed signals using differentially fed-neural network. The motivation of the proposed work is drawn from the significant work being carried out by Manjunath et al. [26, 27] towards formulating the concept of differential-fed neural network. The prime reason for implementing differential fed neural network in our formal verification modeling is basically to reduce the response time and increase the accuracy. The schematic architecture of proposed system is shown in Fig. 1. The proposed model takes the input of generated mixed signals which is trained using multiple types of neural network in order to obtain differential feedback. The contributions of the proposed system are as follow:

- *Cost Effective Formal Verification*: A computational model is presented that possess the capability to perform formal verification on diverse and complex industrial process.
- *Novel Use of Differential Feedback*: This is the first time when a differential feedback is discussed to be applied on signal generation and identification process. It significantly assists in minimizing time for verification as well as increases accuracy rate.
- *Adoption of Mixed Signal Generation*: The proposed model considers generating mixed signal using pseudorandom vector along with randomly added noise.
- *More Applicability on Mixed Signal Circuits*: The complete modeling is carried out considering operational states of mixed signal circuits and thereby its outcome offers more reliability.

The next section discusses about the research methodology adopted for this purpose.

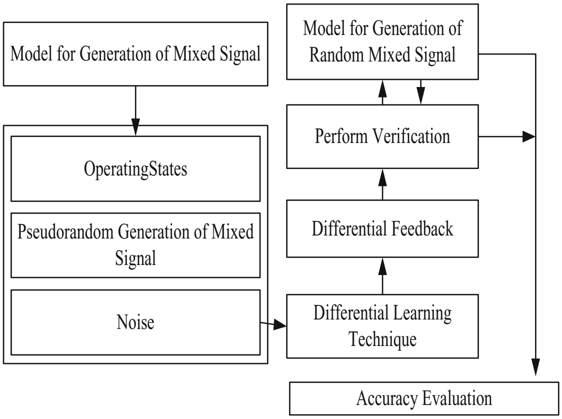

Model for Generation of Mixed Signal
OperatingStates
Pseudorandom Generation of Mixed Signal
Noise
Model for Generation of Random Mixed Signal
Perform Verification
Differential Feedback
Differential Learning Technique
Accuracy Evaluation

**Fig. 1.** Architecture of proposed scheme

## 5 Research Methodology

The design of the proposed technique is carried out using analytical approach, where the prime focus was laid into investigating the effectiveness of proposed formal verification technique for mixed signal. This section discusses about the methodology adopted to design proposed formal verification techniques with respect to various core modules involved:

### 5.1 Module for Generation of Mixed Signal

The study formulates multiple types of mixed signal on the basis of stabilized and un-stabilized state of mixed signal circuits. Both the states are further classified into higher and lower electrical factor for the mixed signal circuit. This is done to ensure that all the feasible forms of mixed signal are considered for analysis.

### 5.2 Module for Training

The generated mixed signal is then subjected to differential learning algorithm considering specific samples and number of neurons. For effectiveness, we consider standard and frequently adopted learning algorithms that can be individually chosen in order to generate differential feedback, which is saved for its usage in formal verification process. The network is finally used to validate the effectiveness of proposed formal verification method.

### 5.3 Module for Formal Verification

Formal verification is carried out in this model where the inputs of randomly generated mixed signals are processed using saved network in order to check the accuracy of verified model. We use the similar model for mixed signal generation that we have used in training module but we randomize the mixed signals for the purpose of giving different inputs to formal verification module. Usage of differentially fed neural network will now assist in reducing the response time while performing verification of the mixed signal. Here the inputs of randomly generated signals are compared with the trained mixed signals using any forms of differential learning approaches. The randomly generated mixed signal is now added with noise and the process of feature extraction is carried out by decomposing the signal using transform-based process. The network that is saved in prior module is now used for identifying the types of the formal model. Transform-based decomposition is applied for preprocessing purpose so that both features and labels of the formal model can be properly identified. It should be noted that we consider the entire design process considering the possible impact of electrical-related factors on the mixed circuit design as well as by adding noise.

This methodology is carried out in order to ensure its maximum applicability on the mixed signal generated from real-time circuits on all the possible states. Consequently, the study outcome is now more permissible to be compared with existing learning approach and proposed differential learning approach in the form of computational model. Although, it is a computational model, but its analysis was carried out on real machines considering all feasible values of samples, neurons, and various optimization principles e.g. damped-least square algorithm. Accuracy and algorithm processing time was also checked and compared with existing learning approach in order to find effectiveness of proposed formal verification. The next section discusses about algorithm implementation.

# 6 Algorithm Description

This section presents the discussion of the algorithms that were developed in order to implement the proposed methodology discussed in prior section. Basically, there are three algorithms designed for this purpose i.e. (i) Algorithm for Generating multiple mixed signals, (ii) Algorithm for Training the Mixed Signals and (iii) Algorithm for validating the mixed signals. The discussions of the algorithms are as below:

## 6.1 Algorithm for Generating Multiple Mixed Signals

This algorithm is mainly meant for generating the multiple forms of mixed signals in order to testify the proposed formal verification model. The steps of the algorithms are as shown below:

The closer look into the algorithm's steps shows that the algorithm takes the input of time *t*, which upon processing yields four different forms of mixed signals i.e. $y_{ms1}$, $y_{ms2}$, $y_{ms3}$, $y_{ms4.}$ We develop a simple pseudorandom vector $\alpha$ for the purpose of incorporating linear time-invariant characteristics of the newly generated signals. Four different types of mixed signal were created for this purpose. The first signal $y_{ms1}$ corresponds to stabilized state of mixed signal circuits with higher current/voltage factor while the second signal $y_{ms2}$ corresponds to stabilized state of mixed signal circuits with lower current/voltage factor. Similarly, the third and fourth signal $y_{ms3}$ and $y_{ms4}$ corresponds to unstabilized state of higher and lower current/voltage factor. An analog-to-digital converter is one of the good examples of the circuit generating mixed signal. Hence, we model our signal generation process in order to map the original state of such circuits where stability and unstability are the common state of assessment of such circuits with respect to voltage or current variation. The algorithm starts by initiating the time variable (Line-1) followed by formulation of pseudorandom vector $\alpha$ (Line-2). The first and second mixed signal i.e. $y_{ms1}$ and $y_{ms2}$ is formulated using step function $\tau$ and time function $t_f$ (Line-3 and Line-4). In our case, the time function tf is considered equivalent to 314*$t$. The equation also considers two relative time $\Delta t_1$ and $\Delta t_2$ that corresponds to ($t$−0.05) and ($t$−0.15). However, Line-3 and Line-4 corresponds to stabilized state of circuit generating multiple signals. In case of unstabilized state of circuit, we upgrade our pseudorandom vectors to include more sub-random vectors e.g. $\alpha_3$, $\alpha_5$, and $\alpha_7$ as shown in Line-5. Line-6 and Line-7 corresponds to unstabilized states of mixed signal generation from the circuits. Hence, this technique results in multiple generations of mixed signals which are found in majority of operational cases. These signals will be now subjected to training followed by formal verification.

```
Algorithm for Training the Mixed Signal
Input: $n$ (No. of samples), $\beta_i$ (network type), $\mu$ (Number of neuron), α (Pseudorandom vector)
Output: net (Network)
Start
1. Input $n$, $\beta_i$, $\mu$
2. For t = $t_{start}$: $t_{end}$
3.     For i=1: $n$
4.         y = $y_{ms}$(t, α)
5.         y = y+0.01*r(size(y));
6.         [Low High] = decompose(y)
7.         Extract Feature, Label
8.     End
9. [net, tr] = train($\beta_i$ ($\mu$), Feature, Label);
10. Save net
11. End
   End
```

```
Algorithm for Generating Multiple Mixed Signals
Input: $t$ (time), τ (Step Function)
Output: $y_{ms1}$, $y_{ms2}$, $y_{ms3}$, $y_{ms4}$ (Multiple Mixed Signals)
Start
1. init t=0
2. Def α= 0.3+0.6*arb[1, 1]
3. $y_{ms1}$[1- α((τ(Δ$t_1$)- τ(Δ$t_2$)))].*sin($t_f$)
4. $y_{ms2}$[1+ α((τ(Δ$t_1$)- τ(Δ$t_2$)))].*sin($t_f$)
5. $\alpha_1 \Rightarrow \sqrt{(1-\alpha_3^2-\alpha_5^2-\alpha_7^2)}$
6. $y_{ms3}$[1- α((τ(Δ$t_1$)- τ(Δ$t_2$)))).(α$_1$.sin($t_f$)+ α$_3$.sin(3* $t_f$)+ α$_5$.sin(5* $t_f$)+ α$_7$.sin(7* $t_f$)]
7. $y_{ms3}$[1+ α((τ(Δ$t_1$)- τ(Δ$t_2$)))).(α$_1$.sin($t_f$)+ α$_3$.sin(3* $t_f$)+ α$_5$.sin(5* $t_f$)+ α$_7$.sin(7* $t_f$)]
End
```

## 6.2 Algorithm for Training the Mixed Signal

The prime purpose of this algorithm is to formulate a new training scheme using neural network considering multiple types of neural network. The steps of the training algorithm are as shown below:

Using the recent work carried out by Manjunath [27], it was said that stochastic process of auto regression was used for differential feedback and was expressed in the form of $y(n + 1) = b_0 y(n) + b_1 y(n) + \ldots\ldots\ldots + a_0 x(n)$, where $a_0$ and $b_0$ are just constants [27]. The realization of attributes of the auto regression was carried out using differential feedback. The study has shown that there was a reduction in training timing using differential feedback. The normalized resultant of 1st order feedback with $y_1$ being 1st order different was represented as $\sum(w_i x_i + b_i y_i)$ that is a separate plane parallel to $\sum w_i x_i$. Hence, there are multiple planes that are parallel to each other as per multiple orders of feedbacks. The similar principle is applied in the proposed algorithm design. This algorithm takes the input of $n$ (No. of samples), $\beta_i$ (network type) and $\mu$ (Number of neuron). It also considers α (Pseudorandom vector) from our previous algorithm of signal generation. The algorithm is implemented considering 3 different types of network $\beta_i$ (i = 3, where i = 1 will corresponds to *feed-forward network*, i = 2 will corresponds to *radial basis function*, and i = 3 will mean *fitting neural network*. The algorithm considers a sampling timing $t$ with start-time $t$ start and end-time $t_{end}$ (Line-2). Considering all the samples of neurons $n$ (Line-3), the computation is carried out for all the mixed signals. The variable $y_{ms}$ is the generalized form of all the four mixed signal considered in our study. It takes the input argument of time $t$ and pseudorandom vector $\alpha$ (Line-4). The input arguments for the 1st and 2nd mixed signal (i.e. $y_{ms1}$ and $y_{ms2}$) are same i.e. time *t* and pseudorandom vector $\alpha$. However, for unstabilized state of signal generation i.e. in $y_{ms3}$ and $y_{ms4}$, we consider different forms of mixed signal). In case of both $y_{ms3}$ and $y_{ms4}$, the input arguments of the mixed signals are *t*, $\alpha$, $\alpha_3$, $\alpha_5$, and $\alpha_7$ where, we consider $\alpha_3 = \alpha_5 = \alpha_7 = 0.15$. In order to map with the real-time scenario, we consider adding random noise to the mixed signal (Line-5). A decomposition of the signal is then carried out using discrete wavelet transform that finally results into decomposition of High-pass filter and Low Pass filter (Line-6). Finally, the feature and labels are extracted. We apply neural network training using 3 different forms of learning schemes (feed-forward, radial basis, and fitness function) (Line-9). Finally, the algorithm results in formation of a network net (Line-9), which will be used further for performing formal verification. Hence, the network *net* is saved with respect to multiple configurations for purpose analyzing various test-cases of formal validation.

## 6.3 Algorithm for Formal Verification

This algorithm is responsible for carrying out formal verification of the mixed signal logic using differentially fed neural network. This process is carried out exclusively for the process of formal verification. However, during the process of formal verification, we like to use different and newly forms of mixed signals types $y_{msr1}$, $y_{msr2}$, $y_{msr3}$, and $y_{msr4.}$

Although, all these signals corresponds to types of mixed signals generated on 4 different stabilized and un-stabilized states of mixed signal generations from the circuits (i.e. $y_{ms1}$, $y_{ms2}$, $y_{ms3}$, and $y_{ms4}$), but, the values are different, which will mean $y_{ms1}$ $y_{msr1}$, $y_{ms2} \neq y_{msr2}$, $y_{ms3} \neq y_{msr3}$, and $y_{ms4} \neq y_{msr4}$. This will also mean that proposed formal verification also offers its input of mixed signal to be verified using any of the three forms of training/learning approaches i.e. *feed-forward network, radial basis function, or fitting neural network.* The primary objective was to ensure preciseness as well as reduce algorithm processing time by implementing differential fed neural network approach. The steps of the algorithms are as highlighted below:

The algorithm considers any one type of mixed signal where input signal $y_{msr}$ is randomly generated such that $y_{msr}$ is a set of newly generated random mixed signals with elements $y_{msr1}$, $y_{msr2}$, $y_{msr3}$, and $y_{msr4}$ (Line-1). Similar pseudorandom vectors (Line-2) and generation process of specific mixed signal (Line-3) is also applied in verification module. We also add random noise as shown in Line-4. We create similar dependencies (e.g. $\alpha_3$, $\alpha_5$, and $\alpha_7$) of $\alpha$ even in Line-3 on the basis of types of signals being used. The newly generated signals are then sampled using 100 kHz of frequency and are further decomposed using discrete wavelet function of 9 level of Daubechies wavelet. This extracted feature will be highly useful for performing formal verification of the recent input signals. This extracted feature resulting from the decomposition if fed directly to the user-defined neural network configuration file, which was saved as an outcome for the second algorithm of training. We have carried out analysis for the implication of all the specific forms of the network (*feed-forward network, radial basis function, or fitting neural network*) on the formal verification of the input mixed signal. The saved network considers this extracted feature of input signal for performing classification. The resultant of this algorithm is to determine the category of input mixed signal to be corresponding with any one of the trained mixed signal. We have also carried out integrated analysis of all the signals and evaluated that accuracy. We also checked the algorithm processing time of each individual learning approach as well as for the entire framework and finally assessed the effectiveness of the proposed formal verification method using comparative analysis.

## 7 Result Analysis

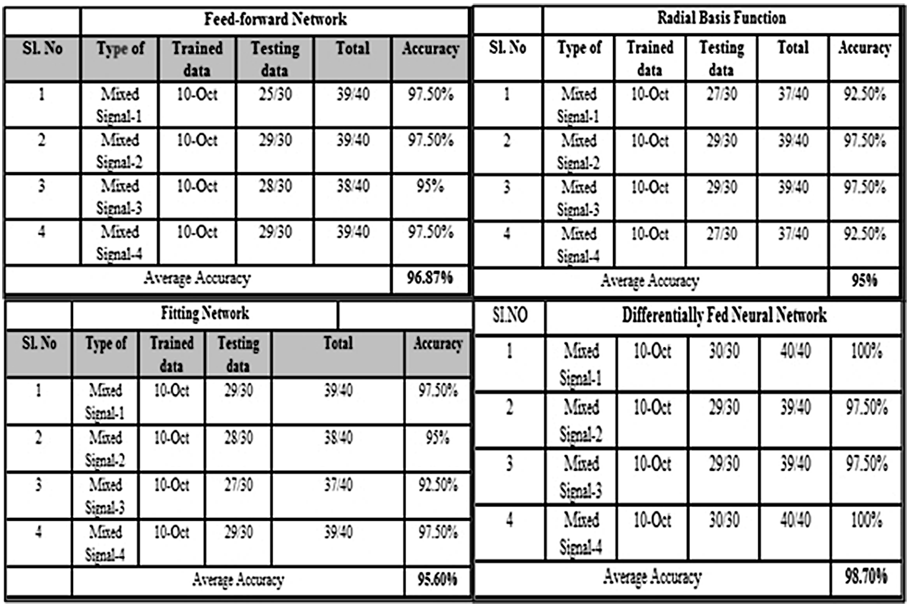
This section discusses about the results being accomplished from the proposed study. As the proposed study targets around the design and development of formal verification of mixed signal; hence, we choose to select accuracy as the prime performance parameters. In order to perform analysis of the proposed study, we select the parameters for the neural network on various test environments. Table 1 shows the comparative performance analysis on 10 trained data of all the four types of mixed signals and 30 test data. Compact outcome of the comparative performance analysis is shown in Fig. 2. The table also shows accuracies being achieved from existing techniques (feed-forward, radial basis function, and fitting function) and proposed differentially fed neural network. Our technique using differentially fed neural network is found with higher rate of accuracy i.e. 98.7% as compared to existing learning approaches of formal verification. The applicability of our approach can be used for formal verification of any mixed signal circuits as it was proven from the algorithm implementation that proposed system can potentially manage data with noise incorporated within it. The analysis was carried out over stabilized and unstabilized state of mixed signal ($y_{ms1}$, $y_{ms2}$, $y_{ms3}$, and $y_{ms4}$).

The analysis of the proposed system is carried out over multiple test environments by varying number of samples, network type, and number of neuron. A complex test environment was designed using input layer using 800 neurons, 2 hidden layers as well as 700 neurons over every layers. It also takes 1 output layer as well as three numbers of neurons. The proposed network was finally trained for multiple numbers of inputs (40 inputs), where every single input consists of multiple numbers of coefficients of wavelet.

We used approximately 2000 iterations of training using 30 extra data that corresponds to four types of generated mixed signal. The initiation of the training for the existing system was carried out using supervised learning algorithm i.e. Scaled Conjugate Gradient applicable for enhancing the operation of feed-forward algorithm in neural network. Neural network toolbox in Matlab is used for applying training function exclusively using *transig* in hidden layers as well as in input layers. The output layer is trained by using *purelin* as the training function.

The tabulated outcome using differentially fed neural network eventually shows that the accuracy of the proposed system is quite higher as compared to that of the existing learning algorithms. We also monitored algorithm processing time which is found to be 0.2077 s for proposed learning system and average of 1.8722 s for the existing learning techniques (feed-forward, radial basis function, and fitting network) considering 20,000 epochs in core i7 processor in windows machine. 100% accuracy is

**Table 1.** Numerical outcome of the comparative analysis



| | Feed-forward Network | | | | |
|---|---|---|---|---|---|
| Sl. No | Type of | Trained data | Testing data | Total | Accuracy |
| 1 | Mixed Signal-1 | 10-Oct | 25/30 | 39/40 | 97.50% |
| 2 | Mixed Signal-2 | 10-Oct | 29/30 | 39/40 | 97.50% |
| 3 | Mixed Signal-3 | 10-Oct | 28/30 | 38/40 | 95% |
| 4 | Mixed Signal-4 | 10-Oct | 29/30 | 39/40 | 97.50% |
| Average Accuracy | | | | | 96.87% |

| | Radial Basis Function | | | | |
|---|---|---|---|---|---|
| Sl. No | Type of | Trained data | Testing data | Total | Accuracy |
| 1 | Mixed Signal-1 | 10-Oct | 27/30 | 37/40 | 92.50% |
| 2 | Mixed Signal-2 | 10-Oct | 29/30 | 39/40 | 97.50% |
| 3 | Mixed Signal-3 | 10-Oct | 29/30 | 39/40 | 97.50% |
| 4 | Mixed Signal-4 | 10-Oct | 27/30 | 37/40 | 92.50% |
| Average Accuracy | | | | | 95% |

| | Fitting Network | | | | |
|---|---|---|---|---|---|
| Sl. No | Type of | Trained data | Testing data | Total | Accuracy |
| 1 | Mixed Signal-1 | 10-Oct | 29/30 | 39/40 | 97.50% |
| 2 | Mixed Signal-2 | 10-Oct | 28/30 | 38/40 | 95% |
| 3 | Mixed Signal-3 | 10-Oct | 27/30 | 37/40 | 92.50% |
| 4 | Mixed Signal-4 | 10-Oct | 29/30 | 39/40 | 97.50% |
| | Average Accuracy | | | | 95.60% |

| SI.NO | Differentially Fed Neural Network | | | | |
|---|---|---|---|---|---|
| 1 | Mixed Signal-1 | 10-Oct | 30/30 | 40/40 | 100% |
| 2 | Mixed Signal-2 | 10-Oct | 29/30 | 39/40 | 97.50% |
| 3 | Mixed Signal-3 | 10-Oct | 29/30 | 39/40 | 97.50% |
| 4 | Mixed Signal-4 | 10-Oct | 30/30 | 40/40 | 100% |
| Average Accuracy | | | | | 98.70% |

```
Algorithm for Formal Verification
Input: $y_{ms}$ (Mixed Signal)
Output: Successful / Failed Verification
Start
For $y=y_{msr}$
    $\alpha=0.3+0.6*arb[1, 1]$
    $y=y_{ms}(gen(t, \alpha))$
    $y = y+0.01*r(size(y));$
decompose y
select net
if (y exists in net)
    Successful verification
else,
    Unsuccessful verification
End
```

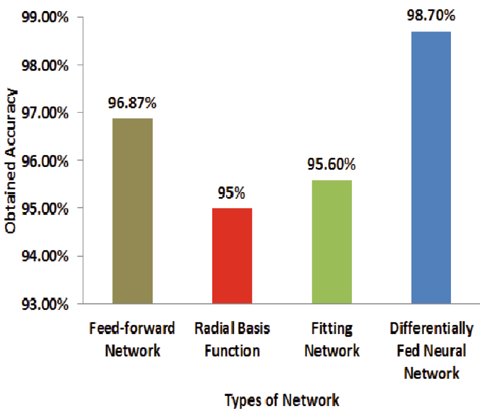


**Fig. 2.** Outcome of comparative analysis

obtained for trained data while 98% of accuracy in test data is obtained for proposed mechanism. Better randomizations, complete coverage on functional blocks, along with adoption of real-numbers in modeling are some of the prime reason behind the outcome of the proposed system.

## 8 Conclusion

The usage rates of mixed-signal circuits are increasing its pace in designing system-on-chip and this process has increased challenges of formal verification to multifold. Inappropriate verification process will cost the higher expenses as well as consume lots of productivity time. As the proposed system is designed over differentially fed neural network, it has exhibited its optimal capacity and performance of performing formal verification of different categories of mixed signals. The entire processing has been carried out with higher accuracy and with lower processing time. Also testified with four possible types of mixed signals, but our framework offers more flexibility to add more number of mixed signals in order to perform formal verification with similar performance exhibits. Our future work will be in the direction of further investigating the categories of errors and perform further minimization of them.

## References

1. Traub, J.F.J.: Formal Verification of Concurrent Embedded Software. BoD – Books on Demand (2016)
2. Zhan, N., Wang, S., Zhao, H.: Formal Verification of Simulink/Stateflow Diagrams: A Deductive Approach. Springer, Heidelberg (2016)
3. Huang, S.-Y., Cheng, K.-T.: Formal Equivalence Checking and Design Debugging. Springer Science & Business Media, New York (2012)
4. Li, L., Thornton, M.A.: Digital System Verification: A Combined Formal Methods and Simulation Framework. Morgan & Claypool Publishers, San Rafael (2010)
5. Scheffer, L., Lavagno, L., Martin, G.: EDA for IC System Design, Verification, and Testing. CRC Press, Boca Raton (2016)
6. Louerat, M.-M., Maehne, T.: Languages, Design Methods, and Tools for Electronic System Design: Selected Contributions from FDL. Springer, Heidelberg (2014)
7. Cerny, E., Dudani, S., Havlicek, J., Korchemny, D.: The Power of Assertions in SystemVerilog. Springer, Heidelberg (2010)
8. Verification with Model Checking. https://github.com/johnyf/tool_lists/blob/master/verification_synthesis.md. Accessed 12 Jan 2017
9. VLSI Professional Network. http://vlsi.pro/formal-verification-an-overview/. Accessed 12 Jan 2017
10. Bailey, B., Balarin, F., McNamara, M.: TLM-driven Design and Verification Methodology (2010). Lulu.com
11. Bailey, B., Martin, G.: ESL Models and their Application: Electronic System Level Design and Verification in Practice. Springer, New York (2009)
12. Almeida, J.B., et al.: An overview of formal methods tools and techniques. In: Rigorous Software Development. Undergraduate Topics in Computer Science, pp. 15–44. Springer, London (2011)
13. Weib, B.: Deductive Verification of Object-oriented Software: Dynamic Frames, Dynamic Logic and Predicate Abstraction. KIT Scientific Publishing, Karlsruhe (2011)
14. Schumann, J.M.: Automated Theorem Proving in Software Engineering. Springer, Heidelberg (2013)

15. Saha, I., Roy, S., Ramesh, S.: Formal verification of fault-tolerant startup algorithms for time-triggered architectures: a survey. Proc. IEEE **104**(5), 904–922 (2016)
16. Alam, Q., et al.: Formal verification of the xDAuth protocol. IEEE Trans. Inf. Forensics Secur. **11**(9), 1956–1969 (2016)
17. Calinescu, R., Ghezzi, C., Johnson, K., Pezze, M., Rafiq, Y., Tamburrelli, G.: Formal verification with confidence intervals to establish quality of service properties of software systems. IEEE Trans. Reliab. **65**(1), 107–125 (2016)
18. Campos, J.C., Sousa, M., Alves, M.C.B., Harrison, M.D.: Formal verification of a space system's user interface with the IVY workbench. IEEE Trans. Human-Machine Syst. **46**(2), 303–316 (2016)
19. Cifuentes, F., Bustos Jimenez, J., Simmonds, J.: Formal verification of distributed system using an executable C model. IEEE Lat. Am. Trans. **14**(6), 2874–2878 (2016)
20. Webster, M., et al.: Toward reliable autonomous robotic assistants through formal verification: a case study. IEEE Trans. Human-Machine Syst. **46**(2), 186–196 (2016)
21. Vidhya, D.S., Manjunath, R.: Research trends in formal verification process for analog and mixed signal design. Int. J. Comput. Appl. **109**(11), 10–15 (2015)
22. Ain, A., Bruto da Costa, A.A., Dasgupta, P.: Feature indented assertions for analog and mixed-signal validation. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **35**(11), 1928–1941 (2016)
23. Lim, B.C., Jang, J.E., Mao, J., Kim, J., Horowitz, M.: Digital analog design: enabling mixed-signal system validation. IEEE Des. Test **32**(1), 44–52 (2015)
24. Yin, L., Deng, Y., Li, P.: Simulation-assisted formal verification of nonlinear mixed-signal circuits with Bayesian inference guidance. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **32**(7), 977–990 (2013)
25. Little, S., Walter, D., Myers, C., Thacker, R., Batchu, S., Yoneda, T.: Verification of analog/mixed-signal circuits using labeled hybrid petri nets. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **30**(4), 617–630 (2011)
26. Manjunath, R., Vasudev, S., Udupa, N.: Differential learning algorithm for artificial neural networks. Int. J. Comput. Appl. **1**(1), 65–70 (2010)
27. Manjunath, R., Gurumurthy, K.S.: System design using differentially fed artificial neural networks. In: TENCON 2002 (2002)

# The Effectiveness of Thai Spoonerism Application

Kunyanuth Kularbphettong(✉) and Tiwa Sreekram

Suan Sunandha Rajabhat University,
1 U-Tong-Nok Road, Dusit, Bangkok, Thailand
{kunyanuth.ku, tiwa.sr}@ssru.ac.th

**Abstract.** In this paper, we present the prototype of Thai Spoonerism Words Application for Bi-Syllable and Tri-syllable based on Android platform by using Thai spoonerism algorithm. This application supports to easily search and find the term of Thai spoonerism terms. To develop this application, an semi-automatic algorithm was applied to build the terms of Thai spoonerism and Thai word segmentation by using longest matching was adapted to enhance the efficiency of this application. The experimental results showed that this prototype can facilitate users to find the efficiency of information searching. To evaluate the preliminary prototype system, Black Box Testing and questionnaires were used to measure system performance and user satisfaction respectively. The results of data analysis by using questionnaires to evaluate user satisfaction were found that specialists and users have satisfied the performances of the system.

**Keywords:** Thai spoonerism · Semi-automatic algorithm · Android platform · Black box technique

## 1 Introduction

Spoonerism is a humorous mistake in speech or deliberate play on words in which corresponding consonants, vowels, or morphemes are switched the first sounds of two or more words [1] and it may lead to misunderstanding and a sense of humor. Thai is the official and the native language of Thailand and Thai spoonerism is widely used in a way of humor and impoliteness in order to significantly reduce immodesty such as using in Li-ke (Thai traditional dramatic performance), choi songs (local songs) and Folk songs.

Thai spoonerism is the way of punning, called กลับคำ "glàp kam/" Thai word play, the first sound of one word is exchanged with the first sound of another word and swapped positions. Monosyllable could not be spoonerism and multi-syllable is split into parts. Therefore, Thai spoonerism is the wisdom of the language that presents the ability of invent the terms of words to demonstrate the humor and avoid incivility.

The advance of mobile technologies has been prevalently acted as a significant device to support in many aspects of life and the mobile application industry will grow tremendously to match demand and keep up with ever-evolving technologies [2]. In this paper, we proposed the Thai spoonerism mobile application by adapting the hybrid

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_41

algorithm to learn how to create Thai spoonerism and give the knowledge to comprehend in spoonerism. This paper is organized as follows. Section 2 discusses previous works and techniques related to Thai spoonerism. Section 3 describes our approach and the processes of knowledge based construction, data preparation, and algorithm. Section 4 presents the experimental results, and shows some examples of the learned suggestion reviews. Finally, Sect. 5 concludes our findings and discussion for future studies.

## 2 Literature Reviews and Related Works

Empirical findings showed that the definition of spoonerism was defined in variety ways like punning, word play, speech error or form of humor [3, 4]. The highest number of syllables in spoonerism was eight syllable and the vowel exchange could create spoonerism by exchanging between the vowel and fixed syllable [5]. Also, there are many ways to create spoonerism by considering consonants, vowel and accent. Kaewrattanapat and Bunchongkien proposed the model to conduct Thai spoonerism word in bi-syllable and the results were shown that the model was satisfaction [6] and Tri-syllable Thai spoonerism algorithm was used initial consonant and cluster word and vowel, intonation mark and final consonant of first syllable and second syllable were transposed [7]. To reach the objectives of this research, mobile technology has also applied to implement and enhance adaptive system. Android platform was used to develop the mobile application to provide management and there is much of research that indicated how to provide requirements for design of a mobile learning. For instance, Schuck et al. [8]. proposed design-based method to implement in smartphone. Moreover, there are many related research that were considered in this project.

## 3 The Methodologies and Experimental Setup

The research aims to develop Thai spoonerism application based on mobile platform and the sample of this project consisted of 5 experts and 50 users. RAD (Rapid Application Development) was applied to implement this application, and user's requirements were analyzed for design processes to indicate user interface in a mobile learning device [9]. Each user was asked to key and test Thai spoonerism and then user's requirements were investigated to close with graphic user's needs. However, the methodologies and experimental set up of this project were as following this:

### 3.1 Thai Spoonerism Algorithm

Spoonerism is composed of one or two in three parts reversed together and there are many factors that effect to spoonerism like initial consonant, vowel sound and intonation tone. According to Kaewrattanapat and Bunchongkien, for the spoonerism of bi-syllables of, the first syllable and second syllable must not have the same initial-sound consonant for example *สวย-แจ่ม* (*suai-chaem*) and *ตาม-หา* (*tam-ha*). If there are same initial-sound consonant, the spoonerism can't be conducted as the results will be only word transposition. In case of Tri-syllable, the each syllable has the first letter as consonant and the second consonant as vowel after that all of which could be analyzed for example *คิด-ถึง-เอ็ง* (*khit-thueng-eng)* and *ปาก-น่า-จุ๊บ* (*pak-na-chup*). Also, the rules to compose spoonerism insists of: reversing vowels using only vowels change and consonants and accents still stay the original position, reversing consonants, reversing accents by changing the position of accents and reversing consonants and vowels.

The example of the rules for bi-syllable of spoonerism is

Syllable1: initial consonant-1, long vowel-1, final consonant-1, intonation marks-1
Syllable2: initial consonant-2, long vowel-2, final consonant-2, intonation marks-2
Syllable1: initial consonant-1, short vowel-1, final consonant-1, intonation marks-1
Syllable2: initial consonant-2, short vowel-2, final consonant-2, intonation marks-2

Moreover, the measurement in term of recall, precision, and accuracy of searching was used to evaluate information retrieval model.

$$Precision = \frac{number\ of\ correct\ terms\ in\ the\ system\ answer}{number\ of\ terms\ in\ the\ system\ answer} \tag{1}$$

$$Recall = \frac{number\ of\ correct\ terms\ in\ the\ system\ answer}{number\ of\ terms\ in\ the\ correct\ answer} \tag{2}$$

$$Accuracy = \frac{number\ of\ correct\ terms\ in\ system\ answer}{total\ number\ of\ correct\ terms\ in\ corrrect\ answer} \tag{3}$$

Also, the results of the proposed algorithm showed the satisfactory in the percentage of accuracy and both in precision and recall.

### 3.2 Experimental Setup

In an overview of the application, spoonerism system insists of 3 parts: input part, display part and evaluate part and user can select the option of spoonerism between bi-syllable and tri-syllable. When user keys the words, the system will firstly separate words by using longest matching algorithm of Thai word segmentation and then the application will translate these words to spoonerism by using Thai spoonerism algorithm and display in both texts and speech (Fig. 1).

## 4 The Experimental Results

In this project, it was divided the results by the research objectives into 2 parts: developing the Thai Spoonerism mobile application and evaluating and testing the application.

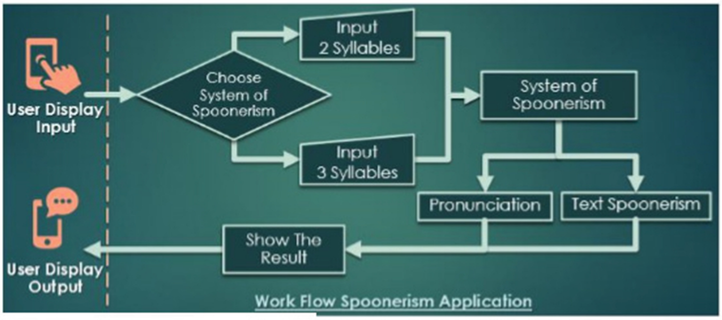


**Fig. 1.** The overview of this application

## 4.1 Developing the Thai Spoonerism Mobile Application

In this section, to develop the mobile application, Fig. 2 was shown the example results of mobile application.

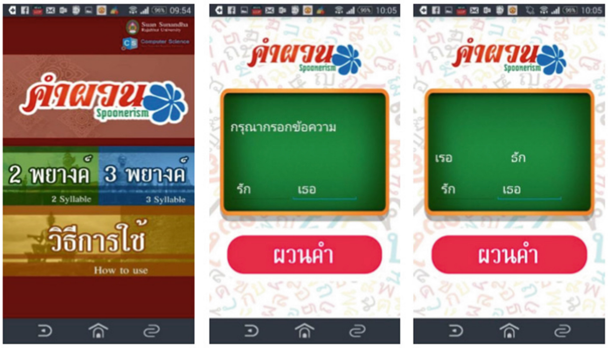


**Fig. 2.** The example results of mobile application

When the user enters the first of this application, users can select the menu to learn and fun with Thai spoonerism. User key the words that he/she needs to translate from the normal term to spoonerism terms. Also, users rate the score of satisfaction ratings of this application.

## 4.2 Evaluating and Testing the Mobile Application

Black box Testing and Questionnaires by experts and users were used to evaluate and test the qualities of this application. Respondents were asked to rate the recommendation results and the rating score was from 1 to 5. Black Box testing is the testing approach that focuses only on the outputs generated in response to selected inputs and execution conditions and the internal mechanism of a system or component is ignored [10]. Black box testing was assessed in the error of the project as following: functional requirement test, Function test, Usability test, Performance test and Security test. The collected data were analyzed by the statistical means and standard deviation (S.D.).

Functional Requirement Test is evaluated the satisfaction on the ability of the system so as to meet the needs of users and functional test was used to evaluate the accuracy of the system. Usability test is a measurement of the suitability of the system. The performance of the system is assessed the processing speed of the system in Performance Test. Finally, Security test was evaluated the security of the system and Table 1 was shown the results of Black box testing. The results showed that Thai spoonerism based on Mobile application was satisfied the requirements of users.

**Table 1.** The results of Black Box Testing

| | Experts | | Users | |
|---|---|---|---|---|
| | $\bar{x}$ | SD | $\bar{x}$ | SD |
| 1. Function requirement test | 4.17 | 0.554 | 4.51 | 0.659 |
| 2. Functional test | 4.15 | 0.610 | 4.28 | 0.718 |
| 3. Usability test | 4.23 | 0.610 | 4.45 | 0.645 |
| 4. Performance test | 4.03 | 0.416 | 4.31 | 0.531 |
| 5. Security test | 4.23 | 0.624 | 4.33 | 0.665 |

## 5 Conclusion and Future Works

In this work, we proposed Thai spoonerism application based on Mobile Application. This system provides more suitable recommendation Thai spoonerism words to users. Thai language is the diversity and flexibility language and it is so difficult to develop the system with precise and accuracy. However, the initial results showed that our approach is successfully generated Thai spoonerism in bi-syllable and tri-syllable for users to learn and comprehend spoonerism. As for the future work, we need to explore more reasonable other technologies and apply more tri-syllable spoonerism algorithm to enhance the performance of this project in the quality and quantity of services to users.

**Acknowledgments.** The authors gratefully acknowledge the financial subsidy provided by Suan Sunandha Rajabhat University.

## References

1. Spoonerism. https://en.wikipedia.org/wiki/Spoonerism
2. UAB businessdegrees Online. The future of mobile application. Accessed http://businessdegrees.uab.edu/resources/infographics/the-future-of-mobile-application/
3. Fromkin, V.A.: Speech Errors as Linguistic Evidence. Mouton, The Hague (1993)

4. Mackay, D.G.: Neuropsychologia **8**, 323–350 (1970)
5. Kaewrattanapat, N., Bunchongkien, W.: The algorithm of semi-automatic thai spoonerism words for bi-syllable. Int. J. Comput. Electr. Autom. Control Inf. Eng. **8**(8), 1430–1434 (2014). World Academy of Science, Engineering and Technology
6. Nookhong, J., Kaewrattanapat, N., Chaiwchan, W., Chaiya, K.: The performance comparison of algorithm of semi-automatic Thai Spoonerism words between bi-syllable and tri-syllable. In: Proceedings International Conference on Business Economic, Social Science & Humanities (BESSH 2016), Osaka, Japan, vol. 277, no. 7, pp. 77–84 (2016)
7. Desuwanna, W.: The Thai Children competence in Spoonerism. doi.nrct.go.th/ListDoi/Download/104333/0981657c7ca0e381b4b260cf7ba5b502?
8. Schuck, S., Aubusson, P., Kearney, M., Burden, K.: Mobilizing teacher education: a study of a professional learning community. Teacher Dev. **17**(1), 1–18 (2013)
9. Javatechig, Rapid Application Development Model. Accessed http://javatechig.com/seconcepts/rapid-application-development-model
10. Laouris, Y., Laouri, R.: Can information and mobile technologies serve to close the gap and accelerate development? In: Proceeding of MLearning, Alberta, Canada, 22–25 October 2006

# Advances in Transformation of MARTE Profile Time Concepts in Model-Driven Software Development

Anna Derezinska(✉) and Marian Szczykulski

Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland
a.derezinska@ii.pw.edu.pl

**Abstract.** UML models can be extended with time concepts from the Modeling and Analysis of Real-Time and Embedded Systems (MARTE) profile. In the Model-Driven Software Development, elements enhanced by stereotypes corresponding to time concepts can be transformed into code and assisted by appropriate library support during an application development and execution. We discuss several issues of the MARTE time concept interpretation and realization in an MDSD approach. Selected solutions were implemented in FXU, a tool for building C# applications based on UML classes and state machines. Realization of the MARTE support was verified in case studies.

**Keywords:** Model transformation · Code generation · UML · State machines · MARTE profile · Time modeling · MDD · MDSD

## 1 Introduction

The main ideas behind Model-Driven Software Development (MDSD) are preparation of comprehensive models of a system, their transformation, and forwarding them to execute [1]. Models can combine structural and behavioral descriptions (e.g. UML classes, components, state machines [2]), with features of a given application domain.

Among MDSD approaches, we focus on creation of a self-contained program application. The main steps of an assumed development process are preparation of models, transformation of models into a code in a general purpose language, and building of an executable application. The target application is based on an automatically generated code, specialized libraries supporting the modeled notions, and an additional code. The application can be run in a standard software development environment. In this paper, we do not discuss direct simulation of models, or executing of some intermediate forms of model transformations [3].

Design and development of domain models can be supported by UML profiles [2]. A profile includes a set of concepts denoted by stereotypes. Model elements are extended using stereotypes and their tagged values that define additional properties. In this way an element meaning can be enhanced while the UML meta-model remain unchanged (in most cases). The Modeling and Analysis of Real-Time Embedded Systems (MARTE) profile belongs to profiles published by the OMG consortium [4].

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_42

This profile has been widely used in system modeling, and UML/MARTE models were transformed into different implementations [5–9]. Some generation tools support MARTE notions, mainly in the Hardware Description Languages: SystemC, VHDL, Verilog. Models using MARTE, especially with detailed description of time behavior, can be applied in different domains. Therefore, we have analyzed transformation of classes and state machines extended by MARTE into a general purpose language. However, there is a lack of detailed analysis of such transformations, especially in the context of state machine models. Moreover, practical realization of an MDSD process required solving of some interpretation issues, e.g. dealing with semantic variation points of UML [2, 10, 11] and selecting working semantics of profile stereotypes.

The main contribution of this paper is presentation of interpretation and transformation of a MARTE subset, namely time concepts specified in MARTE::Time for classes and state machines. The proposed approaches were realized in an extended version of FXU (Framework for eXecutable UML) [12, 13] and verified in cases studies. The FXU tool supports code generation and execution of UML classes and all notions of state machines into C# code. Behavior of classes specified by state machines with MARTE refinements is reflected in a target C# application, which can act as a final implementation or an operational prototype. Thus, we can verify specification ideas, improve productivity and reduce time-to market.

The paper is organized as follows. In the next Section, we present interpretation and transformation issues of MARTE time concepts used in an MDSD process. Implementation of the MARTE support and its verification are briefly described in Sect. 3. We discuss related work in Sect. 4 and conclude the paper in Sect. 5.

## 2 Interpretation and Transformation of MARTE Time

Time concepts have been already presented in the standard OMG profile for Schedulability, Performance and Time (SPTP) [14], which was used in early UML versions (1.x). Simple time notions (*Time*, *TimeExpression, TimeObservation*) have been included in the UML specification since version 2.0 [2]. As the former SPTP profile was not consistent with new UML versions, a new extended OMG profile was developed. The Modeling and Analysis of Real-Time Embedded Systems profile (MARTE) [4, 15] can be used in UML 2.x and SysML models. Detailed concepts of a time domain are specified in a package of the foundation part, called *MARTE::Time*.

With the MARTE profile we can access physical and logical time structure using clocks. They refer directly to a time base on the time model. A clock has a set of units that can be accepted. A clock is an abstract concept specialized as a logical clock or a chronometric clock, which is assigned to a physical time. Time of a logical clock is usually counted in a number of ticks.

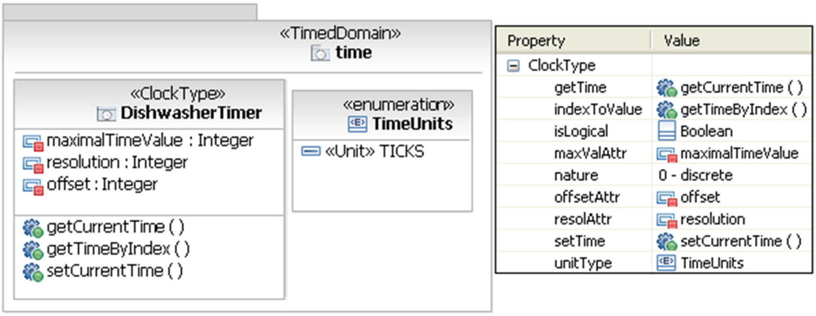
In the following subsections, time concepts in the context of an MDSD approach will be discussed. They correspond to stereotypies from the *MARTE::Time* profile. For each stereotype, we give (i) a short description, (ii) various interpretations with hints to realization of model to code transformation and run-time library, and (iii) a usage example. Examples refer to a case study of a dish washer controller (Sect. 3).

### 2.1 Stereotype *TimedDomain*

The *TimedDomain* stereotype can be assigned to a package treated as a container including definitions of clock types and objects of clocks. Such packages can be nested one in another. If a clock type or a clock object are placed in a package without this stereotype, many interpretations are possible. Lack of the *TimedDomain* stereotype can be ignored, and definitions of clock types and clock objects are allowed in any package. In another approach clock types and objects of clocks are disregarded if they are not located in a package extended with the *TimedDomain* stereotype.

In a practical realization, the above approaches can be combined into a hybrid one. Classes specifying clock types should always be generated, regardless being included in a *TimedDomain* package or not. This solution is motivated by a fact, that clock types can be used for different purposes in a model and their code should always be offered. A more restricted rule is proposed for clock objects, which have to be placed in an adequately stereotyped package. Those objects are used only for time measurement, therefore should always be placed in a package assigned to this domain.

An example of the *TimedDomain* stereotype associated with a package is shown in Fig. 1. The package contains a *Dishwasher Timer* class and an enumeration.



| Property | Value |
|---|---|
| ClockType | |
| getTime | getCurrentTime ( ) |
| indexToValue | getTimeByIndex ( ) |
| isLogical | Boolean |
| maxValAttr | maximalTimeValue |
| nature | 0 - discrete |
| offsetAttr | offset |
| resolAttr | resolution |
| setTime | setCurrentTime ( ) |
| unitType | TimeUnits |

**Fig. 1.** Examples of the *TimeDomain* and *ClockType* stereotypes

### 2.2 Stereotype *ClockType*

The *ClockType* stereotype can be assigned to any class that specifies a type of a clock. There are several tagged values used for detailed specification of a clock type. If these values are fixed, they can be defined directly in a model. Otherwise tags are defined as attribute values and operations in a class. In this way, tagged values can differ in dependence on a clock instance. Tags of *ClockType* have the following meaning:

- *isLogical* - a clock type: *true* for a logical clock, *false* for a physical one (an attribute to get system time is necessary, e.g. *System.DateTime* for C#),
- *nature* - represents the discrete or dense time nature (logical clocks have always discrete time),
- *maxValAttr* – after reaching this value, time is counted from the beginning (logical clock only),

- *setTime/getTime* – operations for changing/reading time value (logical clock only),
- *indexToValue* – an operation to map a time index (number of an event) to a real time value (logical clock only),
- *ResolAttr* – resolution of an associated clock (defined for discrete time only),
- *offsetAttr* – offset of the associated clock expressed in the default time units,
- *unitType* – a set of supported time units.

A class with the *ClockType* stereotype can be transformed into a code as any other class in a program. Its tags are implemented using corresponding methods and fields to store appropriate values. The class is also supplemented with additional data and methods to control a clock behavior in accordance to the *MARTE::Time* specification.

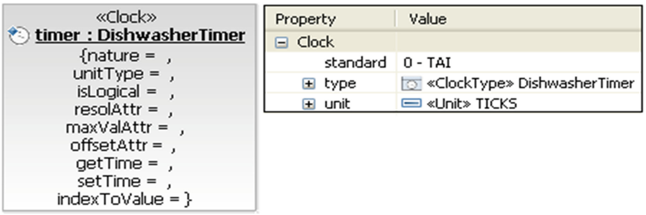
Exemplary application of *ClockType* is shown in Fig. 1. The *DishwasherTimer* class is specified as *ClockType* with a set of tags. Tag names and values are listed in an additional window. An enumeration defines time units accepted by this clock type. Each item should be denoted by the *Unit* stereotype belonging to a package of Non-functional Properties Modeling *MARTE::NFP*s.

## 2.3 Stereotype *Clock*

An instance of a class stereotyped with *ClockType* is labelled with the *Clock* stereotype. This kind of object should be located in an object diagram defined in a package with the *TimeDomain* stereotype. It is used to access time by other elements from the *MARTE::Time*, namely *TimedProcessing, TimedEvent,* and *TimedValueSpecification*.

There are additional properties of a clock. A reference class with the *ClockType* stereotype is defined in the *type* tag. A default time unit used by the clock is given in the *unit* tag. The unit has to belong to a set of units listed in the appropriate clock type. A chronometric clock has a time *standard* specified with a tag. It is equal to one of predefined values from the *TimeStandardKind* MARTE Library [4–Annex D.3.].

Code generation of the *Clock* stereotype requires transformation not only of classes but also of object diagrams. This facility can be restricted to objects that are annotated with the *Clock* stereotype and placed in a package with the *TimeDomain* stereotype. In a class defining a clock type, a static method can be generated for any clock object. The method returns an instance of the clock type specified with parameters defined in an object diagram. Usage of the *Clock* stereotype is illustrated in Fig. 2. A *timer* object is an instance of the *DishwasherTimer* class (Fig. 1).



«Clock»
timer : DishwasherTimer
{nature = ,
unitType = ,
isLogical = ,
resolAttr = ,
maxValAttr = ,
offsetAttr = ,
getTime = ,
setTime = ,
indexToValue = }

| Property | Value |
|---|---|
| ⊟ Clock | |
| standard | 0 - TAI |
| ⊞ type | «ClockType» DishwasherTimer |
| ⊞ unit | «Unit» TICKS |

**Fig. 2.** An instance of *DishwasherTimer* - an example of the *Clock* stereotype

## 2.4 Stereotype *TimedProcessing*

The *TimedProcessing* stereotype can be assigned to any element that has behavior specified by its start and end points, or by a duration time. For example, such a behavioral element can be a whole state machine, an action (*Do, Entry, Exit*) in a state, or an action labelling a transition between states. A stereotyped element is associated with a clock using tag *on*. The stereotype can be refined with several tags: *duration, finish, start*. There are many possible interpretations of specified time behavior in dependence of these tags.

1. *Start* and *finish* tags are specified and a *duration* tag is not. In general, events given in *start* and *finish* stand for beginning and ending points of the behavior. Moreover various interpretation cases are possible:
   (a) After encountering of a start event behavior is started and ended with a finish event. However, behavior can be started and/or ended also in another way. Several solutions can be chosen if a start event happens during an active behavior:
      - behavior cannot be activated once again as it is already in progress,
      - a new behavior instance is launched, e.g. in a separate thread,
      - the behavior is reactivated after its end.

      A *finish* event encountered while the behavior is not active. Then, we can
      - ignore the *finish* event,
      - save the *finish* event in a buffer/queue and wait for activation of the behavior.

      Then, the behavior is ended due to the *finish* event.

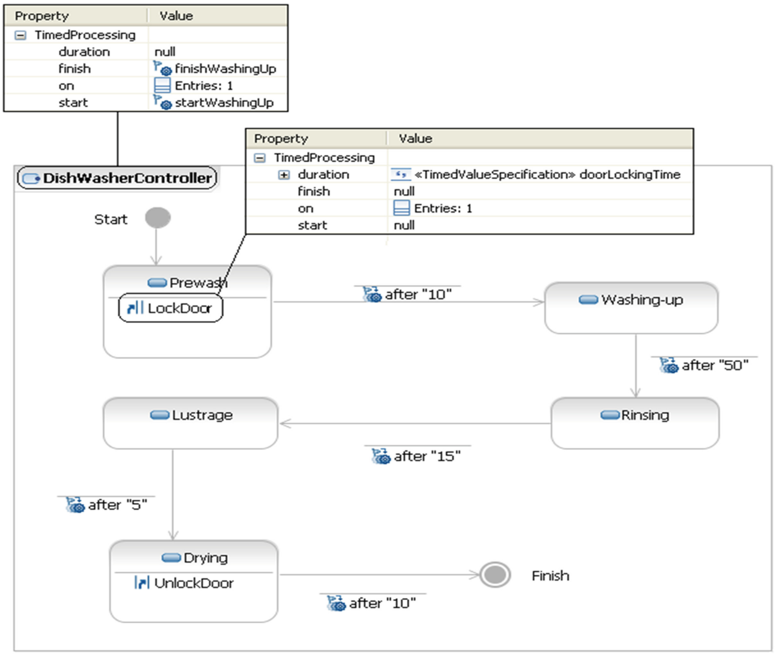
   (b) A behavior can be started and ended only via *start* and *finish* events. Though, not all *start/finish* events can launch/end a behavior.
   (c) Combination of cases (a) and (b), i.e. each occurrence of a *start/finish* event starts/ends a behavior and it is not allowed to activate or deactivate any behavior in another way.
   (d) *Start/finish* events are generated when a behavior starts or ends. This interpretation is opposite to above ones, where the events acted as triggers.
2. A *duration* tag is specified and *start* and *finish* tags are not. In general, a behavior should last for time equal to a *duration* tag value. Time is counted with a clock specified in an *on* tag. However, different interpretations can be used:
   (a) The exact duration time will be forced. If a behavior is longer than stated by the *duration* tag, it is interrupted. Otherwise, if a behavior is due to finish before elapsing of duration time, it is prolonged to be as long as the duration value.
   (b) A behavior can lasts no longer than specified by *duration* value. Otherwise it will be interrupted.
   (c) A behavior should last for an interval equal at least to the *duration* value. Otherwise an error occurs (e.g. an exception is raised) or the behavior end will be postponed to the required moment.

3. All *start*, *finish* and *duration* tags are specified. In this case, we expect a behavior to start in the same moment as the *start* event occurrence, and end at the *finish* event occurrence. But additionally, passing time is restricted by the *duration* tag. If we select the variant when the event *finish* triggers the behavior end, different interpretations are still possible:
   (a) Each *finish* event is ignored during an interval of the *duration* time, which is counted since a *start* event occurrence.
   (b) If a *finish* event occurred and a time interval counted from the *start* event occurrence is shorter than the *duration* value, an error is generated.

Our recommendations depend of a meta-model element to which the stereotype is assigned. In case of a state machine, a *start* event is interpreted as in 1(a), and *finish* as 1(d). When an action in a state or on a transition is concerned, both *start* and *finish* are handled according to 1(d). The *duration* tag will be processed as 2(b) in all cases.

Usage of the *TimedProcessing* stereotype is shown in Fig. 3. The *DishWasher Controller* state machine is specified with this stereotype and its tags: *start* and *finish*. The tags have their events specified. The state machine begins is activity after occurrence of the *start* event. After the end of the behavior, the *finish* event will be launched.

Entry action in the *Prewash* state is also specified with the *TimedProcessing* stereotype. In this case, the *duration* tag is defined by a literal with the *TimedValue Specification* stereotype (Sect. 2.6).



| Property | Value |
|---|---|
| ⊟ TimedProcessing | |
| duration | null |
| finish | finishWashingUp |
| on | Entries: 1 |
| start | startWashingUp |

| Property | Value |
|---|---|
| ⊟ TimedProcessing | |
| ⊞ duration | «TimedValueSpecification» doorLockingTime |
| finish | null |
| on | Entries: 1 |
| start | null |

**Fig. 3.** State machine of *DishWasher Controller* - examples of the *TimedProcessing* stereotype

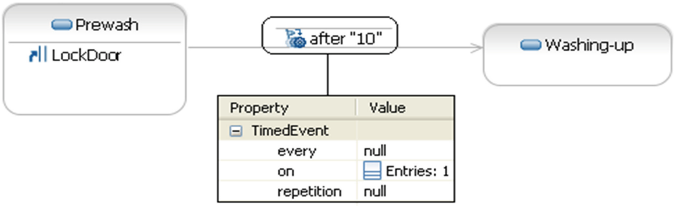
## 2.5 Stereotype *TimedEvent*

Any time event from the *CommonBehaviours::SimpleTime* package of UML can be specified by the *TimedEvent* stereotype. Therefore, additional data of a time event can be specified, or a cyclic time event is created.

Value of a timed event determines when the event is to be generated for the first time. When an event flag *isRelative* is false, its value denotes an absolute time instant presented by the associated clock. Otherwise the event value defines a time between an event instance generation and its entering a queue. Then, the *every* tag denotes a duration time between event occurrences. The number of occurrences is limited by the *repetition* tag. Value of a timed event is defined by a CVS expression (Clocked Value Specification) [4 – Annex C].

Realization of this stereotype requires handling of time events that are placed into appropriate event queues of state machines. The same event can be generated many times, if necessary. Different realizations of time events influence performance of a target application. Variants of time event handling in MDSD were presented in [16].

Several transitions in the dish washer state machine are labelled with time events (Fig. 3). These events are extended with *TimedEvent* stereotypes. An example of a transition and the properties of its event are shown in Fig. 4. Values of tags *every* and *repetition* are empty, as the event is not repeatable. The event occurs after 10 time units in accordance to an associated clock (tag *on*). This time is measured since entering the *Prewash* state.



| Property | Value |
|---|---|
| ⊟ TimedEvent | |
| every | null |
| on | Entries: 1 |
| repetition | null |

**Fig. 4.** A transition of *DishWasher Controller* (Fig. 3) with the *TimedEvent* stereotype

## 2.6 Stereotype *TimedValueSpecification*

This stereotype can be assigned to any value in a UML model (*Classes::Kernel::ValueSpecification*). *TimedValueSpecification* denotes that a corresponding value is interpreted as a time value of a clock referenced by an *on* tag. Meaning of the stereotyped value depends on the *interpretation* tag:

- *Duration* –value of a time interval passing after an event,
- *Instant* – value of a time instant in a given clock,
- *Any* – a duration or instant value in accordance to a TSL (Time Specification Language) expression.

Time expressions are written in TSL, a part of VSL (Value Specification Language) [4–Annex B], therefore realization of the stereotype requires translation of such expressions according to the defined grammar.

## 3 Support for the MARTE Profile in FXU

Framework for eXecutable UML (FXU) was developed as a first tool that supported transformation of classes and state machines into C# code [12]. Its main goal was transforming all notions of state machines into an executable code. Its functionality was enhanced within consecutive versions [13, 16].

In general, FXU consists two parts: FXU Generator that transforms UML class and state machine models into corresponding C# code, and FXU Run-Time Library that implements state machine concepts and is incorporated into a final application. In order to make the tool more flexible and to combine elements of MARTE profile, the tool architecture was refactored. It was made extendable by a set of plug-ins. A plug-in component is responsible for interpretation of elements (stereotypes, tagged values) of a given profile, insertion of appropriate changes of a model and generation of an additional code to realize the profile. Appropriate extensions were also integrated with the Run_Time Library.

Using the new FXU architecture, the tool was facilitated with MARTE code generation. The extended library supports run-time realization of the profile notions according to interpretations given in the previous Section.

FXU with MARTE was used in different case studies. One of them was related to a home alarm system combining features of two models: an intrusion alarm and a fire alarm [17]. Experiences of the model and application development influenced selecting among variants of MARTE stereotype realization.

Examples presented in this paper originate from a case study of a dish washer, mainly its controller. A logical clock is used in the model, therefore an activity time is independent from a physical time. A whole cycle of a dish washer work lasts for 90 logical time units, e.g. assuming 1 min for a unit it makes 90 min. In the state machine of the dish washer controller, movements between consecutive states are realized according to specified time requirements. The washing process starts with the *startWashingUp* event and ends with the *finishWashingUp* event, as stated in the tags of the *TimedProcessing* stereotype of the state machine.

Experimental verification of the FXU with MARTE was also carried out on various models aiming at utilization of all concepts from the *MARTE:Time* profile, in particular: extensive usage of time events, managing of time-driven processes, testing of various clock types, etc. Another case study was based on an example of a 4 stroke engine [18]. The performed experiments confirmed a proper utilization of time concepts in the model-driven application development.

## 4 Related Work

Models with the MARTE profile are widely applied in system modeling and analysis. Therefore, different transformations, mainly into domain targets, were proposed.

UML/MARTE models were used in HW/SW co-design approaches. In [5] models were transformed into SystemC executable used in simulation to verify a target VHDL. Results were applied in an FPGA solution of multimedia embedded systems.

Rapid prototyping of heterogeneous embedded HW/SW systems under consideration of timing and power aspects was presented in [6]. MARTE/UML models were transformed to IP-XACT and further into SystemC. Executable specification was used in an estimation and simulative analysis of timing and power properties of a system.

Generation of the System-Level Architecture Model (S-LAM) from a UML model with the MARTE profile is presented in [7]. Data-parallel applications were designed to be executed on a massively parallel System-on-Chip. The Gaspard2 tool supports transformation of MARTE/UML into OpenCL, a standard for parallel computing [8].

It should be noted, that in the papers discussed above, no information about dealing with state machines and the MARTE interpretation issues are given. State machines were taken into account in [9] where the MODCO transformation tool was presented. However, this approach covers only a small subset of UML state diagram constructs, supporting neither hierarchy nor concurrency.

There are CASE tools that support modeling with UML profiles, including MARTE, e.g. IBM RSA (since v. 7.0), Papyrus, MagicDraw, etc. but only some of them deal also with transformation of MARTE models, like Papyrus.

In the contrary to other approaches, our target is not a domain language, but a general purpose language, namely C#. Moreover, in the code generation and building an application we focus on state machine transformation. We take into account all features of state machines, including complex states with orthogonal regions, history, all pseudostates, etc.

## 5 Conclusions

Modeling of a system with time notions, its automatic transformation and building of an application combined with the support of modelled concepts gives an opportunity to create a well-specified reliable application. Therefore, we discussed transformation variants of MARTE time concepts that were implemented in an MDSD tool. It transforms classes and state machines into C# and supports building an application.

A direction which benefits from the discussed approach is rapid prototyping. A final application can cover control and time-related parts of a system functionality. Detailed modeling of other system features can be cumbersome in an MDSD, and therefore postponed to implementation in a programming language. Taking into account verification purposes, the application can be treated as a conceptual prototype, or an operational prototype for further code extension. Moreover, processing of call and time events specified in state machines can be realized within a target application with a satisfactory performance [16].

## References

1. Liddle, S.W.: Model-driven software development. In: Embley, D.W., Thalheim, B. (eds.) Handbook of Conceptual Modeling, pp. 17–54. Springer, Heidelberg (2011)
2. Object Management Group, OMG Unified Modeling Language (2015). http://www.omg.org/spec/UML/
3. Dominguez, E., Perez, B., Rubio, A.L., Zapata, M.A.: A systematic review of code generation proposals from state machine specifications. Inf. Softw. Technol. **54**(10), 1045–1066 (2012)
4. Object Management Group, UML Profile for MARTE: Modeling and Analysis of Real-Time Embedded Systems. version 1.1. (2011). http://www.omg.org/spec/MARTE/
5. de la Fuente, D., Barba, J., Lopez, J.C., Peñil, P., Posadas, H., Sanchez, P.: Synthesis of simulation and implementation code for OpenMAX multimedia heterogenous system from UML/MARTE models. Multimedia Tools Appl., 1–32 (2016). doi:10.1007/s11042-016-3448-5
6. Grüttner, K., Hartmann, P.A., Hylla, K., Rosinger, S., Nebel, W., Herrera, F., Villar, E., Brandolese, C., Fornaciari, W., Palermo, G., Ykman-Couvreur, C., Quaglia, D., Ferrero, F., Valencia, R.: The COMPLEX reference framework for HW/SW co-design and power management supporting platform-based design-space exploration. Microprocess. Microsyst. **37**, 966–980 (2003)
7. Ammar, M., Baklouti, M., Pelcat, M., Desnos, K., Abid, M.: Automatic generation of S-LAM descriptions from UML/MARTE for the DSE of massively parallel embedded systems. In: Lee, R. (ed.) Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2015. SCI, vol. 612, pp. 195–211. Springer, Cham (2016). doi:10.1007/978-3-319-23509-7_14
8. Wendell, A., Rodrigues, O., Guyomarch, F., Dekeyser, J.-L.: An MDE approach for automatic code generation from UML/MARTE to OpenCL. Comput. Sci. Eng. **15**(1), 46–55 (2013)
9. Coyle, F., Thornton, M.: From UML to HDL: a Model-Driven Architectural Approach to Hardware-Software Co-Design. Information Systems: New Generations Conference (ISNG) (2005)
10. Prout, A., Atlee, J.M., Day, N.A., Shaker, P.: Code generation for a family of executable modelling notations. Softw. & Syst. Model. **11**, 251–272 (2012)
11. Derezińska, A., Szczykulski, M.: Interpretation problems in code generation from UML state machines - a comparative study. In: Kwater, T. (ed.) Computing in Science and Technology 2011: Monographs in Applied Informatics, pp. 36–50. Depart. of Applied Informatics Faculty of Applied Informatics and Math. Warsaw Univ. of Life Sciences (2012)
12. Pilitowski, R., Derezinska, A.: Code generation and execution framework for UML 2.0 classes and state machines. In: Sobh, T. (ed.) Innovations and Advanced Techniques in Computer and Information Science and Engineering, pp. 421–427. Springer, Dordrecht (2007)
13. FXU Framework for eXecutable UML. http://galera.ii.pw.edu.pl/~adr/FXU/
14. Object Management Group, UML Profile for Schedulability, Performance, and Time Specification, version 1.1. (2005). http://www.omg.org/spec/SPTP/
15. Selic, B., Gerard, S.: Modeling and Analysis of Real-Time and Embedded Systems with UML and MARTE. Developing Cyber-Physical Systems. Elsevier (2014)

16. Derezińska, A., Szczykulski, M.: Performance evaluation of impact of state machine transformation and run-time library on a C# application. In: Kobayashi, S.-y., Piegat, A., Pejaś, J., El Fray, I., Kacprzyk, J. (eds.) ACS 2016. AISC, vol. 534, pp. 328–340. Springer, Cham (2017). doi:10.1007/978-3-319-48429-7_30
17. Derezińska, A., Szczykulski, M.: Application of time concepts from the MARTE profile in a model-driven development case study. Przeglad Elektrotechniczny (Rev. Electr. Eng.) **2015** (11), 178–181 (2015)
18. Andre, C., Mallet, F., Simone, R.: Time modeling in MARTE. In: ECSI Forum on specification & Design Languages (FDL), Barcelona, Spain. ECSI, pp. 268–273 (2007)

# Evaluating Suitable Hotel Services in Hotel Booking System Using Expert System

Bogdan Walek[1(✉)], Oldrich Hosek[1], and Radim Farana[2]

[1] Department of Informatics and Computers,
University of Ostrava, 30. dubna 22, 701 03 Ostrava, Czech Republic
bogdan.walek@osu.cz, P15182@student.osu.cz
[2] Institute for Research and Applications of Fuzzy Modeling,
30. dubna 22, 701 03 Ostrava, Czech Republic
radim.farana@osu.cz

**Abstract.** This paper deals with a fuzzy expert system for proposing suitable hotel services in a hotel booking system. The main idea of paper is to propose the most suitable hotel services based on fuzzy expert system and hotel guest preferences. Evaluation of hotel services uses an special fuzzy expert system with a knowledge base and information from a questionnaire with evaluative linguistic expressions filled-in by hotel guests. Proposed expert system was created using Linguistic Fuzzy Logic Controller. In the proposed fuzzy expert system the theory of Natural Fuzzy Logic is applied.

**Keywords:** Hotel booking · Hotel booking system · Fuzzy logic · Natural Fuzzy Logic · Fuzzy expert system · Hotel service · Questionnaire

## 1 Introduction

Nowadays, they are many hotels which offer their accommodation and other services using hotel booking systems. Generally, each hotel has its own website with a registration form or booking system for guests to provide reservation of accommodation. They might use some of professional systems, for example [1, 2].

Hotel booking websites or own booking systems on websites of hotels often provide room reservation and food reservation for their guests.

In this paper we would like to continue in work with expert system for hotel booking system, specially we would like to show technical details of expert system, verification and experimental results. The paper is continuation of paper [3].

## 2 Problem Formulation

Potential guests of specific hotel also plan to visit some interesting places close to the hotel and use interesting services of hotel. They also plan one-day trips and decide which possible activities are interesting. The importance of this issue is highlighted by a number of realized analyses and proposed prediction systems, for example [4–6].

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_43

Current booking systems often lack the possibility to book hotel services or to plan interesting one-day trip or visiting some interesting places.

There are few important functionalities which are missing in current booking systems:

- Showing available services during the process of booking accommodation
- Showing prices of available services
- Showing availability of services in term of guest accommodation
- Evaluating service suitability with respect to the needs of guest
- Showing interesting places to visit
- Showing planning tool to plan one or multiday trip
- Showing other services nearby interesting places to visit (restaurants, relaxation zone, transport services, etc.)

There are few solutions solved in our problem domain – hotels and hotel booking systems, but they don´t solve evaluating suitable hotel services. These solution have been presented in previous works, some of them are also based on expert systems [7–10].

## 3 Problem Solution

Based on the above-mentioned reasons, we propose a fuzzy system for a hotel booking system. The main aim of fuzzy system is to evaluate suitable hotel services shown during booking accommodation in specific hotel. The main parts of proposed fuzzy system were published in paper [3]. The proposed fuzzy system is connected to a simple questionnaire to detect guest preferences [11] and to database of POI (points of interest), activities and events. Fuzzy system is proposed using the knowledge obtained by authors in previous work [12, 13] and visually is shown in Fig. 1.

Main parts of proposed fuzzy expert system are more described in paper [3].

### 3.1 Technical Details of Fuzzy System

Proposed fuzzy system was developed as a web application which represents hotel booking system with these main functionalities:

- Selecting a specific hotel (implemented in a web application)
- Filling check-in and check-out dates (implemented in a web application)
- Selecting a suitable room and type of food (implemented in a web application)
- Filling personal data and questionnaire to detect guest preferences (implemented in a web application)
- Evaluating suitable hotel services using expert system (experimentally verified)
- Showing evaluated hotel services and select the most suitable service by guest (experimentally verified)
- Checking availability of selected hotel services in a specific day and time (in the implementation process).

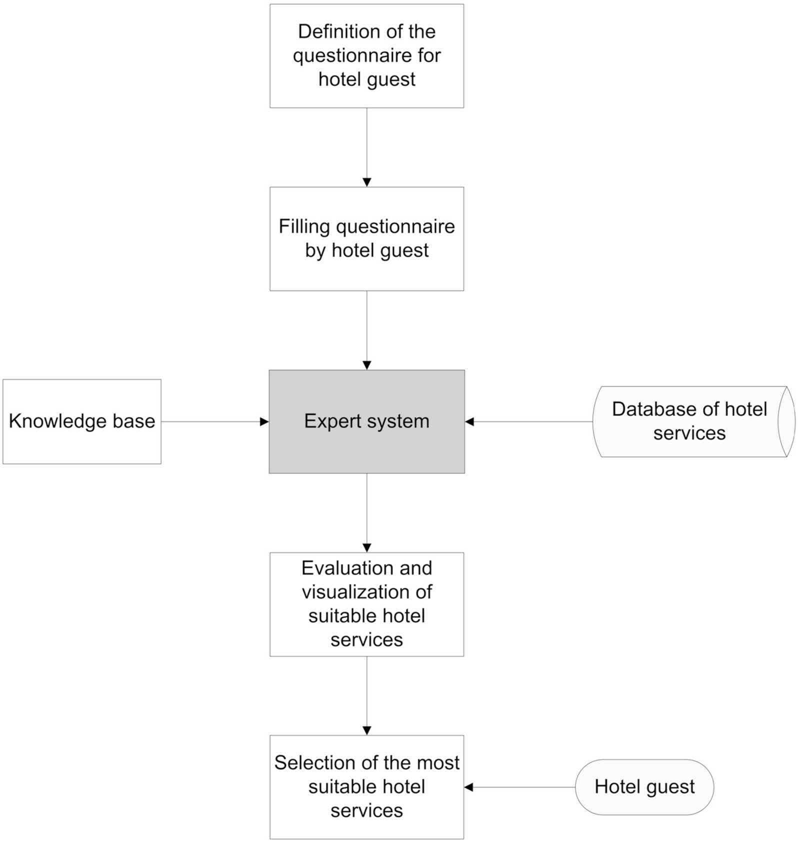

Definition of the questionnaire for hotel guest
Filling questionnaire by hotel guest
Knowledge base
Expert system
Database of hotel services
Evaluation and visualization of suitable hotel services
Selection of the most suitable hotel services
Hotel guest

**Fig. 1.** Proposed fuzzy expert system for evaluating the most suitable hotel services

As we can see in previous list, there are some functionalities which were implemented in a web application. Evaluating suitable hotel services using expert system was experimentally verified and will be integrated into a web application in the future. This part of the fuzzy system is more described in chap. 4 called Verification.
Proposed web application was developed using these technologies:

- HTML5, CSS3, PHP, jQuery
- PHP framework Nette
- MySQL database
- LFL Controller (for creating expert system).

## 3.2 Expert System for Evaluating Suitable Hotel Services

The knowledge base of the expert system consists of IF-THEN which are created by a hotel expert. Based on answers from the questionnaire and knowledge base, the expert system evaluates each hotel service and determines the level of suitability of a particular hotel service.

The structure of the proposed expert system is shown below.
Input linguistic variables:

- A1 – answer of question Q1
- A2 – answer of question Q2
- A3 – answer of question Q3
- A4 – answer of question Q4

Output linguistic variable:

- Level of suitability of specific hotel service

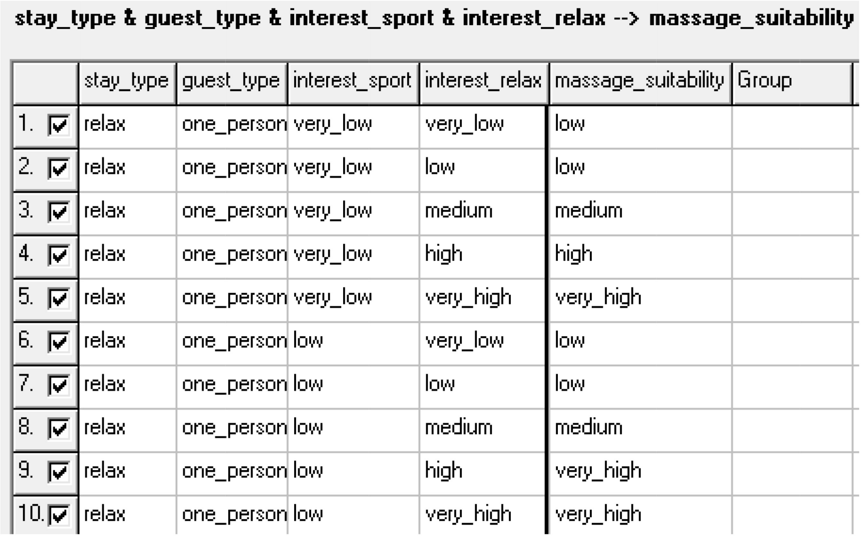
Examples of IF-THEN rules for hotel service called “massage” are shown below:

```
IF(STAY TYPE IS RELAX) AND
(GUEST TYPE IS COUPLE) AND
(INTEREST_SPORT IS HIGH) AND
(INTEREST_RELAX IS VERY HIGH) THEN
(MASSAGE_SUITABILITY IS VERY HIGH)

IF(STAY TYPE IS BUSINESS) AND
(GUEST TYPE IS ONE PERSON) AND
(INTEREST_SPORT IS MEDIUM) AND
(INTEREST_RELAX IS MEDIUM) THEN
(MASSAGE_SUITABILITY IS MEDIUM)

IF(STAY TYPE IS SPORT) AND
(GUEST TYPE IS FAMILY WITH SMALL CHILDREN) AND
(INTEREST_SPORT IS VERY HIGH) AND
(INTEREST_RELAX IS VERY LOW) THEN
(MASSAGE_SUITABILITY IS LOW)
```

The creation of the expert system knowledge base was performed in the LFL Controller. Linguistic Fuzzy Logic Controller is more described in [14].

Part of knowledge base for hotel service called “massage” is shown in the Fig. 2.

The evaluation of the hotel service is a deffuzification process that is realized by the defuzzyfication of fuzzy sets, one of them (medium) is shown in the Fig. 3.



stay_type & guest_type & interest_sport & interest_relax --> massage_suitability

| | stay_type | guest_type | interest_sport | interest_relax | massage_suitability | Group |
|---|---|---|---|---|---|---|
| 1. ☑ | relax | one_person | very_low | very_low | low | |
| 2. ☑ | relax | one_person | very_low | low | low | |
| 3. ☑ | relax | one_person | very_low | medium | medium | |
| 4. ☑ | relax | one_person | very_low | high | high | |
| 5. ☑ | relax | one_person | very_low | very_high | very_high | |
| 6. ☑ | relax | one_person | low | very_low | low | |
| 7. ☑ | relax | one_person | low | low | low | |
| 8. ☑ | relax | one_person | low | medium | medium | |
| 9. ☑ | relax | one_person | low | high | very_high | |
| 10. ☑ | relax | one_person | low | very_high | very_high | |

**Fig. 2.** Part of knowledge base for hotel service “massage”

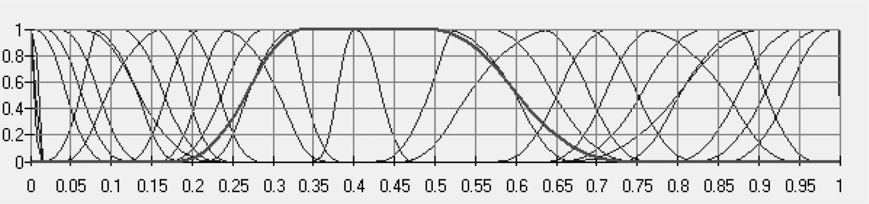


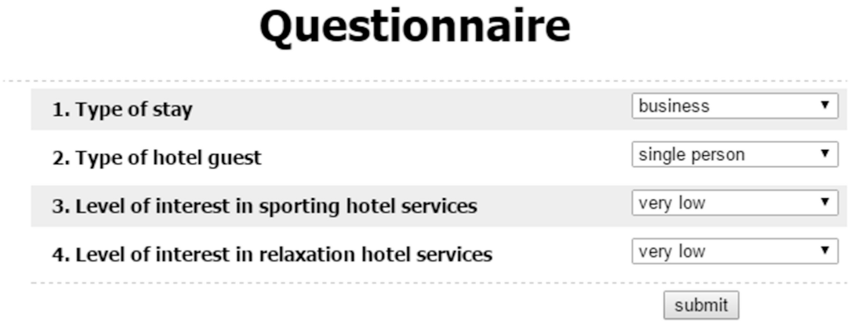
**Fig. 3.** Form of fuzzy set corresponding to the evaluation and defuzzyfication

## 4 Verification

For verification, we chose two types of hotel guests and their preferences verified on three different real hotels and their services. The verification is divided into a few steps and explained in the following subchapters.

### 4.1 Definition of the Questionnaire for Hotel Guest

In the first step, the questionnaire is defined. The prepared questionnaire is shown in Fig. 4.



Questionnaire

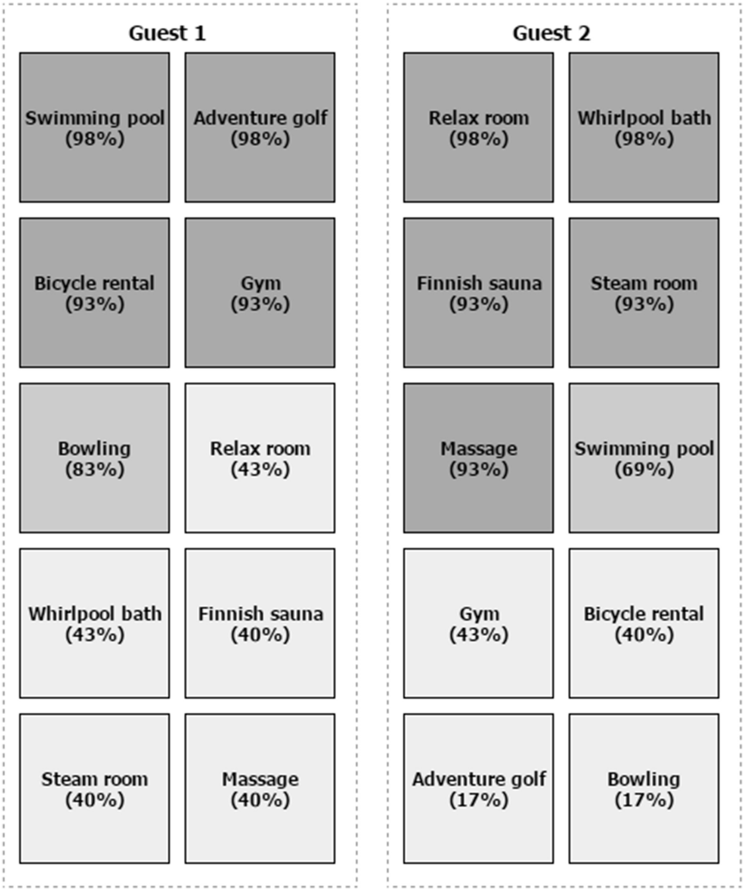
1. Type of stay: business ▾

2. Type of hotel guest: single person ▾

3. Level of interest in sporting hotel services: very low ▾

4. Level of interest in relaxation hotel services: very low ▾

**Fig. 4.** Prepared questionnaire for hotel guest



Guest 1

Swimming pool (98%)
Adventure golf (98%)
Bicycle rental (93%)
Gym (93%)
Bowling (83%)
Relax room (43%)
Whirlpool bath (43%)
Finnish sauna (40%)
Steam room (40%)
Massage (40%)

Guest 2

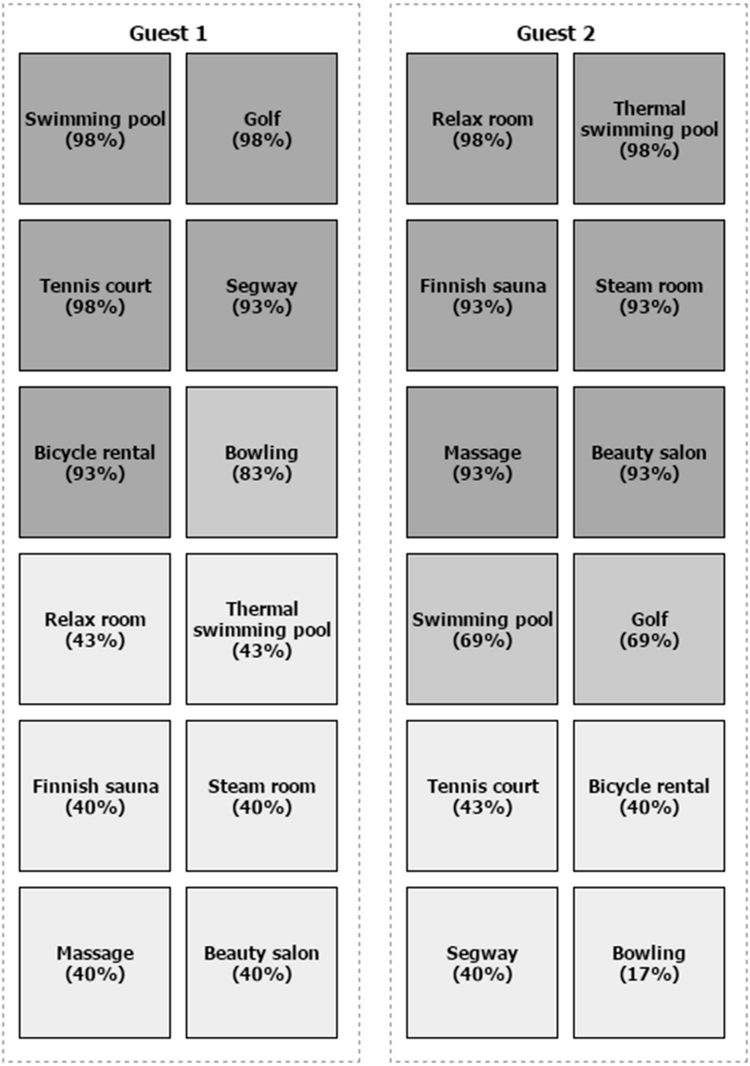
Relax room (98%)
Whirlpool bath (98%)
Finnish sauna (93%)
Steam room (93%)
Massage (93%)
Swimming pool (69%)
Gym (43%)
Bicycle rental (40%)
Adventure golf (17%)
Bowling (17%)

**Fig. 5.** The most suitable hotel services for Hotel 1

## 4.2 Filling Questionnaire by Hotel Guest

In this step, hotel guests fill in the questionnaire. Here are answers of selected hotel guests:
Guest 1:

- Type of stay – business
- Type of hotel guest – single person
- Level of interest in sporting hotel services – very high
- Level of interest in relaxation hotel services – medium.



Guest 1

Swimming pool (98%)
Golf (98%)
Tennis court (98%)
Segway (93%)
Bicycle rental (93%)
Bowling (83%)
Relax room (43%)
Thermal swimming pool (43%)
Finnish sauna (40%)
Steam room (40%)
Massage (40%)
Beauty salon (40%)

Guest 2

Relax room (98%)
Thermal swimming pool (98%)
Finnish sauna (93%)
Steam room (93%)
Massage (93%)
Beauty salon (93%)
Swimming pool (69%)
Golf (69%)
Tennis court (43%)
Bicycle rental (40%)
Segway (40%)
Bowling (17%)

**Fig. 6.** The most suitable hotel services for Hotel 2

Guest 2:

- Type of stay – wellness
- Type of hotel guest – couple
- Level of interest in sporting hotel services - low
- Level of interest in relaxation hotel services – very high.

### 4.3 Evaluation of Suitable Hotel Services

For verification three different real hotels in Beskydy mountains were selected. Hotels are marked as Hotel 1, Hotel 2 and Hotel 3. Hotel 1 offers 5 relaxation hotel services and 5 sporting hotel services. Hotel 2 offers 6 relaxation hotel services and 6 sporting hotel services. Hotel 3 offers 6 relaxation hotel services and 4 sporting hotel services.

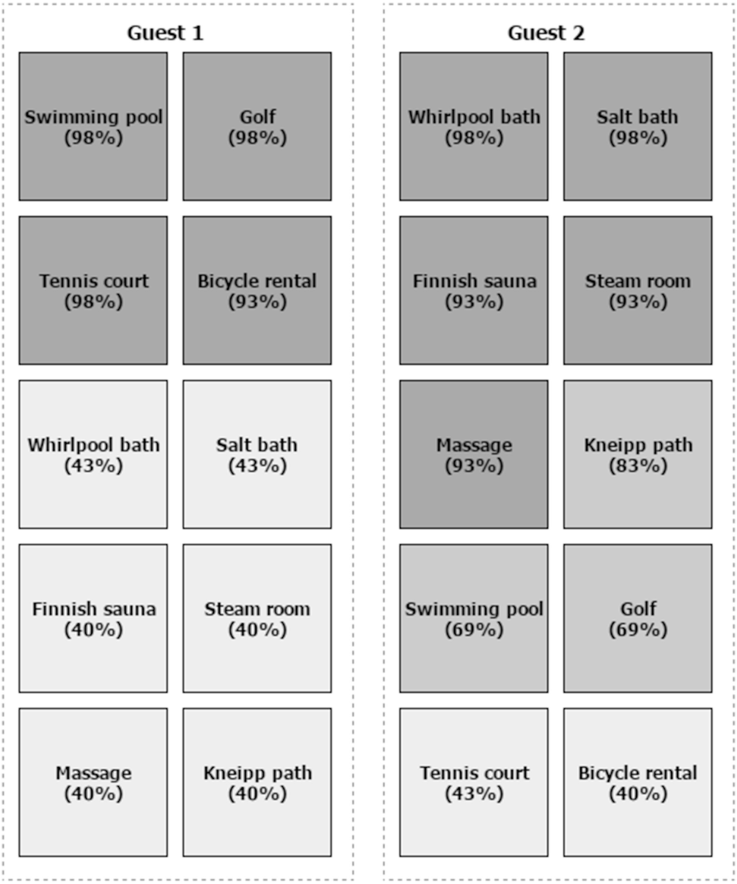


**Fig. 7.** The most suitable hotel services for Hotel 3

The expert system evaluates the hotel services and proposed the most suitable hotel services for each hotel guest based on their preferences (Fig. 5).

For each hotel service, the level of suitability is calculated and shown as a defuzzified value. The hotel services are shown in boxes with colour which determine the level of suitability for a specific hotel service:

- Very high – dark grey color
- High – grey color
- Medium – light grey color (Fig. 6).

As we can see on results, for Guest 1 the sporting hotel services are evaluated with high evaluation, for Guest 2 the sporting hotel services are evaluated with lower evaluation and relax hotel services are evaluated with higher evaluation (Fig. 7).

## 5 Conclusion

The article proposed a fuzzy system for evaluating suitable hotel services in a hotel booking system. The evaluation of hotel services uses an expert system with a knowledge base and information from a questionnaire filled-in by hotel guests. The main parts of the proposed fuzzy system were described. The fuzzy system was verified on specific examples and the results were explained.

In future work, we will focus on verification of the expert system on other hotels and their guests.

**Acknowledgment.** This work was supported by the project “LQ1602 IT4Innovations excellence in science” and during the completion of a Student Grant SGS02/UVAFM/2016 with student participation, supported by the Czech Ministry of Education, Youth and Sports.

## References

1. Nexteam S.r.l. Information Technology. Booking Engine Online hotel reservation system with management availability (2005–2016). Available from Internet: http://www.booking-expert.com/booking-engine.html. Accessed 05 Feb 2016
2. Hospitality Technology Ltd. Hotel Booking Software and Property Management Systems (2009–2010). Available from Internet: http://www.hotec.co.uk/hotec/homepage.aspx. Accessed 05 Feb 2016
3. Walek, B., Hosek, O., Farana, R.: Proposal of expert system for hotel booking system. In: 17th International Carpathian Control Conference ICCC 2016, pp. 804–807 (2016). ISBN 978-1-4673-8606-7
4. Frechtling, D.: Forecasting Tourism Demand: Methods and Strategies. Butterworth, Heinemann (2001)
5. Goldman, P., Freling, R., Pak, K., Piersma, N.: Models and techniques for hotel revenue management using a rolling horizon. Econometric Instituite Report EU 2001-46, Erasmus University Rotterdam, Netherlands (2001)
6. Schwartz, Z., Cohen, E.: Hotel Revenue -management Forecasting. Cornell Hotel Restaurant Adm. Q. **45**(1), 85–98 (2004)

7. Saga, R., Hayashi, Y., Tsuji, H.: Hotel recommender system based on user's preference transition. In: IEEE International Conference on Systems, Man and Cybernetics, SMC 2008, pp. 2437–2442 (2008). ISSN 1062-922X
8. Resnick, P., Varian, H.R.: Recommender systems. Commun. ACM **40**(3), 56–58 (1997)
9. McTavish, C., Sankaranarayanan, S.: Intelligent agent based hotel search & booking system. In: IEEE International Conference on Electro/Information Technology (EIT), pp. 1–6 (2010). ISSN 2154-0357
10. Czekalska, K., Sakowicz, B., Murlewski, J., Napieralski, A.: Hotel reservation system based on the JavaServer Faces technology. In: Proceedings of International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, pp. 470–473 (2008). ISBN 978-966-553-678-9
11. Pokorny, M.: Artificial Intelligence in modelling and control (in Czech). Praha: BEN - technická literatura, p. 189 (1996). ISBN 80-901984-4-9
12. Walek, B., Bartoš, J., Farana, R.: Fuzzy tool for proposal of suitable products in online store and CRM system. In: Corchado, E., Lozano, José A., Quintián, H., Yin, H. (eds.) IDEAL 2014. LNCS, vol. 8669, pp. 433–440. Springer, Cham (2014). doi:10.1007/978-3-319-10840-7_52
13. Bartoš, J., Walek, B., Klimeš, C., Farana, R.: Fuzzy application with expert system for conducting information security risk analysis. In: Proceedings of the 13th European Conference on Cyber Warfare and Security, pp. 33–41 (2014). ISBN 978-1-910309-25-4
14. Habiballa, H., Novák, V., Dvořák, A., Pavliska, V.: Using software package LFLC 2000. In: 2nd International Conference Aplimat 2003, Bratislava, pp. 355–358 (2003)

# The Development of Hybrid Algorithms and Program Solutions of Placement and Routing Problems

Leonid Gladkov(✉), Sergey Leyba, and Nadezhda Gladkova

Southern Federal University, Rostov-on-Don, Russia
{leo_gladkov,nadyusha.gladkova77}@mail.ru

**Abstract.** In the article the algorithm for solving the accommodation problems and trace elements of the digital circuits of computer equipment is offered. Formulation of the problem is presented. The description of the developed hybrid algorithm is shown. The work and function block parameters of fuzzy control algorithm is described. The features of the software implementation of the developed algorithm are considered. Examples of the graphic interface are provides. Experimental studies of the developed algorithms is reviewed.

**Keywords:** Genetic algorithm · Fuzzy logic · Computer-aided design · Optimization · Parallel computing

## 1 Introduction

As a rule design problems are characterized by high computational complexity due to the search of huge number alternative solutions. [1]. Placement and routing problems are the most important problems throughout the lifecycle duration. So, a development of integrated methods for placement and routing problems seems appropriate at the present time. Such methods allow us to take into account constraints and current results during the problems solution [2].

To increase the problem solving effectiveness in terms of automated design of complex engineering systems, which contain a million components, it is useful to employ hybrid models and algorithms [3]. They are based on a combination of various scientific fields such as genetic algorithms, fuzzy systems and neural networks [4–9]. Currently mechanisms of parallelized evolutionary computations are widely used to effective computational resources management [10–12].

Efficient software implementation of the developed algorithms is an important component of creating high-quality, highly productive computing systems. Modern programming languages such as C++, C#, Java make it possible to implement complex algorithms that allow the best use of the potential of the hardware resources.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_44

## 2 Problem Formulation

For short, we use the formulation of the problem, described in an earlier article [10]. For brevity, we use the formulation of the problem in the article.

Let $E = \{e_i \mid i = 1, \ldots, N\}$ denote a set of elements, where $e_i = (l_i, h_i, T_i)$ is an element which should be placed and $N$ is a number of elements. Here $l_i$ is a length of the element, $h_i$ is a height of the element and $T_i$ is a list of pins which can be written as $\mathrm{T_i} = \{\mathrm{t_j}|\mathrm{j} = 1,\ldots,\mathrm{K}\}$. Here $t_j$ is a pin, $K$ is a number of pins in the element. Each pin is described as $t_j = (x_j, y_j)$ where $x_i$, $y_i$ are pin coordinates relative to the base point of the element.

The set of net that connected each element is defined as $\mathrm{U} = \{\mathrm{u_h} \mid \mathrm{h} = 1, \ldots, \mathrm{L}\}$, where $u_h$ is a net, L is a number of nets. The net is defined as $u_h = \{(N_{ek}, N_{ck}) \mid k = 1, \ldots, M\}$, where $N_{ek}$ is a number of element, $N_{ck}$ is a number of pin and $M$ is number of pins connected by the net. It is required to find such elements placement that $V = \{(x_i, y_i) \mid i = 1, \ldots, N\}$, where $(x_i, y_i)$ are coordinates of upper left corner of the *i-th* element. For each net the contact list of connection field needs to be found. $\mathrm{W_h} = \{(\mathrm{x_q}, \mathrm{y_q}) \mid \mathrm{i} = 1, \ldots, \mathrm{Q}\}$, where $Q$ is a number positions through which passes the h-th net.

## 3 Algorithm Description

For simultaneous solution placement and routing problems the parallelized genetic algorithm is used. It supposes a parallel implementation of evolutionary processes for several populations. The synchronization of asynchronous processes are performed in migration points. Migration points are defined by particular asynchronous events which may take place in each evolutionary process. If the event is occurs in the one of processes, the another random selected process is held. After that the migration operator is applied to both populations. The migration operator is transferred and copied individuals from one population to another.

The migration operator is applied to transfer chromosomes between populations. Individuals are selected from a number of chromosomes in populations with the best value of objective function. The selection is based on estimation of unrouting connection in the chromosome. For each placement which described by the chromosome the routing is implemented by the wave algorithm. Then chromosomes with the maximum total number of unrouting connections are copied from the population with the minimum total number of unrouting connections to other population. And the same numbers of chromosomes with the minimum value of the objective function are deleted from the second population.

In each evolutionary process the initial population is defined by the shotgun method. The selection is implemented by the roulette method. In the evolutionary process we apply the single-point crossover operator and multiple-point mutation operator in which the number of genes is proportional to the chromosome lengh. The probability of genetic operators is determined by fuzzy logic controller [4, 6, 13].

In Fig. 1 we show a block diagram of the developed algorithm for two population. In practice a number of population is considerably larger.

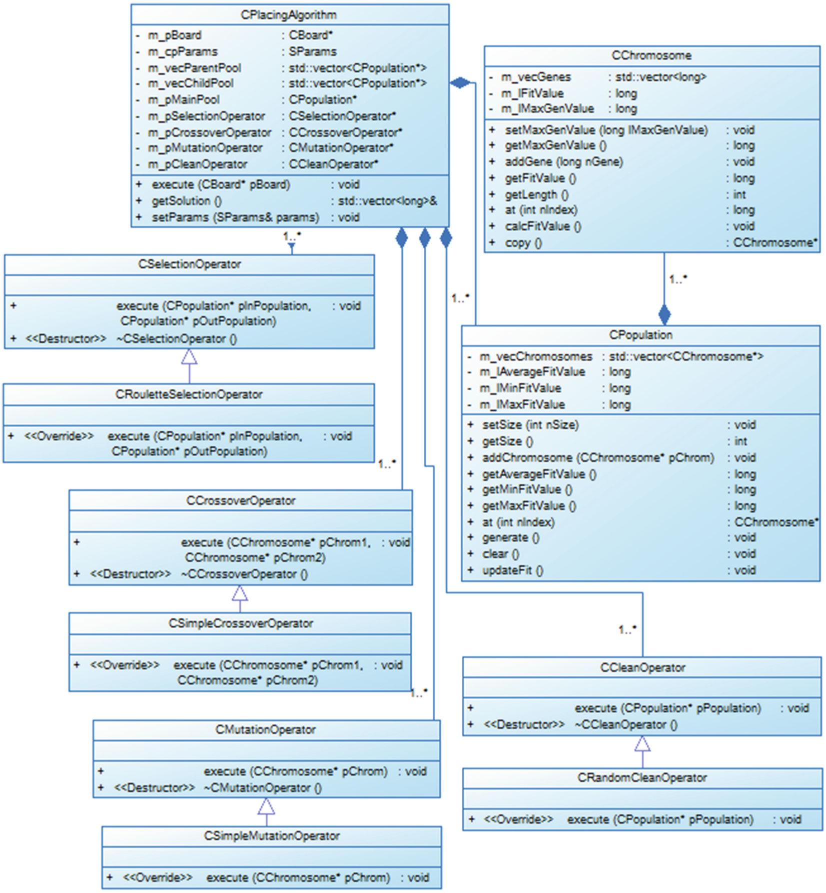

CPlacingAlgorithm
- m_pBoard : CBoard*
- m_cpParams : SParams
- m_vecParentPool : std::vector<CPopulation*>
- m_vecChildPool : std::vector<CPopulation*>
- m_pMainPool : CPopulation*
- m_pSelectionOperator : CSelectionOperator*
- m_pCrossoverOperator : CCrossoverOperator*
- m_pMutationOperator : CMutationOperator*
- m_pCleanOperator : CCleanOperator*
+ execute (CBoard* pBoard) : void
+ getSolution () : std::vector<long>&
+ setParams (SParams& params) : void
CChromosome
- m_vecGenes : std::vector<long>
- m_lFitValue : long
- m_lMaxGenValue : long
+ setMaxGenValue (long lMaxGenValue) : void
+ getMaxGenValue () : long
+ addGene (long nGene) : void
+ getFitValue () : long
+ getLength () : int
+ at (int nIndex) : long
+ calcFitValue () : void
+ copy () : CChromosome*
1..*
1..*
1..*
CSelectionOperator
+ execute (CPopulation* pInPopulation, CPopulation* pOutPopulation) : void
+ <<Destructor>> ~CSelectionOperator ()
CRouletteSelectionOperator
+ <<Override>> execute (CPopulation* pInPopulation, CPopulation* pOutPopulation) : void
1..*
CPopulation
- m_vecChromosomes : std::vector<CChromosome*>
- m_lAverageFitValue : long
- m_lMinFitValue : long
- m_lMaxFitValue : long
+ setSize (int nSize) : void
+ getSize () : int
+ addChromosome (CChromosome* pChrom) : void
+ getAverageFitValue () : long
+ getMinFitValue () : long
+ getMaxFitValue () : long
+ at (int nIndex) : CChromosome*
+ generate () : void
+ clear () : void
+ updateFit () : void
CCrossoverOperator
+ execute (CChromosome* pChrom1, CChromosome* pChrom2) : void
+ <<Destructor>> ~CCrossoverOperator ()
CSimpleCrossoverOperator
+ <<Override>> execute (CChromosome* pChrom1, CChromosome* pChrom2) : void
1..*
1..*
CCleanOperator
+ execute (CPopulation* pPopulation) : void
+ <<Destructor>> ~CCleanOperator ()
CMutationOperator
+ execute (CChromosome* pChrom) : void
+ <<Destructor>> ~CMutationOperator ()
CRandomCleanOperator
+ <<Override>> execute (CPopulation* pPopulation) : void
CSimpleMutationOperator
+ <<Override>> execute (CChromosome* pChrom) : void

**Fig. 1.** The class diagram for placement algorithm

The fuzzy control module is represented as follows:

$$\bar{y} = \frac{\sum_{k=1}^{N} \bar{y}^k \left(\prod_{i=1}^{n} \exp\left(-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right)\right)}{\sum_{k=1}^{N} \left(\prod_{i=1}^{n} \exp\left(-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right)\right)} \tag{1}$$

where $\bar{x}_i^k$ is a centre, $\sigma_i^k$ is a width of Gaussian curve (membership function of fuzzification block), $y^k$ are centers of membership functions of defuzzification block fuzzy sets.

This expression is one of the most popular and frequently used approaches for fuzzy systems realization. Each element is defined by function block (sum, composition, Gaussian function) which allowed create a multilayer network. In this case a neural network contains four layer. At first layer input signals $\mathrm{x_i}$ — arrive, and in output for this signals membership function values are formed. The second layer correspond a rule base and multipliers correspond an output block. The third and fourth blocks realize the defuzzification block [4–7, 13, 14].

To increase the quality of search results expert information includes with evolution process using fuzzy controller which regulate values of factors.

There are input parameters.

$$e_1(t) = \frac{f_{ave}(t) - f_{best}(t)}{f_{ave}(t)} \tag{2}$$

$$e_2(t) = \frac{f_{ave}(t) - f_{best}(t)}{f_{worst}(t) - f_{best}(t)} \tag{3}$$

$$e_3(t) = \frac{f_{best}(t) - f_{best}(t-1)}{f_{best}(t)} \tag{4}$$

$$e_4(t) = \frac{f_{ave}(t) - f_{ave}(t-1)}{f_{ave}(t)} \tag{5}$$

where t is a time interval, $f_{best}(t)$ is the best value of the objective function at the iteration $t$; $f_{best}(t-1)$ is the best value of the objective function at the iteration $(t-1)$, $f_{worst}(t)$ is the worse value of the objective function at the iteration $t$, $f_{ave}(t)$ is an average value of the objective function at the $t$ iteration, $f_{ave}(t-1)$ is an average value of the objective function at $(t-1)$ iteration [11].

As a result we obtain probabilities of crossover, mutation and migration operators.

## 4 Software Features

The development of application architecture was carried out with taking into account modularity and augmentability of the software. During the realization of system's components it is used a coherence decreasing principle which allow to divide placement and routing algorithms, syntax parser and graphical interface. Each system's element can be change to more effective one. The developed algorithm class structure can add new characteristics and behavioral models in the system. For example, genetic operators class inheritance allows to add new classes, which specify data processing algorithm [17, 18].

To save information about PC board layout the authors used the Library Exchange Format (LEF). Library Exchange Format is a specification for representing the physical layout of an integrated circuit in an ASCII format. It contains rules and abstract information about LEF elements and uses in conjunction with Design Exchange Format (DEF) which use for representation of full IS elements placement [19]. Next, there is an example of PC board description with the use of LEF.

```
MACRO ms00f80
  PROPERTY LEF58_EDGETYPE "
    EDGETYPE LEFT 2 ;
    EDGETYPE RIGHT 2 ;
  " ;
  CLASS CORE ;
  ORIGIN 0 0  ;
  SIZE 1.6 BY 2.0 ;
  SYMMETRY X Y R90 ;

  SITE  core ;
  PIN o DIRECTION OUTPUT ;
    PORT
      LAYER metal2 ;
      RECT 0.05 0.500 0.15 1.500 ;
    END
  END o
  PIN a DIRECTION INPUT ;
    PORT
      LAYER metal1 ;
      RECT 1.05 0.500 1.15 1.500 ;
    END
  END a
END ms00f80
```

The example of elements placement on a PC board and description of nets with the use of DEF.

```
COMPONENTS 6;
   - g2278701 ms00f80
      + PLACED (20, 10);
   - g2278702 ms00f80
      + PLACED (20, 40);
   - g2278703 ms00f80
      + PLACED (20, 70);
   - g2278704 ms00f80
      + PLACED (60, 10);
   - g2278705 ms00f80
      + PLACED (60, 40);
   - g2278706 ms00f80
      + PLACED (60, 70);
END COMPONENTS

NETS 2;

– ternarymux_ln49_unr9_z_9_ (g2278701 a) (g2278705 o) (g2278703 a);
– ternarymux_ln49_unr9_z_10_ (g2278704 o) (g2278702 a) (g2278706 o);
```

The placement algorithm is described by a `CPlacingAlgorithm` class. To implement the algorithm calls the `execute` method, which obtain a pointer to a

CBoard class using for PC board layout storing (Fig. 1). The result is processed PC board layout with specified element positions and routed connections.

Algorithm parameters are set by a setParams method with a setParams structure, which contains such fields as a number of chromosomes, a number of iteration, crossover and mutation probabilities and migration frequency. Genetic operators are set by pointers to abstract basic classes defining operators interface. Each pointer can point to the concrete operator realization. Parent and child populations saved in the dynamic memory. Access is provided with the use of pointer vectors. Temporary chromosome buffer is also saved in the dynamic memory.

The routing algorithm is described by a CRoutingAlgorithm class. To perform the algorithm the execute method is called, which obtain a pointer to a CBoard class containing the PC board layout with placed elements (Fig. 2). The result is processed PC board layout with routed connections.

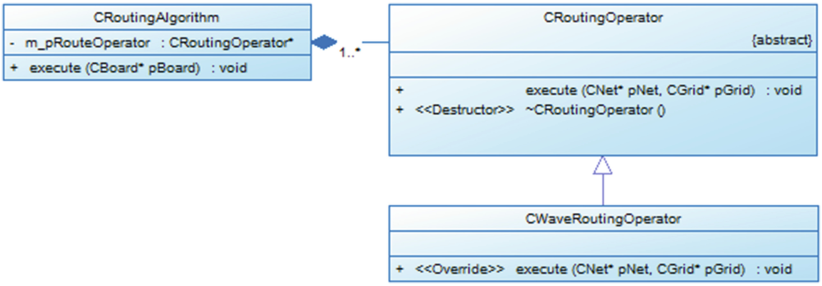

CRoutingAlgorithm
- m_pRouteOperator : CRoutingOperator*
+ execute (CBoard* pBoard) : void
1..*
CRoutingOperator
{abstract}
+ execute (CNet* pNet, CGrid* pGrid) : void
+ <<Destructor>> ~CRoutingOperator ()
CWaveRoutingOperator
+ <<Override>> execute (CNet* pNet, CGrid* pGrid) : void

**Fig. 2.** The class diagram routing algorithm

For each net the routing algorithm is defined by concrete realization of a basic class CRoutingOperator. At current development stage it is used only the wave algorithm which realized in a CWaveRoutingOperator class.

## 5 Graphical User Interface

For graphical interface realization the authors used the Qt 5.6. Qt framework, which represent the cross platform tools for application software development and widely used for graphical interface creation. It was written in C++ and provide a language extension.

Also it contains all basic classes, which can be required for application software development starting with graphical interface elements and finishing with classes for net, databases and XML. The Qt is an object-oriented, easily expandable and supporting a component programming technique.

Let consider main elements of the application graphical interface. An application window consists of a menu, toolbars, workspace and text fiend for output the different background information. Menu contains File and Help items. The item File contains Import and Exit. In the item Import you can download LED and DEF specifications. The item Help contains About and AboutQT by pressing on which it is opened a

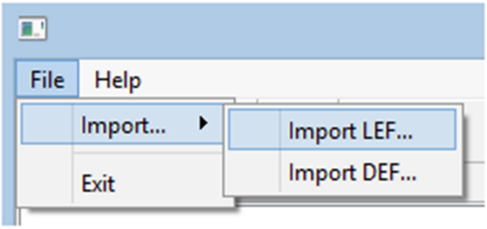


**Fig. 3.** The File menu

window with information about application and information about using the Qt version correspondingly (Fig. 3).

The toolbar contains a button to download the default PC board configuration and buttons to launch placement and routing algorithms.

In the adjustment window you can set such placement parameters as minimal distance between elements and a placement grid step, as well as algorithm parameters such as a number of chromosomes, a number of iterations, crossover and mutation probabilities and migration frequency.

In application workspace a current state of the PC board is drawn by a `QGraphicsView` class. To represent graphic elements there are used classes heritable from a `QGraphicsItem`. PC board elements are represented in a `CGraphicComponent` class, connections – in a `CGraphicNet` class and placement grid – in a `CGraphicGrid` class. All graphic elements are added in a scene. A scene is an object in a class `QGraphicsScene`. A scene are rendered with the use of an object in the `QGraphicsView` class which can be controlled by a transfer matrix. Scaling and rotation of the PC board graphic representations are also realized (Fig. 4). In case of

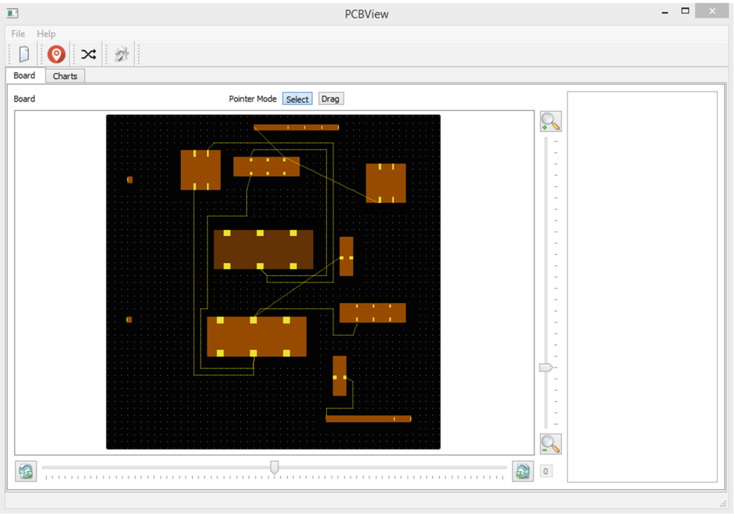


**Fig. 4.** The program window

image magnification you can change visible area by drag transfer. Placed graphical elements also can be dragged for correcting the obtained placement [14].

For graphic representation it is used an extension for the Qt framework – a `QCustomPlot` which is a Qt widget for data visualization. It has no additional dependencies and well documented. This library allows to obtain qualitative graphs and diagrams visual representation and has a high performance to use it in real time systems [15, 16].

To analyze the efficiency of developed algorithms there are used variables graphs of average and minimum values of a placement objective function (Fig. 5). At each iteration it is calculated an average value of an objective function for all populations with evolutionary process. Also it is used graphs of average and minimum value of a routing objective function, values of which there are calculated in migration points.

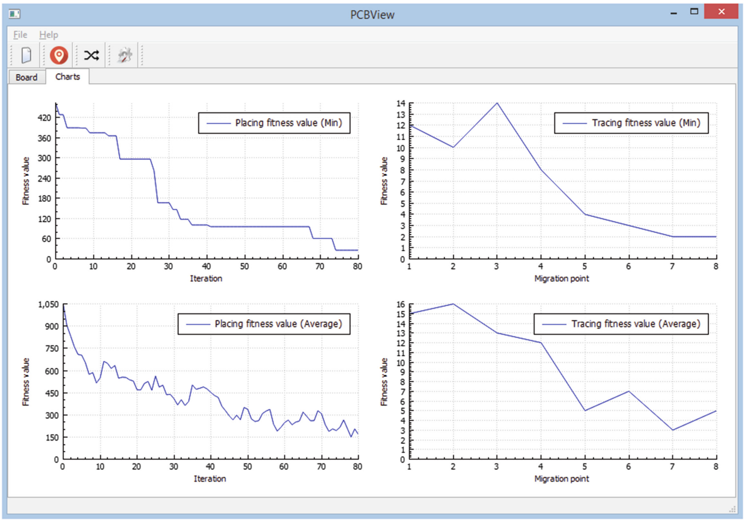


**Fig. 5.** Visualization charts

## 6 Experimental Results

To estimate the algorithm effectiveness we placed and route 300 randomly generated elements and 150 nets with from 2 to 5 pins. Experiments result with different number of parallel algorithm streams is shown in Table 1.

To compare the effectiveness the test problems solved using the FLC and without it are investigated earlier [17]. Table 2 showed that the efficiency of the algorithm with use the controller is much higher than the efficiency of the algorithm without it.

**Table 1.** Experimental results

| Number of streams | Experiment number, % unrouting connections | | | | | Average value of % unrouting connection |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| 1 | 23 | 21 | 24 | 24 | 22 | 22.8 |
| 2 | 16 | 13 | 14 | 14 | 15 | 14.4 |
| 3 | 13 | 14 | 12 | 13 | 11 | 12.6 |
| 4 | 12 | 10 | 13 | 10 | 11 | 11.2 |
| 5 | 14 | 13 | 13 | 14 | 12 | 13.2 |

**Table 2.** Comparison

| Number | Without FLC ($N_{el}$ = 50) | With FLC ($N_{el}$ = 50) | Without FLC ($N_{el}$ = 100) | With FLC ($N_{el}$ = 100) | Without FLC ($N_{el}$ = 150) | With FLC ($N_{el}$ = 150) |
|---|---|---|---|---|---|---|
| 1 | 4585 | 3147 | 29658 | 21296 | 67953 | 48509 |
| 2 | 3870 | 3330 | 31145 | 23582 | 64311 | 51737 |
| 3 | 4245 | 2724 | 28192 | 23145 | 68989 | 50901 |
| 4 | 4056 | 3425 | 31632 | 23481 | 65576 | 50798 |
| 5 | 3774 | 2885 | 29761 | 21844 | 65184 | 48973 |
| 6 | 4896 | 2984 | 28487 | 23148 | 67925 | 49752 |
| 7 | 4129 | 2873 | 31845 | 22946 | 65427 | 52164 |
| 8 | 4812 | 3776 | 29145 | 21941 | 64964 | 48862 |
| 9 | 3981 | 3145 | 29411 | 22157 | 65817 | 50314 |
| 10 | 3876 | 3168 | 30491 | 22981 | 68482 | 50957 |
| Average result | 4222,4 | 3145,7 | 29976,7 | 22652,1 | 66862,8 | 50296,7 |
| Increase quality of solution | 25,6% | | 24,44% | | 24,78% | |

## 7 Conclusion

Results of the experiments showed that the efficiency of the controller is increased after the introduction of the training unit on the basis of an artificial neural network model.

FLC parameters that were used in the study were obtained using a genetic algorithm learning. Training was carried out on the basis of statistical information on the dependence of the FLC parameters and the efficiency of the algorithm placement. This information is collected during the learning process.

We plan to further evaluate the effectiveness of the algorithm by simultaneously solving accommodation problems and tracing. Also, a comparative analysis of the obtained results with known analogues will be held.

**Acknowledgment.** This research is supported by the grant from the Russian Foundation for Basic Research (project # 17-01-00627).

## References

1. Shervani, N.: Algorithms for VLSI Physical Design Automation. Kluwer Academy Publisher, USA (1995). 538 p.
2. Cohoon, L.A., Karro, J., Lienig, J.: Evolutionary algorithms for the physical design of VLSI circuits. In: Ghosh, A., Tsutsui, S. (eds.) Advances in Evolutionary Computing: Theory and Applications. Natural Computing Series, pp. 683–711. Springer, Heidelberg (2003)
3. Gladkov, L.A., Kureichik, V.V., Kureichik, V.M.: Genetic Algorithms. Fizmatlit, Moscow (2010)
4. Michael, A., Takagi, H.: Dynamic control of genetic algorithms using fuzzy logic techniques. In: Proceedings of the Fifth International Conference on Genetic Algorithms, pp. 76–83. Morgan Kaufmann (1993)
5. Lee, M.A., Takagi, H.: Integrating design stages of fuzzy systems using genetic algorithms. In: Proceedings of the 2nd IEEE International Conference on Fuzzy System, pp. 612–617 (1993)
6. Herrera, F., Lozano, M.: Fuzzy adaptive genetic algorithms: design, taxonomy, and future directions. J. Soft Comput. **7**, 545–562 (2003)
7. Liu, H., Xu, Z., Abraham, A.: Hybrid fuzzy-genetic algorithm approach for crew grouping. In: Proceedings of the 5th International Conference on Intelligent Systems Design and Applications, (ISDA 2005), pp. 332–337 (2005)
8. King, R.T.F.A., Radha, B., Rughooputh, H.C.S.: A fuzzy logic controlled genetic algorithm for optimal electrical distribution network reconfiguration. In: Proceedings of 2004 IEEE International Conference on Networking, Sensing and Control, Taipei, Taiwan, pp. 577–582 (2004)
9. Im, S.-M., Lee, J.-J.: Adaptive crossover, mutation and selection using fuzzy system for genetic algorithms. Artif. Life Robot. **13**(1), 129–133 (2008)
10. Rodriguez, M.A., Escalante, D.M., Peregrin, A.: Efficient distributed genetic algorithm for rule extraction. Appl. Soft Comput. **11**, 733–743 (2011)
11. Alba, E., Tomassini, M.: Parallelism and evolutionary algorithms. IEEE T. Evolut. Comput. **6**, 443–461 (2002)
12. Zhongyang, X., Zhang, Y., Zhang, L., Niu, S.: A parallel classification algorithm based on hybrid genetic algorithm. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, pp. 3237–3240 (2006)
13. Gladkov, L.A., Gladkova, N.V., Leiba, S.N.: Hybrid intelligent approach to solving the problem of service data queues. In: Proceeding of 1st International Scientific Conference Intelligent Information Technologies for Industry (IITI 2016), vol. 1, pp. 421–433 (2016)
14. Library Exchange Format. University of Maryland, Baltimore County (2011)
15. Qt Documentation. http://doc.qt.io/qt-5/reference-overview.html
16. QCustomPlot. http://qcustomplot.com/index.php/introduction
17. Gladkov, L.A., Gladkova, N.V., Leiba, S.N.: Electronic computing equipment schemes elements placement based on hybrid intelligence approach. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Prokopova, Z., Silhavy, P. (eds.) Intelligent Systems in Cybernetics and Automation Theory. AISC, vol. 348, pp. 35–44. Springer, Cham (2015). doi:10.1007/978-3-319-18503-3_4

# Applying Verbal Analysis of Decision to Prioritize Software Requirement Considering the Stability of the Requirement

Paulo Alberto Melo Barbosa[1,2(✉)], Plácido Rogério Pinheiro[1],
Francisca Raquel de Vasconcelos Silveira[1,2],
and Marum Simão Filho[1,3]

[1] University of Fortaleza, Fortaleza, CE, Brazil
{placido,marum}@unifor.br
[2] Institute Federal of Ceará, Aracati, CE, Brazil
{alberto.melo,raquel_silveira}@ifce.edu.br
[3] 7 de Setembro College, University of Fortaleza, Fortaleza, CE, Brazil
marum@fa7.edu.br

**Abstract.** The task of selecting and prioritizing requirements should be performed as efficiently as possible. Selecting the most stable requirements for the most important customers to be the first in the line of implementation in a software company can save resources since they are not likely to change. Quantitative methods have been used to solve this type of multi-criteria problem achieving good results. Verbal Decision Analysis can be presented as an alternative to assisting project managers in this task. This paper describes the application of the ZAPROS III-*i* method to classify the software requirements. A comparison is also made with the results obtained by the use of the NSGA-II metaheuristic.

**Keywords:** Software release planning · Multi-objective optimization · Verbal decision analysis · ZAPROS III-*i*

## 1 Introduction

Requirements are critical in the software development process. They provide the basis for estimating costs and effort, as well as allowing the development of development estimates and test specifications [7].

A problem faced by companies developing and maintaining large and complex software systems developed for large and diversified customers is to determine what requirements will be implemented in the next software release [1]. During the software release planning process we can find several constraints, such as: project budget and precedence between requirements [14].

In addition, the original requirements may change throughout software development. This can be caused by bad elicitation of requirements, changes in customer needs, changes in the business environment or company policy [13].

Therefore, Software Release Planning must also consider a new variable: the stability of the requirement. Volatile requirements are considered as a factor that can cause

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_45

great difficulties during software development [5]. However, if the implementations of such requirements are organized so that the most stable requirements are the first to be implemented and the most volatile are the last ones, that is, those that still depend on adjustments between developers and customers, we can have a productivity and satisfaction gain from software development companies and customers.

It is essential that the task of selecting and prioritizing requirements to take effect in the most efficient way possible. Requirements changes are often the main factor in increasing time and cost in software development projects. Therefore, selecting and prioritizing requirements taking into account their degree of stability can increase the effectiveness of the entire software development process.

In the literature, we can find several methods to solve this problem. A common method is the use of metaheuristics that are algorithms used to solve optimization problems. Becceneri [3] say that metaheuristics is algorithmic tool, which can be applied to different optimization problems, with relatively small modifications, in order to make them adaptable to a specific problem. Thus, we can consider metaheuristics as procedures that have generic strategies for escaping from good locations.

Deciding which requirement to prioritize for implementation is typically a decision problem. Routinely, the project manager makes this decision based on his/her experience and knowledge of the project and the requirements involved. We mean that a high degree of subjectivity is present in the decision-making process. This is a suitable scenario for Verbal Decision Analysis (VDA), which consists of an approach based on multi-criteria problem solving through its qualitative analysis [11], that is, VDA methods take into account criterion subjectivity.

Therefore, structured context in this work was the Software Releases Planning, which indicates a methodology that uses Verbal Decision Analysis to be used by software managers as a means to obtain an effective planning taking into account the selection of more priority requirements, having as criterion Stability of Requirements.

Our goal is to obtain a solution that contains an order of requirements to be implemented considering constraints (technical precedence between requirements and resources available to the project) and objectives (maximizing customer satisfaction by selecting the most important requirements for key customers And maximize stability among requirements, initially implementing those with the highest degree of stability). The results of the approach proposed here are compared with Barbosa's methods and results [2] and we find that this approach was similar to the results used in [2]. In Sect. 2 we will see works related to the planning of software releases and the approach proposed by [2]. In Sect. 3 we will see the methods adopted to solve the problem proposed in this paper. In Sect. 4 the results achieved and discussions and in Sect. 5 the conclusions and proposals of future work.

## 2 Related Work

The work published by Bagnall [1] deals with the determination of the requirements that must be executed for the next release of the software. The author predicts that customers have different levels of importance to the company and points out the requirements that have prerequisites and that must be realized in a previous or parallel release that is being implemented. The algorithms applied in this strategy show the obtaining of quick solutions to small problems.

Therefore for Karlsson and Ryan [10], one of the greatest risks faced by organizations developing commercial software is associated with non-fulfillment of users' needs and expectations.

Thus Greer et al. [9] say that defining which release to deliver the requirement is a decision that depends on several variables that relate in complex ways. The different perspectives of the stakeholders and the planning of releases, including effort restraint, are discussed.

However, Barbosa [2] presents a structured approach in multi-objective optimization using metaheuristics for the problem of selection and prioritization of software requirements, considering the inherent characteristics of real projects. Among these characteristics, we considered: (i) stability of the requirement; (ii) costs to implement the requirement; (iii) technical precedence between requirements; (iv) importance of stakeholders to the company and (v) its preferences regarding requirements. The objective of this work was to compare the efficacy of metaheuristics in problem solving through performance measures and to compare the performance of metaheuristics with the result of the implementation and execution of a solution generated from a random algorithm.

The work differential [2] was used as the selection criterion. Unstable requirements to be implemented late to avoid diverting project resources. The work of [2] compared as solutions generated from the execution of the metaheuristics NSGA-II [6], Mocel [7] and a random algorithm. The results showed that metaheuristics were the best methods to reach higher quality solutions.

The task of ordering requirements can be complex and challenging. Many proposals for metaheuristic solutions are found in the research work. However, methods structured in Verbal Decision Analysis are little known in the field.

# 3 Methods

The choice of a multicriteria method among those available, applied to a given context, should be adequate for the characteristics of the problem in question. An important point will be an evaluation of the problem, of the decision objects and the available information. The choice of method should be the result of an evaluation of the chosen criteria, the type and precision of the data, the form of the decision maker's thinking and his knowledge of the problem [21]. It is also emphasized that the direct consequence of possibility of choice between several methods that results can be discordant and even contradictory. The evaluation should not be complicated, since the differences observed are much more related to the diversity of results than to the contradictions and that there are some criteria that allow validating the method chosen [21].

In the opinion of these authors, the methodology of multicriteria support to the decision has several methods that can be applied in the most diverse problems. Therefore, the very choice of a multicriteria decision support method alone is already a multicriteria problem [22].

Therefore, this work proposes to select and prioritize software requirements in the order in which they will be implemented using a VDA method known as ZAPROS III-*i* [18]. The results will be compared to those obtained when using quantitative methods (metaheuristics).

For the problem considered in this work, the application of the ZAPROS III-*i* method came from the acceptance of the method by the decision maker, which meant that the issues that were being presented to the decision maker made sense to him, and he was confident to answer them. In addition to this point, the need to evaluate the acceptance of the data, its properties used by the method, and whether the result supported in the decision-making process was exalted.

For the development of this work, we divided it into three stages: (i) the generation of instances that represent a set of requirements to be implemented, (ii) the classification of these requirements using ZAPROS III-*I*, (iii) the comparison of the results obtained relating these with results obtained through the application of metaheuristic NSGA-II used in [2] and (iv) Finally, we will evaluate the results. The Fig. 1 illustrates this flow.

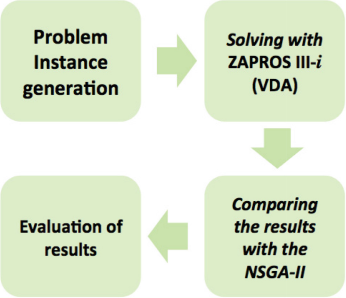

Problem Instance generation
*Solving with* ZAPROS III-*i* (VDA)
*Comparing the results with the NSGA-II*
Evaluation of results

**Fig. 1.** Considered workflow

## 3.1 Instance Generation

Initially an application was developed to generate simulations of problems inherent in the software development process. Each simulation of the problem contains (i) the number of requirements to be implemented, which in this work were fixed in 10 requirements, (ii) the number of customers interested in the project, which in this work was set at 5, (iii) the cost Total of the project that in the experiments was considered between 70% or 80% of the total value needed to implement all the requirements, (iv) the importance of each client to the company (whose value 01 represents the least important customer and 10 Importance), (v) the cost of each requirement (where 10 is the lowest cost and 20 the highest), (vi) the stability level of requirements, ranging from 1 to 10, where 1 means little stable and 10 is very stable, (vii) the technical precedence matrix among the requirements, signaling that a particular requirement should only be implemented after another requirement indicated in that matrix and (viii) the importance of the requirements for the customer that scores from 1 (minimum) to 10 (maximum) a customer's preference for a given requirement. The variations of these situations are represented in Table 1.

**Table 1.** Characteristics of problem simulations

| File name | Number of requirements | Number of clients | Percentage of technical precedence for the requirement | Percentage of budget available for the project |
|---|---|---|---|---|
| I.10.5.10.70 | 10 | 5 | 10% | 70% |
| I.10.5.10.80 | 10 | 5 | 10% | 80% |
| I.10.5.20.70 | 10 | 5 | 20% | 70% |
| I.10.5.20.80 | 10 | 5 | 20% | 80% |

In order to present solid results, four situations with different characteristics were considered in relation to the amount of resource available in the project and the technical precedence among the requirements.

### 3.2 Classification Using ZAPROS III-*i*

Decision-making is an activity that is part of people and organizations' lives. In most problems, to make a decision, a situation is assessed against a set of characteristics or attributes, i.e., it involves the analysis of several factors, also called criteria. When a decision can generate a considerable impact, such as management decisions, and must take into account some factors, the use of methodologies to support the decision making process is suggested, because choosing the inappropriate alternative can lead to waste of resources, time, and money, affecting the company [8].

The ZAPROS III method [12] is an evolution of the ZAPROS-LM one, with the application of the same procedure to elicit the preferences, but with modifications that make it more efficient and more accurate with respect to inconsistencies. Another difference between these methods is that the ZAPROS III is based on the elicitation of preferences around values that represent the distances between the evaluations of two criteria, called Quality Variations (QV), instead of comparing criteria estimates, as in its older version. Besides, it uses the Formal Index of Quality to rank order the alternatives set and, consequently, to minimize the amount of pairs of alternatives to be compared in order to obtain the problem's result [20].

The incomparability of some alternatives, which lead to unsatisfactory results in decision-making models, gave rise to the ZAPROS III-*i* method, which is very similar to ZAPROS III, but presents modifications mainly in the comparison of alternatives process to improve the method's decision power [20].

In such case, the ZAPROS III-*i* method applies (i) the Formal Index of Quality (FIQ) [12], which was used with the purpose of reducing the number of pairs of alternatives to be compared, (ii) the ideas of comparison between alternatives through ordering the values of their quality vectors in ascending order [4] and (iii) the comparison considering all possible alternatives for the problem, which can be used for solving complex decision making process.

Therefore, ZAPROS III-*i* presents a valuable alternative to solve requirements selection problems, since the opinion of the decision-maker is taken into account in this process.

**Application of the methodology.** To rank order the factors that project managers should consider when allocating tasks in distributed software development projects [15], we applied a methodology consisting of four main steps, which are explained next: (1) Identification of the Alternatives; (2) Definition of the Criteria and the Criteria Values; (3) Characterization of the Alternatives; and (4) The ZAPROS III-*i* Method Application.

1. The alternatives considered in this work were the 10 software requirements to be implemented. Each requirement has its own characteristics and will be known ahead.
2. For the purpose of comparison, the criteria adopted in this study were the same as those adopted by [2]. As [2] used quantitative methods and this work adopted the qualitative methodology, the data were converted to the qualitative methodology, where, for example, the Cost criterion represented on a scale of 10 to 20 was discretized into three values, whose values between 10 and 13.3 are represented by criterion A1, values between 13.4 and 16.6 represented by criterion A2 and values between 16.7 and 20 represented by criterion A3. This same methodology was adopted for the other criteria. Thus, the criteria were ranked from the most preferable (A1B1C1D1) to the least preferred (A3B3C3D2), according to Table 2.

**Table 2.** Criteria adopted for the ZAPROS III-*i* classifier

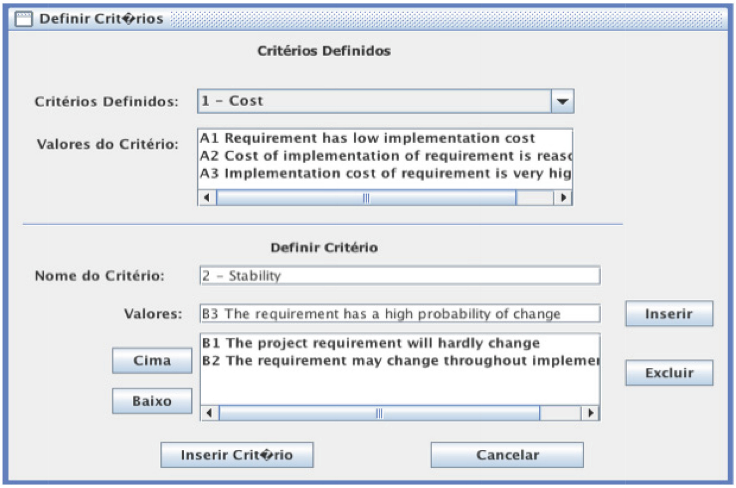
| Criteria | Criteria values |
|---|---|
| **A** Cost | **A1** Requirement has low implementation cost<br>**A2** Cost of implementation of requirement is reasonable<br>**A3** Implementation cost of requirement is very high |
| **B** Stability | **B1** The project requirement will hardly change<br>**B2** The requirement may change throughout implementation<br>**B3** The requirement has a high probability of change |
| **C** Stakeholders | **C1** The stakeholder is influential and very important for the development company<br>**C2** The stakeholder has partial and isolated importance for the development company<br>**C3** The stakeholder is of little importance to the development company |
| **D** Customer requirement value | **D1** The requirement is of great value to the customer and should be prioritized in the process of choosing releases<br>**D2** Requirement has value for the client, but its implementation may be late |

3. The characterization of alternatives was done according to the values contained in the requirements of the work problems of [2]. Thus, considering the definition of criterion presented in the previous item and taking into account the criterion Cost, a requirement that presents Cost 15 in [2] was classified as Cost A2 considering Table 2. In the particular case, as the criteria ‘Stakeholders’ And ‘Customer requirement value’ have more than one customer by punctuating the same requirement, the arithmetic mean between the scores indicated by those clients for the classification represented in Table 2 was considered.

4. After defining and characterizing the alternatives, we moved on to the stage of ordering. At this stage, we applied the ZAPROS III-*i* method to put in order the influencing factors, such that it is possible to establish a ranking of them, how make in [8].

In order to facilitate the decision-making process and perform it consistently, we used the ARANAÚ tool, presented in [16–18]. The tool, which was implemented in Java platform, was first developed in [19] to support ZAPROS III method. In work of [8], was used the updated version to ZAPROS III-*i* method. The use of ZAPROS III-*i* method in the ARANAÚ tool requires four steps, as follows: 1. Criteria and criteria values definitions, 2. Preferences elicitation, 3. Alternatives definition, and 4. Results generation.

The process runs as follows. First, we introduced the criteria presented in Table 2 into the ARANAÚ tool, as shown in Fig. 2.



Critérios Definidos

Critérios Definidos: 1 - Cost

Valores do Critério:
A1 Requirement has low implementation cost
A2 Cost of implementation of requirement is reas
A3 Implementation cost of requirement is very hig

Definir Critério

Nome do Critério: 2 - Stability

Valores: B3 The requirement has a high probability of change

B1 The project requirement will hardly change
B2 The requirement may change throughout impleme

**Fig. 2.** Definition of criteria

Next, the decision-maker decides the preferences. The interface for elicitation of preferences presents questionings that can be easily answered by the decision-maker to obtain the scale of preferences. The process occurs in two stages: elicitation of preferences for quality variation of the same criteria, and elicitation of preferences between pairs of criteria.

The questions provided require a comparison considering the two reference situations [11]. Once the scale of preferences is structured, the next step is to define the problem's alternatives. As mentioned, the quantitative values of [2] were used and converted to qualitative values.

**Overlap of results.** After the introduction of all the data and the answers of the questions, the tests were executed for the four files presented in Table 1. For each simulation of the problem two solutions were extracted: a set of solutions of the NSGA-II execution and a solution of the ZAPROS III-*i*. NSGA-II proposes a set of multi-criteria solutions coming from the Pareto front and the project manager selects the solution closest to his preferences. The ZAPROS III-*i* methodology contains a single solution that was generated from the preferences indicated by the project manager.

## 4 Results

At the end of each execution, the Aranaú [19] tool provided a ranking with the requirements ordered according to their execution priority. In Table 3, we can see the result of this execution for file I.10.5.20.80.

**Table 3.** Ranking generated by the Aranaú tool for the problem file I.10.5.20.80

| Ranking | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Requirements | 6 | 3 | 5 | 7 | 9 | 2 | 1 | 8 | 10 | 4 |

To facilitate comprehension, the results obtained were plotted and compared graphically. In Figs. 3 and 4 we present the results of these executions with the NSGA-II and the ZAPROS III-*i*, where each graph represents a simulation of the problem with its solutions.

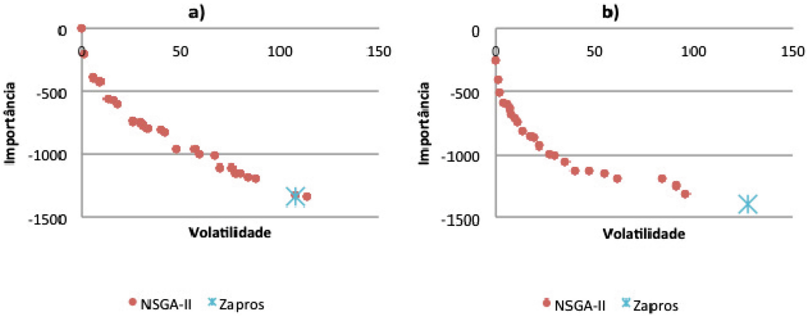


**Fig. 3.** (a) Result of file I.10.5.10.70 and (b) Result of file I.10.5.10.80

The red dots, which form a line of solutions, were generated by the NSGA-II algorithm. The dot in asterisks represents the solution generated by ZAPROS III-*i*. Figures 3 and 4 represent solutions to problems with 10 requirements and 5 clients interested in these requirements. Figure 3(a) and (b) represent problems with 10% technical precedence between the requirements and Fig. 4(a) and (b), 20%. Accordingly, Figs. 3(a) and 4(a) represent 70% of the budget available for the execution of the project, while Figs. 3(b) and 4(b) represent 80%.

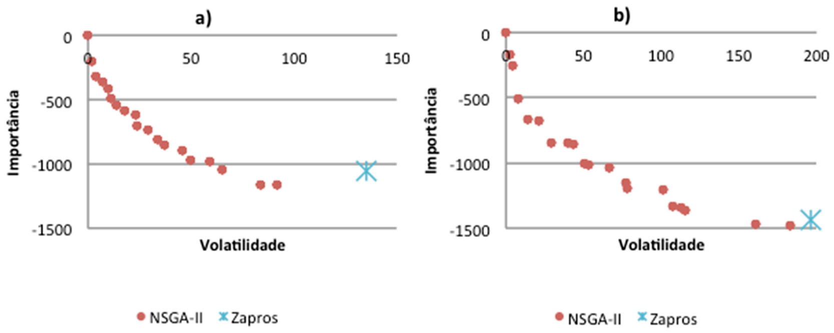


**Fig. 4.** (a) Result of file I.10.5.20.70 and (b) Result of file I.10.5.20.80

The stability criterion was introduced in this work to assist the decision maker in choosing those more stable requirements to be implemented first. However, the solutions showed that the decision maker, in the tests applied, chose to value the importance that the important customer to the company gave to a given requirement. For this reason we see the points of ZAPROS III-i below the chart (valuing the client) and more right prioritizing more volatile requirements. This fact occurred in the four situations shown in Figs. 3 and 4.

In all four cases, we can see that the results generated were very similar. Independent of the different situations, the solution generated by the ZAPROS III-*i* qualitative methodology was close to the solutions generated by the NSGA-II quantitative methods. Considering the differences between these two methods, these results are very satisfactory as they allow VDA to explore other fields beyond existing ones.

## 5 Conclusion and Future Works

One of the premises of multi-objective optimization is to find the set of non-dominated solutions, or Pareto front. This set of solutions can be used by a decision maker to facilitate the choice of the solution that is most appropriate for the proposed problem. The solutions found by the NSGA-II algorithm illustrate this Pareto front in Figs. 3 and 4. As seen, the decision maker has a set of solutions to choose the one closest to his reality.

The results obtained using ZAPROS III-*i* were shown to be very promising, since the generated solution was very close to the Pareto front generated by the NSGA-II algorithm. As in this case this solution was generated from information provided by the decision maker in a qualitative way, we conclude that this solution is what he really expected, since it is structured in real information provided by himself. In Fig. 3(a) the solution generated was equal to one of the solutions found by NSGA-II.

The main contribution of this work is to apply a qualitative methodology structured in ZAPROS III-*i* Verbal Decision Analysis method to order software requirements and compare the solution generated with quantitative methodologies already known for doing this sort of ordering. The Aranaú tool provided support for this work, allowing good performance during testing and execution.

As future work, we can increase the number of requirements to be sorted within the possibilities of VDA. Compare the results obtained by ZAPROS III-*i* with other metaheuristics. Increase the number of criteria to cover other types of problems related to ordering requirements.

**Acknowledgment.** The second author is thankful to the National Counsel of Technological and Scientific Development (CNPq) for the support received for this project.

## References

1. Bagnall, A.J., Rayward-Smith, V.J., Whittley, I.M.: The next release problem. Inf. Softw. Technol. **43**, 883–890 (2001)
2. Barbosa, P.A.M.: An optimal-based approach to the priorization of software requirements, considering the stability of the requirement. Master Thesis – Academic Master in Computer Science, Ceará State University (2013)
3. Becceneri, J.C.: Metaheurísticas e otimização. r.t. lac, inpe. Computers and Operations Research (2007)
4. Christel, M., Kang, K.: Issues in Requirements Elicitation, Carnegie Mellon University, Pittsburgh TR.CMU/SEI-92-TR-12 (1992)
5. Curtis, B., Krasner, H., Iscoe, N.: A field study of the software design process for large systems. Comun. ACM **31**, 1268–1287p (1988)
6. Deb, K., Agrawal, S., Pratab, A., Meyarivan, T.: A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
7. Durillo, J.J., Nebro, A.J., Luna, F., Dorronsoro, B., Alba, E.: jMetal: a java framework for developing multi-objective optimization metaheuristics. Technical report ITI-2006-10, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, Universidad of Málaga, Malaga (2006)
8. Filho, M.S., Pinheiro, P.R., Albuquerque, A.B.: Applying verbal decision analysis to task allocation in distributed development of software. SEKE (2016). doi:10.18293/SEKE2016-181
9. Greer, D., Ruhe, G.: Software release planning: an evolutionary and iterative approach. Inf. Softw. Technol. **46**, 243–253 (2004)
10. Karlsson, J., Ryan, K.: Supporting the selection of Software Requirements. In: 8th Proceedings of the International Workshop on Software Specification and Design (IWSSD 1996), pp. 146–149 (1996)
11. Larichev, O.I., Moshkovich, H.M.: Verbal Decision Analysis for Unstructured Problems. Kluwer Academic Publishers, Boston (1997)
12. Larichev, O.: Ranking multicriteria alternatives: the method ZAPROS III. Eur. J. Oper. Res. **131**(3), 550–558 (2001)
13. Moshkovich, H.M., Mechitov, A., Olson, D.: Ordinal judgments in multiattribute decision. Eur. J. Operational Research **137**(3), 625–641 (2002)
14. Ruhe, G., Saliu, M.O.: The art and science of software release planning. IEEE Softw. **22**, 47–53 (2005)
15. Tamanini, I., Pinheiro, P.R., Machado, T.C.S.: Project management aided by verbal decision analysis approaches: a case study for the selection of the best SCRUM practices. Int. Trans. Oper. Res. **22**, 287–312 (2015). doi:10.1111/itor.12078

16. Tamanini, I., Pinheiro, P.R.: Challenging the incomparability problem: an approach methodology based on ZAPROS. Commun. Comput. Inf. Sci. **14**, 338–347 (2008). doi:10.1007/978-3-540-87477-5_37
17. Tamanini, I., Machado, T.C.S., Mendes, M.S., Carvalho, A.L., Furtado, M.E.S., Pinheiro, P. R.: A model for mobile television applications based on verbal decision analysis. Adv. Comput. Innovations Inf. Sci. Eng. **1**(1), 399–404 (2008)
18. Tamanini, I., Pinheiro, P.R.: Reducing incomparability in multiciteria decision analysis: an extension of the ZAPROS methods. Pesquisa Operacional **31**(2), 251–270 (2011). doi:10.1590/S0101-74382011000200004. (Print)
19. Tamanini, I., Machado, T.C.S., Pinheiro, P.R.: Verbal decision analysis applied on the choice of educational tools prototypes: a study case aiming at making computer engineering education broadly accessible. Int. J. Eng. Educ. **30**, 585–595 (2014)
20. Tamanini, I., Pinheiro, P.R., Machado, T.C.S., Albuquerque, A.B.: Hybrid approaches of verbal decision analysis in the selection of project management approaches. Procedia Comput. Sci. **55**, 1183–1192 (2015). doi:10.1016/j.procs.2015.07.093
21. Bouyssou, D., Marchant, T., Pirlot, M., Perny, P., Tsoukiás, A., Vincke, P.: Evaluation and Decision Models: A Critical Perspective. Kluwer Academic, Boton (2000)
22. Ozernoy, V.M.: Choosing the best multiple criteria decision making method. INFOR **30**(2), 159–171 (1992)

# File Hosting Service Based on Single-Board Computer

Jiri Vojtesek[✉] and Lukas Mlynek

Faculty of Applied Informatics, Tomas Bata University in Zlin,
Nam. T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
vojtesek@fai.utb.cz
http://www.utb.cz/fai

**Abstract.** Single-Board Computers (SBC) are very popular nowadays mainly because of their low price and sufficient performance for basic automation, multimedia, networking etc. tasks. The goal of this contribution was to find appropriate SBC with low price, free software for creation of a personal File hosting service for sharing, distribution and backup of files in small network. There were chosen two candidates from the group of Pi-based SBC, Raspberry Pi 2 and Banana Pi M2 which were then submitted to the performance tests. The open-source ownCloud instance was chosen for the File hosting task. There were also mentioned disadvantages and problems of SBC together with improvements and solutions of these problems.

**Keywords:** Single-board computer · Raspberry Pi · Banana Pi · ownCloud · File hosting service

## 1 Introduction

We can say, that small Single-Board Computers (SBC) are the phenomenon of time. Although the first SBC occurred in 1970s, very fast evolution of this field can be observed in last 10–15 years. Very important point was release of the first version of Raspberry Pi on February 2012 which offered sufficient performance, modifiability and standard USB and HDMI peripherals for 25 $ only [1].

We can find SBC in various tasks. It is ideal solution for home automation, very popular Internet-of-Things (IoT) solutions, multimedia computers (MPC), private cloud servers etc. SBC are also ideal tools for demonstrations, development or educational tasks [2].

SBCs can find in the form of all-in-one solution together with the operation system (OS) which are ready to use for simple everyday tasks. These computers are in the form of small USB or HDMI sticks that can be connected directly to the TV. This work is focused on the second group of SBC without OS and especially on the group of Pi-based SBC [3].

Chosen SBC will work as a File hosting service used for "private" cloud-based synchronization and backup tool.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_46

## 2 Single-Board Computer (SBC)

SBC are defined as a complete computer build in one circuit board. All components like microprocessors, memories, inputs/outputs and peripherals are concentrated on single board which results in small box which can act as a computer but with the less computation power compared to traditional PC. This integration into one board also results in less failure of these systems [3].

The main components and terminology connected with SBC are ARM technology and System on Chip (SoC).

ARM Holding's technology, usually called "ARM technology", uses reduced RISC architecture for building-up microprocessors. As a result, ARM processor is simple with less parts and mainly less demands on the power. We can find this technology in various mobile hardware – mobile phones, tablets, MP3 players, routers, switches etc.

SoC based on ARM is usually the main part of SBC. SoC can be imagined like integration of the main components like CPU, memory blocks, external interfaces etc. into one chip. Typical producers are Altera, Atmel, Broadcom, Intel, Freescale Semiconductor, nVidia, Texas Instruments, AMD etc.

### 2.1 SBC Types

There are several types of SBC but the general division could be on (I.) the SBC with the operating system and (II.) without OS. It is obvious, that the first group is more user-friendly because these systems are debugged and verified for everyday use. They usually act as a USB/HDMI stick and these small computers are connected directly to the television or the computer monitor. On the other hand, they are more expensive due to licensing of the OS etc. Typical members are Google Chromecast, Asus Chromebit, Intel Compute Stick, Lenovo IdeaCentre Stick [5] etc.

The second group of SBC without the OS is an ideal solution for more experienced users due to its options for customization. Another advantage is of course their price which is much lower than in previous case. They are shipped usually in the form of the board with integrated circuits and can be used for home automation and we can find them in the industry too. Typical members of SBC in this group are Arduino [4], Pi-based SBC [2], Cubieboard, Galileo etc.

### 2.2 Pi-Based SBC

Very big and popular group of SBC has in the name "Pi". The increasing popularity of Pi-based SBC is caused by low price and big internet community which produces a lot of documentation, possible applications etc. The first, and one of the famous Pi-based SBC is Raspberry Pi (RPi) that is managed by the Raspberry Pi Foundation [6]. RPi was founded in 2006 by professors at Cambridge and the main idea was to create some simple, cheap tool that can be used in the laboratories for supporting and improving programming skills of students. One of the goals was that the price of RPi does not exceed 25 $. RPi was open

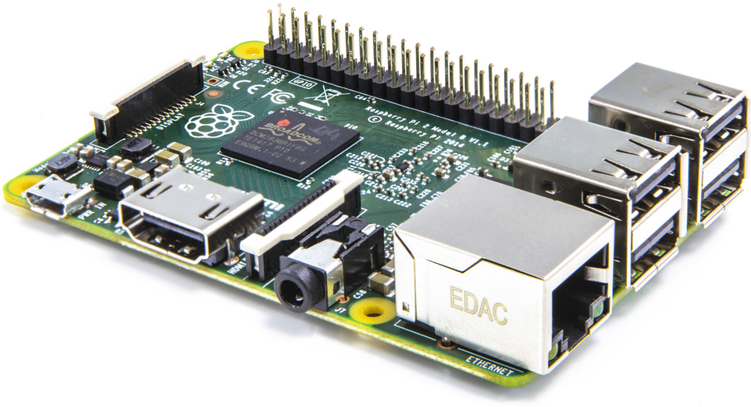


**Fig. 1.** Raspberry Pi 2

the public on February 2012 and initial inquiry was huge. There were sold one million RPi during first year [6].

There can be found several modifications of RPi which profit from the popularity of RPi. One example could Banana Pi (BPi) [7] that was first released on 2013 as a RPi-based SBC with increasing performance and gigabit network interface card (NIC) [8].

Other RPi-based SBC are for example Orange Pi [9], Roseapple Pi [10], NanoPi etc. which are more or less hardware modifications of RPi.

Practical part of this contribution is focused on comparison of two Pi-based SBC – Raspberry Pi version 2 (RPi2) - Fig. 1 and Banana Pi M2 (BPiM2) – Fig. 2 that are easily accessible from the Czech Republic also with the two-year

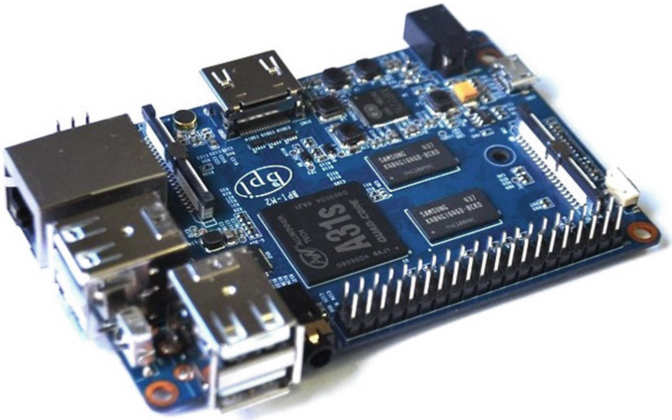


**Fig. 2.** Banana Pi M2

guarantee. Our task will be to find which one of these SBC is more suitable for the private file-hosting service.

### 2.3 Operating System on SBC

Operating System (OS) is very important part of every computer, not only SBC. There is a variety of OS created or modified especially for SBC. They mainly differs in the usage of the SBC – for example OS focused on multimedia is usually called Open Source Media Center (OSMC). In some cases, producers of SBC creates their own OS or modifies existing OS for the use on SBC, for example Raspbian or Bananian used in RPi or BPi respectively. We can find on SBC also adapted versions of traditional Linux distributions – Arch Linux ARM, Ubuntu MATE, Gentoo etc. Microsoft recently released a special version of Microsoft's Windows 10 focused on Internet-of-Things applications. Google's Android OS can be on some SBC. We have a lot of options for choose the right OS that suits our preferences and mainly purposes of SBC.

## 3 Known Problems of SBC and Solution

SBC has several benefits, at it is written in previous chapters. On the other hand, we must mention also some disadvantages and problems connected with the usage of SBC and possible solutions of these problems.

### 3.1 SD Card

One of the main weaknesses of SBC is the file storage that uses usually Secure Digital (SD) (or MircoSD) memory cards mainly because of their dimensions and price. The same card is used not only for the file storing but also for booting and usage of the OS and its grub. Our SBC will act as a File hosting service that definitely means a lot of read/write operations to the SD.

The choice of the right SD card starts with the check of the documentation or web pages of the SBC producer which types of SD are compatible, supported or recommended for this concrete type of the SBC [11]. Big problem is the service time of the SD expressed by the write endurance. This endurance differs by the technology used in concrete SD [12]. There are three main write technologies – Single-Level Cell, Multi-Level Cell and Three-Level Cell. In the first one, Single-Level Cell (SLC), the one bit of information is written in one cell. On the other hand, 2 or 3 bites of information are written in one cell in Multi-Level Cell (MLC) and Three-Level Cell (TLC) technologies. It is obvious, that SLC will have higher write endurance (up to 100 000 erases per block), MLC offers 3 000 – 10 000 erases per block and TLC only 1 000 erases per block. There could be added technologies that prolong the service time of SD like Error Checking and Correcting (ECC) or Wear leveling technologies.

We can mention two possible solutions to this problem. The first one still uses SD for running OS but the programs are optimized for using SD card as less as possible. This could be done for example by decreasing or moving of the swap to the RAM memory or the use of the external HDD connected via USB for the file storing.

The second approach also uses external HDD, but the OS is also stored on this drive. The SD card cannot be fully unmounted but it is used for OS loader or configuration files. This approach reduces rapidly read/write operations on the drive (SD card). Another problem that could occur in this solution is the power supply of the HDD. The maximal power supply 500 mA that runs through the USB in SBS cannot be enough for some types of HDD. This problem could be overcome by the use of the HDD with external power supply, USB hubs with additional supply or SSD disks. Some types of SBC offer also additional programmable increase of the supply power.

### 3.2 Poor Performance

As SBC has in its name "Computer", is it not full equivalent to classical desktop Personal Computer (PC) or notebook which is good to bear in mind. There are still a lot of tasks which need more performance that SBC can offer. Pi-based SBC are mainly focused for the teaching purposes and not for everyday use as a PC with all regular tasks. On the other hand, multimedia or file-hosting services are ideal things that can run on SBC smoothly and sufficiently.

The performance could be affected by the use of appropriate OS. Server application does not critically need graphical interface in OS – text mode is enough and we can save RAM memory for other tasks instead of GUI.

Another option is the overclocking of the CPU. The CPU and other hardware is usually oversized which allows overclocking if the temperature of the hardware does not exceed allowed value. RPi offers overclocking very easily either by the editing of the configuration file *config.txt* on the SD card or by typing a command `sudo raspi-config` in the command line.

We can improve the performance also by the use of the cluster build by more SBCs connected together and performed for one computation item.

## 4 Performance Tests

The practical part is focused on the performance test for parameter verification of the possible SBC systems Raspberry Pi 2 and Banana Pi M2. Results will also help with the choice of the optimal SBC for the task chosen here – e.g. File hosting service.

The performance of the proposed SBC was tested using SysBench tool [15] which tests CPU, memory and work with the database. Tests were done for single-thread and also multi-thread for four nodes each four times and resulting values in tables are their mean values.

Used operating systems, e.g. Raspbian and Bananian, run in the latest available versions and tests were done in the text-mode for better results.

Table 1 shows main parameters of tested SBC.

**Table 1.** Parameters of RPi and BPi

| | Raspberry Pi 2 (RPi2) | Banana Pi M2 (BPiM2) |
|---|---|---|
| CPU | 900 MHz quad-core ARM Cortex-A7, 32 KB Li cache memory, 512 KB L2 cache memory | 1 GHz quad-core ARM Cortex-A7, 256 KB Li cache memory, 1 MB L2 cache memory |
| RAM | 1 GB | 1 GB |
| NIC | 10/100 Mb/s | 10/100/1000 Mb/s |
| OS | Latest Raspbian | Latest Bananian |
| Room temperature | 21 °C | 21 °C |

**Table 2.** CPU performance test

| CPU test | Final time for 1-thread [s] | Final time for 4-threads [s] |
|---|---|---|
| RPi2 | 31.07 | 7.81 |
| BPiM2 | 29.48 | 7.40 |

### 4.1 CPU Test

The SysBench is multiplatform tool that can run in text-mode and tests main parameters of the computer from the performance point of view.

CPU test uses classically prime computation of the set number of primes and the result is a time needed for this computation.

There were done two tests for single-thread and four threads which were run with following two commands (the first one for single-thread and the second one for four threads):

1. `sysbench --test=cpu --cpu-max-prime=2000 run`
2. `sysbench --test=cpu --cpu-max-prime=2000 --num-threads=4 run`

Results in Table 2 show better performance for Banana Pi M2 although results are comparable for both SBCs.

### 4.2 RAM Test

The RAM memory was tested by the sequential read/write operation of 1MB data block again for both single-thread and four threads. Obtained results in seconds are shown in Table 3 and there were obtained using commands

1. `sysbench --test=memory --memory-block-size=1M --memory-total-size=10G --num-threads=1 run`
2. `sysbench --test=memory --memory-block-size=1M --memory-total-size=10G --num-threads=4 run`

Tests of RAM have shown again a bit better results for Banana Pi M2 system for both single-thread and four-threads tests – see Table 3.

**Table 3.** RAM performance test

| RAM test | Final time for 1-thread [s] | Final time for 4-threads [s] |
|---|---|---|
| RPi2 | 1.96 | 1.88 |
| BPiM2 | 1.65 | 1.45 |

### 4.3 Database Performance Test

The last test which uses SysBench tool is the database performance test on the MariaDB database. This type of database was chosen because it was also used in ownCloud File hosting service.

At first, testing database **dbtest** must be created and prepared using command

```
sysbench --test=oltp --oltp-table-size=1000000 --mysql-db=dbtest
      --mysql-user=root --mysql-password=bananapi prepare
```

Then, the performance for one one-tread and four threads were tested via commands

```
         1. sysbench --test=oltp --oltp-table-size=1000000
   --oltp-test-mode=complex --oltp-read-only=off --num-threads=1
--max-time=60 --max-requests=0 --mysql-db=dbtest --mysql-user=root
                 --mysql-password=bananapi run
         2. sysbench --test=oltp --oltp-table-size=1000000
   --oltp-test-mode=complex --oltp-read-only=off --num-threads=4
--max-time=60 --max-requests=0 --mysql-db=dbtest --mysql-user=root
                 --mysql-password=bananapi run
```

Results in Table 4 indicates much better performance for the Banana Pi M2 SBC in both criteria – a number of transactions and a number of read/write requests.

**Table 4.** Database (DB) performance test

| DB test | 1-thread | 4-threads |
|---|---|---|
| *RPi2* | | |
| Number of transactions | 60.2 | 157.4 |
| Number of read/write transactions | 1 143.8 | 1 940 |
| *BPiM2* | | |
| Number of transactions | 99.8 | 869.2 |
| Number of read/write transactions | 1 896.2 | 16 514.8 |

### 4.4 FTP Download Speed Test

As the FHS transfers data through the network, next test is focused file transfer over the Internet with the typical client-server service File Transfer Protocol (FTP). Although FTP is historically one of the first network services, it is still widely used for various tasks – e.g. file distribution over the Internet, uploading files like web pages to the provider's web server etc.

All FTP tests were done in the following conditions:

- Both RPi2 and BpiM2 run in text mode for better performance;
- FTP server was represented by the freeware program CesarFTP;
- FTP server runs on notebook Lenovo THINKPAD SL450 with Intel Core 2 Duo T6570 (CPU frequency 2.1 GHz), 4GB DDR3, Ethernet LAN 10/100 Mbit/s, KINGSTON SV300S37A120G ATA Device - SATA-III
- Operating system in notebook was Windows 7 64bit in the emergency mode with the network services again due to better performance.

There were tested two types of FTP transfers – download of one 512 MB large file and 224 small files with the resulting file size 386 MB and download times are shown in Table 5.

**Table 5.** FTP test

| FTP test | Download of 1 file (512 MB) [s] | Download of 224 files (386 MB) [s] |
|---|---|---|
| RPi2 | 238.3 | 165 |
| BPiM2 | 199.7 | 144 |

It is obvious, that BPi M2 has better results in both test. One 512 MB large file was downloaded nearly 40 s faster than in the RPi2. Also, 224 files were downloaded 21 s faster than SBC RPi2 configuration. Even better results should be obtained with the use of gigabit NIC that BPiM2 offers.

### 4.5 Costs

The last test deals with costs connected with the everyday use of the SBC. As this system is very small with the minimum hardware that needs electric power, costs for electricity consumption are very low.

We have done measurements of the electricity consumption for both SBC configurations and the resulting values of the consumption are shown in Table 6.

Values of energy consumption indicates maximal values around 8 W for the maximal load. We can expect, that mean consumption could be around 3 W in the configuration with SD card or SSD disk as a data storage and 6 W with the use classical HDD with rotation parts as a file storage. It means, that if the SBC runs 24 hour a day, 7 days a week and 365 day a year, the total yearly energy consumption will be 26.28 kWh in the first configuration with SD or SSD disk and 52.56 kWh in the second example with classical mechanical HDD. Both values are very low and acceptable.

**Table 6.** Power consumption tests

| Power consumption test | Consumption of RPi2 [W] | Consumption of BPiM2 [W] |
|---|---|---|
| Running OS | 0–2 (Raspbian with GUI) | ~0 (Bananian without GIU) |
| Maximal load of CPU with 1-thread | 0–2 | ~0 |
| Maximal load of CPU with 4-thread | 2–3 | 3 |
| With connected HDD via USB | 6 | 6–8 |
| With connected SSD via USB | 3 | ~0 |
| Maximal load + connecter HDD via USB | 7–8 | 7–8 |

### 4.6 Results of Testing

All previous tests have shown the winner which is BPiM2 that has better results in all performance tests. RPi performance could be improved by the use of overclocking which is supported by the producer and well described in the documentation. On the other hand, sometimes it is not very safe.

The main advantage of the BPiM2 is gigabit network interface card. The second argument for this Pi-type SBC is that Allwinner A31S SoC which has better performance than Broadcom's SoC chip BCM2836 used in Raspberry Pi 2. Raspberry Pi 2 and also Raspberry Pi 3 uses for USB and LAN SMSC LAN9514 Chip which affects the communication speed if we transfer data from computer network to disk via USB – the data must come through one chip twice. Disadvantage and limitation of BPi2 can be found in the absence of the SATA port which was integrated in the previous version BPi1. This could be solved by the use of the new version BPiM3 which was unreachable during our testing.

## 5 File Hosting Service

In our case, SBC will serve as our own "private" File Hosting Service (FHS) similarly to Dropbox, Google Drive, OneDrive etc. An ideal FHS is secure, fast, has sufficiently drive space for file backup and the last task is to the user-friendly interface. This task affects of course the choice of the right SBC.

Very critical part of the system especially nowadays is security which could be divided into two main fields: (I.) Secure access to the system and its data and (II.) backup of the stored data.

File hosting service gives attention mainly from the hardware point of view on the download/upload speed of the SBC, fast network card and sufficient number of ports for connection of the hard drives. These requirements have shown that Banana Pi M2 is an ideal choice for this task. We will verify the use of BPi M2 by the performance tests mentioned in the practical part.

### 5.1 Used Software for File Hosting Service

Once we have chosen hardware for this solution, we can move to the description of the software equipment installed on SBC for serving a File hosting service.

At first, the Bananian OS [13] created by the producer was used. Advantage of this choice is that this OS is built directly for the use on BPi which means that it is debugged and cleaned from the unused software.

Then, ownCloud [14] was used for File hosting service. OwnClound is distributed on the open-source license and it has good documentation and community in the background. This program together with recommended MariaDB database both placed on external HDD seems to be a good solution.

Here are steps that need to be done for running a ownCloud instance on our SBC:

1. Full update and upgrade of the OS:
   ```
   sudo apt-get install && apt-get upgrade
   ```
2. Restart of the OS:
   ```
   sudo reboot
   ```
3. Installation of the MariaDB database server:
   ```
   apt-get install mariadb-server.
   ```
4. Create database:
   ```
   CREATE USER 'username'@'localhost' IDENTIFIED BY 'password';
   CREATE DATABASE IF NOT EXISTS owncloud;
   GRANT ALL PRIVILEGES ON owncloud.* TO 'username'@'localhost'
   IDENTIFIED BY 'password';
   ```
5. Login into ownCloud via web browser (http://localhost/owncloud).
6. Setup language, manage users etc. can be then done via web interface.

Now we have configured and run our personal instance of ownCloud on our server using SBC BPiM2. Shared folders and other features of the ownCloud are now accessible inside our local network. If we want to access the system from the outside using Internet we need to have public IP address with appropriate security.

The security is also very important nowadays when a cybercrime grows rapidly. The first security condition which needs to be fulfilled is the use of strong passwords not only for administration but also for the user accounts. It is also good to use some security certificate, for example with the SHA-256 hash function which can protect us from the man-in-the-middle attacks. The third security feature is enforce of the HyperText Transfer Protocol Secure (HTTPS) instead of the classical HTTP protocol.

## 6 Conclusion

The goal of this contribution was to show advantage of the SBC in everyday tasks, for example File hosting service. There were described several SBC and their applications in the theoretical part. Although there are various options, there were chosen Raspberry Pi 2 (RPi2) and Banana Pi M2 (BPiM2) because they are very popular and especially RPi2 has great community with a lot of various solutions. There were also mentioned problems which need to be solved in such as choice of the right SD card, boot from the external HDD, overclocking etc. and their solutions.

The practical part is focused mainly on the performance testes for both SBC. There were performed CPU test, RAM memory test, database test and FTP test. The winner in all these tests is Banana Pi M2.

As a result, the Banana Pi M2 was chosen as a hardware solution for the private File hosting service and the software part was done by Bananian OS together with the open-source ownCloud that offers similar features like very famous Dropbox, Google Drive or Microsoft OneDrive services.

Finally, we can recommend this Pi-based SBC as a good option for private File hosting service with good performance results for personal use with low starting and regular costs.

## References

1. Wikipedia: Single-board computer. Wikimedia Foundation, San Francisco (2016). https://en.wikipedia.org/wiki/Single-board_computer
2. Pajankar, A., Kakkar, A.: Raspberry Pi By Example. Packt Publishing, Birmingham (2016). ISBN 9781785286742
3. Pajankar, A.: Raspberry Pi Computer Vision Programming. Packt Publishing, Birmingham (2015). ISBN 9781784395605
4. Schmidt, M.: Arduino: a quick-start guide. Pragmatic programmers. In: LLC 2015 (2015). ISBN 9781941222249
5. O'Donell, L.: Head-To-Head: Lenovo Ideacentre Stick 300 Vs. Intel Compute Stick. http://www.crn.com/slide-shows/components-peripherals/300077274/head-to-head-lenovo-ideacentre-stick-300-vs-intel-compute-stick.htm/pgno/0/1
6. Upton, E., Halfacree, G.: Raspberry Pi User Guide, 4th edn. Wiley, New York (2016). ISBN 978-1-119-26436-1
7. El-Dajani, R.: Banana Pi Cookbook. Packt Publishing, Birmingham (2015). ISBN 9781782174462
8. SINOVOP CO.: ShenZhen [2013–2016] [cit. 2016-02-10]. Dostup-néz. http://www.banana-pi.com/eindex.asp
9. Orange Pi.: Shenzhen Xunlong Software CO, Shenzhen City [2015–2016]. http://www.orangepi.org/
10. Cnxsoft. Roseapple Pi board powered by actions semi S500 comes with 2GB RAM. In: CNXSoft: Embedded Systems News. CNXSoft [2009–2016], Chiang Mai. http://www.cnx-software.com/2015/10/06/lemon-pi-board-becomes-roseapple-pi-gets-an-upgrade-to-2gb-ram/
11. SD Card Benchmarks. In: Raspberry Pi, Cambridge (2016). https://www.raspberrypi.org/forums/viewtopic.php?f=2&t=4076

12. Wikipedia: Flash memory. Wikimedia Foundation, San Francisco (2016). https://en.wikipedia.org/wiki/Flash_memory#Write_endurance
13. Blair, D.: Learning Banana Pi. Packt Publishing, Birmingham (2015). ISBN 9781785283581
14. ownCloud. http://www.owncloud.org/
15. How To Benchmark Your System (CPU, File IO, MySQL) With sysbench. https://www.howtoforge.com/how-to-benchmark-your-system-cpu-file-io-mysql-with-sysbench

# A Business Process Model of Collaborative Approach to Ontology Building

Julia Szota-Pachowicz[1,2](✉)

[1] National Synchrotron Radiation Centre Solaris, Czerwone Maki 98, Kraków, Poland
[2] Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, Prof. Stanisława Łojasiewicza 11, Kraków, Poland
julia.szota@doctoral.uj.edu.pl
http://www.synchrotron.uj.edu.pl

**Abstract.** An ontology forms the basis of a knowledge representation of a system within restricted domain. It provides a vocabulary that refers to concepts in a domain, and determines relationships between concepts that can be shared and reused among intelligent systems. In this paper we describe a collaborative approach to ontology development that distinguishes a domain ontology and an application ontology building stages. The approach has been designed and used while working on building a synchrotron ontology. Presenting the approach as a business process model gives the possibility to identify business activities and actions that an enterprise must perform in order to build an ontology. The model can be integrated with the activities and processes related with software development such as documentation, requirements specification or component reuse.

**Keywords:** Ontology building · Application ontology · BPMN · Business process

## 1 Introduction

Ontologies clarify the structure of knowledge. They capture a shared understanding of a domain, and enable knowledge sharing and reuse [1,2] among intelligent systems and humans. An ontology is defined as a specification of a conceptualization [3]. The conceptualization is an abstract and simplified view of the world, that we are interested in, in order to write an intelligent system that meets the specific objectives. Any software, in order to meet a given requirements must be written with a commitment to a model of the relevant world. Therefore, any intelligent system is committed to some conceptualization, explicitly or implicitly.

### 1.1 Collaborative Ontology Engineering

The process of building an ontology is a complex and time-consuming task. Lots of methods and methodologies for building ontologies have been proposed in the

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_47

literature, some of them are designed to use them in a collaborative environment [1,4]. The complexity of an ontology developing process is a consequence not only of the scope of concepts and their relationships but also the number of participants involved in the process. Development of an ontology that is going to be used in an intelligent system usually requires the collaboration of multiple actors such as domain experts, engineers and others [5]. Building an ontology, especially in order to use it later in a software development process is naturally a team activity, and it may be considered as a part of collaborative software development process [6,7].

In this paper we describe a collaborative approach to ontology development that supports a team to reach consensus through iterative improvements. The approach has been designed and used while working on building a synchrotron ontology [8].

### 1.2 Goals of Research

The purpose of this paper is to present a business process model of the collaborative approach to domain-application ontology building. A business process consists of a set of activities and describes their logical order and dependence in order to perform a business goal [9]. A process model can provide a comprehensive understanding of a process. It gives the possibility to analyze and improve the current process, usually in an enterprise [10]. A building ontology business process can be used to identifying business activities and actions that an enterprise must perform in order to build an ontology. The motivation of presented approach to ontology building using business process model is that a model gives the possibility to take steps towards guidelines that elaborate on how to conduct collaborative ontology building given a domain and application level of a desired ontology. The optimization and further development of ontology building process in a given environment conditions should result in real-world projects. The model can be integrated with the activities and processes related to software development process such as documentation, requirements specification, component reuse, error handling.

The research reported in this paper is driven from the presentation of two stage approach to domain ontology building process as it is a foundation for the implementation of business process mode. The paper explains the aspects of collaborative ontology engineering. Following this section, the results of the analyzis of all the activities that are part of a business process will be described and then a business process model will be presented. The paper will be concluded by some reflections related to presented business process models.

## 2 Building an Ontology

According to the classification of ontologies introduced by Guarino [11] and by van Heijst [12], we can distinguish an application and a domain ontology (taking into account the level of dependence and subject of conceptualization).

An application ontology is more specific and describes all the vocabulary needed to model the knowledge required for a system or application. While a domain ontology is more general and contains the vocabulary related to a generic domain. Because of this feature, a domain ontology is reusable [13].

### 2.1 Domain-Application Ontology Building Approach

A two stage approach to building a domain-application ontology has been designed and used while working on building a synchrotron ontology [8]. The approach has been developed after analysis of existing methodologies, taking into account the collaborative environment in which ontologies can be built. As a result we designed a general approach to ontology building based on the collaborative methodology proposed by Holsapple and Joshih [14] as well as the aspects of Uschold and King's method [15]. The phases defined by Holsapple and Joshih form a backbone of our collaborative approach and the elements of Uschold and King's method are used in different phases. The approach is employed in order to cooperate and reach consensus by a team according to the content of an ontology (Fig. 1).

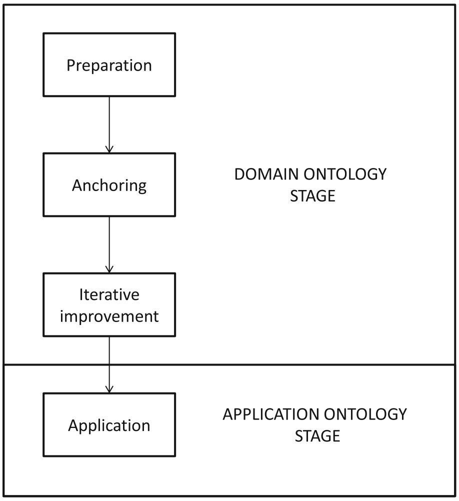


**Fig. 1.** Domain-application ontology building approach [8]

According to [14], the collaborative approach to ontology design is divided into four phases: preparation, anchoring, iterative improvement and application. Our proposed approach complies with the above phases and starts with defining a design criteria [16] and evaluating standards [17], in this step we also identifying existing ontologies that can be merged or used to support a building process in the second phase. The second phase - anchoring includes the development of the initial ontology that will seed the collaborative effort to build a domain ontology in the next phase. Like in an original approach [14], we assume that an ontology that is an output of the anchoring phase do not need to meet all criteria and fulfill all objectives defined in the first phase. The third phase - iterative improvement is the last phase explicitly related to building a domain ontology. An anchoring ontology is incrementally improved to reach the shared and reusable domain ontology using a consensus method. Possible methods that can be used are Delphi technique [14] or Nominal Group technique [18] as these methods are directed at problem-solving and idea-generation. What is important, we do not impose restrictions on what method or tool should be used. It is a good practice to define a method in the first phase, as choosen method may affect the team. However it can be also done separately in the anchoring and iterative improvements phases.

The last phase is the application phase that is the actual usage of the ontology in a specific context [14]. In our approach we use the last phase to build an application ontology based on the domain ontology that is the result of three first phases. The fourth phase includes the whole process of ontology building, once again it consists of preparation, anchoring and iterative improvement phases. At this stage, after iterative improvement phase we additionally define the final evaluation phase that is needed to verify (using an application ontology) does a domain ontology is generic. The final evaluation phase is responsible for evaluation of an application ontology and a domain ontology by comparing concepts and relations and verifying the reusability of a domain ontology. After this process, a domain ontology can be changed, it doesn't change a content of an application ontology as this ontology is being build due to meet a system requirements and capture knowledge needed for a system. We assumed that the team is the same during the whole process - at both stages (a good practice is that people involved in a domain ontology building stage also take part in developing an application ontology, however members of the team may vary).

## 2.2 Collaboration: Achieving Consensus, Roles in a Team

In Ontology Engineering a consensus of ontology content is called the ontological commitment. Consensus can be achieved by collaborative problem solving that allows members of a team to work together to develop an ontology. This mutually acceptable solution is an agreement to use the shared vocabulary in a coherent and consistent manner [19], and constitutes connection between the ontology vocabulary and the meaning of the terms of the vocabulary [11].

According to [13] in ontology engineering we can distinguish three roles: domain experts, knowledge engineers, and ontology engineers. Domain experts have knowledge about developed domain, they understand concepts and are able to specify relations between them. Knowledge engineers are responsible for gaining knowledge from domain experts in order to create a conceptual model of the domain that, in the next step, is used by ontology engineers. Ontology engineers use the appropriate representation language to represent the final ontology and they drive the whole ontology building process by gathering requirements, implementing them and test the resulting ontology [8]. In our collaborative approach, the role of knowledge engineer and ontology engineer is connected which means that one person is responsible both for gaining knowledge from domain experts and driving ontology development process. The reason of such approach is that in the iterative development approach, especially in a collaborative environment the communication is a difficult task. Providing an extra role to collaborative environment makes communication more difficult. Another reason is the usage of a consensus technique during iterative improvement phase, both proposed techniques distinguishes only two main groups of participants: experts and people who drive the process.

In our process we used one main actor - an ontology team that consists of domain experts and ontology engineers. A team may vary in each phase but in order to simplify the process we assume that one team consists of the same members takes part in the whole ontology building process.

## 3 A Business Process Model of Ontology Building Approach

The section describes a business process of the overall domain-application approach to ontology building and includes all main activities and artifacts that are input/output of each phase. Processes are presented using Business Process Modeling Notation (BPMN) that is the global standard for process modeling and one of the most important tool of successful Business-IT alignment.

A business process model is a great source of shared understanding of our approach that can be adjusted to the particular project conditions and requirements. The model can be successfully integrated with the activities and processes related to software development process such as documentation, requirements specification, component reuse or error handling.

BPMN diagrams have been designed using Camunda Modeler which is a free, open source platform for Business Process Management [20].

### 3.1 Main Process

The main process of the domain-application ontology building approach is presented on Fig. 2. The process shows relations between four main phases: preparation, anchoring, iterative improvement and application and marks flow of main artifacts between phases: anchoring domain ontology, domain ontology, application and domain ontology mapping results. All phases are presented as subprocesses because each phase consist of multiple tasks that work together to perform some important part of a total process. The goal of presenting our approach in such way is to give the possibility and flexibility in adjusting particular subprocesses to the project restrictions and team requirements.

The first subprocess represents the preparation phase that affect the other phases. The next step is the subprocess of building an anchoring ontology.

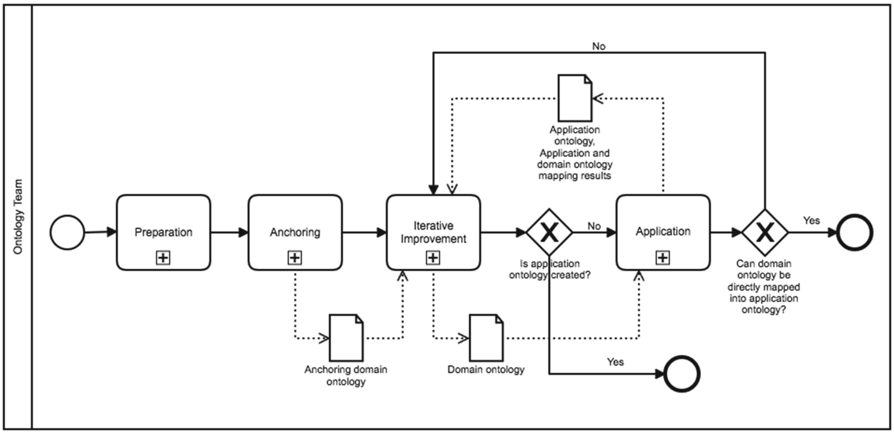

Ontology Team
No
Application ontology, Application and domain ontology mapping results
Preparation
Anchoring
Iterative Improvement
No
Application
Yes
Is application ontology created?
Can domain ontology be directly mapped into application ontology?
Anchoring domain ontology
Domain ontology
Yes

**Fig. 2.** The main process of domain-application ontology building approach with marked flow of main artifacts

The output of this process is the first version of a domain ontology that is used as an input in the iterative improvement process. The result of the iterative improvement process is a domain ontology presented in the language specified in the preparation phase. After third subprocess we modeled an exclusive gateway. Depend on the situation in the project we may finish the ontology building process or continue it leading to the second stage that is responsible for building an application ontology. The main flow is modeled in such way to meet the two scenarios. Firstly, if the only goal in a project is to model one ontology (on a domain, application or other level) the process should be finished after iterative improvement phase. The second scenario that we used also in our project [8] is to manage process of building two ontologies, a domain ontology that can be reused and a specific ontology that is applicable in a system. In this scenario the gateway (is application ontology created?) is related with the input produced in the application phase. The last application phase presents a process of building an application ontology.

The main flow of the application process is presented on Fig. 3 and consists of the same three processes as a main process: preparation, anchoring and iterative improvement. The result of these three phases is an application ontology. For this reason, we say that these three processes are sufficient to build an ontology on any level of conceptualization. The next activity is the final evaluation. The final evaluation process is based on a comparison of the two ontologies (domain and application ontology) by mapping a domain ontology into an application ontology. The mapping process involves the identification of all the concepts that occur in a domain and application ontology and verify do relations between the same concepts. After evaluation step a team should decide does the domain ontology may requires rebuilding. If yes, the process goes back to Iterative improvement phase in the domain ontology stage, the input is the

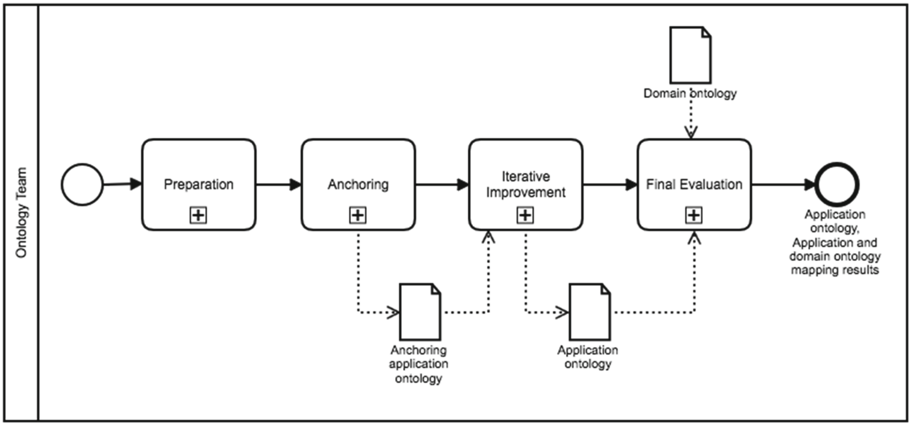


**Fig. 3.** Process of application phase

mapping results and conclusions. Going once again through the domain ontology iterative improvement phase does not need to result in some changes in a domain ontology structure. However we think this step is necessary to maintain a shared and reusable domain ontology.

### 3.2 Business Processes of Preparation, Anchoring and Iterative Phases

The preparation process consists of the following activities shown on the diagram (Fig. 4): define design criteria, evaluating standards, identify existing ontologies, choose consensus method, and ontology language and tools. Each activity produce an output that is later use in anchoring and iterative improvement phases. In this process we assumed that a team is already defined, optionally the step of evaluating and creating a new team (by adding or removing members) can be added as the last activity in this process.

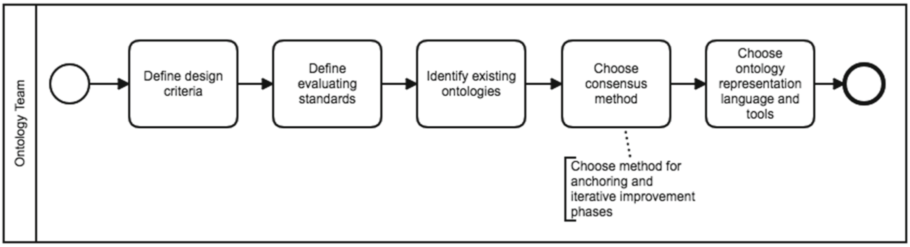

Ontology Team
Define design criteria
Define evaluating standards
Identify existing ontologies
Choose consensus method
Choose ontology representation language and tools
Choose method for anchoring and iterative improvement phases

**Fig. 4.** Process of the preparation phase

The anchoring process starts with the analysis of an input provided by the previous phase. The next activity is gaining knowledge from domain experts that are members of a team in order to create a concept model of an ontology, which is a starting point of a collaboration directly focused on building an ontology. The consensus is reached using the consensus method choosen in the preparation phase. After reaching consensus an ontology is coded and evaluated by ontology engineers. If an anchoring ontology passed evaluation requirements the process is finished. If not, the team again collaborate till reaching consensus (Fig. 5).

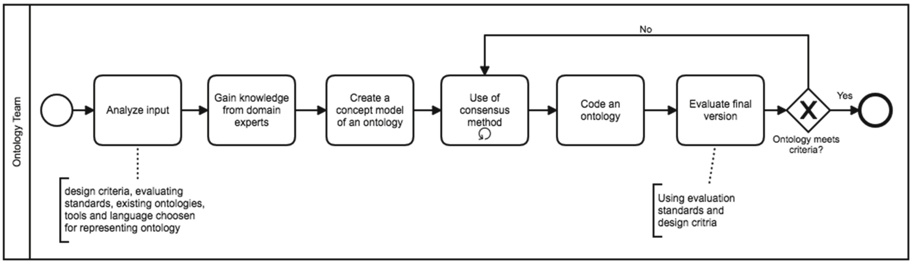

Ontology Team
No
Analyze input
Gain knowledge from domain experts
Create a concept model of an ontology
Use of consensus method
Code an ontology
Evaluate final version
Yes
Ontology meets criteria?
design criteria, evaluating standards, existing ontologies, tools and language choosen for representing ontology
Using evaluation standards and design critria

**Fig. 5.** Process of the anchoring phase

The iterative improvement also starts with the analysis of an input. The main input artifact is the anchoring ontology. After preliminary analysis the consensus method is used in order to develop a shared ontology. Input analysis activity may cover also analysis of the situation in the specific project and the changes in a team or methods may be required. In this process we assumed that a team and consensus method are defined during preparation phase. However, the possible activity after input analysis task is to create a new or rebuild existing team and choose different evaluation method. The next activities are, like in anchoring phase, coding the ontology using specific language(s) and evaluating the final version. The process ends when a shared ontology is properly coded and positively evaluated (Fig. 6).

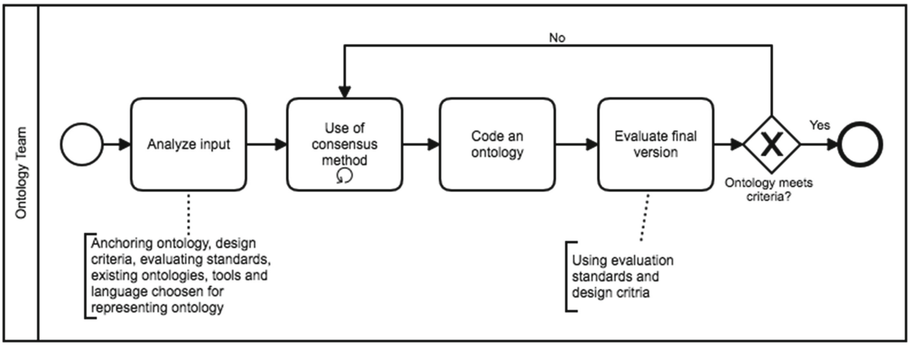

Ontology Team
No
Analyze input
Use of consensus method
Code an ontology
Evaluate final version
Yes
Ontology meets criteria?
Anchoring ontology, design criteria, evaluating standards, existing ontologies, tools and language choosen for representing ontology
Using evaluation standards and design critria

**Fig. 6.** Process of the iterative improvement phase

## 4 Conclusions

In this paper we presented an approach to collaborative ontology building that distinguishes building a domain and an application ontologies. We created a business process models of the approach as well as detailed processes of each phase. Our approach ensures that all participants accept the resulting ontology, being a product of a joint team effort. The same approach can be used to build only one ontology as well as two or more ontologies within the same domain but on a different level of conceptualization. It is also a guideline how to develop many application ontologies based on the same domain ontology that can also be changed during these processes.

Presenting the approach as a business process gives the possibility to identify business activities and actions that an enterprise must perform in order to build an ontology. The model can be integrated with the activities and processes related with software development process such as documentation, requirements specification, component reuse or error handling.

## References

1. Chandrasekaran, B., Josephson, J., Benjamins, V.: What are ontologies, and why do we need them? IEEE Intell. Syst. **14**, 20–26 (1999)
2. Guarino, N.: Formal ontology, conceptual analysis and knowledge representation. Int. J. Hum. Comput. Stud. **43**, 625–640 (1995)
3. Gruber, T.: A translation approach to portable ontologies. Knowl. Acquisition **5**, 199–220 (1993)
4. Fernández-López, M., Gómez-Pérez, A.: Overview and analysis of methodologies for building ontologies. Knowl. Eng. Rev. **17**, 129–156 (2002)
5. Lind, M., Seigerroth, U.: Collaborative process modeling: the intersport case study. In: vom Brocke, J., Rosemann, M. (eds.) Handbook on Business Process Management 1. International Handbooks on Information Systems, pp. 279–298. Springer, Heidelberg (2010)
6. Mistrík, I., Grundy, J., van der Hoek, A., Whitehead, J.: Collaborative software engineering: challenges and prospects. In: Mistrík, I., Grundy, J., Hoek, A., Whitehead, J. (eds.) Collaborative Software Engineering, pp. 389–403. Springer, Heidelberg (2010)
7. Happel, H., Maalej, W., Seedorf, S.: Applications of ontologies in collaborative software development. In: Mistrík, I., Grundy, J., Hoek, A., Whitehead, J. (eds.) Collaborative Software Engineering, pp. 109–129. Springer, Heidelberg (2010)
8. Szota-Pchowicz, J.: Building Synchrotron ontology: the analysis of synchrotron control system in collaborative environment. In: Computer Science, vol. 18. AGH University of Science and Technology Press (2017)
9. Parody, L., Gómez-López, M.T., M. Gasca, R.: Extending BPMN 2.0 for modelling the combination of activities that involve data constraints. In: Mendling, J., Weidlich, M. (eds.) BPMN 2012. LNBIP, vol. 125, pp. 68–82. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33155-8_6
10. Aguilar-Savén, R.: Business process modelling: review and framework. Int. J. Prod. Econ. **90**(2), 129–149 (2004). Elsevier

11. Guarino, N.: Formal ontology in information systems. In: 1st International Conference on Formal Ontology in Information Systems (FOIS 1998), pp. 3–15. IOS Press (1998)
12. van Heijst, G., Schreiber, A., Wielinga, B.: Using explicit ontologies in KBS development. Int. J. Hum. Comput. Stud. **46**, 183–292 (1995)
13. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer, Berlin (2004)
14. Holsapple, C., Joshi, K.: A collaborative approach to ontology design. Commun. ACM **2**, 42–47 (2002)
15. Uschold, M., King, M.: Towards a Methodology for Building Ontologies. In: Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI 1995, pp. 6.1–6.10. Montreal (1995)
16. Uschold, M., Gruninger, M.: Ontologies: principles, methods and applications. Knowl. Eng. Rev. **11**, 93–155 (1996)
17. Gómez-Pérez, A.: A framework to verify knowledge sharing technology. Expert Syst. Appl. **11**, 519–529 (1996)
18. Karapiperis, S., Apostolou, D.: Consensus building in collaborative ontology engineering processes. J. Univers. Knowl. Manage. **1**, 199–216 (2006)
19. Gruber, T., Olsen, G.: An ontology for Engineering Mathematics. In: Fourth International Conference on Principles of Knowledge Representation and Reasoning, pp. 258–269, Morgan Kaufmann Publishers, San Francisco (1994)
20. Camunda Modeler. https://camunda.org/download/modeler/

# The Development of Dynamic Cognitive Interfaces for Multisubject Information Systems (on the Example of Geosocial Service)

A.V. Vicentiy[1,2(✉)], M.G. Shishaev[1,2], and I.V. Vicentiy[2]

[1] Institute for Informatics and Mathematical
Modelling of Technological Processes of the Kola Science Center RAS,
24A, Fersman st., Apatity 184209, Russia
alx_2003@mail.ru, shishaev@arcticsu.ru
[2] Murmansk Arctic State University, Egorova st. 15, Murmansk 183038, Russia
felysite@yandex.ru

**Abstract.** The article describes the methodological basis of the synthesis of cognitive interfaces for multisubject information systems. A definition of the cognitive user interface as well as approaches to its formal assessment are given. Special attention is paid to semantic and perceptual aspects of cognitive interfaces, the concept of pertinence, gestalt and perceptual stereotypes. For practical approbation some possibilities of cognitive interfaces, the web application prototype has been developed. This web application is a geosocial network dedicated to the tourism in the Murmansk region of the Russian Federation. In our prototype to improve the efficiency of perception, we made an attempt to apply some of the principles of gestalt psychology. In the second part of the paper, we describe a prototype implementation and the basic capabilities of the prototype.

**Keywords:** Cognitive interface · Multisubject information system · User mentality model · Mental stereotypes · Gestalt · Geosocial service

## 1 Introduction

Over the past decade, information systems have changed significantly. One of the major changes is the amount of data that modern information systems can store, process and provide to the user. The growth of information in an information systems, the expansion of their functionality have resulted in widespread of large information systems targeted at different categories of users.

To divide users into different categories, it is possible to use a combination of several parameters. For example, such parameters may include gender, age, social status, professional interests, cultural features, etc.

The information systems that support solving of various application problems for different categories of users, called multisubject information systems. The examples of the multisubject information systems are the news sites, reference sites, social networks, geosocial services (also known as geo-social network), and others.

Geosocial networking is one of the types of social networks that bring people together in a virtual space, and creates a new communication environment.

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_48

Geosocial networking creates a virtual social space based on information about a user's location. [1, 2]. A method used to obtain data about a user's location (IP, trilateration or another based) - is irrelevant. Using these data, geosocial networks can unite people according to their interests and needs [3].

Any multisubject systems, including the geo-social network, put forward peculiar requirements for the quality of the user interface (UI). The interface of such systems must be convenient for all users of the multisubject information system. UI must provide a convenient and intuitive mechanism to access and interpreting the information for different categories of users.

In this paper, we will call cognitive interface, an interface that has a special cognitive properties. The cognitive properties of the interface on the one hand should provide an intuitive perspicuity of the interface, on the other hand should contribute to the effective understanding of the information which is transmitted via this interface.

Different categories of users have different ideas about world around, that is, they have different mental models. The formation of user representations about world around (i.e. mental model of the user) can be affected by age, occupation, gender, beliefs, and other factors. Ensure effective static (traditional) interface for different categories of users is challenging. To meet this challenge, we propose to use the dynamic formation of the interface. Such dynamic interface could be adapted to a particular user (or user group) who uses the multisubject information system at the moment.

Some information technology, such as cascading style sheets, tag clouds, and others show the possible ways of forming dynamic interfaces. However, they do not answer the main question: “How to create UI for the multisubject system so that it meets the information needs of particular users?”.

To successfully create such UI is necessary to have clear criteria for the quality of the interface. That is necessary to know the evaluation criteria of the cognitive properties of the interface, which is being developed. And in order to successfully meet the information needs of users, you must have a means of identifying the mental stereotypes that make up the mental model of the user.

We believe that the quality of the cognitive UI is higher when it corresponds better to the user mental stereotypes. In this case, the user needs to make less effort to perceive and understand the information that is transmitted via this interface.

For a formal definition of the quality of cognitive UI is necessary, on the one hand, have a clear formal representation (description) of the information and knowledge that is stored in multisubject information system, and on the other hand, have a formal representation (description) regularities of human perception and ways of interpreting the information.

In this paper we have tried to apply concepts such as cognition, relevance, pertinence, Gestalt, and others to assess the effectiveness of cognitive interface creation.

## 2 Methodological Basis

The development of an effective human-computer interface is a challenging task. Despite the wide opportunities of displaying information and powerful hardware of modern multisubject information systems, the end-consumer of information is a particular person. Exactly he determines the quality of the interface on the basis of their subjective feelings [4, 5].

The main purpose of data visualization in multisubject information systems is to reflect large amounts of data in a convenient form for visual perception by users. If the UI performs the visualization of the data well, the visualization should not contain a large number of elements that are not essential to meet the information needs of the users [6]. Thus, data visualization, it is the representation of the data in the multisubject information system, which provides the most efficient perception, processing and exploring of these data by a human mind [7, 8].

One way to improve the perception of information for end user is the use of display techniques based on cognitive graphics [9, 10]. But this method is not universal and is suitable only for a narrow range of tasks. Therefore it is necessary to look for new methods of effective transfer of information from the multisubject information system to a end user. We believe that one of these methods is the method of developing an interface that allows you to delegate some of the cognitive functions of the human to the interface.

## 2.1 The Model of a Cognitive Human-Computer Interface

Cognitive functions are functions of the human brain that are responsible for the cognition. The cognitive process takes place in the human brain, and includes the implementation of several cognitive functions. For example, the creation of new concepts, processing and linking concepts, construction of estimates and judgments, memory usage, etc. Thus, the cognitive processes use existing knowledge and generate new knowledge [11–13].

Considering the aforesaid, a cognitive interface is an interface that ensures proper and sufficient rapid formation of concepts in the mind of the user.

An important concept for the development of cognitive interface is the concept of "cognitive information". This is information that is generated based on the signals from the environment. Cognitive information is not contained in the environment in a ready kind. To get it, user needs to perform a cognitive process. This process should work according to certain rules in order to establish links between environmental signals and get the new knowledge (i.e., cognitive information). A set of this rules implemented by the human cognitive system (Fig. 1).

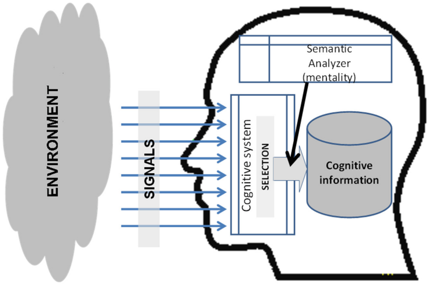


**Fig. 1.** The model of a cognitive process

One of the first model describing the process of human cognition from the perspective of cognitive processes, is a Broadbent's model of early selection [14]. According to this model, the information that comes from the environment in a variety of signals is filtered and is remembered for further processing.

However, J. Gray and A. Wedderburn have proved that the channel selection (selection of information) is carried out taking into account the semantics of the incoming information [15]. Thus, in the cognitive process involved yet another component that provides semantic analysis of incoming information. We will call it as "semantic analyzer" or "mentality" (Fig. 1).

In human-computer interaction, the multisubject information system interface is "the environment". It forms various images that are presented to the user to meet the informational needs. In contrast to the natural environment, the interface generates portions of information and provides these portions to the user in accordance with the rules which are incorporated into this interface. Thus, the interface in this context is a some active entity. Its actions have a specific purpose and carried out in the framework of a managed process.

Our hypothesis can be expressed as follows. If the interface is able to generate signals that are consistent with the user's mentality, the cognitive process is more efficient.

In other words, the user can build a correct mental images (i.e., generate new knowledge) faster and more efficiently than the conventional method of information perception.

For this purpose, is necessary to transfer a part of the selective functions of the user to the information system (Fig. 2). For the implementation of cognitive functions, the information system should have data about peculiarities of user's information perception. The set of such features we call as "user mentality model" or "model of mental stereotypes".

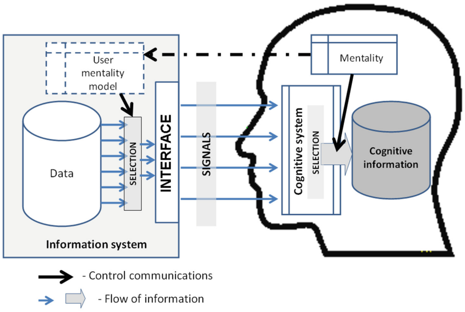


**Fig. 2.** The model of a cognitive interface

Now we can give a more precise definition of what is a cognitive interface. Cognitive User Interface - it is the interface that implements part of human cognitive functions and provides a quick and correct formation of the concepts of the user based on the generated signals. The more human cognitive functions implemented in the interface, the higher the cognitive level of interface.

The cognitive level of interface is limited by the accuracy of user mentality model and the accuracy of semantic model of multisubject information system data. If the model of mental stereotypes and selection procedures are constructed correctly, the information system will give the user first and foremost the most important portions of information, and discard the secondary and minor portions of information. Due to this, the cognitive load on the user of the system is greatly reduced [16, 17]. At high levels of cognitive load, the user can even refuse the use of a multisubject information system. In practical use, often the decisive criterion when choosing a source of the information is, first of all, ease of obtaining information, and not the amount of information that can provide the system [18].

The information visualization method is also very important. If in the process of visualizing data, the stereotypes of visual perception are taken into account, the rate of cognitive process significantly increases [19]. Therefore, in our work we distinguish two parts of cognitive user interface: the semantic part and perceptual part (Fig. 3).

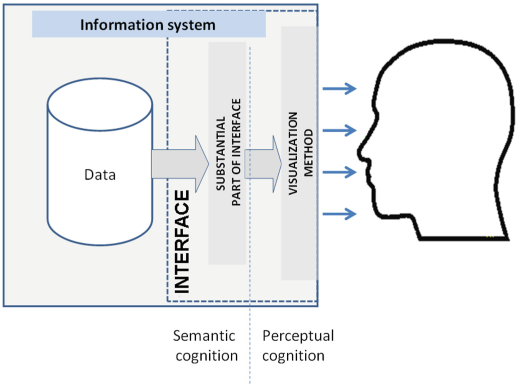


**Fig. 3.** Two formation stages of cognitive user interface

## 3 Results

For practical approbation some of the methodological bases of cognitive interfaces creation, the web application prototype has been developed. This web application is a geosocial network dedicated to the tourism in the Murmansk region of the Russian Federation. Modern tourism is a very information rich activity. There is not a lot of other industries in which the collection, processing, using and transferring of information would have been as important for daily functioning as in the tourism industry. Service in tourism can not be exhibited and reviewed in the point of sale, as consumer or industrial goods. It is usually bought in advance and away from the place of consumption. Thus, the tourism market is almost entirely dependent on images, descriptions, reviews of other users of the service, means of communication and information transfer.

Today the Murmansk region is one of the most attractive destinations for tourism. A feature of the Murmansk region is different directions of tourism, for example, mountain-skiing tourism, geological tourism, mountaineering, historical tourism, fishing tourism, cultural tourism, and others. Due to the great variety of tourism directions, we believe that this geo-social network (implemented in the form of a web application) can be attributed to multisubject information systems.

### 3.1 Implementation of a Web Application Prototype

The web application is divided into two components: the "Client" - works in the browser and "Server" - running on the server. The server part provides access to the data for the client. To develop required to install a local web server (Apache 2.4.20) and the necessary tools to work with it, such as PHP (PHP 7.0.8), MySQL (MySQL 5.7.13) and phpMyAdmin (PHPMyAdmin 4.4.15.7). Also for the the practical implementation of the geosocial service prototype were used different tools and techniques that listed below:

- Open Server - the local server for web-developers;
- Laravel - the PHP-framework with MVC architecture;
- VueJS - the JavaScript-framework with ModelView architecture;
- PHP Storm - the integrated development environment;
- GulpJS - the task manager to simplify the assembly of the project;
- Apache Cordova - the tool for creating mobile applications;
- Material Design Lite - the CSS-framework from Google;
- Vue Debug Extension - the special extension for Google Chrome browser to track application in debug mode;
- Vue-google-maps - the VueJS component to create maps.

To determine the structure of the database application, it is necessary to allocate all the key entities and relationships between them. As a result of the analysis and formalization of the subject area via ER-diagrams, we identified the main entities that must be present in the system: Places, Cities, Markers, Categories, Images, Users, Comments, etc. By normalization ER-diagrams, the relational database structure for geosocial service prototype was created.

One of the important tasks in the development of the prototype is to minimize dependence between the "client" and "server". To solve this problem, it is necessary that HTML is used only in the browser on the client side, and the Web-server provides an interface to obtain the necessary data for the pages. For such a scheme of interaction is necessary to: (1) Determine the objectives and scope of tasks to be solved with the help of generated interface; (2) Determine the server-side API; (3) Choose a communication protocol between the server and client side; (4) Create a protocol based on XML, because most of the modern browsers have built-in support for this language; (5) Create a document that describes the protocol.

Using this model of interaction between the "client" and "server" is possible to make changes to the structural units of the client (browser), without fear of indirect changes to the server-side. This reduces the cost of requests for change processing, because any changes in one structural unit are within the its framework.

Based on the foregoing, the structure of the geosocial service prototype has been developed. It is divided into server and client, and contains the following elements (Fig. 4): (1) A software module for “browser”; (2) A software module for the web-server; (3) A software module for the database; (4) The communication protocol for the “browser” and “web-server” modules; (5) Interface for interaction between “browser” and “web-server” modules; (6) Interface for interaction between “web-server” and “database” modules.

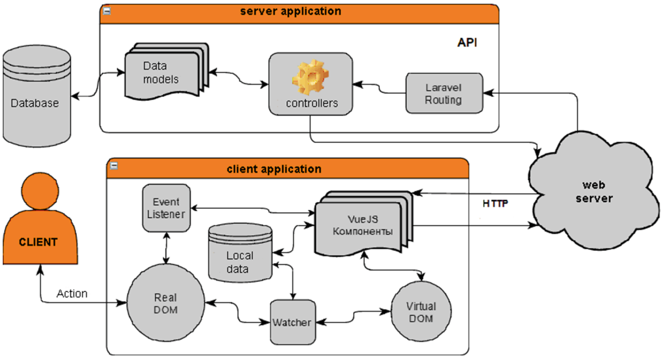


**Fig. 4.** The structure of the web application

### 3.2 Cognitive Elements in the Web Application Interface

When the user requests some information in the multisubject information system, the system finds this information in the database and visualizes it. In other words, the information system creates a visual image of the information that the user requests. The application gives the user the visual image of information using visual human-computer interface. Cognitive interface should facilitate the rapid and correct interpretation of the meaning of the image by the end user. To create a more understandable visual image of the information for the user, the principles of information perception should be taken into account [20]. In our prototype to improve the efficiency of perception, we made an attempt to apply some of the principles of gestalt psychology. Gestalt is a basic concept of gestalt psychology. Gestalt is some stable structure that can not be deduced from its components. The basic principles of gestalt psychology, which can be taken into account when creating visual images [21]:

1. Principle of proximity. Elements arranged close in time or space are perceived together.
2. Principle of common fate. Linking the observed elements in a continuous sequence, or to give them a certain orientation.
3. Principle of similarity. Perception of similar objects as a group.
4. Principle of closure. The aspiration to complete or supplement “incomplete” image.

5. Principle of symmetry. The symmetrical arrangement of elements in the process of formation of an image.
6. The principle of inclusion of B. Keller. The aspiration to perceive only a big figure, and not perceive a small figure (If a big figure included small figure).

The main types of gestalts are perceived almost equally by all people. But, for example, for a group of people who have similar culture, profession, place of residence, education, etc. there may be some principles of information perception that specific to this group. These principles are called "perceptual stereotypes". Such stereotypes are very stable. Perceptual stereotypes (or patterns) often reflect the specifics of a subject area and contribute to the rapid and precise perception of the object in the context of certain application tasks.

Thus, for correct and rapid reception of information, it is important to determine which stereotypes are formed by user perception. In this case it is possible to determine the most efficient method of data mapping.

Figure 5 shows the main page of the prototype of geosocial service. Different markers are displayed on this page. Markers in the system can be of several types, depending on the place category in which the marker is set. For example, for places with category "food", the shopping cart icon is displayed. This allows users, regardless of their existing perceptual patterns, easy to navigate around the map, and immediately understand the types of places that are displayed on the tourist map. It is important to use for marking only the icons that are well and clearly understood by different categories of users (a shopping carts - for the shop, a cup of coffee - for a cafe, a bed - for the hotel, a plane - for the airport, etc.).

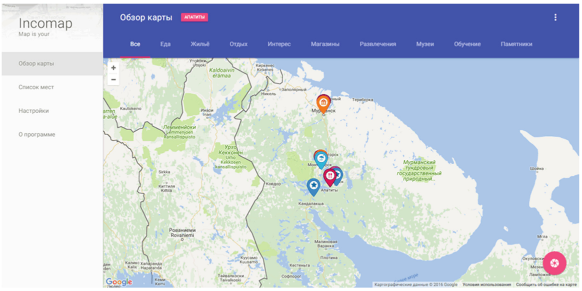


**Fig. 5.** The home page of the geosocial service

Another important characteristic that affects the user's cognitive processes, is the amount of information provided to him. Too much visual information, the presented to the user in a short time, has a negative impact on the perception and cognitive user experience. This property is due to the peculiarities of memory and conscious human attention. The average person is able to hold in its "operational memory" not more than seven unrelated elements [22]. Therefore, the amount of data that is displayed to the user, must be adjusted. The most common approach of this regulation is a data grouping and defining the order of their withdrawal, in accordance with certain criteria. Such mechanisms are implemented in the prototype of geosocial service.

The prototype supports clustering of markers that are displayed in response to a user request (Fig. 6). Clustering of markers allows to combine markers in order to avoid aliasing of markers. In order not to violate the principle of proximity and principle of similarity, clusters are created when in about two square centimeters of the visible image area have generated more than one marker. In order not to violate the principle of proximity and principle of similarity, clusters are created when in about two square centimeters of the visible image area account for more than one marker. In addition, to comply with the principle of "magical number seven" clusters icons are displayed in different colors. If the markers in the cluster are less than 5, then cluster icon is displayed in green; if the markers in a cluster of 5 to 9, the cluster icon appears in yellow; if the markers in the cluster are more than 9, the cluster icon is displayed in red. Thus, the user will spend less mental effort to analyze the visual image.

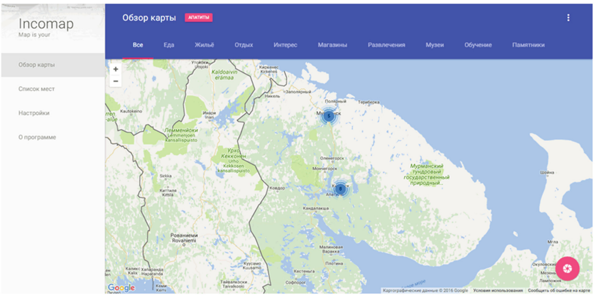
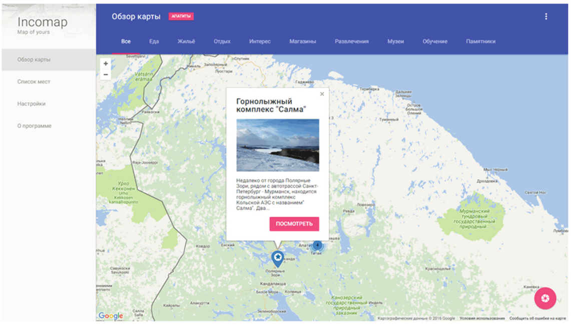


**Fig. 6.** The clustering of markers

Other functions of the prototype are similar to the functions of a typical tourist geosocial service. When choosing a marker, the card appears on the screen, which contains a summary and a link to go to the details (Fig. 7).



Горнолыжный комплекс "Салма"

Недалеко от города Полярные Зори, рядом с автотрассой Санкт-Петербург - Мурманск, находится горнолыжный комплекс Кольской АЭС с названием" Салма". Два...

**Fig. 7.** The place information card

Opening the card, it is possible to see a full description of the place, to share this place with friends and make comments about it. In addition to comments and descriptions, the user will be shown three similar places. The criterion for selection is the category of the selected place.

Also, there are three basic screen resolution for the prototype of geosocial service: 480 px, 840 px, 1025 px. These screen resolutions have been chosen because it is the three basic size of output devices - mobile devices, tablet devices and desktop devices.

## 4 Discussion

In this paper we examined some of the problems of dynamic cognitive interfaces for multisubject information systems. As indicated above, some problems of effective perception and visual information processing for the end user can be solved on the basis of known approaches. For example, such as gestalt psychology.

Accounting of the psychological characteristics of information perception in the development of cognitive user interface improves the efficiency of cognitive processes. However, the mechanisms for implementation of human cognitive functions in cognitive interface are in the early stage of development at the moment.

The key issues of creating the correct and acceptable (in terms of practical implementation) user mentality model still remain unsolved. The problem of accurate representation of the semantic data model, taking into account the heterogeneity and variability of the data in time also has not been solved completely.

At the same time, without claiming to be complete, it is possible to make practical efforts to improve cognitive interfaces for multisubject information systems. For example, using the principles of gestalt psychology and user perceptual stereotypes. Our work in the field of development of dynamic cognitive interfaces for multisubject information systems will continue.

**Acknowledgments.** The research is partially supported by RFBR grant no. 16-29-12878 “The development of the identification methods of the dynamic models with random parameters and their use in Eurasia migration forecasting”.

## References

1. Boyd, D.M., Ellison, N.B.: Social network sites: definition, history, and scholarship. J. Comput.-Mediat. Commun. **12**, 210–230 (2007)
2. Needleman, R., Miller, C.C., Jeffries, A.: Reporters’ Roundtable: Checking in with Facebook and Foursquare. CNET, 3 September 2010. https://www.cnet.com/news/reporters-roundtable-checking-in-with-facebook-and-foursquare/. Retrieved 8 Oct 2010
3. Obar, J.A., Wildman, S.: Social media definition and the governance challenge: an introduction to the special issue. Telecommun. Policy **39**, 745–750 (2015)
4. Bevan, N.: Measuring usability as quality of use. J. Softw. Qual. Issue **4**, 115–140 (1995)
5. Bevan, N.: International Standards for HCI and Usability. Int. J. Hum.-Comput. Stud. **55**(4), 533–552 (2001)

6. Iliinsky, N., Steele, J.: Designing Data Visualizations. O'Reilly, Sebastopol (2011)
7. Bertin, J., Barbut, M.C.: Sémiologie Graphique. Les diagrammes, les réseaux, les cartes. Gauthier-Villars, Paris (1967). 431 p.
8. Oeltze, S., Doleisch, H., Hauser, H., Weber, G.: Interactive Visual Analysis of Scientific Data. Presentation at IEE VisWeek, Seattle (WA), USA (2012)
9. Zenkin, A.A.: Cognitive Computer Graphics. D.A. Pospelov (ed.), M. Nauka (1991). 192 p. (Зенкин А. А. Когнитивная компьютерная графика/ Под ред. Д. А. Поспелова. – М.: Наука., 1991. – 192 с.)
10. Pospelov, D.A.: Artificial Intelligence. Directory. Book 2. Models and Methods, M.: Radio and Communications (1990). 304 p. (Поспелов Д.А. Искусственный интеллект. Справочник. Книга 2. Модели и методы М.: Радио и связь, 1990. — 304 с.)
11. Solso, R.L., MacLin, M.K., MacLin, O.H.: Cognitive Psychology, 8th edn. (September 7, 2007). Allyn & Bacon (2008). 592 p. (ISBN 13: 978-0205521081, ISBN 10: 0205521088)
12. Blomberg, O.: Concepts of cognition for cognitive engineering. Int. J. Aviat. Psychol. **21**, 85–104 (2011). doi:10.1080/10508414.2011.537561
13. Franchi, S., Bianchini, F.: The Search for a Theory of Cognition: Early Mechanisms and New Ideas (Cognitive Science). Rodopi, Amsterdam (2011). 408 p. (ISBN 10: 9042034270, ISBN 13: 978-9042034273)
14. Broadbent, D.E.: Perception and Communication. Oxford Science Publications. Oxford University Press, Oxford (1987). 352 p. (ISBN 10: 0198521715, ISBN 13: 978-0198521716)
15. Gray, J.A., Wedderburn, A.A.: Grouping strategies with simultaneous stimuli. Q. J. Exp. Psychol. **12**, 180–184 (1960)
16. Biernat, M., Kobrynowicz, D., Weber, D.L.: Stereotypes and shifting standards: some paradoxical effects of cognitive load. J. Appl. Soc. Psychol. **33**, 2060–2079 (2003). doi:10.1111/j.1559-1816.2003.tb01875.x
17. Vicentiy, A.V., Shishaev, M.G.: To the question of the development of cognitive interfaces for systems of information management support the development of spatially-distributed systems. Sci. Almanac **5**, 123–127 (2015)
18. Vicentiy, A.V., Shishaev, M.G.: Visualization in scientific and engineering research. Sci. Almanac **4**, 192–196 (2015)
19. Vicentiy, A.V., Shishaev, M.G., Oleynik, A.G.: Dynamic cognitive geovisualization for information support of decision-making in the regional system of radiological monitoring, control and forecasting. In: Proceedings of CSOC 2016 Conference, pp. 483–495, March 2016
20. Kohler, W.: Gestalt Psychology: An Introduction to New Concepts in the Modern Psychology. Liveright Publishing Corporation, New York (1947). 367 p. Gestalt (Psychology)
21. Koffka, K.: Principles of Gestalt Psychology. Routledge, New York (1935). 720 p.
22. Miller, G.: The magical number seven, plus or minus two. Psychol. Rev. **63**, 81–97 (1956)

# Chart Visualization of Large Data Amount

Pavel Pokorný(✉) and Kamil Stokláska

Department of Computer and Communication Systems,
Faculty of Applied Informatics, Tomas Bata University in Zlín,
Nad Stráněmi 4511, 760 05 Zlín, Czech Republic
pokorny@fai.utb.cz, dieres@gmail.com

**Abstract.** The main task of this paper is to describe the form of large data amount obtaining, its processing and a chart visualization. In order to get required information, charts need to have the interactive character with many setting parameters and properties. For these reasons, many visualization libraries exist that simplify the developer's work. In this paper there, most used chart visualization web libraries are described. In the next part, the new software solution was designed. Its development and realization was based on the real data from the industrial environment.

**Keywords:** Visualization · Application · Data · Processing · Charts

## 1 Introduction

Our information age more often feels like an era of information overload. Excess amounts of information are overwhelming; raw data becomes useful only when we apply methods of deriving insight from it. Fortunately, we humans are visual creatures. Few of us can detect patterns among rows of numbers, but even young children can interpret bar charts, extracting meaning from those numbers' visual representations. For that reason, data visualization is a powerful exercise. Processing and the next visualizing data is the fastest way to communicate it to others [1].

For these reasons, data processing and their visualization are hot topics. Data processing represents the collection and manipulation of items of data to produce meaningful information [2]. With the development of more data sources, such as social media platforms, photos, and customer reviews, Big Data has become a concern for small businesses and large corporations alike. Data is coming from all parts of the business like finance, customer service, and sales, and using it effectively helps you gain a competitive advantage [3].

Data collection usually proceeds in the form of online communication, in which the data is stored in the server database. Server applications can usually quickly access to these saved information and filter them for users. The visualization process of this data is then performed by a client application, which is often running in the Web environment as an interactive dashboard.

This contribution briefly describes the process of data processing and visualization in a Web environment using HTML5 and JavaScript technologies. There are encapsulated the common requirements for the solution of this problem, described existing

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_49

common Web libraries used in this filed and designed own solution for chart's data visualization from real traffic monitoring.

## 2 Application's Requirements

This chapter describes specifications of the main requirements for the application that can be used for the chart's visualization process. The core of this application usually are one or more libraries, which solve primary problems that can be universally re-used in different visualization applications. The own application then solve specified user requirements. The following requirements were based on the using that libraries for the real data processing and visualization.

### 2.1 Chart's Output

There are many charts commonly used in exploratory data analysis. The main purpose of charts is to easy understanding of large quantities of data and the relationship between parts of the data. The most used charts are bar, pie and line charts. [4]

In the bar chart, data are rendered in the form of rectangular strips with the different width and height, which is directly proportional to the value of the displayed quantity (Fig. 1). Data values are grouped into categories according to the domain name, which is always a part of the data structure. Categories are distributed uniformly over the axis length in the form of static text string. Each category can contain any number of values that can be grouped together or can be stacked in a single column. A composition of columns brings the advantage of a greater number of values in the category, where is not enough place to display all the values individually.

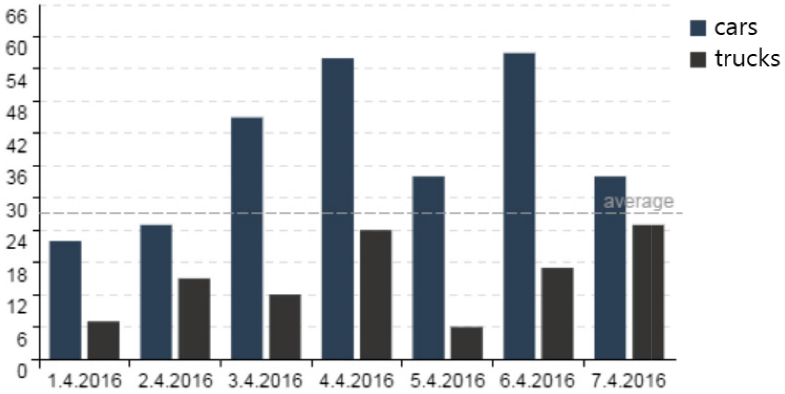


**Fig. 1.** Example of a bar chart

The component of a pie chart displays data in the form of categories of different size circle sector - segment. In the case of drawing multiple data series, data are grouped into categories based on the value of a domain name. For each segment, user can see its value directly or converted to a percentage representation (Fig. 2).

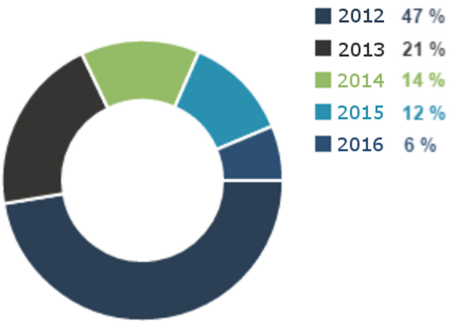


**Fig. 2.** Example of a pie chart

The line chart component is usually used to compare two-dimensional data (Fig. 3). The values are rendered as interactive points with the X and Y positions, which are connected by the line or other simple graphic object and represent the best rate of variables. The structure of the source data can contain multiple numbers of processes that can be drawn together at the same time. A part of the line chart is formed by an intelligent timeline that can be usually adapted to values and a format of the entered data.

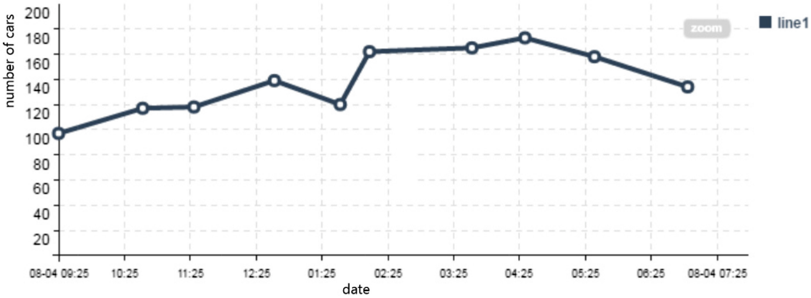


**Fig. 3.** Example of a line chart

### 2.2 Technology, Compatibility and Configuration

Because the library should run in the web environment, the HTML5 and JavaScript languages were selected. These languages are supported in all massively used modern web browsers (Firefox, Chrome, etc.).

HTML5 is a markup language used for structuring and presenting content on the World Wide Web. It is the fifth and current version of the HTML standard. With Cascading Style Sheets (CSS), and JavaScript, it forms a triad of cornerstone technologies for the World Wide Web [5]. Web browsers receive HTML documents from a webserver or from local storage and render them into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document [6].

JavaScript is a high-level, dynamic, untyped, and interpreted programming language. It has been standardized in the ECMAScript language specification. The majority of websites employ it, and all modern Web browsers support it without the need for third-party plugins. JavaScript is prototype-based with first-class functions, making it a multi-paradigm language, supporting object-oriented, imperative, and functional programming styles [7].

The rendering output can be performed via SVG or Canvas. SVG stands Scalable Vector Graphics and specifies own language for describing 2D graphics in the XML technology. This means that every graphics element is available within the SVG DOM (Document Object Model). In SVG, each drawn shape is remembered as an object. If attributes of an SVG object are changed, the browser can automatically re-render the shape. Canvas is rendered pixel by pixel. In Canvas, once the graphic is drawn, the browser forgets it. If its position should be changed, the entire scene needs to be redrawn.

The configuration of all types of graphs must be based on the same names and logic. This unification simplifies its extension for the future. The configuration usually includes definition of all graphical elements (colors, fonts, alignment etc.), definition of size, a reference to the data source (automatic mode) and callback for each event.

### 2.3 Data Processing and Documentation

Each chart should define the input data that is expected. The data format should have the most universal form and should support different variations in dependency of data types. The application should allow the user load, process and visualize new data runtime, if necessary. The source data for the chart can be automatically downloaded at periodic intervals or received based on the configuration or by calling function.

In order to communicate the application with the user, it also should support the reaction on the common events: clicking on a line, column or pie, clicking on the entire chart, filter, settings, and data change (callback).

Used library must contain clear and simple programmer documentation. A simple example with explanations and notes that is included in the documentation for each chart type also simplifies the use of this library, its modification and prospective extensions.

## 3 Web Visualization Libraries

The users can find many JavaScript charting libraries on the internet. These libraries are varied and have different supported chart types, interactivity, databinding, rendering technologies or licenses [8]. In the field of processing and visualization of large data amount, Google Charts, Chart.js, D3 and Flot are most widely used.

The Google Company develops the Google Charts library. Its interface makes easy to render numerical data into a graphical visualization, which can be represented as various type of pre-prepared charts from simple lines to a complex tree structures. Charts uses HTML5 for rendering graphics elements or SVG technics that includes graphical parts in the vector representation directly into the structure of an HTML document independent on the selected platform. Charts also contains implemented the VML (Vector Markup Language) technology that provides backward compatibility with older browsers without support the HTML5 technology [9].

Chart.js is open source simple yet flexible JavaScript library. It can visualize user data in eight different chart types; each of them can be animated and customizable. The next benefit is the mix and match bar and line charts in order to provide a clear visual distinction between datasets. Chart axis types are fully customizable; the user can plot complex sparse datasets on date time, logarithmic or even custom scales with ease. The rendering process is performed by the Canvas technology that offers great rendering performance across all modern browsers and charts are automatically redrawn on window resize in order to get perfect scale granularity [10].

D3 (Data-Driven Documents) is a JavaScript library that is primary used to manipulate documents based on the input data. D3 allows one direct control and manipulation for the native web presentation and its elements, instead of using custom components and abstract interfaces. The D3 library selectively performs to establish data on specific HTML elements and their properties can dynamically make changes based on the entered data. D3 unlike other libraries does not work with pre-defined graphical elements. Here, the user himself creates them dynamically with the help of definitions and their attributes [11].

Flot Charts is an extension for the jQuery library, which separates the functional logic from the HTML structure and facilitates manipulation with the DOM (Document Object Model) element. The Flot library contains ready-made components for the four basic types of charts - bar, line, point and segment. These components can be easily and indefinitely extended and allows one change a wide variety of configuration parameters. The Flot Charts library supports multiple timelines and allows the user insert different data sets into a single view. Thanks to interaction with the jQuery, Flot Charts can also add external components that must not to be a part of an internal implementation and can be easily designed by users. These components are directly inserted into the HTML structure using jQuery selectors and input parameters in the form of a multidimensional array of configuration object and data [12].

## 4 An Example of Customized Solution

On the other hand, they are usually too complex and many of their components cannot be used for specific solutions. Further, some of them do not have open source code, which avoids its adaptation to the unique needs and poor extensibility. In these cases, the own solution can solve these deficits. As example, large data amount was obtained. This data was measured in the real environment of traffic monitoring. This data contains millions of values and we decided to design and realize own solution.

Own solution is performed by a new JavaScript library that implements the selected web technologies and allows the user to create graphical components for data visualization Due to the effort to maintain the greatest optimization, the design of this library was very specific without using partial solutions and third-party libraries. The concept was designed in order to reach the whole requirements described in Sect. 2.

### 4.1 Structure, Technology and Compatibility

The entire library is created only with the help of JavaScript and HTML5 technologies [5, 6]. Program code was based on the object-oriented principles, in order to use their benefits, which object-oriented programming offers [7].

The Canvas technique was selected for the rendering process. Due to this technology, the library is optimized for all modern browsers and Internet Explorer version 9 and above [13].

The Created library is composed from individual components and global static core classes that are common to all graphs, and include methods for general computing and object creation. Individual components are immediately implemented as separated classes, whose instances are associated with the particular Canvas element. The Canvas rendering element is always created automatically while the new chart is created. The basic object includes the static method, which generates required component based on input parameters and places it directly into the content of the website.

All library content is encapsulated in the one object. Source codes of components are available in separated files, which are merged into a single output file during the distribution process. To work with the library, there is just necessary to include the library file into the header or body of an HTML document.

### 4.2 Types of Graphs, Configuration and Data Processing

The created library supports three basic types of charts - pie, bar and line charts. These charts are designed to show specific data from information systems for traffic monitoring. However, each of them can be easily adapted and configured for almost any data.

Bar chart can be generated in two versions - horizontal and vertical. The horizontal variant has category names decomposed on the Y-axis and the X-axis contains generated range of values. The current value is then represented by the width of the bar. The vertical graph has switched axis and the current value is drawn with the help of height.

Pie chart allows set parameters “options.radius” and “options.innerRadius”; with them, the user can set a fixed size of the radius of a circle and its inner part, which is displayed during the rendering process. Both values are entered in proportion to the maximum size that is determined by the dimensions of the parent element. There is not necessary to use the entire range of 360° circle during the rendering the data. Limiting the scope of parameters can be set by “options.startAngle” and “options.endAngle” parameters that are at the beginning initialized to 0 and 360 values.

The line chart component exactly meets the requirements described in the Sect. 2.1. Settings for Canvas rendering of a chart legend are unified for all types of charts. The position of legend can be defined using the “options.legendState” parameter. Then, a template in the form of a text string that contains keywords gives the appearance of the legend. Braces enclose these keywords and they are replaced by the values that correspond to the actual characteristic of the graph before the rendering process.

Data is received in the JSON (JavaScript Object Notation) structure, which provides a universal template for all supported chart types. The data has always two-level structure. The first level contains individual data series, and the second contains data parameters [14]. An example of the structure of two input datasets for a line chart is shown in the Fig. 4.

The configuration structure can be transformed to the constructor that can be dynamically changed while the chart is drawn directly on the auxiliary Options object or can simply create an instance of a chart. This object is constructed for each chart component and contains parameters for rendering the entire graphic elements and overall layout.

All parameters on the Options object are dynamically configurable using predefined values. For each change of these values, the function for redraw the chart is called automatically. Each chart also contains a public method that allows the developer dynamically modify multiple configuration parameters at once.

### 4.3 Interactivity, Animation and Adaptivity

Creating of interactive content on the application environment is performed by capturing events on individual elements. In case of rendering using the Canvas technology, the only single element is used. For this purpose, an event manager and a system of interactive objects were created. The manager controls any interactivity.

The basic element is represented by the class, from which all displayed objects with the interaction can inherit. This class keeps information about the size, settings of the object and location. Any object can register an unlimited number of events that are immediately associated with a specific return functions.

When the event of the mouse cursor is captured on the Canvas element, the specific method for interaction and control with the each available object is called. In case the interaction is positive, this object receives a message that contains the reference of captured event. The object calls associated functions based on its content.

The created library contains two independent managers to support animations. The first manager is represented by a static class, which is used for a frame animation. Frame animation is typically characterized by scrolling of the offset of the source bitmap in periodic intervals. The second manager is used for motion animation and other static class represents it. This class allows one to animate any parameter on the specified object with the selected speed. It also process events of start, end or change of animation. Each animated object is put into a queue, which is automatically updated at each window refresh.

Some elements in the graph do not need to redraw as often as the others. For example, the timeline and the chart layout remains static, but animated data needs to be redrawing itself. Rendering of the content is therefore performed into several bitmap layers, which are separated by the required rendering frequency. Each layer is represented by a separate bitmap virtual Canvas element, which is not actually placed into the page content. After the each level is redrawing, its reference is stored in a global variable and during the chart update, this level is redrawn to the main Canvas.

Adjustment of the Canvas element in HTML5 is possible in two ways. The first adjustment is given by the CSS (Cascading style sheets). Here, the user can dynamically set height and width for each element. Next, an element is adapted to the size of

```
data = {
    "datova_rada_1": [
        {
            "value": "1.4.2016",
            "count": 22,
            "label": "c1"
        },
        {
            "value": "2.4.2016",
            "count": 25,
            "label": "c2"
        }..
    ],
    "datova_rada_2": [
        {
            "value": "1.4.2016",
            "count": 7,
            "label": "n0"
        },
        {
            "value": "2.4.2016",
            "count": 15,
            "label": "n1"
        }..
    ]
}
```

**Fig. 4.** Example of the structure of two datasets for a line chart

the drawing area, but its content remains rendered in the previous size. It causes the result with deformed image data and loss of image quality. For this reason, the second adjustment method is preferred. It works on the principle complete redrawing of data for each change the size of the parent element. After the capturing an event that indicates a size change of the element, the attributes width and height of the container itself are improved and the native Canvas size is modified. The actual Canvas size of chart is defined by its parent element, which is adapted to the window size of the used web browser.

## 5 Conclusion

This contribution describes the processes large data amount obtaining, its elaboration and the chart's visualization. The attention was mainly focused on the web solution, based on the HTML5 and JavaScript technologies.

In this field, many Web libraries are existing. The most widely used JavaScript Libraries are Google Charts, Chart.js, D3 or Flot. They are often used in the real environment and offer many properties and supporting tools in order to the user get the correct required visualization.

Based on the obtained industrial data from real traffic monitoring, the new software chart's solution was designed. Its implementation is strictly platform independent and created library can be better adapted for specific data and achieve maximum optimization for the rendering process. This library was also designed to meet all requirements for the use in modern information systems and environments with output for the next data processing. There is also possible to use the chart's components separately on any platform that supports HTML5. With this technology, the library can be easily extended with other components to customize to a specific system.

**Acknowledgments.** The created application was developed in collaboration with Incinity Company, which develops complex integrated platform for smart cities. This company also provided data source that was processed.

## References

1. Murray, S.: Interactive data visualization for the web. O'Reilly Media, CA (2013)
2. French, C.S.: Data Processing and Information Technology. Thomson, London (2004)
3. Yuk, M., Diamond, S.: Data Visualization for Dummies. Wiley, New York (2014)
4. Graph types - definitions and examples. http://www.typesofgraphs.com/
5. W3C contributors: HTML5 - a vocabulary and associated APIs for HTML and XHTML. http://www.w3.org/TR/html5/
6. Pilgrim, M.: HTML5: Up and Running. O'Reilly Media, CA (2010)
7. Stefanov, S.: Object-Oriented JavaScript. Packt Publishing Ltd., Birmingham (2008)
8. Karavirta, V., Shaffer, C.A.: JSAV: the JavaScript algorithm visualization library. In: Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education, pp. 159–164 (2013)
9. Charts - Interactive charts for browsers and mobile devices. https://developers.google.com/chart/
10. Chart.js - Open source HTML5 Charts for your website. http://www.chartjs.org/
11. D3- Data Driven Documents. https://d3js.org/
12. Flot - Attractive JavaScript plotting for jQuery. http://www.flotcharts.org/
13. Geary, D.: Core HTML5 Canvas – Graphics, Animation and Game Development. Prentice Hall, US (2012)
14. Rischpater, R.: JavaScript JSON Cookbook. Packt Publishing Ltd., Birmingham (2015)

# Software Optimization of Advanced Encryption Standard for Ultra-Low-Power MSP430

Radek Fujdiak(✉), Petr Mlynek, Jiri Misurec, and Janko Slacik

Department of Telecommunications, Brno University of Technology,
Technicka 12, 616 00 Brno, Czech Republic
{fujdiak,mlynek,misurec}@feec.vutbr.cz, slacik.jan@vutbr.cz
http://www.vutbr.cz

**Abstract.** The Internet of Things (IoT) is undoubtedly a current topic for private and public sector. Nowadays, the communication technologies allow to connect even a simplest physical object often with very limited physical resources. The IoT cover among the others also many areas with sensitive data, where the limited devices are also used. The use of these limited devices keep the security issue as a difficult task. The symmetric ciphers are often considered as a best way to encrypt the communication in the limited devices. Despite the fact that there are many hardware optimized solutions, there are still areas where these solutions cannot be used i.e. due to the limiting price or power. This paper focus on a software optimization of the symmetric cipher on limited micro-controller. Two main implementation are introduced. Further, we provide experimental measurements and possible suggestions for time consumption and memory use reduction.

**Keywords:** Advanced encryption standard · Symmetric cipher · Software optimization · Internet of Things · Limited devices

## 1 Introduction

The Internet of Things (IoT) idea is starts in early years of 21st century and over the years its gain big attention from the public and also from the private sector [1]. The IoT is a global interconnected network of physical non-living objects as i.e. computers, sensors, indicators, measuring devices, buildings, cars, robots, wearables, phones, and many others. The inter-networking allows to the devices share and collect various data and information between each other. This might be used for optimization, analysis, safety, security, management, remote controlling, and many other different purposes [2]. Many years was IoT only a vision, but the last years bring various worldwide changes. The new wireless communication technologies are probably the most important changes, which actually allowed to implement the idea of IoT in real environment [3]. Nowadays, it is possible to connect even the simplest devices to the global network. These simple devices are

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_50

**Table 1.** Various key-sizes of different cryptographic algorithms with same security level [8].

| Symmetric key-size [sb] | ECC key-size [sb] | Asymmetric key-size [sb] |
|---|---|---|
| 80 | 160 | 1024 |
| 112 | 224 | 2048 |
| 128 | 256 | 3072 |
| 192 | 384 | 7680 |
| 256 | 512 | 15360 |

often very limited i.e. from the size, power input, memory, price or performance point of view [4]. The IoT cover among the others also many areas with sensitive data [5], where the limited devices are also used. These limiting factors are crucial when it come to the security and these are also keeping the security issue as a difficult task [6]. There are many ways how to solve this difficult task, but over the years the symmetric ciphers showed the best results in the areas with limited physical resources [7]. Table 1 shows the comparison of key-size in secure bits (sb) of different cryptographic algorithms for keeping same security level (same time needed for key break).

The bigger key size brings among the others higher complexity of the sub-operations of each algorithm, higher memory storage requirements, higher performance requirements, more time consumption or higher power consumption. How we can see, the most effective by key-size are the symmetric ciphers. Nowadays, the most used symmetric cipher is Advanced Encryption Standard (AES), thanks to the standardization and many different recommendations [9–11]. AES was established in 2001 by the U.S. National Institute of Standards and Technology (NIST) in FIPS PUB 197 [12]. AES is a standard created from Rijndael cipher (selected 128 b bits block with different key-length: 128 b, 192 b and 256 b). AES shows good performance in hardware and also in software [13]. However, this paper focuses on experimental measurements of AES software implementation on micro-controller with limited physical resources in the real environment. We provide a results summary from these measurements, comparison of two different software implementation, analysis of the results and suggestions for future optimization.

The rest of the paper is organized as follows. Section 2 provides a summary of current work and bring closer the experimental background. Following this research background, the experimental measurements and results are presented in Sect. 3. Last but not least, the Sect. 4 shows the discussion and final conclusion.

## 2 Current Work and Experimental Background

We are focused on the remote data collection in the Smart Grid network as in Fig. 1. This communication provide i.e. remote electricity, water, gas or heat consumption management [14].

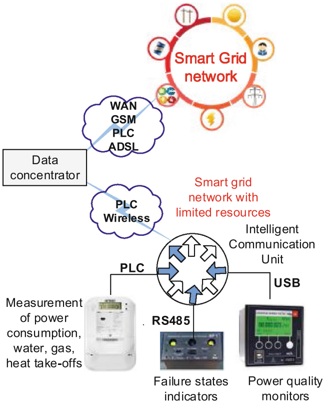

Smart Grid network
WAN
GSM
PLC
ADSL
Data concentrator
PLC
Wireless
Smart grid network with limited resources
Intelligent Communication Unit
PLC
USB
RS485
Measurement of power consumption, water, gas, heat take-offs
Failure states indicators
Power quality monitors

**Fig. 1.** The considered Smart Grid network model for various data collection [14].

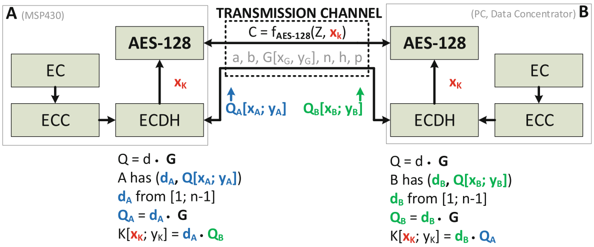


**Fig. 2.** The block diagram of our developed secured communication model [16].

The end-devices (meters, indicators or monitors) are connected via existing connections (power lines, RS485 or USB) to the Intelligent Communication Unit (ICU) [15]. Further, data from the ICU are transferred to the Data concentrator via suitable access communication technology as i.e. wireless technologies or power lines. Finally, the data concentrator is connected to central system (Supervisory Control and Data Acquisition system - SCADA) via suitable transport communication technology as i.e. wide area technology, cellular technology, PLC or others. Our last works were focused on securing the smart grid channel in the part of the network with limited physical resources (Fig. 2).

We developed a secured communication channel, which consists own random number generator (RNG) [17,18]; symmetric key cryptography for online data encryption and a user authentication [14] and public key cryptography for key exchange [16,19]. Finally, the Fig. 3 shows our experimental network[1], where the

[1] Our described secure communication has been tested in an experimental network in CEZ Distribuce, a.s. (one of the biggest Czech power supplier). The MEg40+ Universal energy meter was installed in the power substation Noviny (Velky Grunov area, the Czech Republic). The Data Concentrator was located in Brno, the Czech Republic. The communication distance was approximately 240 km [14].

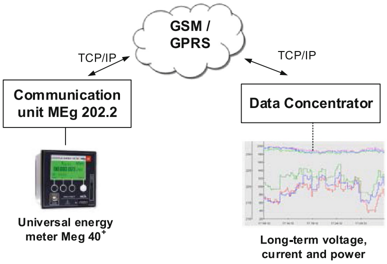

GSM / GPRS
TCP/IP
TCP/IP
Communication unit MEg 202.2
Data Concentrator
Universal energy meter Meg 40+
Long-term voltage, current and power

**Fig. 3.** The block diagram of used network for experimental measurements [14].

measurements were done. There is end-devices an universal measurement unit - MEg40 connected via RS485/USB transferrer to communication unit MEg202.2 (Intelligent Communication Unit from Fig. 1). This communication unit is connected via GSM/GPRS (TCP/IP protocol) to data concentrator from where the data are sent into remote monitoring system.

The communication unit MEg202.2 (Fig. 3) is using the ultra-low-power micro-controller of MSP430 series as a main core for encryption and communication processes. In our measurements, we were using the specific micro-controller MSP430f5438A. This micro-controller has 256 kB FLASH, 16 kB RAM, 32-bit multiplier, high/low frequency crystal (DCO 32 MHz/ACLK 32 kHz) and allow 16-bit operations; more details about technical specification of MSP430 might be found in [20]. We used default DCO frequency $\tilde{1}$ MHz. That means $\tilde{1}$00 ns for one single cycle ($T_{cycle} = 1/f_{CPU}$). $V_{cc}$ was 3000 mV and the operating mode was active mode (AM). The $I_{cc}$ was 300 $\mu$A for our $V_{cc}$ and $f_{CPU}$.

We were working with two implementations of AES-128 cipher. First was our developed C-library (AES method A), which might be found in [21]. First version of this implementation was finished in 2012. However, the final version was made in 2014. That year also came out the second C-library from the manufacturer (AES method B), which might be found in [22]. We provide a comparison of both methods, exact measurements of internal processes of each method, deep analysis of the results and clear conclusion for future optimization.

## 3 Experimental Measurements and Results

First of all, we will shortly introduce the AES algorithm in a high-level description. The AES algorithm operates on 4×4 matrix called the State and it has a four main operations: byte substitution (the S-box is used for replacing each byte in non-linear substitution), rows shift (cycling the last 3 rows of the State), mix columns (combining 4 B in each column) and round key addition (combination of the State and round key block by XOR bit-operation). The algorithm might be described as (for AES-128, AES with 128-bit key) [12]:

1. Expand key (derivation from Rijndael's key schedule)
2. Pre-round (only the round key addition)
3. Rounds 1–9 (using all four basic operations)
4. Rounds 10 (no mix columns).

Further, the Table 2 already shows the first results of the AES algorithm for both implementation (Method A and Method B). The table is divided to the each implementation and separately also for encryption and decryption process. The results represent mean value of multiple measurements.

**Table 2.** Basic results of each AES implementation.

| | # | AES - method A | | AES - method B | |
|---|---|---|---|---|---|
| | | Encryption | Decryption | Encryption | Decryption |
| Byte substitution | Cycles | 448 | 448 | 368 | 375 |
| Shift rows | Cycles | 70 | 70 | 80 | 80 |
| Mix columns | Cycles | 838 | 16259 | 999 | 1568 |
| Add round key | Cycles | 495 | 486 | 338 | 333 |
| Whole iteration | Cycles | 1851 | 17263 | 1806 | 2384 |
| Last iteration | Cycles | 1004 | 1003 | 817 | 815 |
| Summary | Cycles | 19035 | 157743 | 17520 | 26161 |
| FLASH memory | B | 7586 | | 2226 | |
| RAM memory | B | 2034 | | 160 | |

The first measurements already shows significantly higher efficiency of the Method B in the decryption process. The Method A has only one more effective operation the shift rows, but compared with the other ineffective operations this cycle reduction is negligible. The Fig. 4 shows the performance complexity for each operation separately for description and encryption. The graph shows an anomaly in the mix-columns operation for decryption in Method A. This anomaly slow the whole process of encryption and mainly the decryption. However, thanks to these findings there is a possibility for analysis and future optimization. The following text will be more focused on the Mix Columns operation as it is the most demanding operation.

The following code describes the inverted Mix-Columns operation of the AES Method A for decryption process. There are defined basic pre-computation macros: *xtime()* and *Multiply()* for simplifying the final operations; the $a$, $b$, $c$, $d$ are unsigned-char auxiliary variables; and the *state* is two-dimensional (2-D) array variable represents the 4×4 State matrix. If we look to used operations and operators, we can find 1 *for* cycle, 60 multiplying operations, 0 addition operation and 700 bitwise operations (384 shifting operations, 240 logical AND operations, 76 logical exclusive OR operations).

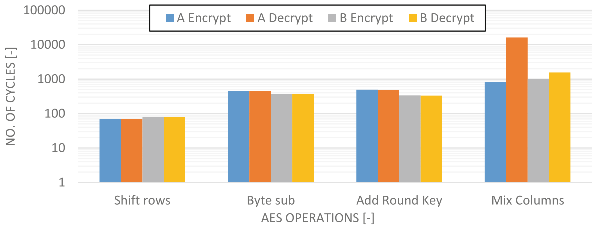


**Fig. 4.** The performance complexity of all operations for each method.

*Inverted Mix Columns (decryption) for AES Method A* [21]

```
...
#define xtime(x)   ((x<<1) ^ (((x>>7) & 1) * 0x1b))
...
#define Multiply(x,y) (((y & 1) * x) ^ ((y>>1 & 1) *
   xtime(x)) ^ ((y>>2 & 1) * xtime(xtime(x)) ^
   ((y>>3 & 1) * xtime(xtime(xtime(x)))) ^
   ((y>>4 & 1) * xtime(xtime(xtime(xtime(x))))))
...
void InvMixColumns() {
...
 for(i=0;i<4;i++) {
...
  state[0][i] = Multiply(a, 0x0e) ^ Multiply(b, 0x0b) ^
   Multiply(c, 0x0d) ^ Multiply(d, 0x09);
  state[1][i] = Multiply(a, 0x09) ^ Multiply(b, 0x0e) ^
   Multiply(c, 0x0b) ^ Multiply(d, 0x0d);
  state[2][i] = Multiply(a, 0x0d) ^ Multiply(b, 0x09) ^
   Multiply(c, 0x0e) ^ Multiply(d, 0x0b);
  state[3][i] = Multiply(a, 0x0b) ^ Multiply(b, 0x0d) ^
   Multiply(c, 0x09) ^ Multiply(d, 0x0e);
...
```

The following code describes the inverted Mix-Columns operation of the AES Method B for decryption process. There are defined basic pre-computation macro: *galois_mul2()* for simplifying the final operations; the *buf1–buf3* are unsigned-char auxiliary variables; and the state variable represents 4×4 the State matrix. However, the Method B is using only one-dimensional (1-D) array for representation the 4×4 matrix. The 2-D 4×4 arrays are transformed to 1-D 16×1 arrays. If we look to used operations and operators, we can find 1 *for* cycle, 0 multiplying operations, 20 addition operation and 40 bitwise operations (10 shifting operations, 0 logical AND operations, 30 logical exclusive OR operations).

*Inverted Mix Columns (decryption) for AES Method B* [22]

```
...
```

```
unsigned char galois_mul2(unsigned char value) {
if (value>>7) {
return ((value << 1)^0x1b); } else
return (value << 1); }
...
for (i=0; i <4; i++) {
...
 buf1 = galois_mul2(galois_mul2(state[buf4]^state[buf4+2]));
 buf2 = galois_mul2(galois_mul2(state[buf4+1]^state[buf4+3]));
 state[buf4] ^= buf1; state[buf4+1] ^= buf2; state[buf4+2] ^=
  buf1; state[buf4+3] ^= buf2;
...
 buf1 = state[buf4] ^ state[buf4+1] ^ state[buf4+2] ^
 state[buf4+3]; buf2 = state[buf4];
 buf3 = state[buf4]^state[buf4+1]; buf3=galois_mul2(buf3);
 state[buf4] = state[buf4] ^ buf3 ^ buf1;
 buf3 = state[buf4+1]^state[buf4+2]; buf3=galois_mul2(buf3);
 state[buf4+1] = state[buf4+1] ^ buf3 ^ buf1;
 buf3 = state[buf4+2]^state[buf4+3]; buf3=galois_mul2(buf3);
 state[buf4+2] = state[buf4+2] ^ buf3 ^ buf1;
 buf3 = state[buf4+3]^buf2; buf3=galois_mul2(buf3);
 state[buf4+3] = state[buf4+3] ^ buf3 ^ buf1;
...
```

The following Table 3 compares both methods in the number of each operations. The FOR cycles are the most demanding operations, but each method use one. The multiplying operation is negligibly slower than addition operation. However, we can see that Method A uses slower operations and it has 40 more of these. The last are the bitwise operations, where again the Method A has 660 more of these operations. We need also consider that the multiplying, addition and bitwise operations are multiplied 4-times because they are in FOR cycle of four rounds and then the differences are even higher. Moreover, the Method B uses only 1-D field, which is also way faster than 2-D field (mostly if we consider the small numbers of cells).

**Table 3.** Comparison of different number of used operations in the both methods.

| Operation | AES method A | AES method B |
|---|---|---|
| FOR cycles | 1 | 1 |
| Multiplying | 60 | 0 |
| Additions | 0 | 20 |
| Bitwise operations | 700 | 40 |

We consider also the automatic code optimization methods from [24]. There are used i.e. unused code, variables and functions reduction; variables in register;

**Table 4.** Results of each AES implementation after automatic code optimization.

| | # | AES - method A | | AES - method B | |
|---|---|---|---|---|---|
| | | Encryption | Decryption | Encryption | Decryption |
| Byte substitution | Cycles | 124 | 124 | 111 | 111 |
| Shift rows | Cycles | 56 | 56 | 48 | 50 |
| Mix columns | Cycles | 591 | 1980 | 344 | 574 |
| Add round key | Cycles | 99 | 99 | 105 | 91 |
| Whole iteration | Cycles | 870 | 2259 | 659 | 869 |
| Last iteration | Cycles | 279 | 279 | 410 | 283 |
| Summary | Cycles | 8561 | 20912 | 6460 | 9357 |
| FLASH memory | B | 5284 | | 2628 | |
| RAM memory | B | 2034 | | 160 | |

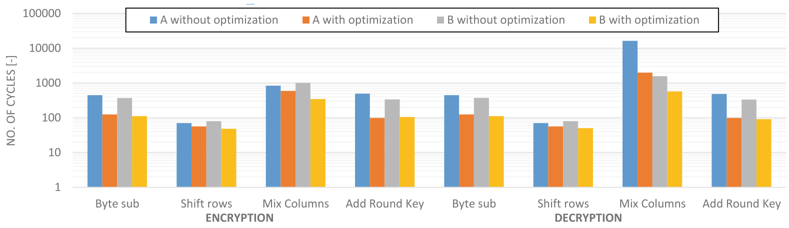


**Fig. 5.** The performance complexity of all operations after optimization.

cycles optimization; in-lines for small functions; and many others. The following Table 4 shows the results of the AES algorithms after automatic code optimization.

The Fig. 5 summarized the comparison of each method and the optimized and non-optimized results for decryption and encryption processes.

## 4 Discussion and Conclusion

We provide experimental results and comparison of two different AES-128 cipher implementation. Further, we provide an analysis of the both software implementation and show the main issues. The implementation Method B shows the most promising results, but it is still possible to optimize it due to the fact of demanding operations. Current works are mostly concentrating to the FPGA or similar implementations with minimizing the time consumption of their algorithm [25, 26]. The comparison value is in this case time (ms), which is not wrong in this are, but for general comparison with other implementations the clock-cycles are missing. Our approach is in specific applied research, where the novelty held in the functional implementation and comparison of different kind of AES implementation on the specific hardware.

The future work should consider to combination of both methods. The possible ways might be just over the most demanding operations via i.e. using only 1-D arrays, cycles reduction, multiplication or addition operations via bitwise operations, and others. The another way how to improve the performance might be to use higher-level of pre-computated parameters, but with these methods the memory requirements is growing.

**Acknowledgement.** Research described in this article was financed by the National Sustainability Program under grant LO1401and by the Technology Agency of Czech Republic project No. TA04021490. For the research, the infrastructure of the SIX Centre was used.

## References

1. Masek, P., et al.: Implementation of true IoT vision: survey on enabling protocols and hands-on experience. Int. J. Distrib. Sens. Netw. (2016)
2. Ruggieri, M., Nikookar, H.: Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems. River Publishess, Aalborg (2013). ISBN 978-87-92982-96-4
3. Palattella, M.R., et al.: Internet of things in the 5G Era: enablers, architecture, and business models. IEEE J. Sel. Areas Commun. **34**(3), 510–527 (2016). ISSN 0733–8716
4. Ometov, A., et al.: Feasibility characterization of cryptographic primitives for constrained (Wearable) IoT devices. In: IEEE International Conference on Pervasive Computing and Communications, pp. 1–6 (2016). ISBN 978-1-5090-1940-3
5. Crabtree, A., Mortier, R.: Personal Data, Privacy and the Internet of Things: The Shifting Locus of Agency and Control. SSRN, Elsevier (2016)
6. Levitt, T.: IoT Governance, Privacy and Security Issues. European Research Cluster on The Internet of Things. European Commission - Information Society and Media (2015)
7. Yadav, S.K., Relan, N., Bhatia, S.J.: On industrial needs of symmetric cryptography. Int. J. Theor. Appl. Sci. **2**(1), 41–44 (2010). ISSN 0975–1718
8. Damien, G.: BlueKrypt v30.2. Cryptographic Key Length Recommendation (2017)
9. European Payments Council: Guidelines on Cryptographic Algorithms Usage and Key Management. EPC342-08 Version 6.0 (2016)
10. European Network and Information Security Agency: Algorithms, Key Sizes and Parameters Report: 2013 Recommendations. Version 1.0 (2013)
11. U.S. National Institute of Standards and Technology: Recommendatin for Key Management - Part 1: General (Revision 3). NIST Special Publication 800–57 Part 1 Revision 3 (2016)
12. U.S. National Institute of Standards and Technology: Announcing the Advanced Encryption Standard (AES). Federal Information - Processing Standards Publication 197, FIPS PUB 197 (2001)
13. Otero, C.T.O., Tse, J., Manohar, R.: AES hardware-software co-design in WSN. In: 21st IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC), pp. 85–92 (2015)
14. Mlynek, P., Misurec, J., Koutny, M., Raso, O.: Design of secure communication in network with limited resources. In: Proceedings of the 4th European Innovative Smart Grid Technologies (ISGT), pp. 1–5 (2013). ISBN 978-1-4799-2984-9

15. Mlynek, P., Misurec, J., Koutny, M., Silhavy, P.: Two-port network transfer function for power line topology modeling. Radioengineering **21**(1), 356–363 (2012)
16. Fujdiak, R., et al.: Efficiency evaluation of different types of cryptography curves on low-power devices. In: Proceedings of the 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 269–274 (2015). ISBN 978-1-4673-9283-9
17. Fujdiak, R., et al.: Random number generator in MSP430 x5xx families. Elektrorevue. **4**(1), 70–74 (2013). ISSN 1213–1539
18. Fujdiak, R., Mlynek, P., Misurec, J., Raso, O.: Cryptography in ultra-low power Microcontroller MSP430. Int. J. Eng. Trends Technol. **6**(8), 398–404 (2013). ISSN 2231–5381
19. Fujdiak, R., Misurec, J., Mlynek, P., Leonard, J.G.: Cryptograph key distribution with elliptic curve Diffie-Hellman algorithm in low-power devices for power grids. Rev. Roum. Sci. Tech. Serie Électrotechn. et Énerg. **61**(1), 84–88 (2016). ISSN 0035–4066
20. Texas Instruments: Mixed Signal Microcontroller: MSP430F5438A-EP. Technical documentation SLAS967A (2014)
21. Mlynek, P., Koutny, M., Misurec, J., Raso, O.: Authentication and encryption DLL library, (Software). http://www.utko.feec.vutbr.cz/~mlynek/dll.html
22. Texas Instruments: Advanced Encryption Standard, (software). http://www.ti.com/tool/AES-128#descriptionArea
23. Fujdiak, R.: Measurement of symmetric cipher on low power devices for power grids. In: Proceedings of the 21st Conference STUDENT EEICT, pp. 556–560 (2015). ISBN: 978-80-214-5148- 3
24. Texas Instruments: MSP430 Optimizing C/C++ Compiler v16.12.0.STS. Technical documentation SLAU132N (2016)
25. Talha, S.U., et al.: Efficient advance encryption standard (AES) implementation on FPGA using Xilinx system generator. In: 6th IEEE International Conference on Intelligent and Advanced Systems (ICIAS), pp. 1–6 (2016)
26. Rao, M., Newe, T., Grout, I.: AES implementation on Xilinx FPGAs suitable for FPGA based WBSNs. In: 9th IEEE International Conference on Sensing Technology (ICST), pp. 773–778 (2015)

# The Effects of Clustering to Software Size Estimation for the Use Case Points Methods

Zdenka Prokopova(✉), Radek Silhavy, and Petr Silhavy

Faculty of Applied Informatics, Tomas Bata University in Zlin,
nam T.G. Masaryka 5555, Zlin, Czech Republic
{prokopova, rsilhavy, psilhavy}@fai.utb.cz

**Abstract.** The main objective of the paper is to present the suitability and effects of several different clustering methods for improving accuracy of software size estimation. For software size estimation was used the Algorithmic Optimisation Method (AOM), which is based Use Case Points (UCP) method. The comparison of K-means, Hierarchical and Density-based clustering is provided. Gap, Silhouette and Calinski-Harabasz criterion were selected as an evaluation criterion for clustering quality. Estimation ability of clustered model is compared on Sum of squared error (SSE). Results shows that clustering improves an estimation ability.

**Keywords:** Clustering · Software size estimation · Use case points · Algorithmic optimisation method

## 1 Introduction

Nowadays a software estimation became a very important and crucial in the process of software system developing. There are exist a lot of methods and techniques for software effort estimation. Many of them are based on historical dataset where linear regression is applied to completed projects for a new estimation obtaining [1]. One of the widespread method is a Use Case Points (UCP) algorithm that was developed by Gustav Karner [2] and is useful for early stage prediction of software effort. The method is based on use case models while a number of use case steps were initially involved in the estimation process. It was developed many modifications of the original principles of UCP, such as use case size points by Braz et al. [3], use case points modified by Diev [4], adapted use case points by Mohagheghi et al. [5], extended use case points by Wang et al. [6], simplified use case points by Ochodek et al. [7], etc. Another modification of UCP has been named as an Algorithmic Optimisation Method (AOM) and was developed by Silhavy et al. [8].

Clustering can be understood as a technique dividing data points into classes (clusters) that share similarity level. The goal of the clustering is to obtain a consistent groups, where similarity brings a better estimation ability of models. The clustering performs better when the similarity within the class and distance between classes are greater [9–11].

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_51

There are many clustering method are under an investigation, which differ in approaches to dividing data into clusters [9, 11]: partitional, hierarchical, exclusive, overlapping, fuzzy, complete or partial clustering.

Partitional clustering divides the set of data points into non-overlapping clusters such as each data point is in exactly one cluster. Hierarchical clustering is creation of a set of clusters that are organized as a tree i.e. clusters have permission to have sub-clusters. Hierarchical clustering can be viewed as a sequence of partitional clustering. Exclusive clustering means that the data points could be placed to a single cluster. In more cases, a data point can belong to more than one cluster and this situation is addressed by overlapping (non-exclusive) clustering. Fuzzy clustering is built on the idea that every data point belongs to every cluster with a membership weight, which is from the interval <0, 1>.

Complete clustering assigns every data point to some cluster. On the other hand, some data points may not belong to defined clusters and in this case, we talk about partial clustering.

Moreover, there are exist several different notions of cluster according to which there are described another types of clusters e.g. well-separated, prototype-based, density-based, conceptual clusters etc. [10, 11]. An idealistic definition of a cluster is specifying that all data points in a cluster must be sufficiently close (or similar) to one another opposite to any data points outside the cluster. Therefore, in well-separated clusters the distance between any two data points in different clusters is larger than the distance between any two data points within a cluster. The prototype can by understood as the most central point for many type of data. In that case, we could refer to prototype-base clusters as center-base clusters. Conceptual clusters we can generally call shared-property clusters i.e. data points within clusters share some property.

## 2 Problem Statement

In this paper, we are focusing to comparison of three clustering technics; K-means, Hierarchical clustering, Density-based clustering [11]. Gap, Silhouette and Calinski-Harabasz [12] criteria will be used for algorithm evaluation. Clustering effect on improving software size estimation accuracy will be evaluated by using AOM algorithm [8].

### 2.1 Research Questions

The research questions answered by this study are as follows:

RQ1: Does clustering improve estimation accuracy?
RQ2: Does it play the role the number of clusters?
RQ3: Are all clustering methods equivalent in the sense of estimation accuracy?
RQ4: Are there any other parameters effecting estimation accuracy in combination with different clustering methods?

### 2.2 Evaluation Criteria

In size or effort estimation, there are several commonly used criteria, which are accepted as a standard evaluation. Therefore, all performed simulations were evaluated according three of these metrics. These are known as follows:

Sum of squared error (SSE)

$$SSE = \sum\nolimits_{i=1}^{n} \varepsilon_i^2, \tag{1}$$

Mean squared error (MSE)

$$MSE = \frac{1}{n}\sum\nolimits_{i=1}^{n} \varepsilon_i^2, \tag{2}$$

Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{\sum\nolimits_{i=1}^{n} \varepsilon_i^2}{n}}, \tag{3}$$

where $n$ is number of observations and $\varepsilon$ is a residual error value.

## 3 Research Methods

### 3.1 Algorithmic Optimisation Method

The AOM [8] is based on standard UCP method [2] and was created by investigating the possibility of increasing accuracy of the estimation process.

Main building blocks in the UCP method are actors and use cases. According to their importance, they are classified as simple, average and complex (weighted by the factors). The sum of weighted actors creates a value called unadjusted actor weights (UAW), and the sum of weighted use cases creates a value called unadjusted use case weights (UUCW). Two other coefficients called technical factors (TCF) and environmental factors (ECF) are used to describe detailed features of the project. Summing UAW and UUCW and then multiplying this value by the TCF and ECF coefficients obtain the number of UCP [2, 3, 6, 7]:

$$UCP = (UAW + UUCW) \times TCF \times ECF \tag{4}$$

An optimisation algorithm is based on application of multiple least squares regression [13, 14] to UCP model described by Eq. (4). The final form of the equation for expressing $\text{UCP}_{\text{AOM}}$ could be written as follows

$$UCP_{AOM} = a_0 + a_1(UAW \times TCF \times ECF) + a_2(UUCW \times TCF \times ECF) \tag{5}$$

where $a_0, a_1$ and $a_2$ are parameters obtained from multiple least square regression applied to set of completed projects (dataset). A detailed presentation of the AOM is described in [8]. For our current purposes we used linear model with intercept.

### 3.2 Clustering Techniques

As was written previous, many approaches divides data points into clusters. For our simulation study we have chosen three of them: partitional, hierarchical, and density-based clustering. Distance measurement takes key role clustering accuracy. Therefore all algorithm were tested with selected types of Euclidean distance and non-Euclidean distance. Euclidian distance is based on the locations of data points in Euclidean space, e.g. squared Euclidean distance and Non-Euclidean measure is based on properties of data points, but not their "location" in a space, e.g. Cosine distance. Selection of distance measurement depends on the kind of clustering data [10].

**K-means.** K-means is one of the oldest and most widely used clustering algorithm. It's a prototype-based, partitional clustering technique that attempt to find a user-specified number of clusters (*k*), which are represented by their centroids. In contrast with hierarchical clustering, k-means clustering operates on actual observations and creates a single level of clusters. For that reason, k-means clustering is often more suitable than hierarchical clustering for large amounts of data [10].

The generic k-means algorithm consist of following steps: Selection of *k* data points as initial centroids (clusters). Each data point is then assigned to the closest centroid and this set creates a cluster. The centroid of each cluster is then recomputed based on the data points assigned to the cluster. Recomputed steps are repeating until the sum of the distances to the centroids is smallest for the chosen number of iterations [11].

**Hierarchical Clustering.** Hierarchical clustering is a second widely used and important group of clustering methods. The main principle is grouping data points over a variety of scales by creating a cluster tree (dendrogram). There are exist two basic approaches of hierarchical clustering [10]:

*Agglomerative*, which starts with data points as individual clusters and at each step merge the closest pair of clusters.

*Divisive*, which starts with one cluster and at each step split a cluster until only singleton clusters remain.

The basic Agglomerative hierarchical clustering algorithm could be described by the next steps: Finding the similarity or dissimilarity between every pair of data points in the dataset, grouping the data points into a hierarchical cluster tree, determination where to cut the hierarchical tree into clusters [11, 15].

**Density-based Clustering.** Density-based clustering works on principle localization of regions with high density and their separation from regions with low density [11]. One of density-based clustering method is using Gaussian Mixture Models. This technique form clusters by representing the probability density function of observed variables as a mixture of multivariate normal densities. Gaussian mixture modelling uses an iterative algorithm that converges to a local optimum. Gaussian mixture modelling may be more appropriate than K-means clustering when clusters have different sizes and correlation within them [15].

### 3.3 Cluster Validation

Cluster validation (evaluation) is essential and should be a part of any cluster analysis. Sometimes the data contains natural divisions that indicate the appropriate number of clusters. Otherwise, when the data does not contain natural divisions, we need to determine the optimal number of clusters. Cluster validation helps to determine clustering tendency, number of clusters and provides accuracy of clustering [12]. There are several approaches to evaluate clusters; depending on different clustering techniques. To determine how well the data fits into a specific number of clusters we can use index values defined by different evaluation criteria, such as Gap, Silhouette or Calinski-Harabasz criterion:

- *Gap criterion* - gap criterion value is used to evaluate the optimal number of clusters:

  $$Gap_n(k) = E_n^*\{log(W_k)\} - log(W_k), \tag{6}$$

  $$W_k = \sum\nolimits_{r=1}^{k} \frac{1}{2n_r} D_r, \tag{7}$$

  where $n$ is the sample size, $k$ is the number of clusters being evaluated, $n_r$ is the number of data points in cluster $r$, $W_k$ is the pooled within-cluster dispersion measurement and $D_r$ is the sum of the pairwise distances for all data points in cluster $r$. Value $E_n^*\{log(W_k)\}$ is determined by Monte Carlo sampling from a reference distribution and $log(W_k)$ is computed from the sample data. The optimal number of clusters occurs at the solution with the largest gap value within a tolerance range [15, 16].
- *Silhouette criterion* - for each data point it is a measure of how similar that data point is to data points in its own cluster, when compared to data points in other clusters. It is defined as follows:

  $$S_i = (b_i - a_i)/max(a_i, b_i), \tag{8}$$

  where $a_i$ is the average distance from the $i$th data point to the other data points in the same cluster, and $b_i$ is the minimum average distance from the $i$th data point to data points in a different cluster, minimized over clusters. The silhouette value ranges from −1 to +1. Data points that are very distant from neighbouring clusters are indicating by value +1, data points that are not distinctly in one cluster or another are indicating by value 0, and data points that are probably assigned to the wrong cluster are indicating by value −1. If most data points have a high value, then the clustering configuration is appropriate, if many data points have a low or negative value, then the clustering configuration may have wrong number of clusters (to many or too much clusters). How well separated the resulting clusters are (the number of clusters are correct), we can determine also by Silhouette graphical technique. Silhouette graph displays a measure of how close each data point in one cluster is to data points in the neighbouring clusters [15, 17].

- *Calinski-Harabasz criterion* - is defined as:

$$CHC_k = \frac{SS_B}{SS_W} \times \frac{(n-k)}{(k-1)}, \tag{9}$$

$$SS_B = \sum\nolimits_{i=1}^{k} n_i \|m_i - m\|^2, \tag{10}$$

$$SS_W = \sum\nolimits_{i=1}^{k} \sum\nolimits_{x \in c_i} \|x - m_i\|^2. \tag{11}$$

$SS_B$ is the overall between-cluster variance, $SS_W$ is the overall within-cluster variance, $k$ is the number of clusters, $n$ is the number of observations, $m_i$ is the centroid of the cluster $i$, $m$ is the overall mean of the sample data, $x$ is a data point, $c_i$ is the $i$th cluster. Expression $\|m_i - m\|^2$ or $\|x - m_i\|^2$ are the norms (Euclidean distance) between the two vectors. Well-defined clusters have a large between-cluster variance ($SS_B$) and a small within-cluster variance ($SS_W$). The optimal number of clusters is the solution with the highest Calinski-Harabasz value [15, 18].

# 4 Experiment

Presented experiment shows how to examine similarities and dissimilarities of specified data points using cluster analysis by the help of Matlab functions for a purpose of improving software size estimation accuracy by using AOM algorithm.

## 4.1 Description and Design

The experiment was evaluated by using dataset, which was collected by the authors from various software companies and was used and analysed in [19]. The plot diagram of the variables (UAW, UUCW, TCF, ECF) from the dataset is illustrated in Fig. 1.

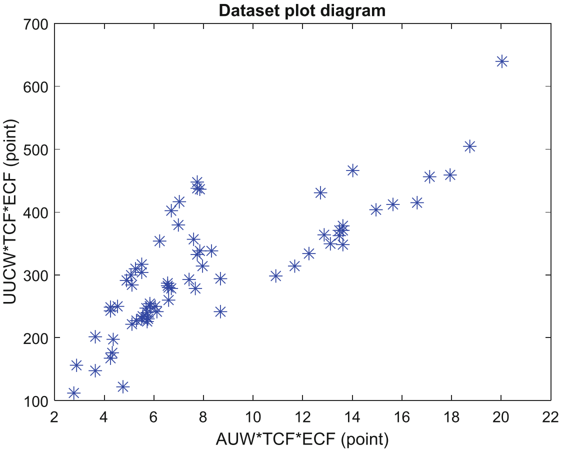


**Fig. 1.** Dataset plot diagram - UAW*TCF*ECF vs. UUCW*TCF*ECF

## 4.2 Simulations and Results

### Clustering

In our experiment, we decided to perform three types of cluster analysis: K-means (Fig. 2), Density-based (Fig. 3) and Hierarchical (Fig. 4) clustering. Criterion values were computed for 2, 3 and 4 clusters. For additional comparison three different distance metrics were used for K-means and Hierarchical clustering:

| | |
|---|---|
| ‘sqeuclidean’ | Squared Euclidean distance |
| ‘cosine’ | One minus the cosine of the included angle between points |
| ‘cityblock’ | Sum of absolute differences |

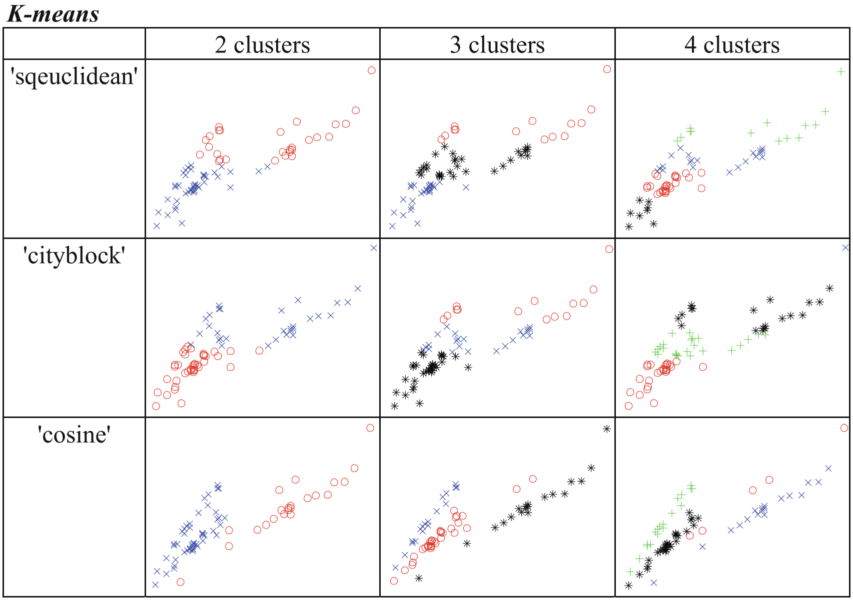


**Fig. 2.** K-means clustering for 2, 3, and 4 clusters using tree distance metrics

### Evaluation

For the evaluation of the clustering solution, we used three different criterion described earlier in the Sect. 3.3:

- *Gap Evaluation* - is used to evaluate the optimal number of clusters according Eqs. (6) and (7). The largest gap values were calculated using K-means and Hierarchical clustering with cosine distances. As we can see on the Figs. 5 and 6, optimal number of clusters is one. Evaluation of Density-based clustering could be calculated only with squared Euclidean distance (other tested distance metrics are not supported) and leads to one cluster again (see Fig. 7).

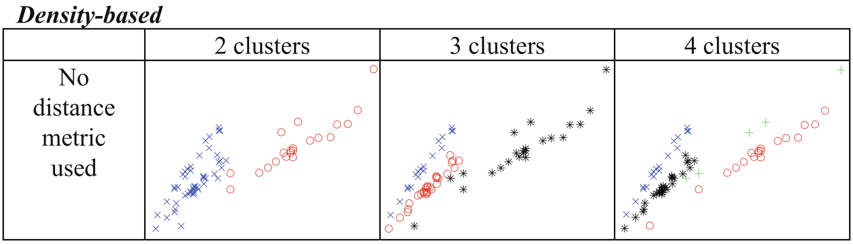


**Fig. 3.** Density-based clustering for 2, 3, and 4 clusters

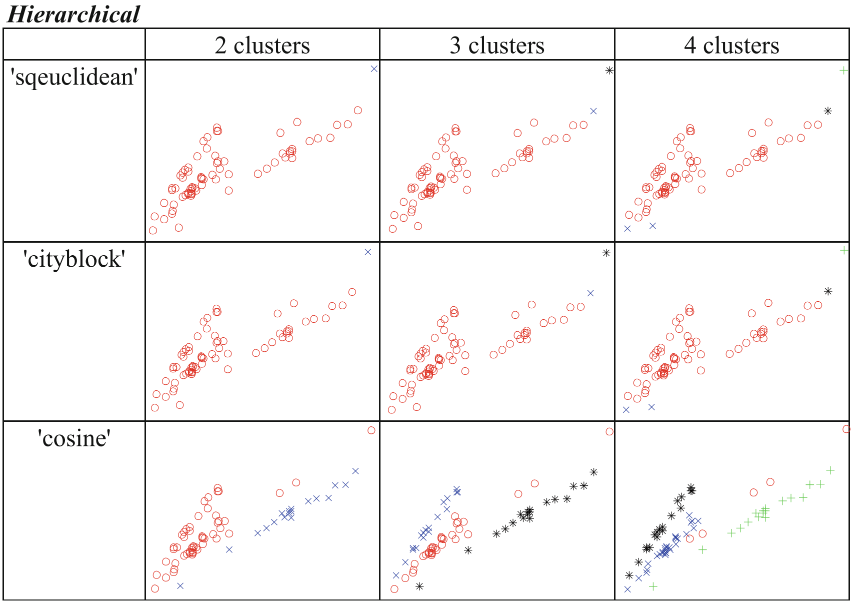


**Fig. 4.** Hierarchical clustering for 2, 3, and 4 clusters using tree distance metrics

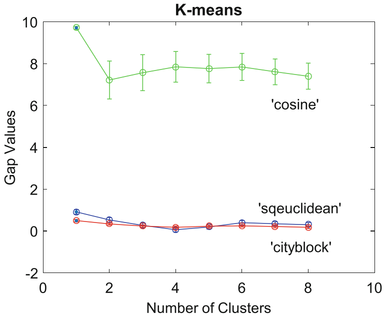


**Fig. 5.** Gap evaluation for K-means clustering

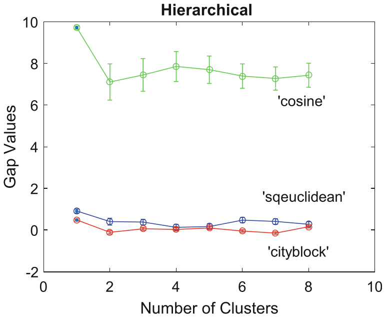


**Fig. 6.** Gap evaluation for Hierarchical clustering

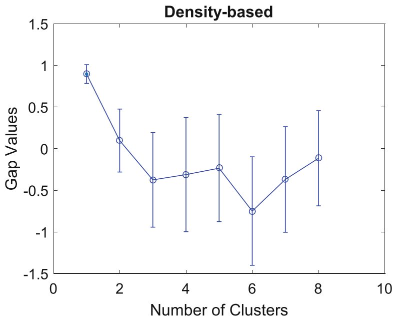


**Fig. 7.** Gap evaluation for Density-based clustering

- *Silhouette Evaluation* – in our experiment, the highest silhouette value was obtained using K-means and Hierarchical clustering with cosine distances (see Figs. 8 and 9). In these cases, optimal number of clusters are four. High silhouette values (greater than 0.93) were obtained. Evaluation of Density-based clustering could be calculated only with squared Euclidean distance and leads to two clusters (see Fig. 10).

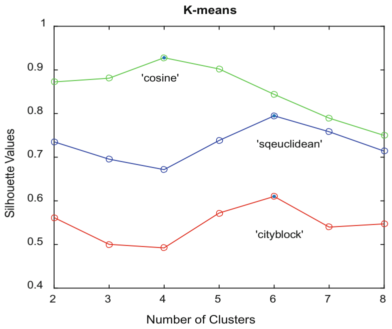


**Fig. 8.** Silhouette evaluation for K-means clustering

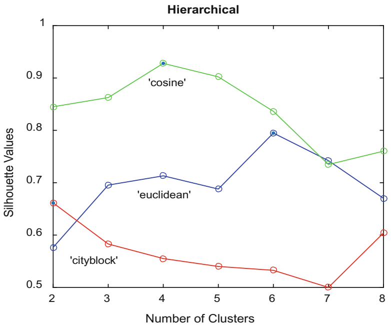


**Fig. 9.** Silhouette evaluation for Hierarchical clustering

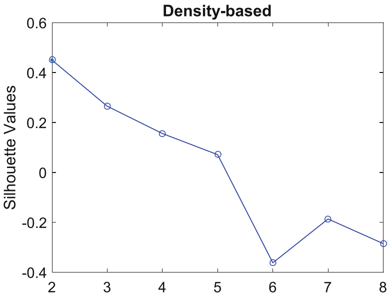


**Fig. 10.** Silhouette evaluation for Density-based clustering

- *Calinski-Harabasz Evaluation* - as with the previous evaluation methods, the optimal number of clusters is connected with the highest Calinski-Harabasz index value. As we can see on the Fig. 11, optimal number of clusters is seven using K-means clustering. For Calinski-Harabasz evaluation is supported only squared Euclidean distance metric.

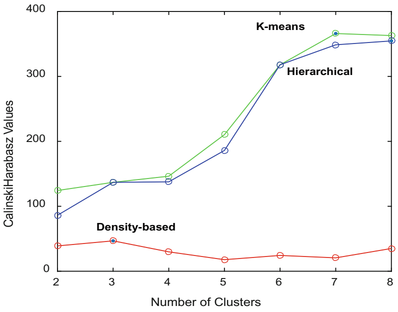


**Fig. 11.** Calinski-Harabasz evaluation using all tested clustering techniques

### Results Comparison

For the comparison purposes, we used three different criterion for three clustering techniques and tree different distance metrics (in short: 'sqeuclid', 'cosine', 'citybl') as we can see in Table 1. Results for the AOM algorithm were calculated according Eq. (5) and evaluated by criteria (1)–(3). For clustered dataset SSE is presented as sum of SSE for individual clusters, MSE and RMSE are presented as mean values of MSE and RMSE for individual clusters.

**Table 1.** Comparison of AOM without clustering and with three clustering technics

| Clustering | Clusters | SSE | | | MSE | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 'sqeuclid' | 'citybl' | 'cosine' | 'sqeuclid' | 'citybl' | 'cosine' | 'sqeuclid' | 'citybl' | 'cosine' |
| AOM | 1 | ***50 182*** | ***50 182*** | ***50 182*** | ***748.9*** | ***748.9*** | ***748.9*** | ***27.4*** | ***27.4*** | ***27.4*** |
| K-means | 2 | 47 902 | 47 336 | 35 917 | 677.8 | 702.5 | 418.0 | 25.2 | 25.9 | 17.9 |
| | 3 | 45 092 | 43 680 | 39 861 | 652.6 | 628.6 | 595.4 | 24.9 | 24.4 | 22.2 |
| | 4 | 40 576 | 41 250 | 32 549 | 593.0 | 660.7 | 444.0 | 23.7 | 24.9 | 18.9 |
| Hierarchical | 2 | – | – | 49 489 | – | – | 525.8 | – | – | 19.6 |
| | 3 | – | – | 39 749 | – | – | 584.9 | – | – | 21.9 |
| | 4 | – | – | 32 549 | – | – | 444.0 | – | – | 18.9 |
| Density-based | 2 | 35 221 | 35 221 | 35 221 | 390.4 | 390.4 | 390.4 | 16.2 | 16.2 | 16.2 |
| | 3 | 32 654 | 32 654 | 32 654 | 544.3 | 544.3 | 544.3 | 21.3 | 21.3 | 21.3 |
| | 4 | 36 779 | 36 779 | 36 779 | 422.1 | 422.1 | 422.1 | 15.4 | 15.4 | 15.4 |

## 5 Conclusion

In this paper, we have presented the simulation study comparing effects of clustering to software size estimation. We have decided to compare K-means, Hierarchical and Density-based clustering techniques with three difference distance metrics: ‘sqeuclidean’, ‘cosine’ and ‘cityblock’. Firstly Gap, Silhouette and Calinski-Harabasz criterion were selected for the evaluation of the clustering algorithm. Secondly clustering algorithm were applied to modified AOM algorithm, which was used for software size estimation. All simulations were calculated by Matlab functions on testing dataset. According the experimental results, we can conclude that:

RQ1: We found that all tested clustering techniques improve estimation accuracy. Table 1 shows that AOM achieved SSE = 50 182, MSE = 748.9, RMSE = 27.4 and for all tested clustering techniques are SSE, MSE and RMSE better (SSE $\in$ <32 549, 47 902>, MSE $\in$ <390.4, 702.5>, RMSE $\in$ <15.4, 25.9>). It means that clustering conclusively improve estimation accuracy.

RQ2: As we can see in Table 1 number of clusters play significant role. Compare e.g. Hierarchical clustering with two clusters where is SSE = 49 489 (MSE = 525.8, RMSE = 19.6) and with four clusters where is SSE = 32 549 (MSE = 444.0, RMSE = 18.9).

RQ3: According experiment results we can conclude that clustering methods are not equivalent. This claim can be substantiated with numeric results in Table 1 (compare differences in SSE, MSE, RMSE) and also visually through Figs. 2, 3 and 4. Table 1 shows that K-means with two clusters and distance metric ‘sqeuclidean’ obtained SSE = 47 902 (MSE = 677.8, RMSE = 25.2), Density-based clustering obtained SSE = 35 221 (MSE = 390.4, RMSE = 16.2) but Hierarchical clustering has produced inappropriate distribution of clusters and AOM model cannot be created.

RQ4: Simulation study shows that estimation accuracy is influenced by distance metric for K-means and Hierarchical clustering algorithm. Effect size of distance metric can be illustrate on K-means algorithm. K-means clustering with two clusters and ‘sqeuclidean’ distance obtained SSE = 47 902 (MSE = 677.8, RMSE = 25.2), with ‘cityblock’ is SSE = 47 336 (MSE = 702.5, RMSE = 25.9) and with ‘cosine’ distance is SSE = 35 917 (MSE = 418.0, RMSE = 17.9.5), as shown in Table 1. Density-based clustering is not effected by distance metrics.

Finally, simulation study can be concluded as follows: as the most suitable for a given dataset appears K-means clustering technique for four clusters with ‘cosine’ distance metric which obtained SSE = 32 549 (MSE = 444.0, RMSE = 18.9) or Density-based clustering for three clusters which obtained SSE = 32 654 (MSE = 544.3, RMSE = 21.3). From the cluster validation point of view, for K-means clustering is best suited Silhouette evaluation (see Fig. 8) and for Density-based clustering Calinski-Harabasz evaluation (see Fig. 11).

## References

1. Jorgensen, M., Shepperd, M.: Systematic review of software development cost estimation studies. IEEE Trans. Softw. Eng. **33**(1), 33–53 (2007)
2. Karner, G.: Metrics for objectory’. Diploma, University of Linkoping, Sweden, No. LiTH-IDA-Ex-9344 21 (1993)
3. Braz, M.R., Vergilio, S.R.: Software effort estimation based on use cases. In: 30th Annual International Computer Software and Applications Conference, COMPSAC 2006 (2006)
4. Diev, S.: Use cases modeling and software estimation. ACM SIGSOFT Softw. Eng. Not. **31** (6), 1–4 (2006). doi:10.1145/1218776.1218780
5. Mohagheghi, P., Anda, B., Conradi, R.: Effort estimation of use cases for incremental large-scale software development (2005). doi:10.1109/icse.2005.1553573
6. Wang, F., Yang, X., Zhu, X., Chen, L.: Extended use case points method for software cost estimation (2009). doi:10.1109/cise.2009.5364706
7. Ochodek, M., Nawrocki, J., Kwarciak, K.: Simplifying effort estimation based on use case points. Inf. Softw. Technol. **53**, 200–213 (2011). doi:10.1016/j.infsof.2010.10.005
8. Silhavy, R., Silhavy, P., Prokopova, Z.: Algorithmic optimisation method for improving use case points estimation. PLoS ONE **10**, e0141887 (2015)
9. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988). Book available online at http://www.cse.msu.edu/~jain/Clustering_Jain_Dubes.pdf
10. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
11. Tan, P. N., Steinbach, M., Kumar, V.: Cluster analysis: Basic Concepts and Algorithms. Introduction to Data Mining (2006). Book available online at http://www-users.cs.umn.edu/~kumar/dmbook/index.php
12. Milligan, G.W.: Clustering validation: results and implications for applied analyses. In: Clustering and Classification, pp. 345–375. World Scientific, Singapore (1996)
13. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis, 5th edn. Wiley, Hoboken (2012)
14. Jorgensen, M.: Regression models of software development effort estimation accuracy and bias. Empir. Softw. Eng. **9**, 297–314 (2004)
15. Mathworks: Cluster Analysis (2016). https://www.mathworks.com/help/stats/hierarchical-clustering.html
16. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. Roy. Stat. Soc. Ser. B **63**(Pt. 2), 411–423 (2001)
17. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. **20**, 53–65 (1987). doi:10.1016/0377-0427(87)90125-7
18. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**(1), 1–27 (1974)
19. Silhavy, R., Silhavy, P., Prokopova, Z.: Analysis and selection of a regression model for the Use Case Points method using a stepwise approach. J. Syst. Softw. **125**, 1–14 (2017)

# Evaluation of Data Clustering for Stepwise Linear Regression on Use Case Points Estimation

Petr Silhavy[(✉)], Radek Silhavy, and Zdenka Prokopova

Faculty of Applied Informatics, Tomas Bata University in Zlin,
nam T.G. Masaryka 5555, Zlin, Czech Republic
{psilhavy, rsilhavy, prokopova}@fai.utb.cz

**Abstract.** In this paper, stepwise linear regression model in conjunction with clustering for effort estimation is investigated. Effect of clustering is compared to Use Case Points model. The 2 to 20 clusters were tested. As shown increasing a number of clusters brings lower prediction errors. More clusters lower a distance between clusters members, which allows to construct more capable stepwise linear regression model.

**Keywords:** Effort estimation · Clustering · Use case points · Stepwise linear regression · Parametric model

## 1 Introduction

At early stage, there are no further information about software projects, which can be used for proper estimation. Project scope is described in limited way and there is no possibility of obtaining valid information, which should be useful for estimation. Therefore, methods based only on analogy approaches can be employed. Parametric effort estimation models are such models. Furthermore, a parametric effort estimation models in which a relationship between depended variable and set of independent variables is descripted. The availability of effort-governing variables is limited.

Linear regression models are often investigated as a solution. Traditional models produce usable solutions, but has some basic limitations. These models are sensitive to historical project data and optimal size of historical data pool have to be investigated. The original Use Case Points (UCP) [1, 2] method is not based on historical projects. Each new estimation is independent [3] and is not influenced by known errors, which were measured in past projects.

## 2 Methods

In this study, there were three datasets involved. Training dataset was obtained random sampling from data, which were collected. Training datasets contains n = 400 data points. Testing Dataset 1 were randomly selected from same dataset and Testing Dataset 2 was based on Ochodek's dataset [5] and on Subriadi dataset [6].

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_52

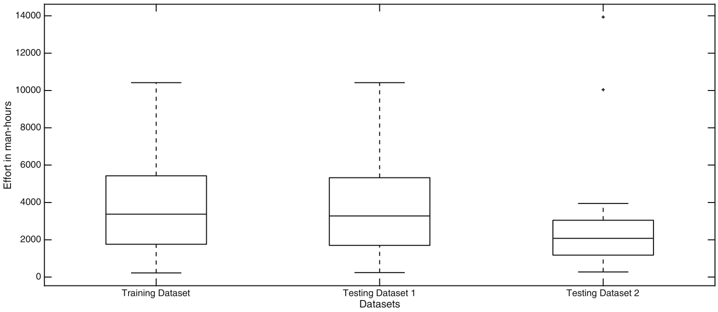


**Fig. 1.** Boxplots of working datasets. Real Effort values of project, which were measured.

This dataset contains 30 data-points. Boxplots of actual effort values in man-hours for all working datasets can be seen in Fig. 1.

### 2.1 Clustering Attributes Definition

UCP equation [1, 5, 7, 8] is consists of several components which are available for clustering. Following attributes were selected for creating 2, 5, 10, 15, 20 clusters:

- Unadjusted Actor Weights (UAW),
- Unadjusted Use Case Weights (UUCW),
- Technical Complexity Factor (TCF),
- Environmental Complexity Factor (ECF).

Using mentioned attributes as features for clustering is advantages because they can be obtained when use case points model is known. Use Case Points model is commonly presented as:

$$\mathrm{UCP} = (\mathrm{UAW} + \mathrm{UUCW}) \times \mathrm{TCF} \times \mathrm{ECF} \tag{1}$$

As described in [4] use method is used for analysis of use case points models. This is represented as UAW and UCW components in Eq. 1. TCF and ECF are used as project characteristics.

### 2.2 Experiment Planning and Hypothesis Formulation

**Experiment Design**

The training dataset was clustered iteratively for 1 to 20 cluster, where only 1, 2, 5, 10, 15 and 20 clusters were considered in our study. If there is only 1 cluster, it means that all data points of training dataset were used.

A clustering techniques k-means is one of the oldest and most widely used clustering algorithm. K-means is based partial clustering that is based on predefined number of clusters (k). Each cluster has a central point, centroid, which is used as base for distance measurement. K-means clustering works on actual observations and creates a single level of clusters. K-means involves a selection of *k* data points as initial centroids. Then a rest of data points are associated by distance measurement to selected centroid. The centroid of each cluster is then recomputed based on the data points assigned to the cluster. Recomputed steps are repeating until the sum of the distances is reached. Many often this is local minima. Working dataset were standardized and then clustered. Then, multiple linear regression in stepwise (MLR) form were applied to each cluster, which implies more optimal estimation equation than UCP equation is used, if all testing dataset is employed.

Later the proper cluster for each testing data point is determine. Actual effort is compared to predicted values which brings residual information, which is used for performance metrics calculation. All metrics are based on all clusters.

### 2.3 Experiment Planning and Hypothesis Formulation

The experiment in our study deals with selecting subsets of project for performing a UCP calculation.

- RQ1: Clustering will bring same level of accuracy as using whole historical data points.
- RQ2: More clusters have a positive impact on estimation accuracy.

We expect better performance of regression model than UCP. This behaviour will be positively influenced by increasing number of cluster. Performance measurement will show significantly better performance for more 20 clusters than for 1 (no clustering) cluster option.

**Adopted Metrics for Performance Measurement**
Residual Sum of Squares (RSS) were adopted as performance metric. Following equations are used for obtaining this performance metric:

$$RSS = \sum\nolimits_{i=1}^{n} \varepsilon_i^2 \tag{2}$$

RSS, which important measurement to evaluate regression models; Firstly, study the influence of increasing number of clusters and secondly compares results with UCP method.

New prediction was performing as follows; Testing dataset is used for model construction, where each data point was classifying into proper cluster and then prediction is obtained by regression model.

## 3 Results and Conclusions

### 3.1 Evaluating Step-Wise Models for Clustered Estimation

As can be seen in Fig. 2 in training phase 20-clusters produce a best option as can be seen from RSS tendency. When more clusters is added, it brings lower RSS. This is clearly visible in comparing to “un-clustered” value. Same in Fig. 3. It can be expected that un-clustered value is lower, but there are many outliners, which has strong influence to mean value (compare to Fig. 2).

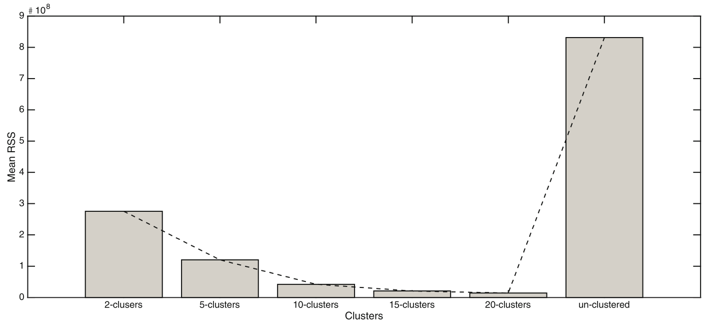


**Fig. 2.** Bar graph of mean RSS for clustering. Tendency can be seen.

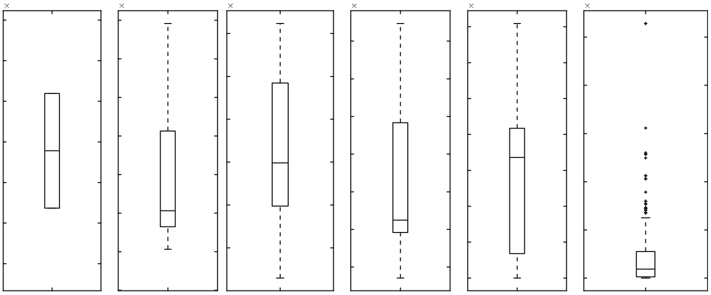


**Fig. 3.** BoxPlot graph for RSS among clusters.

### 3.2 Evaluating Testing Dataset

For each test data-point, the proper cluster was identified. If there are two clusters it brings two linear regression models, obtain in training phase, and vice-versa for more clusters. Each of this model is therefore specifically designed for similar projects from same cluster.

In Fig. 4 can be seen squared error (residual) for each of testing data-point. MLR model, which was obtained by step-wise regression (specific model for each cluster) and UCP equation (Eq. 1) is compared.

In Fig. 5 send dataset is used in same approach for second testing dataset. Both figures (Figs. 4, 5) are based on 20-clusters variants. Which means to construct 20 MLR models. Residuals are calculated by differencing a predicted and know values by using cluster's specific model.

### 3.3 Results Interpretation

According RQ1, we expected that clustering will not improve an estimation accuracy. This assumption can be rejected. As can be seen in Figs. 4 and 5 MLR models which are used for classified projects produce significantly better estimation. The peaks in

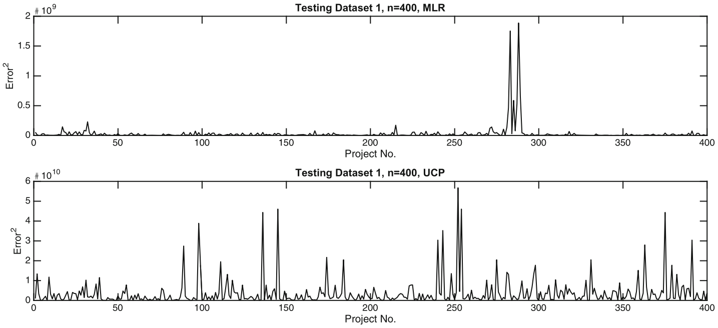


**Fig. 4.** Estimation error squared, testing dataset 1, n = 400

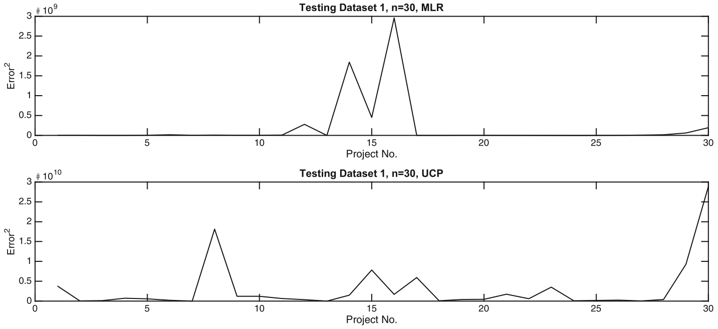


**Fig. 5.** Estimation error squared, testing dataset 2, n = 30

both figures (MLR part) are evidence of an outliners. More specifically this error is caused by inappropriate clustering for those particular projects.

In RQ2 we have presented that more clusters bring positive effect on prediction accuracy. This hypothesis we can accepted. As seen in Fig. 2 clustering makes estimation more accurate (lower RSS value), which is caused by improving similarity between historical (training) projects and new predicted projects.

In future research, more dataset will be tested and more subset selecting methods will be evaluated. The comparison to machine learning methods will be performed too.

## References

1. Karner, G.: Metrics for objectory, Diploma, University of Linkoping, Sweden, No. LiTH-IDA-Ex-9344, vol. 21, December 1993
2. Ochodek, M., Alchimowicz, B., Jurkiewicz, J., Nawrocki, J.: Improving the reliability of transaction identification in use cases. Inf. Softw. Technol. **53**, 885–897 (2011)
3. Nassif, A.B., Ho, D., Capretz, L.F.: Towards an early software estimation using log-linear regression and a multilayer perceptron model. J. Syst. Softw. **86**, 144–160 (2013)

4. Silhavy, R., Silhavy, P., Prokopova, Z.: Algorithmic optimisation method for improving use case points estimation. PLoS ONE **10**, e0141887 (2015)
5. Ochodek, M., Nawrocki, J., Kwarciak, K.: Simplifying effort estimation based on Use Case Points. Inf. Softw. Technol. **53**, 200–213 (2011)
6. Subriadi, A., Ningrum, P.: Critical review of the effort rate value in use case point method for estimating software development effort. J. Theoret. Appl. Inf. Technol. **59**, 735–744 (2014)
7. Robiolo, G., Orosco, R.: Employing use cases to early estimate effort with simpler metrics. Innovations Syst. Softw. Eng. **4**, 31–43 (2008)
8. Wang, F., Yang, X., Zhu, X., Chen, L.: Extended Use Case Points Method for Software Cost Estimation, pp. 1–5 (2009)

# Erratum to: Software Engineering Trends and Techniques in Intelligent Systems

Radek Silhavy(✉), Petr Silhavy, Zdenka Prokopova,
Roman Senkerik, and Zuzana Kominkova Oplatkova

Faculty of Applied Informatics,
Tomas Bata University in Zlín, Zlin, Czech Republic
radek@silhavy.cz

## Erratum to: R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*, Advances in Intelligent Systems and Computing, DOI 10.1007/978-3-319-57141-6

The original version of the book was inadvertently published with swapped first name and family name of authors "Vasilenko Alexandr and Tyrychtr Jan" which have to be corrected to read as "Alexandr Vasilenko and Jan Tyrychtr" in Chapter 11 and "Gruzenkin Denis Vladimirovich, Grishina Galina Viktorovna, Durmuş Mustafa Seçkin, Üstoğlu Ilker, Tsarev Roman Yurievich" which have to be corrected to read as "Denis Vladimirovich Gruzenkin, Galina Viktorovna Grishina, Mustafa Seçkin Durmuş, Ilker Üstoğlu, Roman Yurievich Tsarev" in Chapter 16, respectively.

The erratum book has been updated with the changes.

The updated online version for these chapters can be found at:
http://dx.doi.org/10.1007/978-3-319-57141-6_11
http://dx.doi.org/10.1007/978-3-319-57141-6_16

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6_53

# Author Index

R. Silhavy et al. (eds.), *Software Engineering Trends and Techniques in Intelligent Systems*,
Advances in Intelligent Systems and Computing 575, DOI 10.1007/978-3-319-57141-6