

Tomas Skersys
Rimantas Butleris
Rita Butkiene (Eds.)

Communications in Computer and Information Science

403

Information and Software Technologies

19th International Conference, ICIST 2013
Kaunas, Lithuania, October 2013
Proceedings



Springer

Editorial Board

Simone Diniz Junqueira Barbosa

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),
Rio de Janeiro, Brazil*

Phoebe Chen

La Trobe University, Melbourne, Australia

Alfredo Cuzzocrea

ICAR-CNR and University of Calabria, Italy

Xiaoyong Du

Renmin University of China, Beijing, China

Joaquim Filipe

Polytechnic Institute of Setúbal, Portugal

Orhun Kara

TÜBİTAK BİLGEM and Middle East Technical University, Turkey

Igor Kotenko

*St. Petersburg Institute for Informatics and Automation
of the Russian Academy of Sciences, Russia*

Krishna M. Sivalingam

Indian Institute of Technology Madras, India

Dominik Ślęzak

University of Warsaw and Infobright, Poland

Takashi Washio

Osaka University, Japan

Xiaokang Yang

Shanghai Jiao Tong University, China

Tomas Skersys Rimantas Butleris
Rita Butkiene (Eds.)

Information and Software Technologies

19th International Conference, ICIST 2013
Kaunas, Lithuania, October 10-11, 2013
Proceedings



Springer

المنارة للاستشارات

Volume Editors

Tomas Skersys
Rimantas Butleris
Rita Butkiene

Kaunas University of Technology
Studentu g. 50-313a
51368 Kaunas, Lithuania
E-mail: {tomas.skersys; rimantas.butleris; rita.butkiene}@ktu.lt

ISSN 1865-0929

e-ISSN 1865-0937

ISBN 978-3-642-41946-1

e-ISBN 978-3-642-41947-8

DOI 10.1007/978-3-642-41947-8

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: Applied for

CR Subject Classification (1998): D.2, H.4, H.3, H.2.8, I.2, J.1, K.3, G.1, F.2

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

المنارة للاستشارات

Preface

We are proud to introduce the selection of papers presented during the International Conference on Information and Software Technologies – ICIST 2013. The event formerly known as the IT Conference is now in its late teens – this year marked the 19th iteration of this annual academic meeting.

Organized by the biggest technical university in the Baltic States – Kaunas University of Technology – the ICIST conference went international in 2008. Last year, we were proud to have the proceedings of the conference published by Springer for the first time. This issue of *Communications in Computer and Information Science* series is the result of our continuing cooperation. Combined with the dedication of the conference Program Committee, this fact reflects the relevance and quality of the included papers.

The original mission of the conference remains intact: to bring together practitioners and researchers aiming at the everlasting convergence between business, software, and system requirements as well as the application of new supporting technologies. However, being a rather compact event, we felt the need to focus our attention on accentuating the strengths of our contributors and Program Committee members. Therefore, the topics this year were restructured to encourage the submission of papers in the fields of information systems, business intelligence, software engineering, and IT applications. These topics have already become integral to societies at almost any level, forcing researchers to develop interdisciplinary approaches and to employ multidisciplinary ways of thinking. On the other hand, it is increasingly apparent that academia no longer holds the monopoly of scientific innovation. The need to bring scientists and practitioners together is as pressing as ever. This is exactly why all participants of the conference are encouraged to stay in Kaunas for one more day after the conference attending the full-length advanced industrial tutorials on software development practices by well-known practitioners. The event was co-located with ICIST for the fourth year in a row.

There were more than 60 submissions this year, and 34 were selected for this publication. The papers were reviewed by the Program Committee consisting of 72 professionals (supported by the team of additional reviewers) representing more than 40 academic institutions and four companies from 26 countries. Each submission was reviewed following a double-blind process by at least two reviewers, while borderline papers were evaluated by three or more reviewers.

Finally, we would like to express our gratitude to the Lithuanian State Science and Studies Foundation and Faculty of Informatics of Kaunas University of Technology whose support has driven the conference, making this event and this book possible

September 2013

Tomas Skersys
Rimantas Butleris
Rita Butkiene

Organization

The 19th International Conference on Information and Software Technologies (ICIST 2013) was organized by Kaunas University of Technology and took place in Kaunas, Lithuania (October 10–11, 2013).

General Chair

Rimantas Butleris Kaunas University of Technology, Lithuania

Local Organizing Committee

Rita Butkiene (Chair)	Kaunas University of Technology, Lithuania
Tomas Skersys (Co-chair)	Kaunas University of Technology, Lithuania
Darius Silingas (Industrial Tutorials Chair)	No Magic Europe, JSC, Lithuania
Lina Nemuraite	Kaunas University of Technology, Lithuania
Gintare Bernotaityte	Kaunas University of Technology, Lithuania
Tomas Danikauskas	Kaunas University of Technology, Lithuania
Kestutis Kapocius	Kaunas University of Technology, Lithuania
Jonas Ceponis	Kaunas University of Technology, Lithuania
Mikas Binkis	Kaunas University of Technology, Lithuania
Gytis Vilutis	Kaunas University of Technology, Lithuania

Program Committee

Jan Aidemark	Vaxio University, Sweden
Vassil Alexandrov	University of Reading, UK
Eduard Babkin	Higher School of Economics, Russia
Marko Bajec	University of Ljubljana, Slovenia
Rimantas Barauskas	Kaunas University of Technology, Lithuania
Eduardas Bareisa	Kaunas University of Technology
Ana Šaša Bastinos	University of Ljubljana, Slovenia
Joerg Becker	University of Münster, Germany
J.A. Rodrigue Blais	University of Calgary, Canada
Bernd Blobel	University of Regensburg, Germany
Albertas Caplinskas	Vilnius University, Lithuania
Sven Carlsson	Lund University, Sweden
Joanna Chimiak-Opoka	University of Innsbruck, Austria
Robertas Damasevicius	Kaunas University of Technology, Lithuania
Vitalij Denisov	Klaipeda University, Lithuania

VIII Organization

Kiss Ferenc	Budapest University of Technology and Economics, Hungary
Hamido Fujita	Iwate Prefectural University, Japan
Anna Grabowska	PRO-MED Co. Ltd., Poland
Saulius Gudas	Vilnius University, Lithuania
Remigijus Gustas	Karlstad University, Sweden
Vladimir Hahanov	Kharkov National University of Radioelectronics, Ukraine
Mirjana Ivanovic	University of Novi Sad, Serbia
Alvydas Jaliniauskas	Harland Clarke Digital, USA
Raimundas Jasinevicius	Kaunas University of Technology, Lithuania
Andras Javor	McLeod Institute of Simulation Sciences, Hungary
Hai Jin	Huazhong University of Science and Technology, China
Vacius Jusas	Kaunas University of Technology, Lithuania
Egidijus Kazanavicius	Kaunas University of Technology, Lithuania
Marite Kirikova	Riga Technical University, Latvia
Jerzy Korczak	Wroclaw University of Economics, Poland
Tsvetanka Kovacheva	Technical University Varna, Bulgaria
Dieter Kranzlmuller	Ludwig-Maximilian University of Munich, Germany
Dejan Lavbic	University of Ljubljana, Slovenia
Rob Mark	Queens's University Belfast, UK
Raimundas Matulevicius	University of Tartu, Estonia
Arturas Mazeika	Max-Planck-Institut für Informatik, Germany
Pramod Kumar Meher	Institute for Infocomm Research, Singapore
Lina Nemuraite	Kaunas University of Technology, Lithuania
Dušica Novakovic	London Metropolitan University, UK
Jyrki Nummenmaa	University of Tampere, Finland
Toshio Okamoto	University of Electro-Communications, Japan
Stephan Olariu	Old Dominion University, USA
Tero Paivarinta	Agder University, Norway
Algirdas Pakstas	London Metropolitan University, UK
Marcin Paprzycki	Systems Research Institute, Polish Academy of Science, Poland
Michael Petit	University of Namur, Belgium
Henrikas Pranevicius	Kaunas University of Technology, Lithuania
Abhijit Ray	ST Electronics (Training & Simulation Systems) Pte. Ltd., Singapore
Dalius Rubliauskas	Kaunas University of Technology, Lithuania
Rok Rupnik	University of Ljubljana, Slovenia
Giedre Sabaliauskaite	Singapore University of Technology and Design
Marco Sajevo	Palermo University, Italy

Rimantas Seinauskas	Kaunas University of Technology, Lithuania
Darius Silingas	No Magic Europe, Lithuania
Kulwinder Singh	University of Calgary, Canada
Ilmars Slaidins	Riga Technical University, Latvia
Janis Stirna	Stockholm University, Sweden
Darijus Straszunas	Norwegian University of Science and Technology, Norway
Giancarlo Succi	Free University of Bozen-Bolzano, Italy
Aleksandras Targamadze	Kaunas University of Technology, Lithuania
Laimutis Telksnys	Vilnius University, Lithuania
Peter Thanisch	University of Tampere, Finland
Babis Theodoulidis	University of Manchester, UK
Sofia Tsekeridou	Athens Information Technology, Greece
Raimund Ubar	Tallinn Technical University, Estonia
Olegas Vasilecas	Vilnius Gediminas Technical University, Lithuania
Radu Adrian Vasii	Politehnica University of Timisoara, Romania
Damjan Vavpotic	University of Ljubljana, Slovenia
Benkt Wangler	University of Skovde, Sweden
Stanislaw Wrycza	University of Gdansk, Poland
Zheyang Zhang	University of Tampere, Finland
Antanas Zilinskas	Vilnius University, Lithuania

Additional Reviewers

Linas Ablonskis	Kaunas University of Technology, Lithuania
Tomas Blazauskas	Kaunas University of Technology, Lithuania
Rita Butkiene	Kaunas University of Technology, Lithuania
Lina Ceponiene	Kaunas University of Technology, Lithuania
Kestutis Driaunys	Vilnius University, Lithuania
Kestutis Kapocius	Kaunas University of Technology, Lithuania
Antanas Lenkevicius	Kaunas University of Technology, Lithuania
Virginija Limanauskiene	Kaunas University of Technology, Lithuania
Audrius Lopata	Vilnius University, Lithuania
Antanas Mikuckas	Kaunas University of Technology, Lithuania
Gytenis Mikulenais	Kaunas University of Technology, Lithuania
Alfonsas Misevicius	Kaunas University of Technology, Lithuania

Co-editors

Tomas Skersys	Kaunas University of Technology, Lithuania
Rimantas Butleris	Kaunas University of Technology, Lithuania
Rita Butkiene	Kaunas University of Technology, Lithuania

Table of Contents

Information Systems

Analysis of Control System with Delay Using the Lambert Function	1
<i>Irma Ivanovienė and Jonas Rimas</i>	
The Quality Management Metamodel in the Enterprise Architecture . . .	11
<i>Jerzy Roszkowski and Agata Roszkowska</i>	
Ontology Matching Using TF/IDF Measure with Synonym Recognition	22
<i>Marko Gulić, Ivan Magdalenić, and Boris Vrdoljak</i>	
Moving Averages for Financial Data Smoothing	34
<i>Aistis Raudys, Vaidotas Lenčiauskas, and Edmundas Malčius</i>	
Knowledge Transfer in Management Support System Implementation . . .	46
<i>Bartosz Wachnik</i>	
Collective Intelligence Utilization Method Based on Implicit Social Network Composition and Evolution in the Scope of Personal Learning Environment	57
<i>Genadijus Kulvietis, Andrej Afonin, and Danguole Rutkauskiene</i>	
Automation of Upgrade Process for Enterprise Resource Planning Systems	70
<i>Algirdas Laukaitis</i>	
Business Process Flow Verification Using Knowledge Based System	82
<i>Regina Miseviciene, Germanas Budnikas, and Dalius Makackas</i>	
Web-Based Analytical Information System for Spatial Data Processing	93
<i>Viacheslav Paramonov, Roman Fedorov, Gennagy Ruzhnikov, and Alexandr Shumilov</i>	
System Architecture Model Based on Service-Oriented Architecture Technology	102
<i>Tarkan Gurbuz, Daina Gudoniene, and Danguole Rutkauskiene</i>	
Towards the Combination of BPMN Process Models with SBVR Business Vocabularies and Rules	114
<i>Eglė Mickevičiūtė and Rimantas Butleris</i>	

Process for Applying Derived Property Based Traceability Framework in Software and Systems Development Life Cycle.....	122
<i>Saulius Pavalkis and Lina Nemuraite</i>	
Developing SBVR Vocabularies and Business Rules from OWL2 Ontologies	134
<i>Gintare Bernotaityte, Lina Nemuraite, Rita Butkiene, and Bronius Paradauskas</i>	
Exploring Key Factors of Pilot Projects in Agile Transformation Process Using a Grounded Theory Study.....	146
<i>Taghi Javdani Gandomani, Hazura Zulzalil, Abdul Azim Abd Ghani, Abu Bakar Md. Sultan, and Khaironi Yatim Sharif</i>	
Incompleteness in Conceptual Data Modelling	159
<i>Peter Thanisch, Tapio Niemi, Jyrki Nummenmaa, Zheyong Zhang, Marko Niinimäki, and Pertti Saariluoma</i>	
Semi-supervised Learning of Action Ontology from Domain-Specific Corpora	173
<i>Irena Markievicz, Daiva Vitkute-Adzgauskiene, and Minija Tamosiunaite</i>	
Business Intelligence for Information and Software Systems	
Speech Keyword Spotting with Rule Based Segmentation	186
<i>Mindaugas Greibus and Laimutis Telksnys</i>	
Business Intelligence Maturity Models: Information Management Perspective.....	198
<i>Alaskar Thamir and Babis Theodoulidis</i>	
Modified Stochastic Algorithm for Mining Frequent Subsequences	222
<i>Loreta Savulioniene and Leonidas Sakalauskas</i>	
On Two Approaches to Constructing Optimal Algorithms for Multi-objective Optimization	236
<i>Antanas Žilinskas</i>	
Recognition of Voice Commands Using Hybrid Approach.....	249
<i>Vytautas Rudžionis, Kastytis Ratkevičius, Algimantas Rudžionis, Gailius Raškinis, and Rytis Maskeliūnas</i>	
Estimation of the Environmental Impact on the Accuracy of Signal Recognition	261
<i>Gintarė Čeidaitė and Laimutis Telksnys</i>	

Software Engineering

Automated Method for Software Integration Testing Based on UML Behavioral Models	272
<i>Dominykas Barisas, Eduardas Bareiša, and Šarūnas Packedvičius</i>	
Computational Algorithmic Generation of High-Quality Colour Patterns	285
<i>Alfonsas Misevičius, Evaldas Guogis, and Evelina Stanevičienė</i>	
Design of Visual Language Syntax for Robot Programming Domain	297
<i>Ignas Plauska and Robertas Damaševičius</i>	
Testing Stochastic Systems Using MoVoS Tool: Case Studies	310
<i>Kenza Bouaroudj, Djamel-Eddine Saidouni, and Ilham Kitouni</i>	
Two Scale Modeling of Heterogeneous Solid Body by Use of Thick Shell Finite Elements	322
<i>Dalia Čalnerytė and Rimantas Barauskas</i>	
Development in Authentication of AODV Protocols to Resist the Attacks	334
<i>Ahmad Alomari</i>	
Evaluation of Open Source Server-Side XSS Protection Solutions	345
<i>Jonas Ceponis, Lina Ceponiene, Algimantas Venckauskas, and Dainius Mockus</i>	
Minimization of Numerical Dispersion Errors in Finite Element Models of Non-homogeneous Waveguides	357
<i>Andrius Krisciunas and Rimantas Barauskas</i>	
Novel Method to Generate Tests for VHDL	365
<i>Vacius Jusas and Tomas Neverdauskas</i>	
Empirical Analysis of the Test Maturity Model Integration (TMMi)	376
<i>Kerli Rungi and Raimundas Matulevičius</i>	
Behavior Analysis of Real-Time Systems Using PLA Method	392
<i>Dalius Makackas, Regina Miseviciene, and Henrikas Pranevicius</i>	
Measuring the Performance of Process Synchronization with the Model-Driven Approach	403
<i>Vladislav Nazaruk and Pavel Rusakov</i>	
Author Index	415

Analysis of Control System with Delay Using the Lambert Function

Irma Ivanovienė and Jonas Rimas

Department of Applied Mathematics, Kaunas University of Technology,
Studentu 50, LT-51368, Kaunas, Lithuania
irma.ivanoviene@yahoo.com, jonas.rimas@ktu.lt

Abstract. The mathematical model of the mutual synchronization system, having ring form structure and composed of n ($n \in N$) oscillators, is investigated. The mathematical model of the system is the matrix differential equation with delayed argument. The solution of the matrix differential equation with delayed argument is obtained applying the Lambert function method. Using obtained solution, the transients in the system are examined. The results of calculations, received by the Lambert function method, are compared with the results, obtained by the exact method of consequent integration.

Keywords: synchronization system, differential equations, delayed arguments, Lambert function.

1 Introduction

The control systems find application in various engineering equipments including the networks of transmitting and distributing of the information. Usually control systems are being investigated applying their mathematical models. More exact analysis of systems demands the use of the more complicated mathematical models. Often the delays of the signals, transferred along the control system, must be included into these models. The delays make the investigation of the model more complicated. Within last decade many works in which analytical investigation of systems with delays is executed by a method based on the use of Lambert functions are published (see [1], [7] and their references). In all these works practical application of this method is limited to systems of differential equations with delayed argument the order of which not exceeds three. In the given work we apply the Lambert function method to investigate the synchronization system described by linear systems of differential equations up to fifteenth order. Relative errors of the received results are estimated using the solutions obtained by the exact method of consequent integration (method of “steps”) [2].

2 Formulation of the Problem

Consider the multidimensional control system with delays and with ring form structure, which mathematical model is the matrix differential equation with delayed argument [3], [9], [10], [11]

$$\begin{aligned} x'(t) + B_1 x(t) + B_2 x(t - \tau) &= z(t), \\ x(t) &= \phi(t), \quad t \in [-\tau, 0]; \end{aligned} \quad (1)$$

here $x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$ is the desired vector function, T (here and in what follows) denotes the operation of transposition, τ is a constant time delay, $\phi(t)$ is a vector valued preshape (initial) function, $z(t)$ is a free term (continuous function depending on the initial conditions), κ is a coefficient, B_1 and B_2 are $n \times n$ ($n \in \mathbb{N}$) numerical matrices ($B_1, B_2 \in R^{n \times n}$),

$$B_1 = \kappa I, \quad (2)$$

$$B_2 = \frac{\kappa}{2} B, \quad (3)$$

$$B = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 1 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & 0 & 1 \\ 1 & 0 & \dots & 0 & 1 & 0 \end{pmatrix} \quad (4)$$

($I \in R^{n \times n}$ is the identity matrix, matrix $B \in R^{n \times n}$ outlines the structure of the internal links of the system). As an example of a control system, described by the equation (1), the mutual synchronization system of the communication network, composed of n oscillators and having ring form structure, can be pointed out [10] (Fig. 1). In this case the symbol $x_i(t)$ in (1) stands for the normalized phase (that is the phase divided by 2π) of the i -th oscillator. In what follows, the normalized phase we will name, simply, phase. We shall assume, that

$$x_i(t) = \begin{cases} \int_0^t f_i(\xi) d\xi + x_{0i}, & \text{if } t > 0, \\ f_{0i} t + x_{0i}, & \text{if } t \leq 0; \end{cases} \quad (5)$$

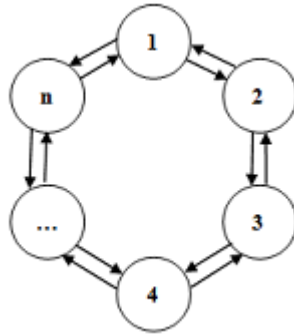


Fig. 1. The scheme of internal links of the system

here $x_{0i} = x_i(0)$ ($i = \overline{1, n}$) is the initial phase of the oscillations of i -th oscillator, $f_i(t)$ is the frequency of the i -th oscillator, f_{0i} is the own frequency of the i -th oscillator (the frequency of the i -th oscillator, when the control signal is disconnected). The meaning of (5) is following: the control signal to the i -th oscillator at time $t = 0$ is connected up. Before this time moment the i -th oscillator works with its own frequency f_{0i} . Taking into account (5), we get following expressions for the initial vector function $\phi(t)$ and the free vector $z(t)$ of (1):

$$\phi(t) = (\phi_1(t), \phi_2(t), \dots, \phi_n(t))^T, \tag{6}$$

$$\phi_i(t) = f_{0i}t + x_{0i}, \quad i = \overline{1, n}, \quad t \in [-\tau, 0], \tag{7}$$

$$z(t) = (f_{01}, f_{02}, \dots, f_{0n})^T. \tag{8}$$

3 Solution of the Matrix Differential Equation with Delayed Argument

Applying the Lambert function method (see [1]), the solution of (1) on the interval $[0, +\infty)$ can be expressed as follows:

$$\begin{aligned} x(t) &= \sum_{k=-\infty}^{\infty} e^{S_k t} C_k + \int_0^t \sum_{k=-\infty}^{\infty} e^{S_k(t-\xi)} C'_k z(\xi) d\xi = \\ &= \lim_{N \rightarrow \infty} \sum_{k=-N}^N e^{S_k t} C_k + \int_0^t \left(\lim_{N \rightarrow \infty} \sum_{k=-N}^N e^{S_k(t-\xi)} C'_k z(\xi) \right) d\xi; \end{aligned} \tag{9}$$

here C_k is a $n \times 1$ coefficient matrix - column computed from the given preshape function $x(t) = \phi(t)$, $t \in [-\tau, 0]$, which is an initial state of delay differential equation (1), C'_k is a $n \times n$ coefficient matrix computed from the given free term $z(t)$ of the matrix differential equation (1) (procedures of calculation of these matrices are explained in [1]). In the computations of the solution $x(t)$ we shall use the approximate expression, obtained from (9) at fixed and finite N .

4 Comparing the Lambert Function Method with the Method of Consequent Integration (Method of “Steps”)

The solution of inhomogeneous matrix delay differential equation (1) is presented by the infinite functional series (see (9)), which determines the exact solution. In the real calculations we apply the approximate formula, obtained from (9) with finite N ($2N + 1$ indicates the number of branches of the Lambert function, which are used in calculations of the solution). We shall investigate the rate of convergence of the approximate solution to the exact solution with increasing N . For this purpose we shall apply the exact solution, obtained by the method of

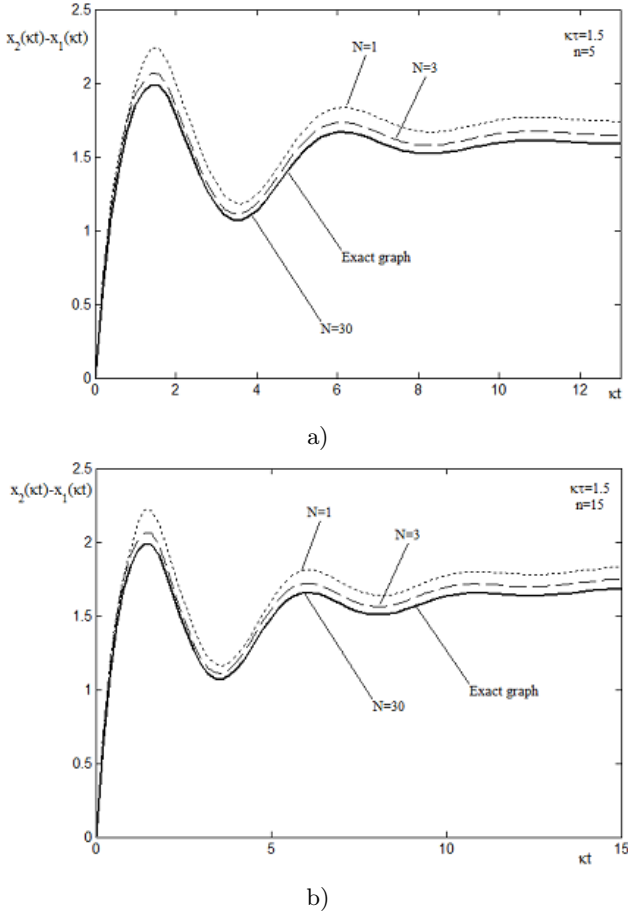
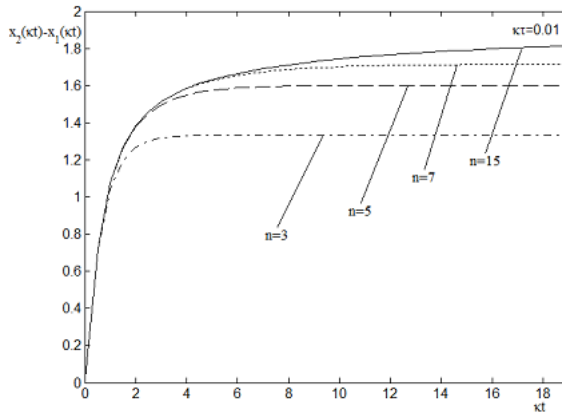
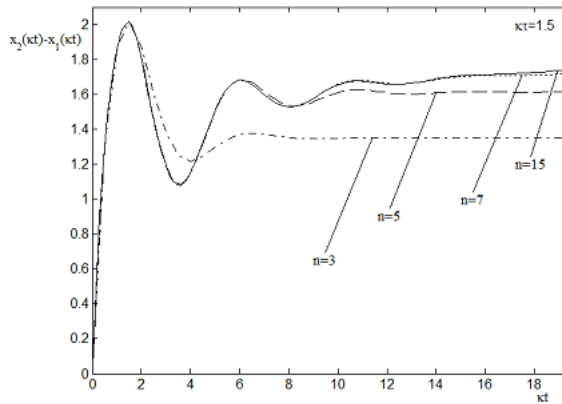


Fig. 2. Graphs of the step responses $h_{i1}(\kappa t)$ at different values of N

consequent integration (see A.7 in the Appendix A). The phase difference $x_2(t) - x_1(t)$ of mutual synchronization system with a structure of a ring, computed by the method of consequent integration (the exact method) and by Lambert function method with different values of N , are presented in the Fig. 2. As we see from this figure, increasing N the approximate solution approaches the exact solution. For all calculations we choose initial phases x_{0i} ($i = \overline{1, n}$) equal to 0.3, the ratio $\frac{f_{01}}{\kappa}$ equal to 2001 and $\frac{f_{0i}}{\kappa}$ ($i = \overline{2, n}$) to 1999. The maximal relative errors δ_{\max} obtained for $\kappa t \in [0, +\infty)$ using different values of N are presented in the Table 1, when $n = 5$. As we see from the table, for $N = 50$ the maximal relative error is not greater than 0.01 (with increase of N the maximal relative error decreases). Dependence from n of this relative error is insignificant. Such accuracy is sufficient for practical applications.



a)



b)

Fig. 3. Graphs of the phase difference $x_2(t) - x_1(t)$

Table 1.

N	1	3	30	50
δ_{\max}	0.1359	0.0678	0.0068	0.00381

5 Results of Calculations

The transients in the synchronization system were investigated applying derived formulas. Some results of calculations are presented in Fig. 3, 4 as graphs of step responses. For the calculation of the phase differences $x_i(t) - x_j(t)$ we have applied the approximate formula, obtained from (9) with $N = 50$ (this means that we have used 101 branches of the Lambert function in the computations).

With such N the relative error is not greater than 0.01 for any κt on the base

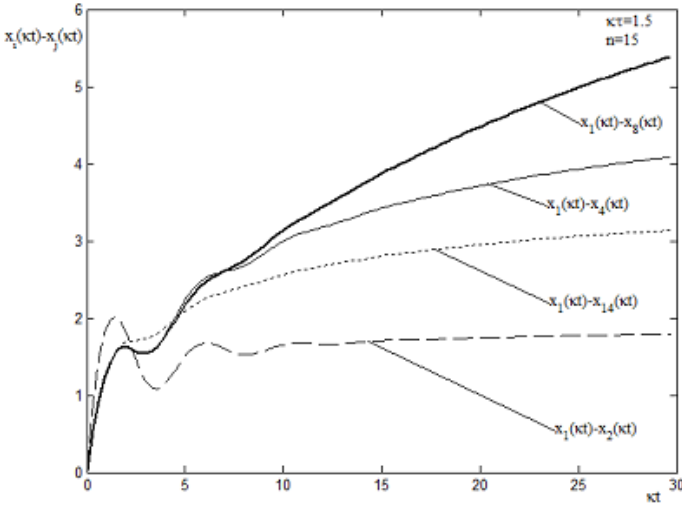


Fig. 4. Graphs of the phase differences $x_i(t) - x_j(t)$

of the 4-th section. So the graphs of the step responses, presented below, are sufficiently accurate (in the presented figures these graphs practically coincide with the exact ones). In Fig. 3 the graphs of the phase difference $x_2(t) - x_1(t)$ are given at different values of $\kappa\tau$ and for different numbers of oscillators in the synchronization system. From the figure we see that the duration of transients in the synchronization system depends on number n of oscillators in the synchronization system and on magnitude of the product $\kappa\tau$. With increase of n and $\kappa\tau$ the duration of transients in the system tend to increase. The transients get oscillatory features if $\kappa\tau > 1.5$. In Fig. 4 the graphs of the phase differences $x_i(t) - x_j(t)$ between oscillations of different oscillators of the synchronization system are presented. From the figure we see that duration of transients of the phase difference $x_i(t) - x_j(t)$ depends on the distance between i -th and j -th oscillators in the ring structure.

6 Conclusions

1. The Lambert function method is used for computing transients for the synchronization system. It is shown that using 101 branches of the Lambert function (taking $N = 50$) in calculations of phase differences $x_i(t) - x_j(t)$ the relative error is not greater than 0.01 for any $\kappa\tau$ and practically does not depend on the number of oscillators in the system.
2. The Lambert function method has the advantage in comparison with a method of consequent integration (method of “steps”), as time of calculation of the step response by this method does not depend on delay size,

whereas time of calculation of the step response by means of a method of consequent integration is in inverse proportion to the delay size.

3. The method of research of dynamics, used in the presented work, can also be applied to other control systems, described by the linear matrix differential equations with delayed arguments and with commuting coefficient matrices.

References

1. Asl, F.M., Ulsoy, A.G.: Analytical solution of a system of homogenous delay differential equations via the Lambert function. In: Proceedings of the American Control Conference, Chicago, IL, pp. 2496–2500 (2000)
2. Bellman, R., Cooke, K.L.: Differential-difference equation. Academic Press, New York (1963)
3. Bregni, S.A.: Historical perspective on telecommunications network synchronization. IEEE Communications Magazine, 158–166 (1998)
4. Corless, R.M., Gonnet, G.H., Hare, D.E., Jeffrey, D.J., Knuth, D.: On the Lambert W function. Advances in Computation Mathematics, 329–359 (1996)
5. Horn, R., Johnson, C.: Matrix Analysis. Cambridge University Press (1995)
6. Jarlebring, E., Dammb, T.: The Lambert W function and the spectrum of some multidimensional time-delay systems. Automatica 43(12), 2124–2128 (2007)
7. Yi, S., Nelson, P.W., Ulsoy, A.G.: Time-delay systems (analysis and control using the Lambert W function). World Scientific, New Jersey (2010)
8. Yi, S., Ulsoy, A.G.: Solution of a system of linear delay differential equation using the matrix Lambert function. In: Proceedings of the American Control Conference, Minneapolis, MN, pp. 2433–2438 (2006)
9. Kuo, B.C.: Automatic control systems. Prentice-Hall (1995)
10. Rimas, J.: Investigation of the dynamics of mutually synchronized systems. Telecommunications and Radio Eng. 32, 68–79 (1977)
11. Rimas, J.: Analysis of multidimensional delay system. In: MTNS 2004: Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems, Leuven, Belgium, pp. 1–5 (2004)
12. Rimas, J.: On computing of arbitrary positive integer powers for one type of odd order symmetric circulant matrices-I. Applied Mathematics and Computation 165, 137–141 (2005)

A Appendix: Finding Exact Expression of Phase Differences by Method of Consequent Integration (Method of “Steps”)

The exact solution of (1) on the interval $[0, (L+1)\tau]$, $L = 0, 1, 2 \dots$ using method of consequent integration can be expressed as follows (see, for example, [10])

$$\begin{aligned}
 x(t) &\doteq \sum_{k=0}^L \left(\frac{\kappa}{2}\right)^k \frac{e^{-p\kappa\tau}}{(p+\kappa)^{k+1}} B^k Z(p) = \\
 &= \frac{1}{(p+\kappa)} Z(p) + \sum_{k=1}^L \left(\frac{\kappa}{2}\right)^k \frac{e^{-p\kappa\tau}}{(p+\kappa)^{k+1}} B^k Z(p), \quad 0 \leq t < (L+1)\tau;
 \end{aligned}
 \tag{A.1}$$

here $x(t) = (x_1(t), x_2(t), \dots, x_n(t))^T$ is the desired vector function (components of this vector function are phases of oscillations of the oscillators), T denotes the operation of transposition, \doteq is a sign linking the function with its Laplace transform, κ is a coefficient, τ is a delay, $B \in R^{n \times n}$ is matrix defined by (4), p is a complex variable, $z(t) \doteq Z(p)$, $Z(p) = (Z_1(p), Z_2(p), \dots, Z_n(p))^T$ is function depending on initial conditions (we assume that initial conditions are defined by (5)). Using (A.1) we get

$$x_i(t) - x_j(t) \doteq \frac{1}{(p + \kappa)} (Z_i(p) - Z_j(p)) + \sum_{k=1}^L \left(\frac{\kappa}{2}\right)^k \frac{e^{-p\kappa\tau}}{(p + \kappa)^{k+1}} ([B^k Z(p)]_i - [B^k Z(p)]_j), \quad 0 \leq t < (L + 1)\tau; \quad (\text{A.2})$$

here $[]_k$ is the k -th component of the vector $[]$. From the initial conditions (see 5) follows:

$$Z_i(p) = \frac{f_{0i}}{p} + \varphi_{0i} + \kappa \left(\frac{f_{i-1,i+1}}{p^2} - \frac{f_{i-1,i+1}\tau}{p} + \frac{\varphi_{i-1,i+1}}{p} - \frac{f_{i-1,i+1}}{p^2} e^{-p\tau} - \frac{\varphi_{i-1,i+1}}{p} e^{-p\tau} \right), \quad i = \overline{1, n}; \quad (\text{A.3})$$

here $\varphi_{i-1,i+1} = \frac{\varphi_{0i-1} + \varphi_{0i+1}}{2}$, $\varphi_{0,2} = \frac{\varphi_{0n} + \varphi_{02}}{2}$, $\varphi_{n-1,n+1} = \frac{\varphi_{0n-1} + \varphi_{01}}{2}$, $f_{i-1,i+1} = \frac{f_{0i-1} + f_{0i+1}}{2}$, $f_{0,2} = \frac{f_{0n} + f_{02}}{2}$, $f_{n-1,n+1} = \frac{f_{0n-1} + f_{01}}{2}$, φ_{0i} and f_{0i} are the initial phase and the own frequency of the i -th oscillator ($i = \overline{1, n}$). Let $I \in R^{n \times n}$ is the identity matrix. Then we have

$$Z_i(p) = [Z(p)]_i = [IZ(p)]_i = \sum_{k=1}^n [I]_{ik} [Z(p)]_k; \quad (\text{A.4})$$

here $[]_{ik}$ is the ik -th component of the matrix $[]$. Using (A.4), we write

$$Z_i(p) - Z_j(p) = [IZ(p)]_i - [IZ(p)]_j = \sum_{k=1}^n ([I]_{ik} - [I]_{jk}) [Z(p)]_k = \sum_{k=1}^n Z_k(p) (\delta_{ik} - \delta_{jk}); \quad (\text{A.5})$$

here

$$\delta_{ik} = [I]_{ik} = \begin{cases} 1, & i = j, \\ 0, & i \neq j \end{cases} \quad (\text{A.6})$$

is the Kronecker delta. Applying the inverse Laplace transform to right hand side of (A2), we obtain the exact expression of phase differences

$$x_i(t) - x_j(t) = \sigma_{ij}(t) + S_{ij}(t), \quad 0 \leq t < (L + 1)\tau, \quad i, j = \overline{1, n}; \quad (\text{A.7})$$

here

$$\sigma_{ij}(t) = \sigma_{ij,1}(t) + \sigma_{ij,2}(t) + \sigma_{ij,3}(t) - \sigma_{ij,3}(t - \tau) - \sigma_{ij,4}(t - \tau),$$

$$\sigma_{ij,1}(t) = \alpha_{ij,1} e^{-\kappa t} 1(t),$$

$$\sigma_{ij,2}(t) = \beta_{ij}(t) (1 - e^{-\kappa t}) 1(t),$$

$$\sigma_{ij,3}(t) = \alpha_{ij,4} (\kappa t - 1 + e^{-\kappa t}) 1(t),$$

$$\sigma_{ij,4}(t) = \alpha_{ij,3} (1 - e^{-\kappa t}) 1(t),$$

$$\alpha_{ij,1} = \sum_{m=1}^n \varphi_{0m} (\delta_{im} - \delta_{jm}),$$

$$\alpha_{ij,2} = \sum_{m=1}^n \frac{f_{0m}}{\kappa} (\delta_{im} - \delta_{jm}),$$

$$\alpha_{ij,3} = \sum_{m=1}^n \varphi_{m-1,m+1} (\delta_{im} - \delta_{jm}),$$

$$\alpha_{ij,4} = \sum_{m=1}^n \frac{f_{m-1,m+1}}{\kappa} (\delta_{im} - \delta_{jm}),$$

$$\beta_{ij}(t) = \alpha_{ij,2} + \alpha_{ij,3} - \kappa t \alpha_{ij,4},$$

$$S_{ij}(t) = S_{ij,1}(t) + S_{ij,2}(t) + S_{ij,3}(t) - S_{ij,3}(t - \tau) - S_{ij,4}(t - \tau),$$

$$S_{ij,1}(t) = \sum_{k=1}^L \frac{1}{2^k} a_{ij,1}(k) \frac{\kappa^k (t - k\tau)^k}{k!} e^{-\kappa(t-k\tau)} 1(t - k\tau),$$

$$S_{ij,2}(t) = \sum_{k=1}^L \frac{1}{2^k} b_{ij}(t, k) \left(1 - \sum_{r=0}^k \frac{\kappa^r (t - k\tau)^r}{r!} e^{-\kappa(t-k\tau)} \right) 1(t - k\tau),$$

$$S_{ij,3}(t) = \sum_{k=1}^L \frac{1}{2^k} a_{ij,4}(k) \left(\kappa(t - k\tau) - (k+1) + \sum_{r=0}^k \sum_{s=0}^r \frac{\kappa^s (t - k\tau)^s}{s!} e^{-\kappa(t-k\tau)} \right) 1(t - k\tau),$$

$$S_{ij,4}(t) = \sum_{k=1}^L \frac{1}{2^k} a_{ij,3}(k) \left(1 - \sum_{r=0}^k \frac{\kappa^r (t - k\tau)^r}{r!} e^{-\kappa(t-k\tau)} \right) 1(t - k\tau),$$

$$a_{ij,1}(k) = \sum_{m=1}^n \varphi_{0m} (d_{im}(k) - d_{jm}(k)),$$

$$a_{ij,2}(k) = \sum_{m=1}^n \frac{f_{0m}}{\kappa} (d_{im}(k) - d_{jm}(k)),$$

$$a_{ij,3}(k) = \sum_{m=1}^n \varphi_{m-1,m+1} (d_{im}(k) - d_{jm}(k)),$$

$$a_{ij,4}(k) = \sum_{m=1}^n \frac{f_{m-1,m+1}}{\kappa} (d_{im}(k) - d_{jm}(k)),$$

$$b_{ij}(t, k) = a_{ij,2}(k) + a_{ij,3}(k) - \kappa t a_{ij,4}(k),$$

$$d_{ij}(k) = [B^k]_{ij}, \quad i, j = \overline{1, n}.$$

Expressions of $d_{ij}(k)$ for odd n are found in [12]. Taking into account that B and B^k are n -th order circulant matrices, we have [12]

$$B^k = (d_{ij}(k)) = \text{circ}_n \left(d_{11}(k), d_{12}(k), \dots, d_{1 \frac{n+1}{2}}(k), d_{1 \frac{n+1}{2}}(k), \dots, d_{12}(k) \right),$$

$$d_{1j}(k) = \frac{1}{n} \sum_{m=1}^{\frac{n+1}{2}} l_{n-2m+2} \lambda_{n-2m+2}^k T_{j-1} \left(\frac{\lambda_{n-2m+2}}{2} \right), \quad j = 1, \overline{\frac{n+1}{2}},$$

$$d_{1j}(k) = d_{1 \overline{n-j+2}}(k), \quad j = \overline{\frac{n+3}{2}, n};$$

here

$$l_s = \begin{cases} 1, & s = n, \\ 2, & s \neq n. \end{cases}$$

$\lambda_s = -2 \cos\left(\frac{s\pi}{n}\right)$, $s = 1, 3, 5, \dots, n$ are the eigenvalues of the matrix B , $T_i(x)$ is the i -th degree Chebyshev polynomial of the first kind ($T_i(x) = \cos(i \arccos(x))$, $-1 \leq x \leq 1$).

The Quality Management Metamodel in the Enterprise Architecture

Jerzy Roszkowski¹ and Agata Roszkowska²

¹ Management Systems Consulting, Poznańska 28/1 Street, 93-134 Łódź, Poland
office@mcs-roszkowski.com.pl

² Baden-Württemberg Cooperative State University Stuttgart, Faculty of Technology,
Jägerstraße 56, 70174 Stuttgart, Germany
agata.roszkowska@bshg.com

Abstract. The paper presents the methodology for determining, management, simulation and optimization of the quality of an enterprise architecture based on defined by the author of two metamodels: classes and processes for quality management of this architecture. The second of them (the process metamodel) of quality management developed in BPMN has undergone simulation and optimization using ARIS Business Process Simulator. The results of this simulation and optimization are presented in the article. The presented research method developed by the author, and the results are related to and are a creative extension of the following ISO standards: ISO/IEC 24744 [1], ISO/IEC 12207 [2], ISO / IEC 42010 [3] and the well-known methodologies: "A Framework for Information Systems Architecture". - Zachman, J. A., and TOGAF: "The Open Group Architecture Framework" [12].

Keywords: process simulation and optimization, metamodel, enterprise architecture, quality.

1 Initial Information

1.1 What Is a Quality?

The quality concept in this article for the development process of the enterprise architecture is understood as a set of measurable and immeasurable characteristics of the product required by the customer (customer, product). Product quality is monitored at all stages of its manufacture, especially in the so-called checkpoints. Control points in the architecture development are the points where we get the different stages of the manufacturing cycle of the architecture products (e.g. documentation, analysis, design,). Detailed information concerning modeling and building of quality system in the software development in the special case of business intelligence systems was published by one of the authors in [7] and [9].

1.2 Area and Purpose of Interest

The problem of architectural quality has so far been considered marginal. Elements of this issue can be found in project management (e.g., product quality methodology Prince 2). Here the quality of the architecture is considered by the quality of one of the products of the manufacturing cycle software (design of the system architecture). The problem of quality in the manufacturing cycle of IT is often regarded as testing and determining the quality of the software through various testing techniques [7]. However, the software is only the end (final) product of a development cycle. The quality and usefulness of the application decides at an earlier stage of the design of its architecture. The quality of architecture is clearly a key element for enterprise architecture [5]. Individual systems (business software) and processes, requirements, technical components that make up the architecture of the enterprise, even if they have the quality but they are not making the quality and optimal the total architecture. As a result of the experience of one of the authors in the TELCO sector can be considered general lack of techniques and methods for testing the quality and optimization of complex and integrated corporate IT environments built on the concepts of enterprise architecture [8], [10].

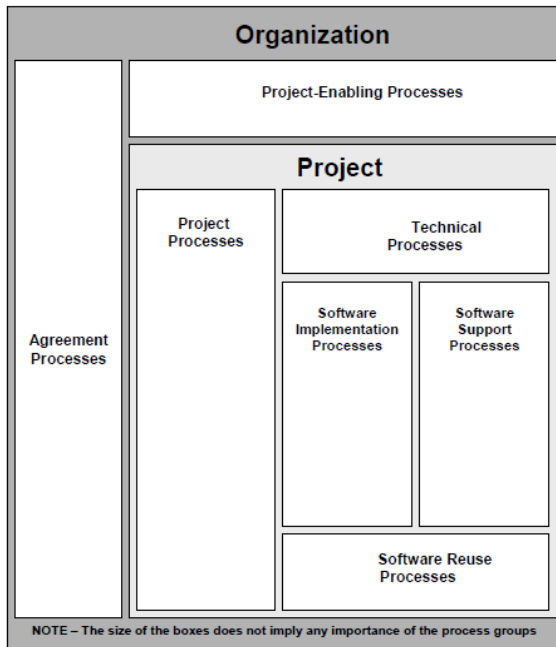


Fig. 1. Life Cycle Process Groups (Source: Systems and Software Engineering Software Life Cycle Processes - ISO/IEC 12207)

The result is growing additional costs which increase with the complexity of the IT environment in addition to the cost of projects which to introduce changes in the integrated environment. The increasing complexity of the IT environment reduces business results related to investments (individual projects) [10].

This problem does not solve the TOGAF ADM (Architecture Development Method) [12] in particular, the problem of architecture management is considered there only as change management. Is completely omitted the issue of quality management architecture. The purpose of this paper is to fill this gap in the conceptual level by introducing appropriate meta-models in terms of quality enterprise architecture. Acceptable metamodel introduced here is the first step to understanding the problem, resulting in the development of measurable methods and tools for measuring the quality of the architecture.

The International (Software Life Cycle Processes - ISO/IEC 12207) [2] groups the activities that may be performed during the life cycle of software system into eight process groups. Each of the life cycle processes within those groups is described in terms of its purpose and desired outcomes and list activities and tasks which need to be performed to achieve those outcomes. These life cycle process groups are depicted in Figure 1. The object of the study (modeling and optimization) is the area: Software Support processes. According to the standard ISO/IEC 12207 [2], this area consists of the following subprocesses:

- A. Software Documentation Management Process
- B. Software Configuration Management Process
- C. Software Quality Assurance
- D. Software Verification Process
- E. Software Validation Process
- F. Software Review Process
- G. Software Audit Process
- H. Software Problem Resolution Process

The subject of optimization the integrated model of these processes with the exception of process B and H belonging to the area of quality management.

2 Architecture Quality Management Metamodel

2.1 What Is a Metamodel ?

According to ISO standard ISO/IEC 24744 (A Metamodel for Development Methodologies) [1], a metamodel is the specification of the concepts, relationships and rules that are used to define a methodology. A *methodology* is defined as the specification of the process to follow together with the work products to be used and generated, plus the consideration of the people and tools involved, during the development effort. A methodology specifies the process to be executed, usually as a

set of related activities, tasks and/or techniques, together with what work products must be manipulated (created, used or changed) at each moment and by whom, possibly including models, documents and other inputs and outputs. In turn, specifying the models that must be dealt with implies defining the basic building blocks that should be used to construct these models. Any metamodel consists from *elements*. An *element* is a simple component of a methodology. Usually, methodology elements include the specification of what tasks, activities, techniques, models, documents, languages and/or notations can or must be used when applying the methodology. Methodology elements are related to each other, comprising a network of abstract concepts. Typical methodology elements are Capture Requirements, Write Code for Methods (kinds of tasks), Requirements Engineering, High-Level Modelling (kinds of activities), Pseudo-code, Dependency Graphs (notations), Class, Attribute (kinds of model building blocks), Class Model, Class Diagram, Requirements Specification (kind of work products), etc.

2.2 Class Metamodel for the Quality Management in the Enterprise Architecture

This model is shown below in Figure 2. Meaning of various class concepts are explained in Table 1.

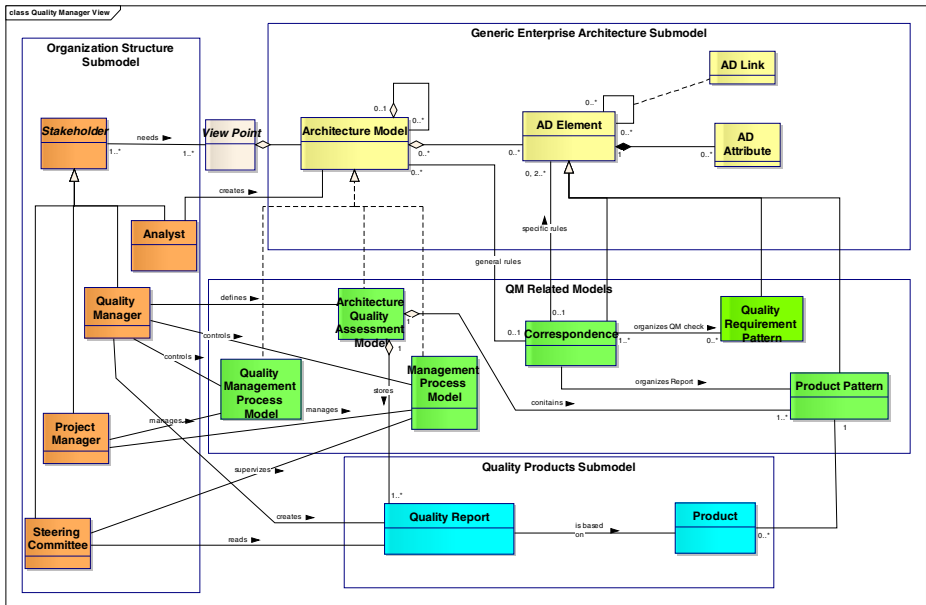


Fig. 2. Metamodel of classes for the topic “Architecture quality management”



Table 1. Definition of classes for the topic “Architecture Quality Management”

Seq. no.	Object name	Definition
Organization Structure Submodel		
1	Analyst	Person(s) responsible for the definitions of the model architecture and its components.
2	Quality manager	Person responsible for the definitions of the quality requirements and and quality control
3	Project manager	Person responsible for the quality requirements and the quality management
4	Stakeholder	Abstract project stakeholder
5	Steering Committee	Group of stakeholders responsible for the strategic management of the project.
Generic Enterprise Architecture Submodel		
6	Architecture Model	Object representing architecture model equivalent to the architecture repository.
7	AD Element	Architecture Description Element - any type of UML/BPMN Element. It builds Architecture Model.
8	AD Link	Architecture Description Link - any kind of link between AD Elements.
9	AD Attribute	Architecture Description Attribute-any type of AD Element Attribute.
QM Related Models		
10	Quality Management Process Model	BPMN submodel of Management Process Model concerning the quality area
11	Architecture Quality Assessment Model	Contains all QM related elements
12	Management Process Model	BPMN general model for the Project Management
13	Correspondence	Defines correspondence between AD Elements and Quality Patterns
14	Quality Requirement Pattern	Definition of Quality Requirements Algorithms
15	Product Pattern	Organizes report from parts
Quality Products Submodel		
16	Report used by Steering Committee	Report used by Steering Committee
17	Product	Quality Management Process Product

The class objects in Figure 2 are grouped into four different groups, showing the four business areas (views) of quality topic, namely:

- Organization Structure Submodel.
- Generic Enterprise Architecture Submodel.
- QM Related Model.
- Quality Products Submodel.

Stakeholders are here the people responsible for the various stages of the manufacturing cycle, in this case producing architecture and its quality control (project manager, analyst, architect, quality manager), and not those of the business because the production of architecture is not the final product as opposed to the application of IT. These are the so-called internal customers.

2.3 Preliminary Process Metamodel for the Quality Management in the Enterprise Architecture Development Process

According to the metamodel definition given in the ISO standard ISO/IEC 24744 [1] methodology allows the construction of the process metamodels.

This process model is presented as the BPMN model and is shown below in Figure 3. Meaning of various processes concepts are explained in Table 2. According to the BPMN modeling methodology the process objects in Figure 3 are placed grouped into four different “LINES” that are groups, showing the organizational units responsible for these processes.

Table 2. Definition of processes for the topic “Architecture Quality Management”

Seq.no.	Process name	Definition
1	Define Quality Requirements	Defines all initial Quality Requirements for the products.
2	Define Quality Process & Product Patterns	In this process the pattern for each product of architecture development is defined.
3	Iteration Management	The process of the responsibility of the Project Manager, involving coordination of architectural production process and quality control of its products.
4	Define Correspondence - AD Model to Product Report Patterns	The process of checking the quality requirements which involves checking architecture models with the product standards.
5	Create AD Elements	All kind of operations on AD Elements
6	Quality Validation	Creates QM recommendation for iteration and product revision.
7	Iteration Acceptance	Product acceptance by the Steering Committee.
8	Create Final Report	Preparing report concerning the project closing.

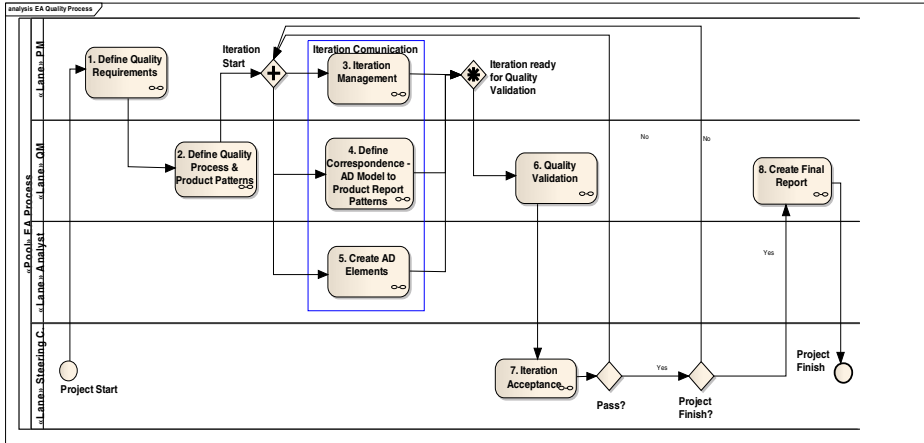


Fig. 3. Preliminary process metamodel for the topic “Architecture Quality Management”

2.4 Verifying the Completeness and Redundancy of the Classes Model in Relation (versus) to the Process Model

Comparison and verification is performed using the technique of “relationship matrix”. The relevant dimensions of this matrix are the „list classes” created from the class metamodel and “list process” created from the processes metamodel. The process of being in relationship with the class is marked by “X”.

Table 3. Relationship matrix “Processes versus Classes”

	Analyst	Create Final Report	CRUD AD Elements	Define Correspondence - AD	Define Quality Process &	Define Quality Requirements	EA Process	Iteration Acceptance	Iteration Management	Iteration ready for Quality	Iteration Start	Message1	Pass?	PM	Project Finish	Project Finish?	Project Start	QM	Quality Validation	Steering Committee
AD Attribute	X																			X
AD Element	X																			X
AD Link	X																			X
Analyst	X	X																		
Architecture Model																				
Architecture Quality Assessment Model																		X	X	
Correspondence																		X	X	
High Level Quality Report								X											X	
Management Proces Model														X						
Project Manager									X	X				X						
Quality Management Process Model				X															X	
Quality Manager		X	X	X	X	X	X	X	X	X	X	X	X	X				X	X	
Quality Report																			X	X
Quality Report Pattern				X															X	
Quality Requirement Pattern				X															X	
Stakeholder																				
Steering Committee		X						X	X			X	X	X	X	X	X	X		X
View Point																				

3 Simulation and Optimization of Quality Management Process Metamodel

3.1 Simulation Assumptions and Input Data

Simulation and optimization was performed by using the tool “Aris Business Process Simulator” available in the “ARIS Business Architect”. Figure 3 shows the preliminary model preliminary. Model is due to the requirements of simulation tools have to be transformed. Because of size and usefulness during interpretation of simulation results, completed BPMN diagram is shown on Figure 5. Based on BPMN model and process parameters, 10 process executions have been performed. On-line observed process simulation (animation) provides numerical results of the current simulation status. Every function from business process model parses process instances and dynamically provides information about current status of simulation. Figure 4 shows ARIS way of presentation progress of simulation. Every function is surrounded by numbers, presenting simulation results, as described on Figure 4.

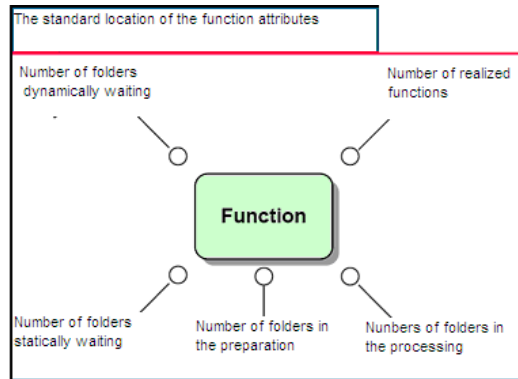


Fig. 4. The meaning of the individual numbers, describing simulation progress

The meaning of the various descriptions in Figure 4 is as follows:

- Number of folders dynamically waiting – number of instances of processes, waiting for execution due of human resource absence (eg. in the same time, needed resources are busy because of handling other processes).
- Number of realized functions – number of instances, which has been carried out and completed.
- Number of folders statically waiting – number of instances of processes that cannot be done; they are not waiting for resources, but for trigger from decision point XOR or parallel AND flow.
- Number of folders in the preparation – number of instances of the process, which are in preparation status (some processes require preparation time to run them).
- Number of folders in the processing – number of instances of processes, which are currently in processing status.

3.2 Business Process Simulation and Optimization

Business process simulation and optimization has been performed for the example as number of iterations, where the main goal is concerned to select optimal duration time of every step in the process and selecting the optimal value of process time for each processes in such way to perform 10 process instances executions and complete them in just one year (time of the project). Each instance of the process simulation is randomly activated in sequence during the event input. Starting a new instance of the process by the initial event is equivalent to running a quality management process for the new (next) of the product. It follows that the number of instances generated by the event input is equal to the number of products formed during the project.

Table 4. Input data before and after optimization

No. of the process	Name of the process instantiation	Processing time for the optimization (ddd:hhh:mm:sec)	Processing time after the optimization (ddd:hhh:mm:sec)	The share (%) of time in the total time
	Number of the products instances/year	10		
1	Define Quality Requirements	0010:00:00:00	0006:00:00:00	14,20%
2	Define Quality and product Patterns	0014:00:00:00	0007:00:00:00	16,60%
3	New Product Activation (Preparation time- organizing the team and tools)	0000:05:00:00	0001:00:00:00	2,40%
4	Iteration Management	0002:00:00:00	0001:00:00:00	2,40%
5	Define Correspondence AD Model to Product Report Patterns	0005:00:00:00	0005:00:00:00	11,80%
6	Create AD Elements	0010:00:00:00	0015:00:00:00	35,50%
7	Quality Validation	0010:00:00:00	0006:00:00:00	14,20%
8	Product and Iteration Acceptance	0005:00:00:00	0001:00:00:00	2,40%
9	Create Final Report	0001:00:00:00	0000:05:00:00	0,50%
	TOTAL	0057:05:00:00	0042:05:00:00	100,00%

Simulation and optimization study consists of several steps, some of which are repeated in an iterative cycle. At the beginning, UML activity diagram, obtained with some exploring methods from workflow system database, has been converted to BPMN form. Some additional model parameters, which are also obtained with exploring methods, are placed in to the simulation BPMN model. Then the simulation process has been carried out and first analysis has detected three sources of “bottlenecks”:

- **“bottlenecks” associated with the allocation of resources** – human resources are insufficient; they are invoked by the currently processing instance but resources are busy. Therefore, they cannot be used until the end of processing another process execution. As the result, on the simulation model are created process instances called *“dynamically waiting”* for the execution until the release of resources. The optimization solution is to increase resources in order to eliminate *“dynamically waiting”* processes.
- **“bottlenecks” associated with function processing time (function execution time)** – exists, if possibility to increase the resources has been already exhausted and the only solution is to reduce the execution time (but according to business possibilities). As the result, on the simulation model are created process instances called *“statically waiting”*. Optimization solution is to reduce the implementation time of such processes, in order to eliminate *“statically waiting”* processes.
- **“bottlenecks” associated with downtime (which means to stop the activation process instance)** – it results as elongation of total duration of the simulation. The solution optimizes “bottlenecks” described above in order to reduce downtime.

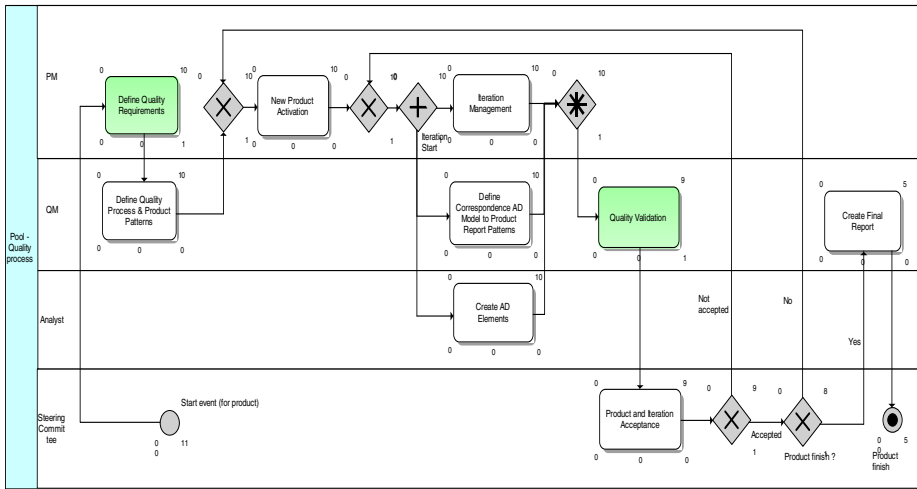


Fig. 5. Final metamodel of processes for the topic “Architecture quality management” after the optimization

4 Summary

Introduced by the author metamodels for classes and processes concerning quality management in the manufacturing cycle of the software allow you to define the methodology of this process. In this sense it is an extension of ISO standards given in [1], [2], [3] and TOGAF™ [12]. The simulation and optimization of the process model can determine the ratio between the duration times of individual elementary processes in the metamodel concerning quality management. Optimal durations of

processes eliminate the so-called bottlenecks in the process. The simulation results bring important guidance for managers in the construction process allowing the project plan to build an optimal plan consisting the right proportions between the various stages and between elementary processes in the project and optimal resource allocation. The work is a generalization of another author's work [7], which involved the construction of a quality management system but without the optimization and only for business intelligence systems.

References

1. ISO/IEC 24744: Software Engineering - a metamodel for Development Methodologies
2. ISO/IEC 12207: Systems and Software Engineering – Software Life Cycle Processes
3. ISO/IEC 42010: Systems and software engineering – Architecture description
4. Gawin, B., Roszkowski, J.: Extension the Capacity of the Cellular Network – Process Simulation and Optimization. In: Proc. of the 9th International Conference on Perspectives in Business Informatics Research, Rostock University, Germany (2010)
5. Goikoetxea, A.: A Mathematical Framework for Enterprise Architecture Representation and Design. International Journal of Information Technology and Decision Making 3(1) (2004)
6. Kasprzak, T.: Praca zbiorowa (Collected work) Modele referencyjne w zarządzaniu procesami biznesu (Reference models in business process management DIFIN (2005)
7. Roszkowski, J., Kobyliński, A.: Quality Management Reference Models for Business Intelligence Class Systems. In: The 8th International Conference on Perspectives in Business Informatics Research, October 1-2. Kristianstad University College, Sweden (2009)
8. Roszkowski, J.: The Simulation of Processes in the Integrated Computer Environment in the TELCO Sector. In: Proc. of the 7 Symposium: Modeling and Computer Simulation, Łódź, Poland, College of Computer Science (2010)
9. Roszkowski, J., Kobyliński, A.: Modele referencyjne zarządzania jakością w systemach klasy biznesu (Quality Management Reference Models for Business Intelligence – Class Systems). Roczniki Kolegium Analiz Ekonomicznych (Annals of the College of Economic Analysis), Poland, Warsaw, SGH 19/(2009)
10. Roszkowski, J., Kobyliński, A.: The Simulation of Software Processes in the Integrated Computer Environment in the Case of Telco Sector. In: Proc. of the 17th International Conference on Information and Software Technologies, Kaunas, Lithuania (2011)
11. Scheer, A.W.: Business Process Modeling. Springer, Heidelberg (2000)
12. TOGAF™ Version 8.1.1 Enterprise Edition. Van Haren Publishing (2000) ISBN: 9789087531737

Ontology Matching Using TF/IDF Measure with Synonym Recognition

Marko Gulić¹, Ivan Magdalenić², and Boris Vrdoljak³

¹ Faculty of Maritime Studies, University of Rijeka,
Studentska ulica 2, HR-51000 Rijeka, Croatia
marko.gulic@pfri.hr

² Faculty of Organization and Informatics, University of Zagreb,
Pavlinska 2, HR-42000 Varaždin, Croatia
ivan.magdalenic@foi.hr

³ Faculty of Electrical Engineering and Computing, University of Zagreb,
Unska 3, HR-10000 Zagreb, Croatia
boris.vrdoljak@fer.hr

Abstract. Ontology matching is an important process for integration of heterogeneous data sources. A large number of different matchers for comparing ontologies exist. They can be classified into element-level and structure-level matchers. The element-level matchers compare entities ignoring their relations with other entities, while the structure-level matchers consider these relations. The TF/IDF (term frequency / inverse document frequency) measure is useful for specifying key terms weights in documents. In our matching system we use the TF/IDF measure for comparing documents that store data about ontology entities. However, the TF/IDF does not take synonyms into account, and it may occur that the terms that describe two entities the best are synonyms. In this paper we propose a matcher that combines the TF/IDF measure with synonym recognition when determining key term weights, in order to improve the results of ontology matching. Evaluation of the matcher is performed on case study examples.

Keywords: ontology matching, TF/IDF, synonym recognition, ontology integration.

1 Introduction

An ontology is an explicit specification of a conceptualization [1]. Ontology defines a set of entities and relations between them in a way that both humans and machines understand it. Ontologies are expressed in an ontology language. Web Ontology Language (OWL) [2] is one of the most popular languages and it is recommended by W3C organization.

A key challenge in the integration of heterogeneous applications is specifying a semantic mapping between different ontologies. These ontologies describe similar domain of interest, but describe the same entities with various terminology (e.g. Car and Automobile) and build different structure (by using object properties) between these entities.

In ontology matching, a matcher is a method that defines the correspondences between two entities. The main classification of matchers classifies matchers into element-level and structure-level matchers. Element-level matchers compare entities ignoring the relations with other entities, while structure-level matchers compare entities considering these relations with other entities within ontologies. As each ontology is unique, the efficiency of any basic matcher depends on the implementation of the ontologies that are matched [5]. For example, let us suppose that a matcher that compares comments of the entities is included in the matching system and the comments of the entities are created in the first ontology, but not in the second one. In this case, the result of the matcher will be poor.

A large number of ontology matching systems ([6], [8], [9], [10], [11], [12], [13], [14]) include a matcher that uses the TF/IDF (term frequency / inverse document frequency) measure [16]. The TF/IDF is a measure for specifying term weights in textual documents. A textual document must be created for each entity of an ontology in order to compare them by using TF/IDF. The document consists of a set of terms related to the entity. These terms can be annotations of entity, entity's name, description, name and range of data properties related to entity, relations with other entities (especially generalization and specialization). Therefore, the document of an entity can contain a lot of information related to the element and structure level and this is the advantage of the matcher that uses the TF/IDF measure. Thus, two entities of different ontologies can be compared although some information about an entity does not exist in one ontology (e.g. entities of the first ontology do not have comments). The key term of the certain document in the TF/IDF measure is one that appears many times within that document and rarely in other documents. If two documents share the same key terms, the similarity between entities will be high. Matching between two documents starts with determining term weights in each document. Then, all the terms from all documents are put into the vector of all terms. Comparison of vectors is calculated with the cosine similarity [17] which is the scalar product of two vectors. The higher the scalar product, the greater the similarity between documents.

If the key terms within two compared documents are synonyms (the different terms of the same meaning) or different terms derived from the same basic morphological form (e.g. playing and play), the similarity between entities can be decreased, because TF/IDF measure does not recognize either synonyms or different terms derived from the same basic morphological form as the same term while determining term weights in the documents. In the rest of the paper, we use the name "family terms" to relate to different terms derived from the same basic morphological form. In other words, TF/IDF measure recognizes two terms as the same term only if they are identical. For example, two documents that are representing a concept of automobile have synonyms for their labels. First document has label *car* and the second document has label *automobile*. These terms are very important within the documents because these terms define exactly a kind of vehicle that the documents describe. It can also occur, for instance, that the first document has a label *entity that describes cars* and the second document has a label *car*. The TF/IDF measure would not recognize terms *car* and *cars* as similar.

Synonyms are usually discovered with usage of thesaurus. Thesaurus is a special kind of lexicon that contains relations between terms. The most popular thesaurus is WordNet [18]. In this paper we propose a matcher that uses the TF/IDF measure and also recognizes synonyms during determining key terms weights in order to improve the results of matching. The main purpose of this paper is to introduce the synonym recognition into the TF/IDF measure rather than comparing existing synonym recognition methods. When some terms are synonyms, they are all replaced with one of these terms in order to be recognized as similar in the TF/IDF measure.

“Family terms” can be resolved with stemming. The stemming is the process for converting terms into basic morphological form. It takes some time and the conversion quality depends on the stemming algorithm that is used. If two compared terms are “family terms”, our matcher will also recognize them as synonyms. Therefore, using the WordNet, the appearance of both synonyms and “family terms” can be resolved in order to improve the TF/IDF measure.

The paper is organized as follows. In Section 2 we discuss related work. In Section 3 basic terminology of ontology matching and TF/IDF measure are introduced. Our matcher that uses TF/IDF measure and recognizes synonyms as well as “family terms” during determining key term weights is presented in Section 4. In Section 5 evaluation of our matcher is performed on case study examples and a comparison with matcher that uses standard TF/IDF measure is given. Finally, the conclusion is given in Section 6.

2 Related Work

In this section, we are going to present some of the ontology matching systems that use the TF/IDF measure for comparing documents of entities that contain several characteristics of entities. These systems did not resolve the recognition of synonyms in these documents. The matchers that recognize synonyms, proposed in [6], [9] and [10], are run separately from other matchers. Thus, they cannot improve the results of certain matcher that use the TF/IDF measure. Of course, separate matchers that recognize synonyms can be always executed regardless of the other matchers. In this paper our focus is on improving the matcher that uses the TF/IDF measure for comparing entities by introducing synonym recognition in the TF/IDF measure.

FALCON AO [8] combines two basic matchers that are included in the language part of the system: matchers that compare names of entities and the statistical analysis of text that describe certain element. The second language matcher uses the TF/IDF measure for determining key terms of the entities documents. The document of an entity consists of the terms extracted from the entity's names, labels and comments of certain entity and its neighbors. The authors do not take into consideration synonym recognition in the TF/IDF measure, which in our opinion would improve the results.

H-Match [10] first executes the element-level language matcher. Language matcher compares the names of entities with the WordNet thesaurus. The similarity of two entities of different ontologies is defined considering their relation within WordNet. The most similar entities are those whose names are synonyms. The hypernyms and

hyponyms are less similar than synonyms. The similarity depends on the distance of two names. It is not the same if the hypernyms or hyponyms are directly related or they are related through one or more terms between them.

ASCO [9, 11] has three basic string matchers: the first one compares names, the second one compares labels and the third one compares virtual documents of every entity. The virtual document contains terms from label, name, comments and data of the neighbors. The weights for every term are calculated with the TF/IDF measure. The WordNet is used for recognizing synonyms in the first and the second matcher. In the third matcher, authors do not deal with synonyms.

In MapSSS [13] author proposes a matcher that will compare the meaning of the entity's name using the Google search engine. This matcher will replace a matcher that uses the WordNet for recognizing synonyms and other relations between two terms. Advantage of this method is the availability of the jargon terms that users use. The idea is very interesting for future research.

AgreementMaker [14, 15] is a matching system that uses element-level and structure-level matchers. It has three matchers that compare strings. One matcher uses the TF/IDF measure and compares virtual documents of entities that are a bag of terms, like previous matchers. The data from instances of the entity is also in the entity document. However, it does not deal with synonyms.

Yam++ [6] consists of a large number of element-level and structure-level matchers. One matcher uses the WordNet for comparing the basic morphological form of every term. The authors propose three types of virtual documents [7] that are compared using the TF/IDF measure. Individual profile contains label, name and comment of entity. Semantic profile of entity contains its individual profile and individual profiles of its neighbors. External profile contains data from instances of entity.

In Prior+ [12] authors propose three basic matchers and one of the matchers is a matcher that uses the TF/IDF measure. Apart from the standard data in the document (name, label, comment, data of neighbors), the document contains the names of relations and their range and domain. The recognition of synonyms is not examined here.

3 Ontology Matching Terminology and TF/IDF Measure

As it is stated before, ontology is a shared understanding of some domain of interest [1]. Definitions of the most important terms related to our research are defined in the following [3]. Matching is the process of finding relationships or correspondences between entities of different ontologies. Alignment (A) is a set of correspondences between two ontologies o and o' . The alignment is the output of the matching process. Correspondence $\delta (e_i, e_j')$ is the relation holding between entities of different ontologies.

There are many data sources that are not structured and have some form of textual document. Key terms in these documents must be somehow determined in order to compare these documents. TF/IDF measure defines the weights of the terms within documents that are compared. This measure is widely used in the area of information retrieval and it is also very useful for ontology matching. There are several preparing

steps before determining weights with TF/IDF measure, such as tokenization (splitting terms), removing stop words (prepositions, conjunction) and stemming.

The first step of this measure is to make a vector of all terms that occur in the documents. Every document gets its own vector of weights for each term because the weight of each term must be calculated for each document. The weight of term i in the document j is calculated with the equation:

$$w_{ij} = TF_{ij} \times IDF_i, \quad (1)$$

where the TF_{ij} is the term frequency of term i in the document j and the IDF_i is inverse document frequency of appearance of term i within all documents. The equation for the TF_{ij} is:

$$TF_{ij} = f_{ij} / \max_z f_{zj}, \quad (2)$$

where f_{ij} is the number of appearance of term i within the document j and $\max_z f_{zj}$ is the maximal appearance of particular term within the document j . The IDF_i is calculated with the equation:

$$IDF_i = \log(N/n_i), \quad (3)$$

where N is the number of all documents and n_i is the number of documents in which the term i exists. The lower the n_i , the higher the IDF_i and it means that term i describes well the document j . When the weights of all terms in every document are determined, the vectors of weights can be compared in order to find similar documents. The similarity between two documents is calculated with the cosine similarity that compares the weights of terms in these two documents:

$$\cos(c, s) = \sum_i^k w_{ic} * w_{is} / (\sum_i^k \sqrt{(w_{ic})^2} * \sum_i^k \sqrt{(w_{is})^2}) \quad (4)$$

In figure 1 the example of determining similarity with TF/IDF measure and cosine similarity is explained.

After the calculation of the TF/IDF measure and the cosine similarity, it can be seen that the terms *car* and *automobile* were not recognized as synonyms therefore the similarity between the entities *car* and *automobile* is the same as the similarity between *truck* and *automobile*. In section 4 we propose the recognition of synonyms while executing TF/IDF measure in order to improve the results of similarity.

4 TF/IDF Measure with Synonym Recognition

In the previous section it can be seen that standard TF/IDF measure with cosine similarity did not recognize terms *car* and *automobile* as the same terms. Therefore, the result of comparing the document₁ and the document₁' was poor although these documents describe the same entity in different ontologies. In this work, we propose a matcher that uses the TF/IDF and recognizes synonyms within documents. All the synonyms (e.g. *car* and *automobile*) are converted into the same term (e.g. *car*) therefore the TF/IDF measure will deal with them as the same term. Also, stemming, as the

preparation step for the TF/IDF measure, is resolved during the recognition of synonyms by using the WordNet. In figure 2 the example of determining similarity with TF/IDF measure and synonym recognition is explained.

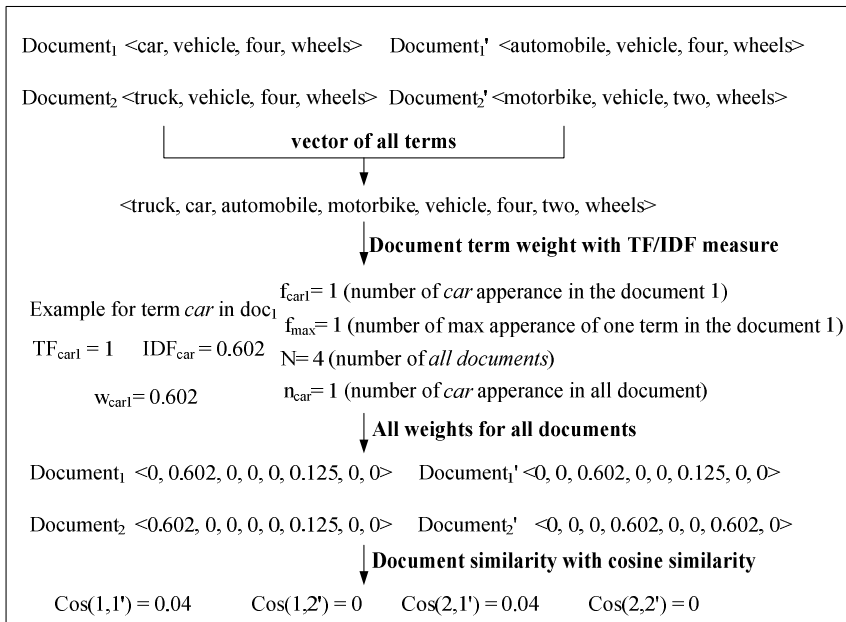


Fig. 1. TF/IDF measure and cosine similarity

First, the synonyms will be determined. The algorithm of recognizing synonyms in the documents of entities consists of the following steps:

1. Find all synonyms within a document and replace them with one common term
2. For each specific ontology: find all synonyms within all documents and replace them with one common term
3. Find all synonyms within all documents of two ontologies and replace them with one common term

As already mentioned, the focus of this paper is not to compare the existing synonym methods, but to introduce synonyms into the TF/IDF measure. We have chosen WordNet as the most used tool for synonym recognition. There are several methods that compare relation between two terms based on WordNet [19]. However, we need a method that only finds synonyms while comparing two terms. Hence, we implemented three methods for synonym recognition that can be used for our TF/IDF measure. WordNet contains the relations of synonyms and hypernyms between terms.

In our TF/IDF measure, two sets of terms, a set of synonyms and a set of hypernyms are assigned to every term in the document. A set of hypernyms includes only the hypernyms that are directly associated with the current term within WordNet. We call these sets a bag of synonyms and a bag of hypernyms. In the first measure

(Synonym TF/IDF 1), if two terms have an identical term within the bags of synonyms, the terms are synonyms. The second measure (Synonym TF/IDF 2) extends the first measure including the recognition of synonyms with the bag of hypernyms. If two terms have an identical term within the bags of synonyms, or one compared term is identical with at least one term in the bag of hypernyms of another compared term, the terms are synonyms. The third measure (Synonym TF/IDF 3) includes the previous two measures and extends recognition of synonyms comparing separately two bags of hypernyms. If two terms have an identical term within the bags of hypernyms, the terms are synonyms.

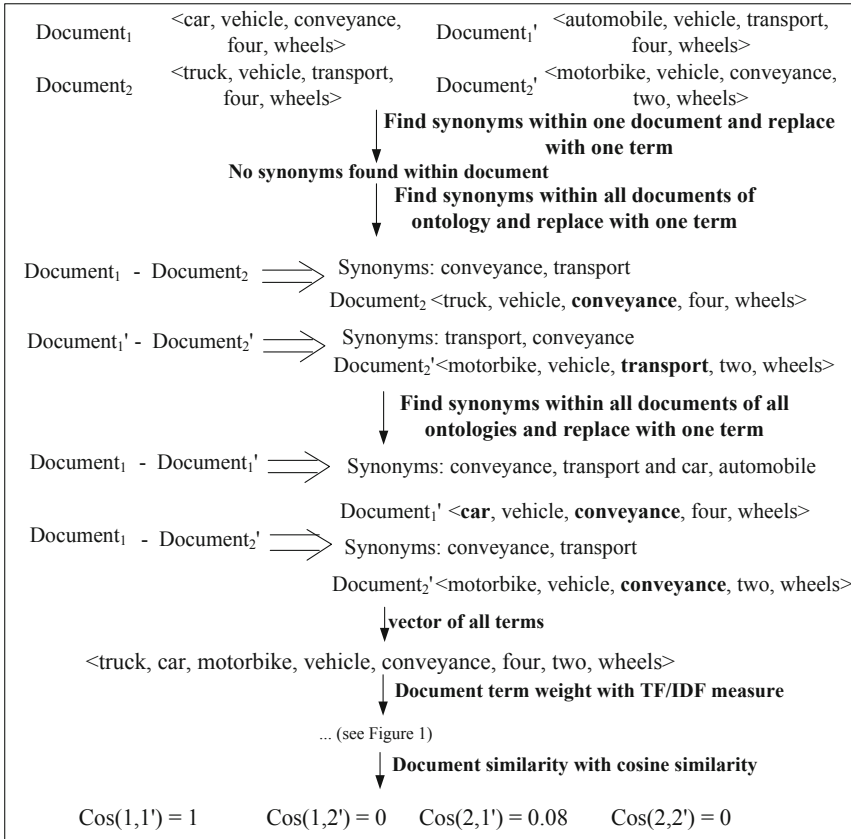


Fig. 2. TF/IDF measure with synonym recognition and cosine similarity

As we stated before, we assume that only a matcher that accurately determines synonyms is appropriate for our TF/IDF measure with synonym recognition. False synonyms would deteriorate the results of matching. Hence, we suppose that the Synonym TF/IDF 1 measure will be better than the Synonym TF/IDF 2 and 3 because it determines synonyms more accurately. The tests that confirm this assumption are discussed in Section 5.



The first step of recognizing the synonyms is to search synonyms in each particularly entity. In the example in figure 2, we used Synonym TF/IDF 1 measure for recognizing synonyms. There are no synonyms found within each particular document.

Next, it is examined whether the synonyms exist within all documents of each ontology. The order of documents, in which the searching of synonyms is examined, is not important. Therefore, all synonyms will be replaced with the term of first document, then with the term of the second document etc. In the first ontology, the document₁ and the document₂ have synonyms *conveyance* and *transport* because they have the same terms in their bag of synonyms. In the document₂ *transport* is replaced with *conveyance* hence the document₂ has terms <truck, vehicle, conveyance, four, wheels>. The synonyms *transport* and *conveyance* exist in the documents of the second ontology hence the document₂' is changed and the terms are <motorcycle, vehicle, transport, two, wheels>.

Next, the synonyms within all documents of ontologies have to be found. Two pairs of synonyms (*car-automobile* and *conveyance-transport*) are found within the document₁ and the document₁'. After replacement, the document₁' has terms <car, vehicle, conveyance, four, wheels>. Synonyms *conveyance* and *transport* are found within the document₁ and the document₂'. The document₂' now becomes <motorcycle, vehicle, conveyance, two, wheels>. Next, the document₂ has to be compared with the document₁' and the document₂'. There are no synonyms found because the term *transport* is replaced with term *conveyance* in the previous step (comparison of the document₁ with the document₁' and the document₂'). The vector of all terms is <truck, car, motorbike, vehicle, conveyance, four, two, wheels>. The weights, calculated with TF/IDF measure, of the terms in the documents are: the document₁ <0,0.301,0,0,0,0.125,0,0>, the document₂ <0.602,0,0,0,0,0.125,0,0>, the document₁' <0,0.301,0,0,0,0.125,0,0>, the document₂' <0,0,0.602,0,0,0,0,0.602,0>. After the calculation of cosine similarity, the results are: $\cos(11') = 1$, $\cos(12') = 0$, $\cos(21') = 0.08$ and $\cos(22') = 0$. Our matcher has determined that the document₁ and the document₁' are more similar than the document₂ and the document₁'. The similarity was equal before when the standard TF/IDF measure was executed.

5 Evaluation of Matcher That Uses the TF/IDF Measure with Synonym Recognition

To test the alignments considering different aggregation methods we took one pair of ontologies: onto100.owl and onto205.owl. We used the corresponding alignments between these ontologies to compare the results. These ontologies and corresponding alignments are used at the OAEI (Ontology Alignment Evaluation Initiative) campaign and can be found in [20]. OAEI initiative has been established for evaluation of increasing number of ontology matching systems. It offers a yearly evaluation event where the implementations of matching systems are tested through the controlled experimental evaluation. Ontologies onto100.owl and onto205.owl are developed for testing the matching system when the synonyms exist within ontologies. A software

implementation of a matcher that uses standard TF/IDF measure and our matcher that uses TF/IDF measure with synonym recognition has been developed and evaluated.

After the correspondences between entities are obtained, the best correspondences must be chosen for the final alignment (figure 3).

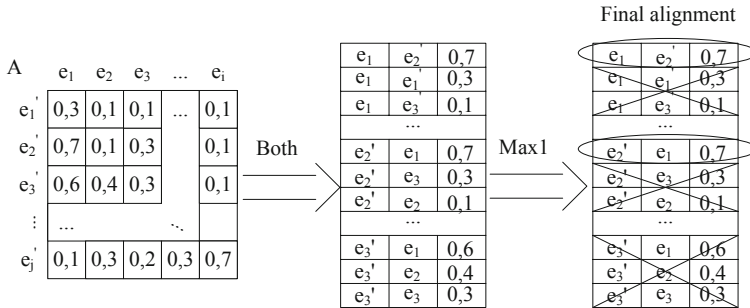


Fig. 3. Process of the final alignment

Several procedures of determining final alignment are explained in [4]. First, the direction of selection of candidates for alignment (from one ontology to another, or vice versa) should be decided. If we choose the option Both, all the elements e_j' from ontology O' that are associated with an element e_i from ontology O (and vice versa) will be selected and all this correspondences will be ranked by the value of correspondence. Second, the method of selection of appropriate correspondences for the final alignment should be selected. If we choose the option Max1, only the best correspondence of an element e_i or e_j' will be chosen. We also use the threshold that is a minimal value of selected correspondence for the entrance of the final alignment. The threshold will emphasize the difference in results between the standard TF/IDF matcher and the TF/IDF matcher with recognition of synonyms.

Evaluation measures that we use are precision and recall. Precision is the ratio of correctly found correspondences over the total number of returned correspondences. Recall is the ratio of correctly found correspondences over the total number of expected correspondences. For class entities, we insert the name of class, label, comment and the labels of subclasses and superclasses in the class document. For property entities, we insert the name of property, label and comment in the property document.

First, we compare results of the matcher with the standard TF/IDF and our matcher that uses the TF/IDF measure with three different synonyms recognition (Synonym TF/IDF 1, 2 and 3). We tested matchers with several thresholds: 0.5, 0.6, 0.7, 0.8 and 0.9. The results of Recall can be seen in figure 4.

The results of Recall show that our matcher that uses Synonym TF/IDF 1 measure has the best results for every threshold value. The difference between the Synonym TF/IDF 1 and the Standard TF/IDF is increased as the threshold is increased. These results indicate that the found correspondences are also higher with our Synonym TF/IDF 1 measure. The higher correspondences will improve the efficiency of the matching system because the higher and more accurate correspondences will increase the total alignment as the result of aggregating alignments of all matchers.

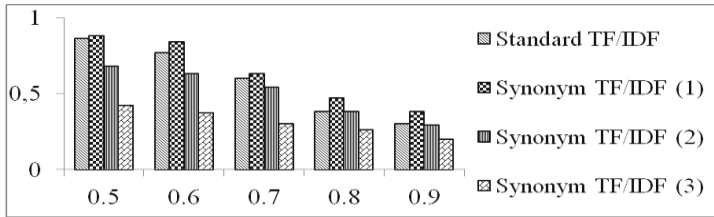


Fig. 4. Measure Recall for the matcher with standard TD/IDF and the matcher using TF/IDF with three different synonyms recognition for 5 values of threshold

In figure 5 the values of the Precision are presented. The results of Precision shows that our Synonym TF/IDF 1 measure has slightly better results. Here, the difference is not significant because every matcher with TF/IDF measure finds accurate correspondences as a result of richly described documents which are compared. Hence, these matchers find a small number of incorrect correspondences between entities.

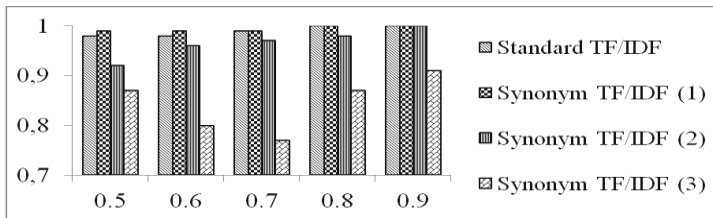


Fig. 5. Measure Precision for the matcher with standard TD/IDF and the matcher using TF/IDF with three different synonyms recognition for 5 values of threshold

In figure 6 the average correspondences obtained by two best compared matchers are displayed. Synonym TF/IDF 2 and 3 measures are not compared because of their poor values of Recall and Precision. These values have proven that synonyms recognition must be very accurate. Hence, we propose only the Synonym TF/IDF 1 measure for comparing ontologies with TF/IDF measure that uses synonym recognition.

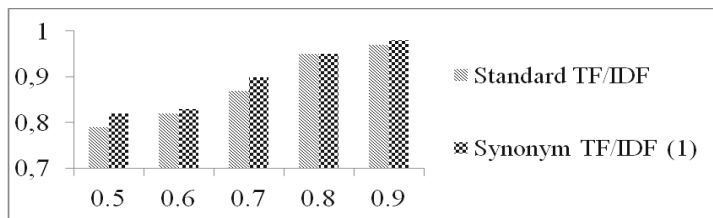


Fig. 6. The average correspondences obtained by matchers with TD/IDF

Apart the Recall and the Precision are higher for our Synonym TF/IDF 1 measure, the average found correspondences have also higher values for matcher using TF/IDF with synonym recognition. Therefore, the correspondences between two entities are

determined more accurate with our matcher. As stated before, this is important accordingly matcher with TF/IDF will be one of the matchers in the matching system and its correspondences will be aggregated with correspondences of other matchers.

6 Conclusion

In this paper we proposed a matcher that compares document entities with TF/IDF measure and recognizes the synonyms between these documents in order to improve matching results. The developers of ontology matching systems usually implement the matcher with TF/IDF measure for comparing documents of entities because these documents contain a lot of information (based on the element and structure level) that describe entities and always produce relevant results.

If the synonyms are key terms of some documents, the similarity results between these documents would be lower because the standard TF/IDF measure does not recognize synonyms in these documents. To the best of our knowledge, our matcher is the only matcher that recognizes synonyms within TF/IDF measure. In order to improve the results of the standard TF/IDF measure, it is important to accurately recognize these synonyms.

The tests have proven that our matcher produces better results for evaluation measures Recall and Precision than the matcher that uses standard TF/IDF measure. It also correctly found correspondences between entities of different ontologies with higher value than the standard TF/IDF measure.

In our future work we will introduce some other methods for recognizing synonyms within TF/IDF measure, like comparing the meaning of the entity's name using the Google search engine, and introducing the closest hyponyms and hypernyms of certain term. We will then integrate this improved matcher into our ontology matching system, which we are going to evaluate in the next OAEI campaign.

References

1. Gruber, T.R.: A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2), 199–220 (1993)
2. Antoniou, G., van Harmelen, F.: *A Semantic Web Primer*. MIT Press, Cambridge (2004)
3. Euzenat, J., Shvaiko, P.: *Ontology matching*. Springer, Heidelberg (2007)
4. Do, H., Rahm, E.: COMA – a system for flexible combination of schema matching approaches. In: 28th International Conference on Very Large Data Bases (VLDB), pp. 610–621. VLDB Endowment, Hong Kong (2002)
5. Gulić, M., Vrdoljak, B.: Specifying parallel composition of matchers for ontology matching by using genetic algorithm. In: 34th MIPRO International Convention, pp. 953–958. IEEE, Opatija (2011)
6. Ngo, D., Bellasene, Z., Coletta, R.: YAM++ results for OAEI 2011. In: Euzenat, J., Shvaiko, P., Heath, T., Quix, C., Mao, M., Cruz, I. (eds.) *ISWC International Workshop on Ontology Matching*. CEUR-WS, vol. 814, pp. 228–236. CEUR-WS.org, Bonn (2011)

7. Ngo, D., Bellahsene, Z., Coletta, R.: A generic approach for combining linguistic and context profile metrics in ontology matching. In: Meersman, R., Dillon, T., Herrero, P., Kumar, A., Reichert, M., Qing, L., Ooi, B.-C., Damiani, E., Schmidt, D.C., White, J., Hauswirth, M., Hitzler, P., Mohania, M. (eds.) OTM 2011, Part II. LNCS, vol. 7045, pp. 800–807. Springer, Heidelberg (2011)
8. Jian, N., Hu, W., Cheng, G., Qu, Y.: Falcon-AO: Aligning ontologies with Falcon. In: Euzenat, J., Shvaiko, P., Ehrig, M., Stuckenschmidt, H. (eds.) K-CAP Workshop on Integrating Ontologies. CEUR-WS, vol. 156, pp. 87–93. CEUR-WS.org, Banff (2005)
9. Qu, Y., Hu, W., Cheng, G.: Constructing virtual documents for ontology matching. In: 15th International World Wide Web Conference, pp. 23–31. ACM Press, Edinburgh (2006)
10. Castano, S., Ferrara, A., Montanelli, S.: Matching ontologies in open networked systems: Techniques and applications. *Journal on Data Semantics V*, 25–63 (2006)
11. Bach, T., Dieng-Kuntz, R., Gandon, F.: On ontology matching problems - for building a corporate semantic web in a multi-communities organization. In: 6th International Conference on Enterprise Information Systems (ICEIS), vol. 4, pp. 236–243. ICEIS Press, Porto (2004)
12. Mao, M., Peng, Y., Spring, M.: An adaptive ontology mapping approach with neural network based on constraint satisfaction. *Journal of Web Semantics, Science, Services and Agents on the World Wide Web* 8(1), 14–25 (2010)
13. Cheatham, M.: MapSSS Results for OAEI 2011. In: Euzenat, J., Shvaiko, P., Heath, T., Quix, C., Mao, M., Cruz, I. (eds.) ISWC International Workshop on Ontology Matching. CEUR-WS, vol. 814, pp. 184–190. CEUR-WS.org, Bonn (2011)
14. Cruz, I., Stroe, C., Caci, M., Caimi, F., Palmonari, M., Palandri Antonelli, F., Keles, U.C.: Using AgreementMaker to align ontologies for OAEI 2010. In: Euzenat, J., Shvaiko, P., Giunchiglia, F., Stuckenschmidt, H., Mao, M., Cruz, I. (eds.) ISWC International Workshop on Ontology Matching. CEUR-WS, vol. 689, pp. 118–125. CEUR-WS.org, Shanghai (2010)
15. Cruz, I., Palandri Antonelli, F., Stroe, C.: Efficient Selection of Mappings and Automatic Quality-driven Combination of Matching Methods. In: Euzenat, J., Shvaiko, P., Giunchiglia, F., Stuckenschmidt, H., Noy, N., Rosenthal, A. (eds.) ISWC International Workshop on Ontology Matching. CEUR-WS, vol. 551, pp. 49–60. CEUR-WS.org, Chantilly (2009)
16. Salton, G., McGill, M.H.: *Introduction to Modern Information Retrieval*. McGraw-Hill, New York (1983)
17. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*. Addison-Wesley, Boston (1999)
18. Miller, G.A.: WordNet: A Lexical Database for English. *Communications of the ACM* 38(11), 39–41 (1995)
19. Lin, F., Sandkuhl, K.: A Survey of Exploiting WordNet in Ontology Matching. In: Brämer, M. (ed.) *Artificial Intelligence in Theory and Practice II*. IFIP AICT, vol. 276, pp. 341–350. Springer, Heidelberg (2008)
20. Ontology Alignment Evaluation Initiative, <http://oaei.ontologymatching.org>

Moving Averages for Financial Data Smoothing

Aistis Raudys, Vaidotas Lenčiauskas, and Edmundas Malčius

Vilnius University, Faculty of Mathematics and Informatics,
Naugarduko st. 24, LT-03225 Vilnius, Lithuania
aistis@raudys.com, edmundas.malcius@gmail.com

Abstract. For a long time moving averages has been used for a financial data smoothing. It is one of the first indicators in technical analysis trading. Many traders debated that one moving average is better than other. As a result a lot of moving averages have been created. In this empirical study we overview 19 most popular moving averages, create a taxonomy and compare them using two most important factors – smoothness and lag. Smoothness indicates how much an indicator change (angle) and lag indicates how much moving average is lagging behind the current price. The aim is to have values as smooth as possible to avoid erroneous trades and with minimal lag – to increase trend detection speed. This large-scale empirical study performed on 1850 real-world time series including stocks, ETF, Forex and futures daily data demonstrate that the best smoothness/lag ratio is achieved by the Exponential Hull Moving Average (with price correction) and Triple Exponential Moving Average (without correction).

Keywords: moving average, smoothing, filter, time series, smoothness, lag, hull, exponential, TRIX.

1 Introduction

Moving averages are one of the key tools used to analyse financial time series. In a nutshell, moving average is simple weighted sum (mean) calculated over selected historical price range. Financial data usually is noisy, if we choose to represent today's price as mean of today's price and 2 days before, all ups and downs will be averaged. Using more historical prices (increasing period), we can achieve more smoothed price that would show the trend, despite the price fluctuations. Let's define x_i as a price value at the time i . Let $X = x_1, x_2, \dots, x_p$ and p is the time series length. So most simple moving average at the time t would be

$$ma_t^n = \frac{1}{n} \sum_{i=1}^n x_{t-i} \text{ or } ma_t^n = \sum_{i=1}^n x_{t-i+1} w_i,$$

where $w_i = \frac{1}{n}$, $i = 1, \dots, n$, and integer n determines the averaging window width.

Moving averages are heavily used to show the trend in the noisy data. At the same time the smoothing plays role of regularisation, wide known in statistical data analysis. Smoothing is often improving stability of conclusion/predictions as well. As period (the window width), n , of moving average is increased, more noise can be filtered

from financial data, better smoothness is achieved. But sometimes price fluctuation is a trend reversal - not noise. Since moving average combines historical prices and a current price to get filtered price, for some time, depending on the period, it will show previous trend instead of a new one. This is called a lag. Raw price has zero lag, and n -bar moving average has $n/2$ bar lag. The most important question is how to reduce the lag, which causes missed trade opportunities or false trades, and keep reasonable smoothness remains unsolved problem, which immense armada of moving averages (ways to chose and techniques) try to solve.

In [1], the problems that moving averages in financial data suffer are well explained. The main one is that nondeterministic unknown process is generating financial time series. But to filter and smooth this data, we can use one of many defined moving average processes, which fit differently from time to time, and by definition can't be perfect. Still a lot of effort is thrown to invent and upgrade moving averages to get better results. Also complexity varies: from simple linear moving averages to higher order processes and neural networks.

One of approaches is to dynamically adjust period of moving average. Authors in [2] use reinforced learning in their described trading strategy to alter period of moving average on the fly. All trading system is able beat the market by about 50 percentage points, according to authors. In [4] it is also claimed, that using variable period moving averages is possible to achieve profit, even during financial crisis. In [9] authors profiled investor risk using multitude of factors. The idea of adaptive moving averages has been extensively discussed in [3] and some trading strategies involving adaptive element has been assessed in [16].

Artificial neural networks are widely used in time series analysis and forecasting. In [5] authors use recursive Elman neural network to calculate moving average. The average is later used in proposed stop loss – maximum return (SLMR) trading strategy. Authors claim big success due to optimizations by joining a SLMR trading strategy with a moving average calculation inside an Elman neural network.

Smoothing (blurring of the images or time series) was considered in statistical data analysis. Actually it is some sort of regularization. A degree of optimal smoothing depends on the velocity of time series changes: the more frequent are the changes the smaller window with (lag) should be chosen. In Parzen window multidimensional nonparametric features input density estimation used for classification purposes Marina Skurikhina [17] compared 13 functions of smoothing window shape (Gaussian, trapezoidal, three angle, rectangular etc.). She found and that most of the smoothing functions were approximately equally effective. The Gaussian shape appeared the best. The rectangular shape appeared to be the worst one. The main effect of smoothing was obtained due to correct choose of lag - the window width. For financial data analysis we also have a number of diverse methods, however, there is a lack of comprehensive practical overview in the literature of moving averages for financial data smoothing, particularly paying attention to criterion "smoothens vs. lag ratio". Authors try to fill this gap by analyzing numerous moving averages, on numerous instruments, their types and time frames.

The paper is organised as follows. In the next section we present evaluation methodology, next we describe moving averages and methods. In the following sections we present data description, taxonomy, experiments results and conclusions.

2 Evaluation Methodology

We evaluate two opposite properties of moving averages: smoothness and lag. Traders want optimal filter so trend following would be nice and one could avoid whipsaw trades. Usually moving averages take one parameter – period p of past prices to use. As period increases, lag grows and edges smoothen. In Fig. 1, we see two different period moving averages. When trend changes, red one responds very slowly, value is far from real price. But the line is smooth. Blue follows price more aggressively, stays close to the price, but is bumpier. Not all bumps represent reverse of trend.

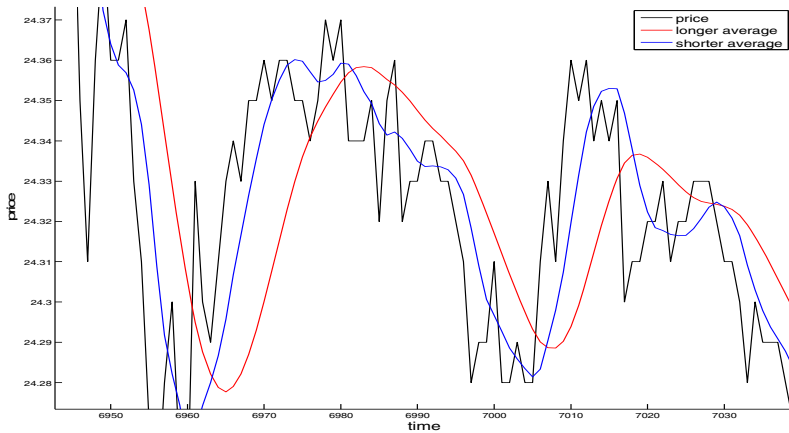


Fig. 1. Two MA (red and blue). Red one is long period MA hence more lagging and smoother, while blue is shorter period and less lagging and bumpier.

In this study we will investigate smoothness and lag in more detail. Assume we have prices $X=(x_1, x_2, \dots, x_p)$, here p is the length of time series. Moving average of length n at time t would be $ma(t, n) = \frac{1}{n} \sum_{i=0}^{n-1} x_{(t-i)}$. We define lag as a distance between current bar and moving average at that point, then at time t lag is: $lag_t = |x_t - ma(t, n)|$. For entire dataset, average lag is calculated like: $lag_T = \frac{1}{p} \sum_{t=1}^p lag_t$, here p is the number of data points in time series, x_i is the data point at pos i and ma_i is moving average of the n period at the position i . Lag estimate tells how moving average is lagging behind the price. Now smoothness needs to be measured. Assume we have moving average as vector of values (ma). First, we calculate how value changed from previous one:

$$dif_i = \begin{cases} i > 1, ma_i - ma_{i-1} \\ i = 1, 0 \end{cases}$$

Now $dif = \{dif_i\}, i = 1 \dots p$ vector holds values of how much moving average changed during each time period. If some value is negative, it indicates a negative change of direction. If growth is not constant, it creates bumps. We define smoothness as average of difference changes:

$$smo = \frac{1}{n - 1} \sum_{i=2}^p |dif_i - dif_{i-1}|$$

Using mean of such vector, we average bumpiness/smoothness of our moving average. As we defined lag and smoothness we can calculate some moving averages, increasing period from 2 to 30, and evaluate their smoothness and lag. Results are visualised in Fig. 2. Having short period moving average is very bumpy – high smoothness value. But it follows price very well and the offset is low. As the period increases, average becomes very smooth, but lag increases. Alternatively, smoothness and lag can be plotted against each other, as in Fig. 3 (lower values are better).

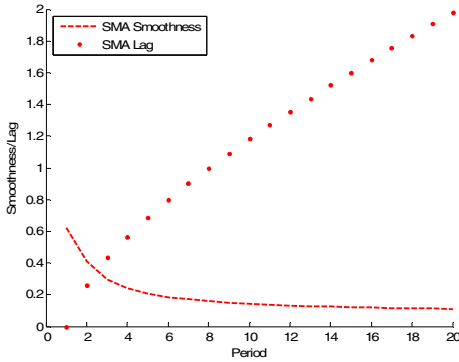


Fig. 2. Smoothness and lag of simple moving average can be plotted against each other (lower values are better)

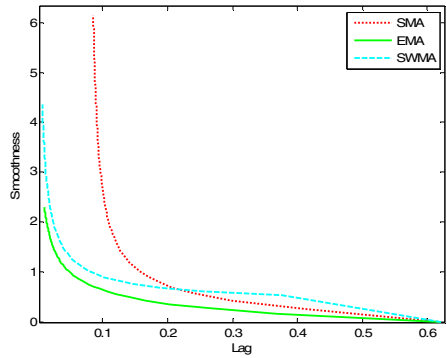
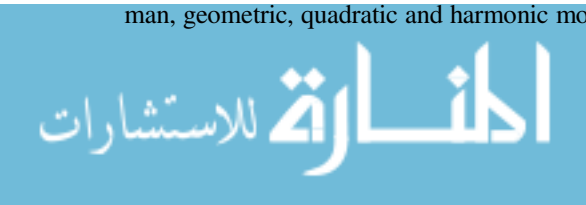


Fig. 3. In this figure we demonstrate how smoothness and lag are related. As MA period increases smoothness value is decreasing (smaller-better) and increases the lag.

Same situation is here, as lag grows, smoothness gets better and vice versa. Three different moving averages are plotted in Fig. 3, each calculated changing period from 2 to 30 over same dataset. One average looks slightly better. These two simple measurements can help to answer question, which moving average has best smoothness/lag ratio?

3 Moving Average Methods

In this paper we analyze 19 most popular moving averages (simple, exponential, weighted, sinus weighted, Spencers, median, Tilson, Hull, double exponential, TRIX/triple exponential, Ehlers, zero lag, Butterworth, Mesa, Savitzky-Golay, Kaufman, geometric, quadratic and harmonic moving average). In various sources one could



find even more, but in most cases same average comes under different names. Proprietary moving averages (like Jurik Moving Average) are discarded from this paper.

Simple moving average (SMA) is well known and widespread. It gives equal weights to all past prices and by definition is just average of them. Although very simple, it can solve serious problems. It will be used as a benchmark to compare against other averages. For its simplicity, formula is discarded.

Exponential moving average (EMA) gives exponentially diminishing weights to all past prices. This moving average is very well known and used, therefore formula is not included.

Weighted moving average (WMA) gives arithmetically diminishing weights for past prices, depending on length of the average.

Sinus weighted moving average (SWMA) is a weighted average, based on motivation, that price fluctuates following some unknown wave. As model, Sine wave is used to adjust price weights. SWMA is calculated using formula:

$$SWMA_n(X) = \frac{\sum_{n=1}^m \sin\left(n \frac{180}{6}\right) X_n}{\sum_{n=1}^m \sin\left(n \frac{180}{6}\right)}$$

Where m is period of moving average, X is list of prices with X_0 the most recent one.

Spencers 15 point moving average (SpMA) is another version of WMA used by actuaries. It is fixed 15 position mean with weights 3, -6, -5, 3, 21, 46, 67, 74, 67, 46, 21, 3, -5, -6, -3. The problem with this average is high lag.

Double exponential moving average (DEMA) is whole different from described above. It is composite moving average and uses other moving averages to get the result [11]. In case of DEMA, the EMA is used. Also, DEMA is adaptive - it employs some mechanism to adapt to price swings dynamically. DEMA uses trick to get better smoothness by running moving average on itself. But this operation increases lag, so to counter this technique called twicing is used. It takes difference between price and moving average to adjust itself, making DEMA adaptive. Formula:

$$DEMA_n(X) = EMA_n(X) + EMA_n(X - EMA_n(X))$$

where n is length of moving average and X is the prices.

Triple exponential moving average (TRIX) is similar to DEMA but uses exponential moving average three times:

$$TRIX_n(X) = EMA_n(EMA_n(EMA_n(X)))$$

Zero lag moving average (ZMA) sounds like a perfect moving average [9]. But the only thing without lag is the price, which this adaptive and composite moving average uses to correct itself. In a nutshell, ZMA ads portion of price to EMA to counter lag, while giving up some smoothness. Formula (n – period, X – prices

$$\text{if } n > 1, \beta = 0,2$$

$$ZMA_n(X) = \alpha * (X_{n-1} + \beta * (X_n - ZMA_{n-1}(X_{n-1}))) + (1 - \alpha) * ZMA_{n-1}(X_{n-1})$$

Tilson moving average (TMA) is also known as T3. It is both composite and adaptive. It is build using EMA [11]. To make notion more readable, formula is decomposed. First one describes generalized DEMA average introducing parameters n and v . For Tilson moving average, v is 0,7. If v would be 1, then GD would be DEMA. To improve smoothness of TMA, moving average is applied over again.

$$TMA_n(X) = GD_n(GD_n(X))$$

$$GD_n(X) = (1 + v) * EMA_n(X) - EMA_n(EMA_n(X)) * v, \text{ where } v = 0,7$$

Hull moving average (HMA) is composite moving average made from composing WMA of various period lengths [12]. Formula:

$$HMA_n(X) = WMA_{\sqrt{n}}(2 * WMA_{\frac{n}{2}}(X) - WMA_n(X))$$

Exponential Hull moving average (EHMA) is exactly the same as Hull MA but Exponential MA is used instead of Weighted MA:

$$EHMA_n(X) = EMA_{\sqrt{n}}(2 * EMA_{\frac{n}{2}}(X) - EMA_n(X))$$

Ehlers moving average (EhMA) is another adaptive moving average [8]. To use it, data must be first detrended subtracting SMA (of the same period as EhMA) from the price. Then EhMA coefficients are recalculated for each position, based on quadratic distance. This makes EhMA computational expensive with large periods over bigger datasets. Formula (X – detrended prices, n – period of EhMA) is gives for detrended prices, after applying EhMA result is obtained adding SMA back to it:

$$A = SMA_n(X); X = X - A; c_i = \sum_{m=1}^n (X_i - X_{i-m});$$

$$EhMA'_n(X) = \frac{\sum_{i=0}^{n-1} c_i X_i}{\sum_{i=0}^{n-1} c_i}; EhMA_n(x) = EhMA'_n(x) + A$$

Butterworth moving average (BMA) came from analogue circuits' era [8]. Very well known there, works for trading as well. Formula (n – period, X - prices, i – current bar) to calculate current bar $BMA(i)$:

$$BMA(i) = (X_i + 2 * X_{i-1} + X_{i-2}) + a_1 * BMA(i - 1) + a_2 * BMA(i - 2);$$

$$\beta = 2.415 * \left(1 - \cos\left(\frac{360}{n}\right)\right); \alpha = -\beta + \sqrt{(\beta^2 + 2 * \beta)};$$

$$c_0 = \frac{\alpha^2}{4}; a_1 = 2 * (1 - \alpha); a_2 = -(1 - \alpha)^2;$$

Mesa moving average (MAMA) uses Hilbert transform to make EMA adaptive. Because of Hilbert transform this moving average has complex formula, only main parts will be given. By definition MAMA is EMA with variable alpha: $MAMA(i) =$

$\alpha * Price + (1 - \alpha) * MAMA(i-1)$, where $\alpha = FastLimit/DeltaPhase$. $FastLimit$ is the upper bound of α and $DeltaPhase$ is the rate of change of the Hilbert Transform homodyne discriminator. The α value is kept within the range of $FastLimit$ and $SlowLimit$. This moving average focused on very short periods and tries to show cycles in them [3 p. 737].

Savitzky-Golay moving average (SGMA) is polynomial smoother [15]. Given last n prices, it tries to fit k level polynomial over them using MSE. Then polynomial value is used as filtered value. SGMA has two parameters: n – period, and k – level of polynomial to fit.

Kaufman moving average (KAMA) is adaptive one, which alters alpha of EMA using smoothing constant C to achieve addictiveness [3 p.731]. Formula (n – period, X – prices, X_i – past price i bars back):

$$ER = \frac{|X_i - X_{i-n}|}{n * \sum_{i=1}^n |X_n - X_{n-i}|}; C = (ER * (0,6667 - 0,0645) + 0,645)^2;$$

$$KAMA(i) = KAMA(i - 1) + C * (X_i - KAMA(i - 1))$$

KAMA adjust alpha using efficiency ratio of the market. It is ratio between direction and volatility. Constants 0,6667 and 0,0645 represent adaptiveness range from 2 to 30 bars of EMA alpha value. These constants are suggested by author, so we will keep them.

Chande's variable index dynamic average (VIDYA) follows same concept as KAMA. VIDYA, however, uses relative volatility to adjust smoothing constant [3 p.736]. Formula (s – constant, representing 9 bar EMA smoothing constant, C – closing prices, i – current time, C_n – prices of recent n bars, C_m – prices of longer historic period $m > n$):

$$VIDYA(i) = k * s * C_i + (1 - k * s) * VIDYA(i - 1)$$

$$s = \frac{2}{9+1} = 0.2, k = \frac{stdev(C_n)}{stdev(C_m)}$$

Other types of moving averages

Median moving average (MeMA) isn't weighted average, as by definition it is just median of a price range. So when calculating moving average, it just takes median element of a frame as average of frame. Formula is not included.

Geometric moving average (GMA) represents a growth function in which a price change from 50 to 100 is as important as a change from 100 to 200 [3 p.20]. Formula ($a_1..a_n$ – prices, n -period): $GMA = (a_1 * a_2 * ... * a_n)^{\frac{1}{n}}$

Quadratic moving average (QMA) is made from well known error estimator [3 p.21]. Formula (a – price, n - period) $QMA_n = \sqrt{\left(\frac{\sum a^2}{n}\right)}$

Harmonic moving average (HaMA) is time weighted mean, not biased towards higher or lower values as in the geometric mean [3 p.21]. Formula (n – period, a_i - prices): $HaMA = \frac{n}{\left(\sum_{i=1}^n \left(\frac{1}{a_i}\right)\right)}$

4 Taxonomy of Methods

Moving averages can be categorised into several groups depending on their behaviour. Most of the categorisation is based on properties of the weights. Some however are weightless (like Median MA) and cannot be assigned to any category. We characterised MA by three properties: positive/negative weights, look-back period and adaptiveness.

First categorisation is based on weights positivity. Weights can be positive only or positive and negative. If MA has negative weights it tries to reduce lag by correcting itself. This improvement also introduces overshooting on trend reversals. This can be seen from impulse response diagram. Each historical data point is weighted with positive weight and summed up afterwards. Other group of moving averages weights higher recent history and subtracts older history. This way it reduces lag but overshooting phenomenon appears on trend reversals.

Data dependant adaptive moving averages changes their behaviour based on the data, thus their smoothness and lag varies. Impulse response diagram may not correctly reflect their weighting scheme.

Fixed length/infinite length – look-back period. Some moving averages use exactly the same number of historical data points to calculate smoothed value (Simple MA, weighted MA). The other group references all the values to the beginning (Exponential MA). Latter has a problem of MA calculated on different length of data may not be the same.

5 Data Description

In this study we used solely real-life data. No synthetically generated time series or processes [6] have been used. The aim of this paper is to empirically evaluate large number of MAs on large-scale real-life data. We used daily Stock data, ETF data, Futures data and foreign exchange (Forex) data. Initially we planned to use intraday data (1 min, 5 min, 60 min frequency) but we faced overnight and weekend non trading hours problem, then a big gap in the data can occur. MA in such cases may behave inadequately so we decided to restrict to daily data for the current research. We obtained time series data from Tradestation Securities. Summary of our data is presented in Table 1. For each time series we preformed experiments on all time series, where we selected best moving average for each of them.

Table 1. Summary of the data used in the analysis process

	Stocks	ETF	FOREX	Futures	Total
Instruments	630	1092	34	94	1850
Days	4524	5105	4379	5111	5117
Start	2001-01-01	1999-05-19	2001-05-21	1999-05-20	1999-05-20
End	2013-05-22	2013-05-10	2013-05-17	2013-05-17	2013-05-22
Total days	1990795	1518000	82157	282580	3873532

Stock data is a series of prices stock was traded on the exchange. Historical stock prices are adjusted at the point they pay a dividend. Next day after dividend payment historical data is moved down by the amount of dividend stocks paid. Hence some stocks that paid unusually big dividend at some point in the history may have negative price. In our study we selected the most popular and liquid stocks. We filtered stocks with highest trading volume and with the recent price above 10 USD.

Exchange Traded Funds (ETF) are instruments traded on the main stocks exchanges and representing some index or other investable assets. They cover most of the investment universe: Equity, Bond/Fixed Income, Commodity, Currency, Alternative, Inverse instruments, Leveraged instruments and Real Estate across the globe. The excellent source for more information on ETF is <http://etfdb.com/>.

Foreign exchange market (Forex) is probably the most liquid market in the world. It trades trillions of dollars every day. It is decentralised market where every broker trades separately and synchronises prices between each other in real-time. We used all major currency pairs traded by typical currency broker. The data we used came from Tradestation Securities.

Futures are vanilla derivative instruments traded in regulated futures exchanges such as CME, EUREX, ICE, etc. We included only US and European futures in this study. Future it is a contract to buy or sell specific underlying instrument at specific price at some point in the future. Futures usually have an expiration date on a monthly or quarterly basis. Hence long and continues data is composed of multiple contracts by sticking them together and adjust the difference at the sticking point - moving the history up or down depending on the price difference at the point of joining.

6 Experiments

In this paper we empirically compared various MA on real world datasets. To compare two moving averages we used Simple MA as a benchmark. We selected 5, 10, 21, and 63 periods as a benchmark periods for smoothness. These are most common periods representing a week, two weeks a month and a quarter.

$$S_n^{SMA} = \text{Smoothness}(SMA_n)$$

where $n=(5,10,21,63)$. At mentioned periods, we measured smoothness of SMA and selected other MA with the same or better smoothness. For example for Exponential MA (EMA):

$$S_g^{EMA} \leq S_n^{SMA}$$

$$g = \arg \min_g S_g^{EMA} \leq S_n^{SMA}$$

$$S_{SMA5}^{EMA} \leq S_5^{SMA},$$

then we measured their lag

$$L_{SMA5}^{EMA} = \text{lag}(EMA_g)$$

In Table 2, we present a relationship between SMA periods and periods of other MA. It can be used as a reference to select desired smoothness. This is very useful reference as authors were not able to find any literature that contains such reference. Winner selection was performed using voting. For each stock, ETF, forex or future and for each smoothness level (SMA equivalent $n=5,10,21,63$) we selected the best (smallest lag) MA as a winner. Later we counted the wins and selected the most often winning MA as a final winner for the category. More information can be seen in the Table 2. So for example, we have 1000 stocks in our database, for each stock we have 4 smoothness levels (SMA equivalent $n=5,10,21,63$) so in total we can have 4000 winners. For each out of 19 MA we selected a winner in each smoothness level.

Table 2. Corresponding periods of a MA that has similar smoothness and lag to that of SMA

No.	Title	By Smoothness				By Lag			
		P5	P10	P21	P63	P5	P10	P21	P63
1	Simple	5	10	21	63	5	10	21	63
2	Butterworth	50	62	28	59	52	55	19	57
3	Double exponential	10	19	31	49	15	32	82	277
4	Exponential	5	10	16	24	8	17	40	151
5	Hull	12	17	28	39	12	28	67	301
6	Sine weighted	4	7	11	17	3	9	23	83
7	Spencers 15 point	2	5	10	15	1	2	2	61
8	T3	4	6	9	12	5	9	19	80
9	Triangular	5	9	14	20	2	6	15	54
10	Chande's variable index	5	10	17	30	8	17	41	180
11	Weighted	6	10	16	26	8	18	40	134
12	ZERO lag	5	12	25	70	8	25	244	162
13	Geometric	5	10	22	94	5	9	23	60
14	Exponential Hull	8	14	19	29	14	30	78	342
15	Median	19	61	50	146	2	8	21	78
16	Harmonic	5	11	22	89	5	11	23	60
17	TRIX	2	3	4	5	3	5	10	39
18	Ehlers`					13	30	66	303
19	Savitzky-Golay	67	124	208	299	30	80	168	396

7 Results

We present results in the Table 3 below. Table is composed of 4 parts, each for different type of time series. Rows represent different moving averages and columns represent 4 smoothness levels equivalent of SMA $p=5,10,21,63$. The number in the table indicates how many times that moving average had smallest lag in comparison to other ones (note that smoothness is the same). Last column "Tot." Summarises win count. We sorted the list with highest win count at the top of the table.

As can be seen winning algorithms are Exponential Hull and TRIX. TRIX is the leader between stocks and EHMA is everywhere else. For Futures, Forex and ETF TRIX is the second best algorithm.



Table 3. Results of winning moving averages**Futures**

Name	P5	P10	P21	P63	Tot.
Exp. Hull	13	37	39	42	131
TRIX	31	30	16	20	97
Dbl. exp.	11	3	2	1	17
Butterwo.	0	0	2	4	6
ZERO lag	2	1	0	1	4
Exp.	3	0	0	0	3
T3	0	0	1	0	1
Weighted	0	1	0	0	1

Stocks

Name	P5	P10	P21	P63	Tot.
TRIX	270	235	287	280	1073
Exp Hull	2	277	291	231	801
Butterworth	0	2	15	101	118
ZERO lag	22	0	1	0	23
Weighted	6	0	0	0	6
Double exp.	5	0	0	0	5
Hull	0	1	2	2	5
Exp	4	0	0	0	4

Forex

Name	P5	P10	P21	P63	Tot.
Exp. Hull	12	15	16	15	58
TRIX	4	11	10	12	37
ZERO lag	7	1	2	1	11
Exp.	6	0	1	1	8
Double exp.	0	3	2	2	7
Butterwo.	0	0	0	1	1
Chande v. i.	0	0	0	1	1
T3	0	0	1	0	1

ETF

Name	P5	P10	P21	P63	Tot.
Exponential Hull	48	131	223	227	659
TRIX	52	85	81	113	331
Double exp.	68	24	22	22	136
ZERO lag	27	15	8	8	58
Exp.	19	8	4	8	39
T3	1	3	2	0	6
Butterworth	0	0	0	1	1

8 Conclusions

In this paper we compare 19 the most popular moving averages used in practical trading and determine the most suitable according to the criteria “smoothens vs. lag ratio”. We performed large-scale study by testing all the MAs on 1850 real-world daily time series from following domains: Stock, ETF, Futures and Forex. We compared all MA at 4 different smoothness levels equivalent of a simple MA 5, 10, 21 and 63 days and selected the best one for each category and each time series. Finally we counted which one won most of the time. Two best moving averages identified: Exponential Hull Moving Average (EHMA), next followed by a Triple Exponential Moving Average (TRIX). EHMA uses a correction term to reduce lag and is different in that from TRIX. Correction term subtracts older history to reduce lag of the moving average but introduces “overshooting” behaviour in trend reversals. For stocks TRIX showed the best results as stocks tend to be more volatile and have frequent trend reversals where correctionless MA is more accurate. For all other time series EHMA was the winner. All other methods are far behind the two winners.

We also created a reference table where we link different moving averages to the smoothness of Simple MA. The other table references lag to a SMA lag. This can be used by practitioners trying to replace one MA with other one with the same lag or the same smoothness.

For the future work we already did preliminary research. Our aim is to create a tailor-made moving averages for specific time-series that would have lowest lag for a given level of smoothness. We plan to create two versions, one with positive only weights and other with positive-negative weights (i.e. with correction term). We estimate that

tailor-made moving average will have better smoothness and lag characteristics than current winners EHMA and TRIX. Our inference supports conclusions in [17] where classification accuracy was used as a performance criterion. We also plan include other criteria in the analysis: forecasting accuracy, maximum profit (in case of trading system), minimum risk or similar criteria.

Acknowledgments. This work was supported by the Research Council of Lithuania under grant MIP-018/2012.

References

1. Hamilton, J.D.: Time series analysis, vol. 2. Princeton University Press, Princeton (1994)
2. Tan, Z., Quek, C., Cheng, P.Y.K.: Stock trading with cycles: A financial application of ANFIS and reinforcement learning. *Expert Systems with Applications* 38(5) (2011)
3. Perry, J.: Kaufman, New Trading Systems and Methods, 4th edn. John Wiley & Sons (2005)
4. Ni, Y.-S., Lee, J.-T., Liao, Y.-C.: Do variable length moving average trading rules matter during a financial crisis period? *Applied Economics Letters* (2012)
5. Marques, N.C., Gomes, C.: Maximus-AI: Using Elman Neural Networks for Implementing a SLMR Trading Strategy. In: Bi, Y., Williams, M.-A. (eds.) KSEM 2010. LNCS, vol. 6291, pp. 579–584. Springer, Heidelberg (2010)
6. Ruseckas, J., Gontis, V., Kaulakys, B.: Nonextensive Statistical Mechanics Distributions And Dynamics of Financial Observables From The Nonlinear Stochastic Differential Equations. *Advances in Complex Systems* 15(suppl. 1) (2012)
7. Jurgutis, A., Simutis, R.: An investor risk profiling using fuzzy logic-based approach in multi-agents decision support system. In: *Proceedings of the 17th International Conference on Information and Software Technologies*, Kaunas (2011)
8. John, E.: *Cybernetic Analysis for Stocks and Futures*, pp. 213–227. John Wiley & Sons (2004)
9. John, E.: *Rocket Science for Traders*, 245 pages. John Wiley & Sons (2001)
10. Kirkpatrick, C.D., Dahlquist, J.R.: *The Complete Resource for Financial Market Technicians*, pp. 39–50. Financial Times Press (2006)
11. Tillson, T.: Smoothing Techniques For More Accurate Signals. *Stocks & Commodities* 16, 33–37 (1998)
12. Hull, A.: Hull moving average, http://www.justdata.com.au/Journals/AlanHull/hull_ma.htm
13. John, E.: *Cybernetic Analysis for Stocks and Futures*, pp. 213–227. John Wiley & Sons (2004)
14. John, E.: *Rocket Science for Traders*. John Wiley & Sons (2001)
15. Person, P.-O., Strang, G.: Smoothing by Sawitzky-Golay and Legendre filters, <http://persson.berkeley.edu/pub/persson03smoothing.pdf>
16. Ellis, C.A., Parbery, S.A.: Is smarter better? A comparison of adaptive, and simple moving average trading strategies. *Research in International Business and Finance* 19(3), 399–411 (2005)
17. Skurichina, M.: Effect of the kernel functional form on the quality of nonparametric Parzen window classifier. In: Raudys, S. (ed.) *Statistical Problems of Control*, vol. 93, pp. 167–181. Institute Mathematics and Informatics, Vilnius (1991) (in Russian)

Knowledge Transfer in Management Support System Implementation

Bartosz Wachnik

Warsaw University of Technology, Faculty of Production Engineering, Warsaw, Poland
bartek@wachnik.eu

Abstract. Knowledge transfer during management support system implementations is one of the key elements of project success. The scope of this article is to present the results of research on the methods of completing knowledge transfer for specific implementation phases in selected groups of IT projects consisting in implementing ERP, CRM, BI and DMS class IT systems. The results may be interesting for researchers specialising in the subject of IT project implementation and for practitioners completing IT projects.

Keywords: knowledge transfer, project, management support systems.

1 Introduction

Knowledge transfer during management support system implementations is one of the key elements of project success. Identifying the conditioning for effective knowledge transfer within the mutually acceptable transaction amount is an important activity area for practitioners and theoreticians. The author's research concentrates on knowledge transfer from external consultants implementing projects both to key and end users of ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), BI (Business Intelligence) and DMS (Document Management System) class management support systems. Within the completed research, the following areas of conditioning are analysed: organisation of trainings, the use of technological tools, knowledge transfer methods depending on the completed project phase. The research combined two methods. At the first stage, a questionnaire method was used, extended by expert workshops and at the second stage, the action research method was applied. The initial chapters discuss the role of knowledge transfers in IT projects, research assumptions and the methods used. Then, research results are presented, which may contribute to better understanding of suppliers' and customers' logics of action in completing a task as important as knowledge transfer during a management support IT project.

2 The Role of Knowledge Transfer in Management Support System Implementation

The project of adapting a standard ERP, CRM, BI or DMS management support IT pack is an activity limited by time and financial budget, undertaken in order to create

a unique configuration which, supporting enterprise strategy, will allow the company to achieve a temporary competitive edge. The final product of this project will be a programmed and configured information system, functioning within a specific range, not only supporting the users in completing planning functions, filing and reporting economic events in the company, but also catalysing changes in the enterprise's functioning.

Implementation methodology covers the methods of organising the implementation team, dividing the project into separate tasks and milestones, defining responsibility for specific implementation areas, recommendations linked to the way of completing project tasks, among other things knowledge transfer, and recommendations linked to the completion of technological tasks in an information system. Analysing management support system implementation methods recommended by the producers of ERP, CRM, BI and DMS applications, we can identify the following phases that are most crucial to the project: training key users, functional analysis, system adaptation and system configuration acceptance testing and its possible regulations, developing post-implementation documentation and process work instructions, training end users and system launch. Completing a management support system implementation project, it is worth noting that knowledge transfer between consultants and employees is a significant component of knowledge management [6] and a significant success factor in IT project realisation [2]. Subject literature [7] divides knowledge transfer into two categories:

- Codified, where knowledge transfer applies to formalised knowledge included in project documentation and programming.
- Personalised, where interaction between training participants and their experience is what is important.

Other classification [1], based on the criterion of knowledge absorption method, divides knowledge transfer into the following categories:

- Exploration-oriented. This category is dominated by inductive reasoning, i.e. a bottom-up approach, learning based on reasoning and testing the system, learning from mistakes and learning through teaching others.
- Instruction-oriented. This category is dominated by deductive reasoning, i.e. a top-down approach, learning through planned studies of complete materials and available study cases.

As part of the research conducted, the author proposed the following categorisation of knowledge during its transfer.

Formal, found in the provided documentation of an IT system or in the completed project documentation, e.g. system settings documentation. This type of knowledge is strongly codified and based on knowledge included in documentation and knowledge repositories.

Informal, based mostly on a specialist's professional experience and their particular skills of transferring this knowledge, seen for example in creating simple descriptions of difficult problems or finding examples clearly illustrating a given issue. In

this case, it is very difficult to separate the specialist's expertise from their unique skill to explain some difficult problems.

The research will analyse the combinations of existing knowledge transfer categories, i.e. exploration-oriented, instruction-oriented and the categories of knowledge, i.e. formal and informal.

3 Research Assumptions and the Methods Used

The choice of subject matter stemmed from the belief that the issues linked to knowledge transfer during management support system implementation have a significant impact on the realisation of project goals despite a lack of unequivocal agreement amongst researchers [4]. Article research goals are:

- Identifying the conditioning for effective knowledge transfer in IT projects consisting in the implementation of ERP, CRM, BI and DMS class IT systems.
- Recommendation of assumptions concerning effective data transfer in IT projects consisting in the implementation of ERP, CRM, BI and DMS class IT systems.

The research was conducted in companies located in Mazovia in the period between 2009-2012. The enterprises qualified for the study by meeting the following criteria: a number of employees between 80 and 100, own IT department, a minimal income of 40 mln zloty.

The enterprise group included companies with both Polish and foreign capital and a wide autonomy in their IT strategy realisation. The selected companies achieved good or average results in their industry – so they are neither leading nor marginal companies. An important methodological assumption was reaching people directly or indirectly engaged in the implementation of management support IT projects. The respondents were company owners, directors, members of the board, financial directors or IT directors. The study combined two scientific methods defined in the following way:

Phase 1. Conducting questionnaire research amongst 20 enterprises that completed 28 IT projects in the 2009-2010 period. The questionnaire research consisted in posing one question concerning problems encountered due to knowledge transfer during the completion of four project phases and conducting analytical workshops with chosen respondents. IT projects analysed consisted in implementing an ERP system - 15 projects, a CRM system - 5 projects, a BI system - 3 projects and a DMS system - 5 projects. On the basis of conclusions, recommendations for changes in knowledge transfer organisation during management support system implementation projects were devised.

Phase 2. Having completed the questionnaire research, the author carried out an action research. The method used promotes forming theory through practice and speculations how to improve the practice. The author used this method to contribute to the existing knowledge and help solve practical problems. Action research is a useful way of conducting research for practitioners who work on improving the understanding of

their activity. In the proposed study, the recommended changes concerning knowledge transfer were implemented in nine chosen projects of management support system implementation carried out in 2011 and 2012. The modification of data transfer realisation method was aimed at increasing the effectiveness of completing this task with the assumption of bearing the optimal cost linked to completing this tasks as part of a specific methodology milestones. On this stage, the following number of specific system implementations was completed, accounting for recommendations from Phase 1: ERP – 4, CRM – 2, BI – 1, DMS – 2. After the completion of the above mentioned projects, questionnaire research was carried out again, revealing changes in the evaluation of data transfer by project groups in companies implementing IT systems. Thus, conclusions will serve as practical recommendations for swift and effective knowledge transfer organisation during all the phases of management support system implementation methodology in all the countries.

4 Study Results

28 IT projects, analysed at the first stage of the research, had the following characteristics:

- Average implementation project budget – EUR 63 400.
- Average project realisation time – 4 months.
- Average number of key users in customer project group – 3 people.
- Average number of end users in customer project group – 21 people.
- Average percentage of expenditure linked to the completion of tasks concerning formal knowledge transfer (e.g. trainings, completing documentation) on the scale of implementation budget – 19%.
- Average percentage of expenditure linked to project management on the scale of the whole implementation budget – 12%.
- Average percentage of expenditure linked to knowledge transfer during the first year after system launch on the scale of implementation budget – 9%. The expenditure was mainly linked to the purchase of trainings for new system users.
- 43% of projects ended according to schedule and budget, and the goals have been met; 18% of projects ended according to schedule, the budget has been exceeded and the goals have been met; 21% of projects did not end according to schedule and budget, but the goals have been met; 18% of projects did not end according to schedule and budget and the goals not been met.

Among 28 studied projects, 93% of respondents pointed out after the implementation that knowledge transfer is, apart from a correctly performed functional analysis, the most crucial task in a management support system implementation project. Table 1 presents the percentage distribution of respondents' answers to the question about problems identified by them during the completion of knowledge transfer in four phases of management support system implementation project.

Table 1. Research results presenting the problems identified during knowledge transfer during four phases of a management support system implementation project

Training key users		Functional analysis		System adaptation, acceptance tests including regulation.		Project documentation		Training end users	
Lack of process approach to the functionality.	82 %	Lack of support in identifying risk factors by key users.	68 %	Insufficient way of transferring knowledge concerning acceptance tests and regulation by key users.	75 %	Lack of post-implementation documentation.	89 %	Lack of process approach to the functionality.	93 %
Wrongly configured training database, resulting in a lack of possibility to practise the completion of business processes.	64 %	Lack of expert proposals for industry-specific functional solutions.	64 %	Lack of test scenario patterns required by keys users.	64 %	Process work instructions describe functionalities and processes in the information system in too general terms.	32 %	Lack of sufficient exercise during trainings.	71 %
Lack of experience transfer in using an IT system for industry-specific problems.	57 %	Functional analysis documentation is too general.	57 %	The level of key users' engagement in testing the configured system was too low.	46 %	A limited range of subject work instructions, not matching the range of the implemented functionality.	25 %	Incorrectly adjusted training database.	50 %

Table 1. (continued)

Lack of documentation for the standard system version.	54%	Insufficient knowledge transfer in functional analysis performance by the external consultant.	32%			Process work instructions developed by external consultants adapted to a standard information system, not to the customer's needs.	18%	Mistakes appearing in the configured system result in a lack of fluidity in the completed information system exercise.	43%
Low communication skills of the trainer.	43%	Lack of project documentation that increases the effectiveness of obtaining customer data and information to structure data for analysis.	18%					The external trainer's explanations of the issues were too complicated.	25%
External consultants lack knowledge about the customer's business activities, resulting in a lack of practice in the completion of system business processes.	43%	External consultants unwilling to transfer full and complete system configuration knowledge.	14%					The number of end users participating in one training was too high.	14%

Source: Own study

Analysing the characteristics of the projects and collected answers points us towards the following conclusions:

- A grave error committed during formal and informal knowledge transfer during trainings was discussing system functionalities in general terms, without referring to any specific business processes in the enterprise. It is worth underlining that end users stressed that they had been informed how to generally implement a given system function but without getting details necessary to complete a specific business process implemented in their enterprise. Additionally, trainers did not spend enough training time on exercises dedicated to completing processes within the system. It is worth noting that during the trainings end users encountered errors in insufficiently tested systems which resulted in impeding the transfer of formal and informal knowledge aimed at exploration.
- Insufficient transfer of formal and informal knowledge oriented mostly at exploration in the phase linked to system adaptation, acceptance tests and system regulation. What is important in the category of knowledge transfer aimed at exploration, is the independent work of the key consultant, who should absorb knowledge through: performing many system tests based on test scenarios they completed, reasoning and teaching others. Respondents have pointed out that the interest of key users in the process of testing the configured system was too low, which resulted in superficial testing. The main reason was a conflict between performing everyday duties at work and tasks linked to the information system implementation project.
- Insufficient transfer of accumulated expert knowledge including formal and informal knowledge concerning the functionality design in a given industry, especially in the two first project phases, i.e. training key users and the functional analysis. Additionally, a standard system did not have the functionality specific for a given industry. Consultants could not recommend system configurations for particular business processes, e.g. in the area of logistics and production, that could entail a process innovation and thus a competitive advantage.
- Lack of sufficient formal and informal knowledge transfer within project management that would present basic project management components, i.e. cause and effect relationship between project phases, risk management policy, documents used during project implementation, e.g. functional analysis template, risk log template, test scenarios template.
- Within the completion of four project phases, formal knowledge transfer problems occurred. Implementation companies did not provide information system documentation, i.e. the description of standard functionality in Polish or they did not perform project documentation as part of implementation – i.e. post-implementation documentation and process work instructions. During in-depth analytical workshops, the respondents pointed out that some foreign software producers did not have a Polish version of standard package documentation. The respondents indicated that in order to lower the project price during contract negotiations, they had given up on designing project documentation, i.e. a post-implementation project description of system configuration and process work instructions.

Based on the presented conclusions, the author developed the following recommendations that were implemented in the second phase of the research.

- The customer (information system recipient) should organise working time for key users and motivate them so that they are able to absorb formal and informal knowledge through exploration-oriented approach that demands greater intellectual engagement and dedication, and in most cases more time than the instruction-oriented approach.
- The customer should develop and hand over a list of business processes to the system supplier that would serve as a basis for training key users. Trainings should show how to complete a given business process and discuss the parameters of system configuration. Trainings for key users should be carried out exclusively on the basis of process approach, i.e. discussing the implementation of specific system processes. Trainings should be conducted on the basis of a tested database that has been successfully used for acceptance tests of all the implemented processes. On the basis of respondents' suggestions, trainings for end users should be divided in a proportion of 20/80, 20% of trainings being lectures and 80% trainings linked to the implementation of specific business processes.
- The customer should choose a software supplier who is experienced in implementing systems in a given industry or who offers vertical solutions dedicated to that given industry. Currently, an increasing number of ERP, CRM, BI and DMS IT packs expand them with additional functionalities required by chosen enterprise groups. Thus, suppliers offer a preliminarily adapted IT pack implemented by consultants specialising in a given branch.
- The customer should organise a formal knowledge transfer through completing documentation of adapted system configuration and process work instructions. Documentation may be completed by a key user as an additional task accompanying acceptance tests. Completing documentation by key users is an example of an informal knowledge transfer through the exploration-oriented approach.
- The customer should make sure that the project manager has executed the following provisions between the supplier and the customer of the completed project, which may be defined by contractual provisions regulating the completion of implementation services:
 - Guaranteeing a formal knowledge transfer within the scope of project management, especially implementation methods.
 - Training participants evaluate the trainer's performance on the basis of a questionnaire. A positive evaluation may result in accepting the completion of a given task or phase.
 - The maximum number of users participating in trainings is 12.

The second phase of research consisted in implementing recommendations in 9 IT projects and completing questionnaire research that indicated mistakes and problems in knowledge transfer in each of the four projects phases. The respondents implemented recommendations during: the selection of a standard software package, the selection of implementation service provider, drawing up a contract defining the completion of implementation services, the organisation of knowledge transfer during

the implementation, project management organisation. 9 IT projects analysed in the second phase had the following characteristics:

- Average implementation project budget – EUR 59 500.
- Average project duration – 4 months.
- Average number of key users in the customer project group – 3.
- Average number of end users in the customer project group – 28.
- Average percentage of expenditure on the completion of tasks concerning formal knowledge transfer (e.g. trainings, documentation design) on the scale of implementation budget – 15%.
- Average percentage of expenditure on project management on the scale of the whole implementation budget – 9%.
- Average percentage of expenditure on knowledge transfer within a year after system launch on the scale of the implementation budget – 5%. The expenditure was mostly linked to the purchase of trainings for new system users.
- 56% of projects ended according to schedule, budget and met the goals; 22% of projects ended according to schedule, the budget was exceeded and the goals were met; 11% of projects did not end according to schedule and budget but the goals were met; 11% of projects did not end according to schedule and budget and the goals were not met.

Below, research results concerning the incidence and character of problems identified in knowledge transfer, in four phases of management support system implementation project, are presented.

- Training key users phase:
 - Lack of documentation for the standard system version – 67% of cases
- Functional analysis phase:
 - Lack of branch-specific expert proposals for functional solutions in an IS from the consultant conducting functional analysis despite branch-specific functionalities in the system – 33% of cases
- System adaptation and acceptance test (including regulation) phase:
 - The level of key users' engagement in the process of configured system testing was too low – 33% of cases
- Project documentation phase:
 - The level of key users' engagement in developing post-implementation documentation and process work instructions was too low – 44% of cases
- Training key users phase:
 - The external trainer's explanations of the issues were too complicated – 56% of cases

5 Final Conclusions

The study results from the period 2010-2012 presented in this study let us formulate the following key conclusions. First of all, 93% of respondents indicated from the expert perspective that knowledge transfer, apart from a correctly performed functional

analysis, is the most important project task during a management support system implementation. This declaration is confirmed by in-depth ex-post research, presented in Table 2.

Table 2. Comparative analysis of implementation evaluation during the first and second phase of study

Characteristics of studied projects	Projects analysed in Phase 1.	Projects analysed in Phase 2.
Projects ended according to schedule, budget and met the goals	43%	56%
Projects ended according to schedule, the budget was exceeded and the goals were met	18%	22%
Projects did not end according to schedule and budget but the goals were met	21%	11%
Projects did not end according to schedule and budget and the goals were not met	18%	11%

Source: Own study

In light of the study results, it is important to design a coherent method of performing knowledge transfer in order to optimise the cost of completing this task in specific project completion phases. It is noteworthy that the knowledge transfer cost during four phases in the first study phase amounted to 19% and after the implementation of recommendations 15%, with a simultaneous decrease of expenditure linked to knowledge transfer after system launch on the scale of the whole implementation budget from 9% in the first phase to 5% in the second phase.

Secondly, the organisation and the quality of knowledge transfer depends on the system provider's direct experience in knowledge transfer, the implementation methods used and the potential delivery of a vertical system that can provide the system with branch-dedicated functionalities. Implementation consultants' expertise and implementation experience from similar enterprises are particularly noteworthy, as they will help develop a better suited system configuration, which will increase the chances of achieving a higher process innovation.

Thirdly, it is important for the customer to have a guarantee of a formal knowledge transfer linked to the documentation of the information system they purchased. Additionally, it is important for the transaction of implementation services purchase to cover the completion of post-implementation documentation and optional performance of process work instructions. If the customer does not decide to purchase services linked to the performance of process work instructions, they should ensure their completion by using their own resources, e.g. key users.

Then, it is necessary to ensure that trainings are focused on learning the completion of processes, with the prevalence of exercise and completion of specific system tasks, not in the form of lectures describing available system functionalities in general terms.

Finally, the study has shown that in order to achieve a high level of knowledge transfer, key users from the project group have to have a high level of engagement in gaining knowledge in each project phase. To achieve that, it is important for managers to boost key users' motivation and organise their work to minimise the conflict between their current, everyday duties and the implementation tasks.

To sum up, we need to stress that there is a competence gap amongst theoreticians and practitioners of business informatics when it comes to the questions of knowledge transfer during management support IT project completion. The competence gap means: a limited number of publications containing practical tips and describing experiences, a lack of expert advice concerning the organisation of knowledge transfer as part of implementation methodology proposed by producers and a lack of awareness among knowledge transfer decision makers. The author hopes that the study results presented in the publication will help achieve two goals, i.e. they will show the specific character and the role of knowledge transfer during management support IT projects.

References

1. Bostrom, R., Olfman, L., Sein, M.K.: The Importance of Learning Style in End-User Training. In: Emery, J.C. (ed.) *MIS Quarterly* 1990, vol. 14(1), pp. 101–119. Society for Information Management and The Management Information Systems Research Center Minneapolis, Minneapolis (1990)
2. Gallivan, M.J., Spitler, V.K., Koufaris, M.: Does Information Technology Training Really Matter? A Social Information Processing Analysis of Coworkers Influence on IT Usage in the Workplace. In: Zwass, V. (ed.) *Journal of Management Information Systems* 2005, vol. 22(1), pp. 153–192. M.E. Sharpe, Inc., Armonk (2005)
3. Gill, J., Johnson, P.: *Research Methods for Managers*. Sage Publications, London (2002)
4. Haines, M., Goodhue, D.: Implementation Partner Involvement and Knowledge Transfer in the Context of ERP Implementation. *International Journal of Human-Computer Interaction* 2003 16(1), 23–38 (2003)
5. Koskinen, K.: Knowledge Management to Improve Project Communication and Implementation. *Project Management Journal* 2004 35(2), 13–19 (2004)
6. Kumar, J.A., Ganesh, L.S.: Research on Knowledge Transfer in Organisations – a Morphology. *Journal of Knowledge Management* 2009 13(4), 161–174 (2009)
7. Lech, P.: Knowledge Transfer Procedures From Consultants to Users in ERP Implementation. In: Turner, G. (ed.) *The Electronic Journal of Knowledge Management* 2011, vol. 9(4), pp. 318–327. Academic Publishing International Ltd. (2011)

Collective Intelligence Utilization Method Based on Implicit Social Network Composition and Evolution in the Scope of Personal Learning Environment

Genadijus Kulvietis¹, Andrej Afonin², and Danguole Rutkauskiene²

¹ Vilnius Gediminas Technical University, Vilnius, Lithuania
genadijus_kulvietis@gama.vtu.lt

² Kaunas University of Technology, Kaunas, Lithuania
{andrej.afonin,danguole.rutkauskiene}@ktu.lt

Abstract. Personal Learning Environment (PLE) is an emerging concept in learning technology field. PLE allows users aggregate content from distributed Web 2.0 services in one place and arrange it in a way that is convenient for a learner. Despite the fact, that PLE operates with social software and is a type of social media, social networking component is used very poorly. A new model of “networked knowledge” utilization in the scope of PLE is presented in this paper. The exclusive feature of this model is learners’ aggregated and generated data analysis and digital identity development based on both sources. Another exclusive feature is constant digital identity update, depending on constantly evolving learners’ implicit network. Such evolution ensures continuous implicit network update along with changing learners’ interests.

Keywords: Personal learning environment, social networking analysis, virtual communities, social software, Web 2.0.

1 Emerging Concept of Personal Learning Environment

The notion of Personal Learning Environment (PLE) appeared as a result of discussion among experts in different fields regarding the future of Virtual Learning Environments, which are the main tool for nowadays learning support [1]. Virtual Learning Environments were seen as a fenced garden without any connection with other virtual environments, which are used by students for information collection and results dissemination [2]. On the opposite, Personal Learning Environments were rather seen as platforms for content aggregation from different contexts where learning takes place, such as home, workplace or educational institution [2]. However, there is still no commonly accepted definition of what is a PLE. There are several groups of researchers that have different vision of what PLE is.

Some researchers see a PLE as a predefined set of software tools, which are used by learners to organize their learning process. Thus, Mark van Harmelen from Manchester University defines PLE as a single learner’s e-learning system, which

provides access to different e-learning resources and/or personal or virtual learning environments used by students and teachers [3]. Other researchers use PLE as a metaphor to describe modern student's online activity and environment. Graham Attwell's definition of PLE refers not only to software tools, but also to peripheral devices, that could ensure learning continuity outside the institution boundaries, such as mobile phones, laptops or portable music players [4]. Despite the fact that explicit definition of PLE is still under consideration, still a common feature could be highlighted – personal learning environment passes the control of learning process to the learner himself.

PLE design and implementation is a topic of hot discussions as well. Nial Scatter [5] distinguishes researchers to three groups with their own perspectives and functionality vision. According to the first group, PLE has to be implemented as a desktop application and serve as intermediate node between learner and online services [6]. In their perspective, PLE is a learner's owned software application, which communicates with distributed educational web services and databases on service oriented bases. The second initiative group's vision is that the PLE construction is based only on Internet browser, using either separate online services, or integrated online environments, that aggregates different kind of information from distributed, mostly Web 2.0, services, such as blogs, wiki, social bookmarking, multimedia sharing and others services, that enable students' collaboration and organizational activities. This group has most successors. Third group of researchers state that personal learning environment is not only a piece of software, but the complex infrastructure, which combines both software applications and distributed web services and technical equipment, and the main goal is to propose suitable teaching and learning methods for successful infrastructure exploitation and focus more on use cases and learning scenarios [4], [7].

It is important to mention that differently than virtual learning environment PLE is seen as environment that leads person through all his life. It means that learner's social network is going to be different at different periods of time and is constantly changing along with learning topics and interests. Learner's social network evolution together with leaning preferences change has to be taken into consideration.

PLE is a self-directed and self-controlled learning environment with social media background, which aggregates information from distributed, mostly Web 2.0, services, and allows organizing received information in the way that matches the learner's needs in the most sensible way.

This paper presents a new approach on how to use Web 2.0-based collective intelligence in the scope of PLE. The definition of PLE has been introduced earlier on; next, a brief introduction to social media and hidden social structures in the background, types of relationships and their building principles. Section three introduces proposed method and explains its working principles in detail. Section four describes how the method was implemented in practice and shows the results that have been achieved. A case study finalizes the whole article.

2 Social Media Connections and Their Roles

The previous analysis of PLE concept unveiled that due to its nature, PLE is a type of social media. In order to understand the nature of ongoing processes, an analysis on social media is required. There are two major types of social media [8]:

1. Social networks
2. Online communities

In the beginning it's important to clear out what are main differences between these social structures. Everyone has their social networks (whether online or offline) (Fig. 1 A)). Social networks consist of friends, family, co-workers and people they are acquainted with. Social networking sites didn't create social networks. They are simply making these networks visible and help maintaining them. The most important difference between social networks and online communities is how people are held together on these sites. People are held together by pre-established interpersonal relationships, such as classmates, friends, co-workers, etc., on social network sites. Connections as these are made to last. People join social networking sites to maintain old relationships and establish new connections as well [8].

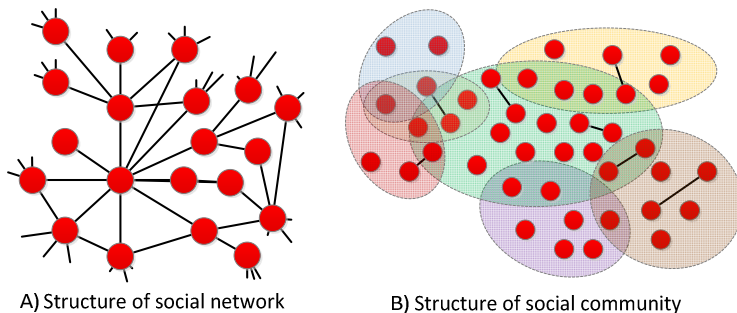


Fig. 1. Social media structures

Unlike social networks, communities are held together by common interests (Fig. 2 B)). It can be a mutual hobby, a common project or a goal, the way of life or a profession. People participate in online communities, because some members feel they can contribute to the community with their experience, while others feel they can benefit from being there. Being a part of online community doesn't require pre-established connections; actually it even doesn't require any interpersonal interaction. It is common for an individual to be a part of more than one community. Moreover, communities can overlap and are often nested [8].

Examples of the structure of social network and online community are presented in Fig. 1. Individuals are shown as red nodes in these pictures, and lines between those nodes represent relationships, that people establish between each other. However, the nature of these relationships is slightly different. A relationship in social network indicates, that two people are members of the same social structure and have established connection there: it could be family, friends, co-workers, etc. However,

it's hard to say without additional metrics, on how useful this relationship is to both sides, how strong it is, it is constant or happened only once. Relationships in online communities, on the other hand, are built in the same field and are linked by the same interest. Relationships of such type are more relevant for educational purposes as they represent the same field of interest.

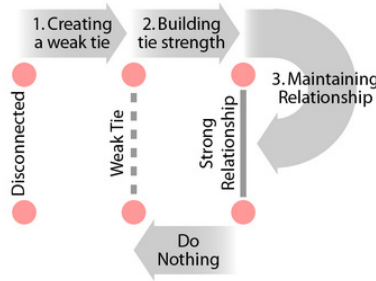


Fig. 2. Lifestyle of Relationship [9]

Every relationship has similar lifecycle. The lifecycle of every relationship consists of three stages (Fig. 2):

- 1) creating the weak tie: the first step of any relationship;
- 2) building up the tie strength: transformation of weak ties into strong relationships;
- 3) maintaining the relationship: preventing strong relationships from eroding and reverting back to weak ties [9].

A weak tie could be created both in social networks and in online community. The formation of weak ties between two people depends on their desire to connect, the amount of communities they share in common and the network distance between them. But tie strength predominantly is built in communities. What builds strong relationship within communities is the combination of frequent engagements, deep interactions, and the time spent together. If relationships are well developed, they become a part of person’s social network. So, communities are needed for transforming weak ties into strong ones, and social networks are for maintaining and sustaining these relationships [10][11].

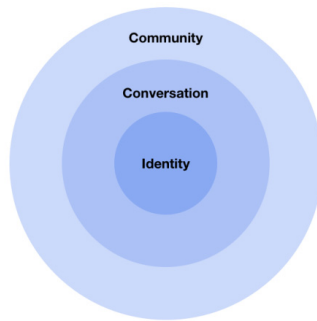


Fig. 3. Social Design Diagram by E. Fisher [12]

An approach of utilizing personal social network is proposed by Facebook social evangelist Eric Fisher and is called Social Design Strategy [12]. Social design consists of three core components: identity, conversation and community, in other words, the person himself, the other people and the conversations between the person and the other people. In the diagram (Fig. 3), identity is put to the center, conversation is in the middle and community is on the outside. Conversation is a media that serves as glue between the identity and the community. The conversation is the way people express their identities to the community and receives feedback from it.

Fisher [12] proposes to start from the center and work the way out, during the process of designing a social product. That is, to allow people to create their identity, talk about it and build community over the time.

However, over the time, he proposes to take the reverse approach and work from the outside in. That is to utilize the community, define new types of conversations and to perform further identity updates.

3 “Identity-Network-Proposal” Model

The analysis of relationship development in social networks and online communities, as well as analysis in social design strategy allows defining general a model of collective intelligence utilization. A general method is to construct digital identity, create weak ties with other members, turn weak ties to strong relationships and maintain these relationships. This section presents the potential of proposed method of PLE’s collective intelligence for hidden network composition and its’ further utilization for learning purposes.

The problem, in the scope of PLE, is weak ties establishment and their conversion into strong relationships, as PLE is a single persons’ environment. Nevertheless, PLE by its design nature aggregates data mainly from distributed Web 2.0 services, meaning that social network or community could be established on distributed services side. Proposed model allows overcoming this shortage and using collective intelligence potential, accumulated in social software services, in a scope of single person’s environment.

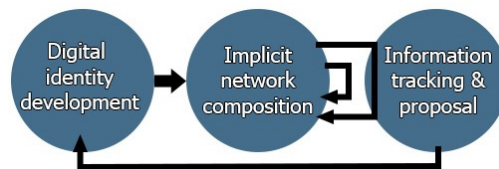


Fig. 4. Identity-network-proposal model

General model (Fig. 4) working principle is as following: the first step is to develop the digital identity of the person. In order to do that, the method proposes to separate and analyze 2 sources of information: users’ aggregated content (source of knowledge) and users’ generated content (reflection on learning process). Similar

digital identities are created to all PLE platform users. The map of digital users' identities is created after the first step. The second step is finding users with similar digital identities and mapping them to each other. This step composes artificial communities that are based on users' interests, thus creating weak ties between users. The third step is turning weak ties into strong relationships. In order to do this, users are prompted with other users' operated content. A constant monitoring of user's activity is performed and logged. If users get interested in proposed content (clicks proposed data for further information, adds to favorites, etc.) the weak tie between these two users is labeled as strong relationship. At the same time, users' digital identities and connections between them are updated with new information.

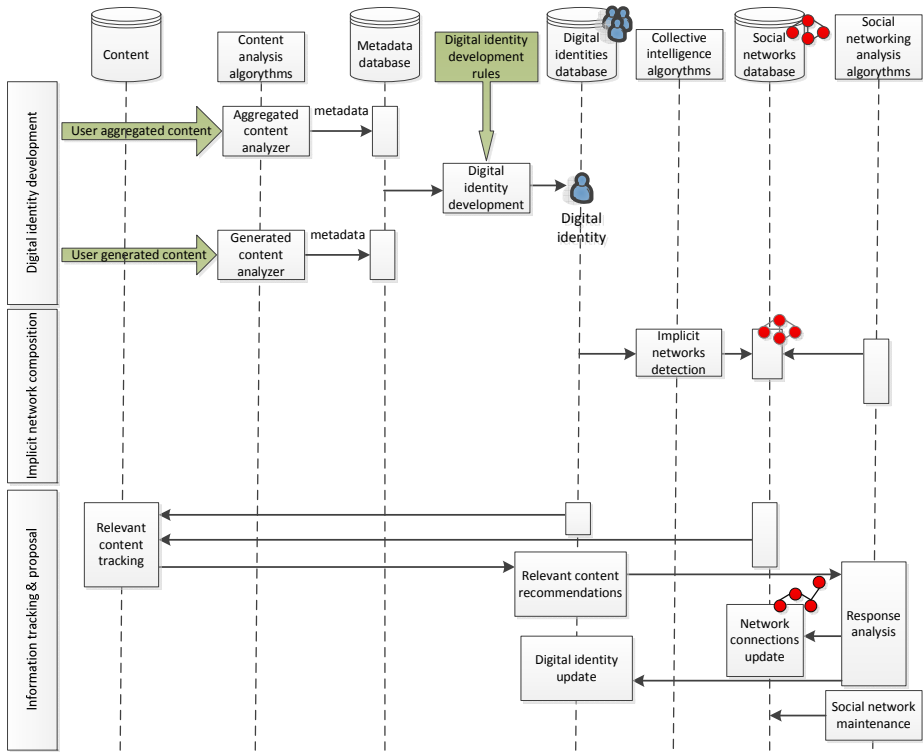


Fig. 5. Identity-network-proposal model

A more detailed model analysis is presented below. Model graphical representation is shown in Fig. 5. As it is presented, model is divided to three main steps. First step is users' digital profile development, based on his learning material preferences. Proposed method distinguishes two types of content: learners' consumed data, which is used for learning and produced material, which represents reflection on gained knowledge.

Next step is to use these digital profiles to find similar users, and to build kind of online communities according to users' similarities (but still no user interaction

between each other). And the last step is to use these communities to update its members with relevant information, which is created by other network members. At this stage online communities are turned into implicit social networks according to interactions that are made between these users. This stage is responsible for keeping up to date networks connections relevance. At the same time learners' digital profile is being constantly updated, thus ensuring persons' profile evolution according to his changing learning preferences.

Step 1. Digital identity development.

Main source of users' information is distributed Web 2.0 services, aggregated in the scope of PLE platform [13]. Every aggregated Web 2.0 service item usually comes with metadata that is called tags. Tagging is an inexpensive and easy way of using the wisdom of the crowd and making resources visible and sortable [15].

Tag is a metadata about the element that allows working with service data in a more convenient way. A set of separate tags is called a tag cloud. Usually, tags in a tag cloud are visualized in different sizes, meaning that the tag with bigger size was used more often. The structured list view (Fig. 6) with tags and their usage density shows a clear picture of users' interests. At this stage, all platform users are merged to common matrix with used keywords and their usage rate (Table 1).

Table 1. Common Users' Interest Matrix

	User1	User2	User3	User4	User5	User6
web 2.0	6	5	1	2	6	3
education	3	4	9	5	3	2
technology	5	6	8		5	3
software	6			2		
.net	5					
learning		4	2			
python		4				
management			6			
hr				4		
programming					5	2

In order to develop a more explicit users' profile, the presented method proposes the usage of two types of metadata. The first type of the metadata comes in a form of tags pinned to Web 2.0 services elements.

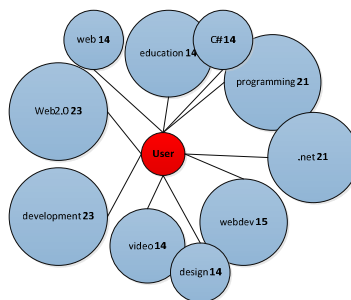


Fig. 6. Tag list

Generally, it is users' aggregated content: links from social bookmarking services, podcasts, vodcasts and "youtube" type videos, and other structured information. In learning processes this is kind of data, that learner consumes for building his knowledge base. Second part of the metadata comes from users' generated content. This is data, that learner produces as a reflection of his learning process. At this stage, the method proposes to analyze and extract metadata from users' reflections on learning activities that they post in their blogs and wikis. The aggregated type of metadata corresponds to knowledge gained during the learning process. As reflection information comes as a text (blogs, wiki, etc.) this information is analyzed and another set of metadata is generated. Both types of metadata (aggregated and generated) are combined in a common user's interest matrix with the same weight (Table I). This weight could be set to different value, thus defining which type of data, either consumed or produced, gives more authority building complex digital identity. Metadata weight ratio is defined by digital identity development rules. At this stage all users' digital profiles are stored to digital identities database and later are used to find learners with similar identities.

Such approach allowed defining more explicit user's profile, which not only combines consumed, but also created contents, that correspond gained knowledge and reflection during the learning process.

Step 2. Implicit network composition.

This stage is responsible for the weak ties composition. At the beginning there is no activity between users, thus there is no possibility to define these ties upon their actions. Therefore, weak ties between the users are defined using collective intelligence algorithms. In this case, an algorithm is used, that calculates Pearson correlation (1) [17] between all users.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad (1)$$

The result of calculations is shown in Table 2.

Table 2. Pearson Corelation Matrix

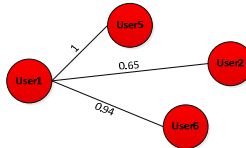
	User1	User2	User3	User4	User5	User6
User1	1	0.6546	-0.826	-1.0	1.0	0.9449
User2	0.6546	1	0.1705	-1.0	0.6546	0.866
User3	-0.826	0.1705	1	1.0	-0.8260	-0.596
User4	-1.0	-1.0	1.0	1	-1.0	-1.0
User5	1.0	0.6546	-0.8260	-1.0	1	0.6882
User6	0.9449	0.866	-0.596	-1.0	0.6882	1

The results of Pearson correlation algorithm illustrate that the biggest coefficient and, accordingly, biggest similarity have user pairs (*User4, User3*), (*User1, User6*), (*User2, User6*), (*User5, User6*) and (*User1, User2*), and the smallest similarity is between users (*User1, User4*), (*User2, User4*), (*User4, User5*), (*User4, User6*), (*User1, User3*) and (*User3, User6*).

Table 3. Pearson Corelation Matrix

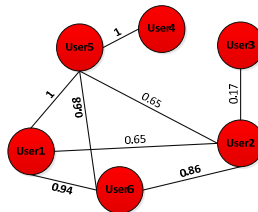
User	Similarity coefficient
User5	1.0
User6	0.9449111825230654
User2	0.6546536707079769

Based on Pearson coefficient calculations, a set of users with similar preferences is evaluated for every user. In this step model detects implicit networks and store them to social networks database. A set of similar users to *User1* is presented in Table 3.

**Fig. 7.** User1 similarity graph

Data on Table III show users with similar interests as *User1* are *User5* (1.0), *User6* (0.94) and *User2* (0.65). It means that there are weak ties between *User1* and *User5*, *User6* and *User2*.

Such matrixes are calculated for every platform user.

**Fig. 8.** Total similarity graph

After this step weak ties are established between all users with similar interests.

Step 3. Information tracking and proposal.

The last step is responsible for converting weak ties into strong relationships. To do this, an appropriate user is prompted with information, operated by another user from similarity set. Relevant content is selected based on users' closeness in implicit network and on their digital identities.

Table 4. User1 strong relationships table

User1	User5	User6	User2
	11	8	5

If the user responds to proposed information (clicks a link, saves to favorites, etc.), the weak tie get additional weight (gets +1 point) and is turned into strong

relationship. Response analysis updates relationship between users with appropriate values and recalculates responded user's digital identity values. Such response is treated like social interaction between users of the same social network.

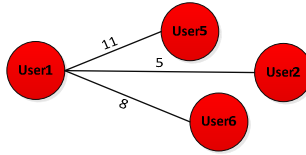


Fig. 9. User1 networks' weighted relationships graph

Relationships with bigger weight are considered more valuable and represent better social closeness. This weight affects information flow that is prompted for user later. Also, at this stage user's digital identity information is appended with new metadata according to his shown interests. This updated information is used to discover new weak ties. Digital identity update allows constant users' digital identity evaluation. It means that users' digital identity will be constantly updated with his preferences and learning direction change and help to discover new connections according to newly appeared interests in implicit network.

4 Method Evaluation

Big players like Google, Facebook or eBay use collective intelligence utilization approach in their products. Nonetheless, their methods are not published and are held as commercial secrets. Such companies publish only general guidelines, like social design strategy, which was overviewed in the second chapter on this paper.

On the other hand, collective intelligence utilization methods and social networking analysis algorithms are not applied in online education systems so far. That is why there are no legitimate numbers to compare with.

The proposed method could be implemented in any PLE platform. For the proof of the concept, method was implemented and tested in open source PLE platform "Droptings" [16]. Principal schema of method implementation is illustrated in Fig. 10.

As Fig. 10 illustrates, selected for the proof of the concept "Droptings" PLE platform is a widget based aggregation platform. As the method suggests, widgets are divided to two groups: ones used to aggregate content (social bookmarking, YouTube videos, etc.) – learning material consumption, and those, which are used for reflection on learning process (blogs, wiki). Each group has separate analysers. Aggregated data analyser extracts tags from Web 2.0 items and passes them to the digital profile agent. Generated content analyser scans user's generated text information and extracts keywords from there, and passes them to the profile agent. The profile agent is responsible for digital identity storage in digital identity database after applying digital identity development rules. The relationships manager analyses profile

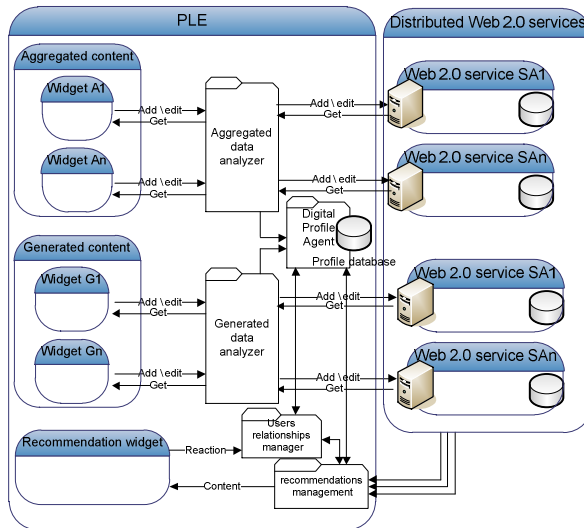


Fig. 10. Method implementation in a scope of PLE

information at the beginning and afterwards finds similar users according to their interests. That is how weak ties are established. Next these implicit connections are used for peered users activity tracking and content proposition. The recommendation manager uses these ties to find potentially useful data and propose it to the user in recommendation widget. If proposed information is useful for the user, and he clicks proposed link, appropriate information is send to the relationships manager. The relationship manager adds addition weight to that relationship and turns it into strong connection, meaning, that from now on appropriate learner will see more information from his peer as they become members of the same implicit social network. At the same time recommendation manager sends appropriate information to digital profile agent to make certain changes to learners' profile according to committed interaction. Weak ties expire after predefined time, if they do not get additional weight.

In order to evaluate proposed approach, the survey was made among users, to find out, if this method allows students to discover useful information. Survey results revealed that 31% of users found such prompted help 'extremely useful', 57% stated that it was 'reasonably useful' and only 12% stated that it was 'useful occasionally or not useful at all'.

5 Conclusion and Future Work

A new model of collective intelligence usage in the scope of personal learning environment is presented in this paper. An analysis of PLE concept allows defining common platform structure and the kind of data that is used in such environments. Following analysis of social media showed, what kind of ties and relationships are established there and how they can be used. Based on previous information, a general

method of collective intelligence usage in the scope of PLE was proposed. Despite the fact, that PLE is designed for single persons' use, the proposed model allows defining weak ties, that are appropriate for online communities, and turning them to strong relationships that are specific to social networks, and using these connections for users learning. Further interaction maintenance allows network evolution along with changing user's interests. The exclusive features of this method is both aggregated and generated information source analysis, corresponding to persons' knowledge gaining and reflection on learning process, and digital identity along with social networks' relevance update, which stands for users' learning focus change during (lifelong) learning process.

The proposed method was implemented in the scope of "Droptings" PLE platform. Evaluation survey showed that the majority of users (88%) found such prompted help useful in their activities (31% - extremely useful, 57% - reasonably useful), and only 12% didn't gain any additional value.

The future work is maximizing usefulness of proposed material, which can be used for persons' learning. Next step is to change +1 (or like based) prompted content evaluation system to a grading system (it could be grades from 1 till 5), and thus enhance recommended content relevance.

References

1. Wilson, S., Liber, J., Johnson, M., Beauvoir, P., Milligan, C.: Personal Learning Environments: Challenging the dominant design of educational systems. In: Tomadaki, E., Scott, P. (eds.) *Innovative Approaches for Learning and Knowledge Sharing, EC-TEL*, pp. 173–182 (2006)
2. Attwell, G.: Supporting Personal Learning in the Workplace. In: PLE Conference, Barcelona (2010)
3. van Harmelen, M.: Personal Learning Environments. In: Sixth International Conference on Advanced Learning Technologies, ICALT 2006 (2006), 0-7695-2632-2/06
4. Attwell, G.: Personal Learning Environments. The Wales-Wide Web portal (June 1, 2006), web access: http://www.knownet.com/writing/weblogs/Graham_Attwell/entries/6521819364 (retrieved: May 2013)
5. Sclater, N.: Web 2.0, Personal Learning Environments and the Future of Learning Management Systems (Research Bulletin), vol. (13). EDUCASE Center for Applied Research, Boulder (2008)
6. Wilson, S., Liber, O., Johnson, M., Beauvoir, P., Milligan, C.: Personal Learning Environments: Challenging the dominant design of educational systems. In: Tomadaki, E., Scott, P. (eds.) *Innovative Approaches for Learning and Knowledge Sharing, EC-TEL 2006*, pp. 173–182 (2006)
7. Conole, G., de Laat, M., Dillon, T., Darby, J.: Students Experiences of Technologies. Final Project Report, JISC, London, UK (2006), Web access <http://www.jisc.ac.uk/publications/reports/2006/lxpfinalreport.aspx> (retrieved May 2013)
8. Wu, M.: Community vs Social Network (June 6, 2010), Web access <http://lithosphere.lithium.com/t5/Building-Community-the-Platform/Community-vs-Social-Network/ba-p/5283> (retrieved May 2013)

9. Wu, M.: How do people become connected: Community vs Social Network 2 (June 13, 2010), Web access <http://lithosphere.lithium.com/t5/Building-Community-the-Platform/How-Do-People-Become-Connected-Community-vs-Social-Networks-2/ba-p/6620> (retrieved May 2013)
10. Wu, M.: From weak ties to strong ties: Community vs Social Network 3 (June 22, 2010), web access <http://lithosphere.lithium.com/t5/Building-Community-the-Platform/From-Weak-Ties-to-Strong-Ties-Community-vs-Social-Networks-3/ba-p/6834> (retrieved May 2013)
11. Wu, M.: Maintaining the strong tie: Community vs Social Network 4 (June 30, 2010), web access <http://lithosphere.lithium.com/t5/Building-Community-the-Platform/Maintaining-the-Strong-Ties-Community-vs-Social-Networks-4/ba-p/6854> (retrieved May 2013)
12. Fisher, E.: Social Design Strategy (May 3, 2011), web access <http://fishofthebay.com/posts/social-design-strategy> (retrieved May 2013)
13. Afonin, A.: Distributed Social Bookmarking Web Service Architecture. SOAP vs iCamp Feedback, Informatics in Education 10(2), 149–162 (2011) ISSN 1648-5831
14. Rutkauskienė, D., Patašienė, I., Targamadžė, A.: Advanced online tools for increasing study effectiveness in the Lithuanian Distance Education Network. In: Lifelong Learning in Higher Education: Networked Teaching and Learning in a Knowledge Society: Proceedings of the EADTU Annual Conference 2008, September 18-19, pp. 1–10. European Association of Distance Teaching Universities, Poitiers (2008)
15. Kanter, B.: Tagging for Collaboration and Knowledge Sharing (May 3, 2007), presentation web access <http://www.slideshare.net/kanter/tagging-for-collaboration-and-knowledge-sharing> (retrieved May 2013)
16. DROPTINGS, Open source platform, <http://droptings.codeplex.com> (retrieved May 2013)
17. Segaran, T.: Programming Collective Intelligence. O'Reilly Media, Inc. (2007) ISBN-10: 0-596-52932-5, ISBN-13: 978-0-596-52932-1

Automation of Upgrade Process for Enterprise Resource Planning Systems

Algirdas Laukaitis

Vilnius Gediminas Technical University , Sauletekio al. 11,
LT-10223 Vilnius-40, Lithuania
algirdas.laukaitis@vgtu.lt

Abstract. This paper presents a framework for semi-automatic process of enterprise resource planning (ERP) system upgrade. We suggest to change currently accepted practice of manual upgrade process when domain expert-programmer works through all localizations and transforms them manually to the new version of ERP system. The core idea for this framework is to induce the software code transformation patterns from completed upgrade projects and then to refine these patterns by using knowledge of ERP upgrade expert. These patterns lets us to increase productivity of upgrade process by improving automatic code alignment and annotation and by providing code transformation to the new version of ERP system. The price for these improvements is a requirement for upgrade expert to move from traditional 4/GL ERP programming language to stochastic meta-programming language which is used to describe code alignment and code transformation patterns.

Keywords: ERP system upgrade, code alignment, rules induction, knowledge representation, automatic code generation.

1 Introduction

The Process of enterprise resource planning software upgrade is often highly complex and requires many hours of work from programmers and domain experts. Many Commercial Off-The-Shelf (COTS) ERP products with the ability to extend, like Microsoft Dynamics Nav [16] or Microsoft Dynamix AX [6], include high level programming languages (4GL) as an option for standard version modification. These programming languages allow for rapid customizations of the products at a very affordable price. But after several years of constant adaptation, these localized products can contain hundreds and thousands of code lines that reflects local customer requirements.

On the other hand, the company that provided Off-The-Shelf ERP system develops a new standard version to reflect technological innovation in the business systems market. This new version of ERP system can contain many lines of code that describes business logic in the new standard version. At some point a business company decides to update its ERP system with the new standard version. It faces a problem how to move all localizations to this new version. Traditionally this is done by going manually through all localization code units and

moving them one by one to the new version. It is important to emphasize that the customization projects are usually carried and implemented by a certified channel of value added resellers (VAR) worldwide and that ERP localization process is highly decentralized. All that adds additional level of complexity when it comes to the ERP system upgrade project and development of a universal ERP upgrade tool [11].

In this paper we present a framework for the development of this universal ERP upgrade tool. Particularly we focus on the following process:

1. We have the standard version of ERP system code written in some high-level programming language (we refer to it as OldBase) and we have a customized ERP system code written in the same language (we refer to it as OldCustom).
2. As technology changes over time, ERP systems providers are developing a new version of the system (we refer to it as NewBase). At some moment a business company that runs its customized old version (OldCustom) of ERP system is faced with a dilemma of changing its old version system (OldCustom) by moving localizations to the new one (we refer to it as NewCustom).
3. Then we compare the OldBase and OldCustom ERP systems code files and collect all changes that has been done in OldCustom system. After that, we compare the OldBase and the NewBase ERP systems code files and gather all changes that has been done in the NewBase system.
4. We annotate all code differences by patterns and dependency graph information.
5. Then we generate the new version of customized ERP system and we document all changes that we done in the NewCustom version.
6. Programmer-domain expert reviews automatic code generation results, and if there is need for modification, he develops new patterns and regenerates new ERP version.

Often it is an expensive and long term process. Our study shows that on average there is more than 4000 differences between any pair of ERP system versions if the time span between them is several years.

It is important to note that there is several areas of research that relates to the results presented in this paper. First of all there is general software merge area. As pointed out by Mens [15] software merging remains complicated and error-prone process because software components that are involved in merging process depends on both the syntax and semantics of these elements. Nevertheless, many available software merge tools like Unix *diff* or *diff3* are based on textual merge techniques [9], [10] without consideration of software syntax and semantics. Several methods that considers syntax in software merging can be found in [3] and methods for semantic merge of programs can be found in [8]. There has been number of research projects that tried to consider domain independent syntax and semantics dimensions for software merging [14]. But all these approaches often doesn't suit well for ERP systems upgrade because ERP systems merge require deep knowledge of the domain area for which we need an effective process to develop rich code pattern base.

Software clone detection is another area of research that is related to ERP upgrade [19]. Various clone detection techniques can be used in ERP upgrade to compare two code segments that are semantically identical but syntactically different. Source code search using program patterns is yet another area of research that is similar to software clone detection [18] [13]. We used ideas from these areas of research to define code alignment patterns in our framework.

The rest of the paper is structured as follows. In section 2 we describe a general process of ERP systems upgrade that reduces the cost and time of the business system upgrade projects and increases the quality of the final system (i.e. NewCustom). In section 3 we present more details on ERP system patterns and knowledge representation. In section 4 we present evaluation of suggested method. Finally, concluding remarks are provided.

OldBase	OldCustom	NewBase
a)		
<pre>{ 7600; ;Base Calendar Code ;Code10 ;TableRelation="Base Calendar"; CaptionML=ENU=Base Calendar Code }</pre>	<pre>{ 7600; ;Base Calendar Code ;Code10 ;TableRelation="Base Calendar"; CaptionML=ENU=Base Calendar Code } { 50000; ;Use ADCS ;Boolean } }</pre>	<pre>{ 7600; ;Base Calendar Code ;Code10 ;TableRelation="Base Calendar"; CaptionML=ENU=Base Calendar Code } { 7700; ;Use ADCS ;Boolean ;CaptionML=ENU=Use ADCS }</pre>
b)		
<pre>IF Code <> CurrentCode THEN SendForm(1) ELSE Process; CLEAR(DOMxmlin); END;</pre>	<pre>// Backwards IF Code = CurrentCode THEN Process ELSE SendForm(1); CLEAR(DOMxmlin); END;</pre>	<pre>IF Code <> CurrentCode THEN BEGIN SendForm(1); MiniformHeader.Code := 'TEST'; END ELSE Process; CLEAR(DOMxmlin); END;</pre>

Fig. 1. Two examples of 4/GL code segments in ERP system. Upper row represents data description and second row represents code description.

2 Processes of ERP Systems Upgrade

The old process of ERP system upgrade was quite simple. An expert of ERP migration receives localized version of the system (OldCustom) and a request to migrate that system to the new one (NewBase). He compares all three versions (OldBase, NewBase, OldCustom) by using some standard code comparison and merging tool (like kdiff3) and collects the set of differences between different ERP system versions. Then, an expert of ERP upgrade goes manually one by one through all pairs of code differences and decides whether to move them to the new version of ERP system (NewCustom). It is important to note that

the standard code comparison software uses longest common subsequence (LCS) algorithms and comparison is done by comparing code lines. More sophisticated software compares code syntax trees but in the ERP domain usually comparison by syntax tree is done only partially. These tools select sets of objects (like tables, forms, reports, code units) and compare objects with the same name using LCS algorithms.

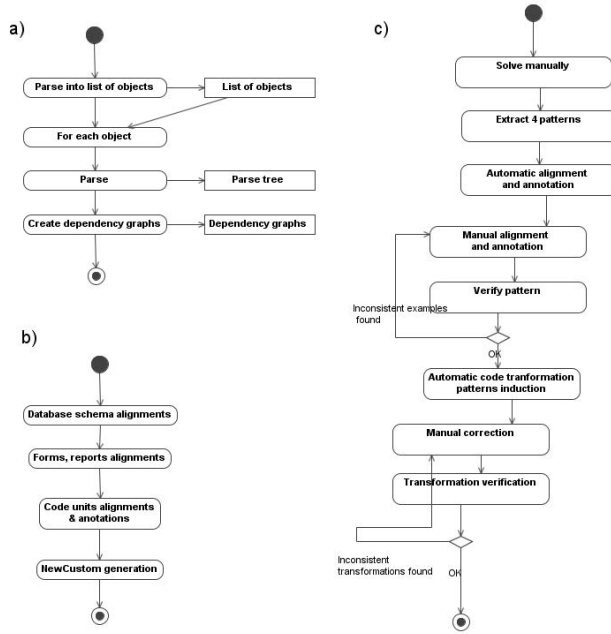


Fig. 2. General processes of ERP system upgrade

In order to understand old ERP upgrade process and suggested new one let us look at two simple examples from Microsoft Dynamics NAV [16]. In the Figure 1 we can see typical situation when there is changes in data structures (first row) and we see typical situation when there is changes in program code (second row).

From the first row in the example we can see that localized version of ERP system (OldCustom) contains new field $\{ 50000; ;Use ADCS ;Boolean \}$ in the table *OBJECT Table 14 Location*. As the rule of thumb, the new fields in the data structure from localized version are moved to the new customized version. But in this particular example we can see that ERP vendor created (in NewBase) the new field in the table with the same name but different ID. In this case upgrade expert must merge these fields by moving all references in OldCustom version to the new field in the NewBase version. It can be difficult to do that manually and to avoid programming bugs from such manual transformation. This example can be solved automatically by using a pattern that matches fields with identical name in the data structure description and which comes with method to replace the references to the field.

The second example deals with changes in code that describes business logic. We can see that even in this simple case it can be confusing to reconcile what changes has been done and how they must be transformed to the NewCustom. Traditional merge software will be of no use. Solution that an expert will produce will be analysis of conditions for each statement in OldCustom and insurance that in NewCustom these statements are called on the same conditions. Additionally, expert will move all new statements in NewBase to NewCustom if they do not conflict with business logic after OldCustom statements are moved. Our solution for this particular example is based on special representation of the program code and pattern with the method to evaluate logical expressions. This special representation of the program code means that we transform program code to the form where each statement has its own conditional expression i.e we remove all blocks of IF...ELSE...END and build equivalent code where each callable statement has own IF... expression.

We can see that even these simple examples can take significant amount of time to transform them into new version of ERP system. There can be hundreds of complex differences between two code versions and the whole update process can take several month to complete it. In the Figure 2 we present the process that we developed during this research project and which, if implemented as we describe in this paper, reduces upgrade project time and increases generated NewBase code quality.

In the figure the diagram a) shows the basic steps that we use to preprocess initial code of ERP system. As we can see the basic idea at this step is to process code into list of objects. Each ERP system has such objects as tables, forms, reports, triggers, code units etc. At this step we just split code into the list of such objects and then compare pairs of objects by requiring that name and type of object must match. Additionally we parse each object into parse tree and create for each object dependency graph.

The diagram b) shows four major processes from our framework. We can see that at the first stage we are handling database schema alignments and the NewBase generation. After that all forms and reports are aligned and generated. Then we align code units and finally we generate new code units (procedures, functions, triggers etc.)

The diagram c) shows the final stage in our framework. This process is responsible for manual inspection of automatic alignments and generation. The main idea that is implemented in this framework is that during manual inspection we review differences manually and modify patterns in a such way that automatic alignments and NewCustom generation brings desirable results. The following list presents more detailed description of this diagram elements:

Solved manually/automatically. In order to create code patterns we select completed upgrade project (solved manually) which means that we prepare 4 versions of ERP system : OldBase, NewBase, OldCustom, NewCustom. If we do not have NewCustom version then we create it automatically with current version of our system (solved automatically).

Extract object quadruple. We extract 4 objects with the same name from OldBase, NewBase, OldCustom, NewCustom.

Automatic alignment and annotation. We run current version of our alignment procedure. We start with comparison of programm code as sequence of words. Then, we use a set of heuristic rules to refine alignments and to annotate code lines with a semantic information.

Manual alignment and annotation. We review alignments.

Verify pattern. We verify patterns.

Automatic code transformation patterns induction. We run pattern induction algorithm .

Manual correction. An expert corrects patterns .

Transformation verification. We run pattern verification algorithm.

3 Patterns and Knowledge Representation

ERP system upgrade algorithms are closely related to the format of knowledge representation and an efficient knowledge representation relates to the domain it represents. We can see from the section above that our domain of interest (i.e. ERP upgrade) is closely related to the more general question of automatic software generation from informally or semi-formally defined business system requirements. We know that in order to get formal structure (i.e. computing program) by using an algorithms (i.e Turing machine that always halts within a finite time) we need to define the input by some formal language. Then, these statements defines some general theoretical aspects of our task. We have as a formal input 4GL program code (i.e. OldBase, OldCustom and NewBase) and we need as an output a formal program in the same 4GL programming language. It seem that at least theoretically it is possible to write the algorithm which transforms such input to such output. On the other hand changes in ERP versions like OldCustom and NewBase are based on requirements that are not observable or are informal. It means that in general the task of ERP upgrade is undecidable and we must to seek for suboptimal solution.

Then we have the problem of how to transform ERP upgrade expert knowledge and competed ERP upgrade projects code to the set of formally defined examples. Additionally, we require to reuse algorithms and software that already has sound representation in the field of artificial intelligence.

All these requirements and considerations that we defined for the knowledge and programm code representation language suggest to use several forms of 4GL code representation [12]. In this paper we suggest to build all knowledge base on a set of patterns where each pattern uses one of the following structures of programming code representation.

1. Programming language abstract syntax tree (see Figure 3).
2. Programming language dependency graph(see Figure 4).
3. Programming language sequential string of lines or multi-aligned tree as the set of paired strings(see Figure 5). It allows us to transform our knowledge base in such way that for learning algorithms it will appear as simple string of symbols and it means that we can reuse existing machine learning algorithms.

4. Programming language sequential string of conditional calls. This representation of the program is a transformation of a computer program where all conditional branching is removed and instead of it each callable statement has separate conditional sentence. A simple example below demonstrates it by presenting original and transformed code.

```

IF variable1=1 THEN          IF variable1=1 THEN
  Call proc1                 Call proc1
  IF variable2=1 THEN        IF variable1=1 AND variable2=1 THEN
    Call proc2               Call proc2
  ELSE                       IF variable1=1 AND NOT variable2=1 THEN
    Call proc3               Call proc3
  END IF
ELSE                          IF NOT variable1=1 THEN
  Call proc4                 Call proc4
END IF
    
```

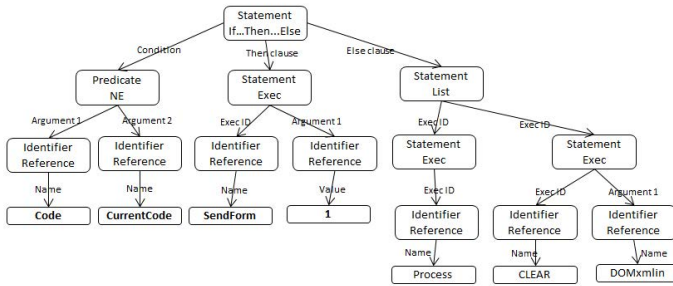


Fig. 3. Example of abstract syntax tree

A set of code patterns is one of the core elements in our framework. All code alignments and code transformation algorithms use the patterns to match code segment and transform it to the new one. Then, the pattern language in our framework is an extension by abstraction of the source code programming language. We use a set of wildcard symbols, predicates and constants as substitutes for programming language elements in order to define our pattern language. This approach is very general and we can use it to derive any pattern for programming code match and transformation. What differentiates our approach from other pattern languages like in [18] is that we do not write specifications to define patterns before matching programming code. Instead we use machine learning and online learning to derive series of templates from competed migration projects. Additionally, for the derived pattern to move to knowledge base we require that it be verified by the migration expert.

All ERP systems that we considered use imperative programming languages in order to localize system. In such languages there is syntactic entities as control block, variables , functions etc. We introduce several operators that transform



concrete programming language code into alignment and code transformation pattern. The following list of operators was used in this project.

1. Substitution of wildcard.
2. Substitution of predicate.
3. Removal.
4. Order invariant.

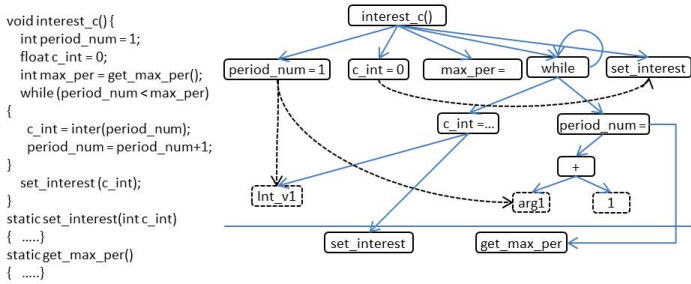


Fig. 4. Example of programming code dependency graph

In order to better understand how pattern of the code are generated let's look at the figure 5. The first two rows of paired strings $Line_{f_1}1 \dots Line_{f_1}7$ and $Line_{e_1}1 \dots Line_{e_1}5$ represents an ordinary programm code i.e. two versions of the ERP system. On this level of abstraction we have standard task of comparing two versions of ERP program code. The tasks is to find differences in code and align semantically and syntactically equal segments of code. Usually this task is approached by using dynamic programming and sequence alignment algorithms such as Needleman-Wunsch algorithm [17] or Smith-Waterman algorithm [20].

The pattern from this example can be generated by taking any aligned segment of OldBase, NewBase, OldCustom, NewCustom code and just memorizing it by putting it in our knowledge base.

Next, we take the same code example and generate an abstract representation of ERP program code $Line_{f_k}1 \dots Line_{f_k}6$ and $Line_{e_k}1 \dots Line_{e_k}5$. This is done using the following procedure. We take a word, line or segment of program code and decide to replace or not to replace it by the wildcard or constant from knowledge base. For example we can replace all the objects descriptions in the code file by the labels like '*Table:Currency*', '*Form:Transaction*' etc. We can take one of the parsers and replace some code by its documentation or some logical formal representation. We can take a language parser and replace all code lines in parse subtree by the subtree label. Even in the simple case when we consider only single word replacement by one possible label there is 2^n possible unique string pairs. Which one to choose we model as the hidden parameter.

Another hidden parameter that we introduced in our model is alignments themselves. In the figure 5 they are represented by the labels (a, b) . Label a represents direct mapping of code segment to code segment in two versions of the

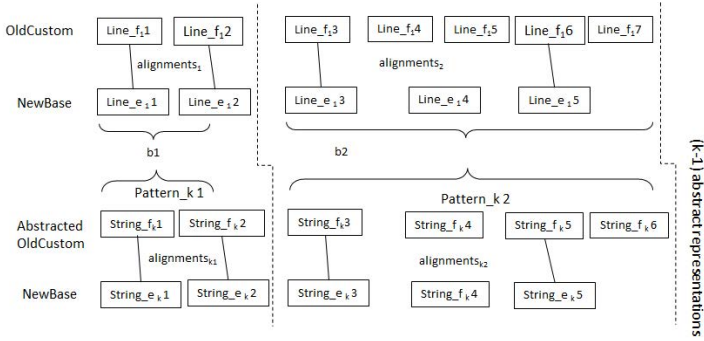


Fig. 5. Alignments within two versions of the ERP systems code

program code and labels b represents splitting boundaries between code segment chunks. The parameters b can have another interpretation. As we mentioned earlier we don't know exact mechanism on how decoding is done. Then we can only guess the model structure which is used by ERP system upgrade expert. Then in the broad sense parameters b represents mechanism of transformation given that we learned transformation patterns base i.e. parameters a . Another important fact about parameter b is that it incorporates our knowledge about context which ERP system upgrade expert takes into account. And again, it is a hidden parameter which we model.

Then, we can choose some model from machine learning theory which we believe can be used in the decoding process i.e. final code generation from e to f , if we assume that we accumulated sufficient amount of ERP system upgrade projects code and learned parameters a with high accuracy. In this paper we chose to investigate a simple noisy channel model [2], [4]. We use noisy channel model to align two versions of ERP systems code. From the theoretical learning we know that given enough data we can use greedy approach to learn such models [1], [5].

4 Evaluation

In this paper we suggested the framework for semi-automatic generation of ERP system programming code and in this framework we suggested two main algorithms for ERP 4/GL code alignment and code generation. Then we formulate two questions about performance of these algorithms: 1. How accurate is suggested alignments algorithm when we compare it with longest common subsequences algorithm [10] used in most version control systems? 2. How accurate is suggested NewBase generation algorithm?

In order to answer these questions we incrementally selected 19 Microsoft Dynamics NAV completed projects for the evaluation of suggested algorithms. Each project has 4 files (OldBase.txt, OldCustom.txt, NewBase.txt and NewCustom.txt). Each file contained all data structures and business logic code from

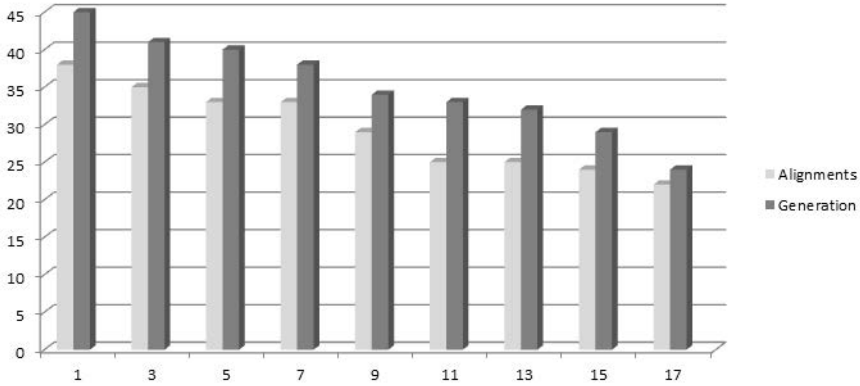


Fig. 6. Evaluation results

one version of Microsoft Dynamics NAV system. On average each file has size of 4.4 megabytes and about 67000 code lines. Due to the amount of data we decided to verify alignments were there is conflict between our alignment algorithm and standard software merge tool kdiff3. In order to verify our NewBase generation algorithm we selected one project that has been manually completed by an expert and we compared it with automatically generated version. In the figure 6 we can see the results.

The following expression has been used to estimate error rate:

$$AlignmentsError = \left(\frac{A_{mismatched}}{N} \right) * 100$$

where N is the total number of alignments or generated code segments with errors, $A_{mismatched}$ number of mismatched alignments between our alignment algorithm and standard software merge tool kdiff3.

5 Conclusion

We presented general framework that enables us to build a business system upgraded tool. Presented framework focuses on ERP systems that enables localization through the use of 4/GL language but we think that some ideas in this framework can be applied to other information systems. The framework presented in this paper can increase quality of programming code in the new version of generated ERP system and decreases time of upgrade projects, but in order to achieve this, developers must invest their time to build pattern base that is used to align and transform code from one ERP system version to another.

Additionally, we think that we were able to suggest a method for using machine learning approach for information systems engineering. The basic idea behind this method is that machine learning is used to detect and sort code patterns and then domain human-experts refines some of the patterns for the final use to generate new version of information system.

Our preliminary theoretical analysis shows that it is impossible to have fully automated solution of this upgrade process unless we put more constraints on localization language. Nevertheless it is possible to solve significant amount of conflicts with the framework suggested in this paper and the rate can be increased by constantly updating knowledge base from the new upgrade projects.

In order to solve ERP upgrade automatically we need formal software specifications as it was suggested in [21]. But in practice this is not the case, we don't know any ERP system that has a complete formal specification.

References

1. Angluin, D., Smith, C.H.: Inductive Inference: Theory and Methods. *ACM Computing Surveys* 15(3), 237–269 (1983)
2. Baum, L.E.: An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities* 3, 1–8 (1972)
3. Buffenbarger, J.: Syntactic software merging. In: Estublier, J. (ed.) *ICSE-WS 1993/1995 and SCM 1993/1995*. LNCS, vol. 1005, pp. 153–172. Springer, Heidelberg (1995)
4. Dempster, A.E., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39(B), 1–38 (1977)
5. Gold, E.M.: Language identification in the limit. *Information and Control* 10(5), 447–474 (1967)
6. Ehrenberg, M.: *Microsoft Dynamics AX, A New Generation in ERP* (2011)
7. Gold, E.M.: Language identification in the limit. *Information and Control* 10(5), 447–474 (1967)
8. Horwitz, S., Prins, J., Reps, T.: Integrating Noninterfering Versions of Programs. *ACM Transactions on Programming Languages and Systems* 11(3), 345–387 (1989)
9. Hunt, J.W., McIlroy, M.D.: An algorithm for differential file comparison. *Computer Science Technical Report 41*, Bell Laboratories (1975)
10. Hunt, J.W., Szymanski, T.G.: A fast algorithm for computing longest common subsequences. *Commun. ACM* 20(5), 350–353 (1977)
11. Laukaitis, A.: Automation of Merging in ERP Revision Control. *Information and Software Technologies Communications in Computer and Information Science* 319, 1–14 (2012)
12. Laukaitis, A., Vasilecas, O.: Multi-alignment templates induction. *INFORMATICA* 19(4), 535–554 (2008)
13. McMillan, C., Hariri, N., Poshyvanyk, D., Cleland-Huang, J., Mobasher, B.: Recommending source code for use in rapid software prototypes. In: *34th International Conference on Software Engineering (ICSE)*, pp. 848–858 (2012)
14. Mens, T.: *A Formal Foundation for Object-Oriented Software Evolution*. PhD thesis, Vrije Universiteit Brussel - Faculty of Science - Departement of Computer Science - Programming Technology Lab (August 1999)
15. Mens, T.: A state-of-the-art survey on software merging. *IEEE Transactions on Software Engineering* 28(5), 449–462 (2002)
16. Microsoft Corporation. *Microsoft Dynamics NAV* (2012)

17. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3), 443–453 (1970)
18. Paul, S., Prakash, A.: A framework for source code search using program patterns. *IEEE Trans. Softw. Eng.* 6(20), 463–475 (1994)
19. Roy, C.K., Cordy, J.R.: A survey on software clone detection research. Technical Report. Queens University at Kingston (2007)
20. Smith, T.M., Waterman, M.S.: Identification of Common Molecular Subsequences. *Journal of Molecular Biology* 147, 195–197 (1981)
21. Zaremski, A., Jeannette, M.W.: Specification Matching of Software Components. *ACM Transactions on Software Engineering and Methodology* 6(4), 333–369 (1997)

Business Process Flow Verification Using Knowledge Based System

Regina Miseviciene, Germanas Budnikas, and Dalius Makackas

Kaunas University of Technology, Faculty of Informatics, Studentu 56-443, Kaunas, Lithuania
{regina.miseviciene,germanas.budnikas,dalius.makackas}@ktu.lt

Abstract. Analysis of business process flows presented in this paper constitutes three main activities: representation of business flows by AND/OR graphs, their transformation to Prolog clauses and verification in Prolog environment using created knowledge based system where deadlock and endless loop properties are defined. An ordering process verification example is used for illustration of the approach proposed.

Keywords: verification, business process, AND/OR graphs, Prolog.

1 Introduction

Multiple activities in product design, engineering and manufacturing can be defined and maintained as a process. The processes are defined as collections of linked, structured activities or tasks that will implement a particular organizational goal. The process definition is one of the most complicated parts. Different systems use several techniques to define the processes. Most of them are traditionally workflow-oriented. Papers on business workflow verification usually focus on representations like UML Activity Diagrams [13], Business Process Modeling Notation [4], Business Process Execution Language [16], Petri nets [8], Event-driven Process Chains [5], etc. However, process overall definition using aforementioned representations can become too composite in case of complex domains.

At the same time, definition of processes can be done much easier – in a form of written rules [10]. Such a process workflow definition may be fully readable and interpretable by the people involved into a problem domain.

This paper presents a novel approach of business process workflow verification using knowledge based system. The approach combines benefits of graph notation for representation of business workflows and knowledge -based techniques for their verification. The approach can be seen as a possible alternative solution additionally to the already existing modeling and verification techniques. Verification is performed using created knowledge based verification system (KBS) in Prolog [2]. Created KBVS interface allows user involved in problem solution to interfere, find and correct mistakes during verification process. A main benefit of the presented approach is that it provides a simple verification technique and does not require complex tools and related background knowledge on workflow formalization.

Following the introduction, the paper presents summary of existing business process notions and tools. Next, a graph notation for representation of business workflows is introduced. Section 4 describes implementation issues of knowledge based system used for the verification task. The suggested approach is illustrated by an example in section 5 where an ordering business process is represented by graph notation and verified. Conclusion summaries the approach proposed.

2 Business Process Modeling Tools

Business processes are important for organizations because they pursue some specific goals. Usually these three kinds of business processes are defined:

- management processes – that are responsible for operation management;
- operational processes – the ones that constitute the core business and create the primary value flow;
- supporting processes – the ones that support management processes.

Business processes are characterized as sets of linked and ordered activities or tasks responsible for an achievement a defined objective. The process flows usually are modeled in order to eliminate possible ambiguities especially in complex problems. A workflow model is used to represent business processes. There exists a wide variety of workflow modeling notations, e.g. Business Process Modeling Notation [4], UML [13], Petri nets [8], and flowcharts [15].

Business Process Modeling Notation (BPMN) can be used for workflow representation. BPMN provides a graphical notation for specifying business processes. It also provides a mapping between the graphics of the notation and the underlying constructs of execution languages, particularly Business Process Execution Language (BPEL) [20].

Unfortunately, BPMN models are not intended to be directly executed [4]. They need to be refined before the execution, for example into the Business Process Execution Language, where actions of business processes are specified. Yet Another Workflow Language (YAWL) that is considered as an alternative to BPEL could be seen other example [20].

Another XML Process Description Language (XPDL) [19] is designed to exchange the process definition, both the graphics and the semantics of a workflow business process. XPDL has been projected purposely to collect the full set of BPMN diagram attributes. Unlike BPEL language XPDL is not an executable programming language. The XPDL describes XML schema used for characterization of specifying the declarative part of workflow in a business process.

Other widespread modeling language is Unified Modeling Language (UML) activity diagrams. They are widely used in modeling of business processes. UML has potential applicability in business process modeling [13].

Business processes can be modeled using Petri nets [8]. Petri nets belong to mathematical modeling language used for the defining business workflows. The above presented workflow notations can be translated into a representation equivalent to Petri nets. Many transformations from workflow representations into Petri nets were published (e.g. BPMN, EPC [8]).

Business process workflows often can be visualized with a flowchart (or Data Flow Diagrams) [15] as a sequence of activities. Flowcharts are used in analyzing, designing, documenting and managing a process.

However, when processes become complicated, the overall definition of a process using BPMN, UML, Petri nets or flowchart notations can become complicated too.

At the same time, definition of processes can be done verbally as a set of business rules. Such a workflow definition can be done much easier and can be easily read and interpreted by people involved into problem solution [11]. In [12], [17] business rules are used for presentation and verification of workflows. Authors in [12], [7] present comparison between different business process modeling languages and business rules representation language sets. Articles [5], [9], [10] demonstrate application of graph notation for representation of business workflows. A transformation of BPMN and UML notations to directed graph is discussed in [14].

In this paper we demonstrate novel business process modeling approach using combined notation: firstly a business process is determined as AND/OR graph, then it is implemented using Prolog.

3 Business Process Flow Verification

Verification is concerned with determining whether a business process demonstrates certain required behaviors.

Many different formal approaches and techniques can be applied to verification. In this paper we refer only to verification of properties (described as goals to reach or conditions to meet) that systems states have to satisfy. Reachability analysis, deadlock-freeness analysis, and generic temporal logic properties are typical properties that are important to verify [4].

In this paper we solve a reachability problem using search algorithms on a graph. Fig. 1 explains the problem. All paths must be analyzed having the start node S and the goal node G in the reachability verification. Every path p_j from the source vertex S to the goal node G can be decomposed into $S \xrightarrow{p_{sj}} X_j \xrightarrow{p_{jg}} G$, $j = 1, 2, 3$. It is impossible to reach the goal vertex G from the start node via intermediate node X_3 . This leads to violence of the reachability property.

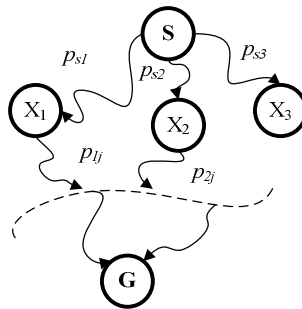


Fig. 1. Illustration of the reachability problem

The reachability problem can be solved using path-finding algorithms in a graph. For example, Dijkstra and Bellman-Ford algorithms [3] use weighted graphs and find the only one shortest path from the source vertex to the goal. In order to find all paths, the algorithms must be modified running a single-source shortest path algorithm $|X|$ times (there X is a set of the graph vertices), once for each vertex as the source (for example, Floyd-Warshall algorithm [3]).

For business process verification we have presented our earlier modified algorithm [11] that uses a weightless breadth-first search for the directed graphs. The algorithm applies recursive solution of breadth-first search to each vertex and computes the shortest-distance in bottom-up style. Bottom-up style analyses a graph starting from the goal node. The search algorithm finds all reachable paths from the start node S to the goal node G . Our search algorithm can perform analyses of graphs with cycles. Main idea of the proposed verification algorithm is presented in figure 2.

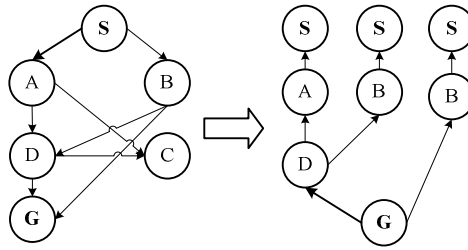


Fig. 2. Main idea of the proposed verification algorithm

For presentation of business process flows we will formally define AND/OR graph that have already been successfully applied in Artificial Intelligence (AI).

The structure of business process is represented by a AND/OR graph $G = (X, W)$:

- A set of nodes $X = \{x_1, x_2, \dots, x_n\}$ expresses finite set of tasks in business process. The node may be of two types – OR/AND;
- A set of directed edges $W \subseteq X \times X$ portrays a flow of the tasks;
- A set $X^s \subseteq X$ is a finite set of *start* tasks, $X^g \subseteq X$ is a finite set of *goal* tasks.
- A *path* is a sequence of nodes $p = x_1, x_2, \dots, x_{k-1}, x_k$ and all nodes are distinct for $i = 1, 2, \dots, k-1, k$. A path is *cyclic* if $x_i = x_j$.

The graph is characterized by the following concepts [1, 15]:

- *Environment* of the node $x \in X$ is a set of nodes adjacent to it and denoted by $E(x) = \{u \in X : \{x, u\} \in W\}$. Two nodes $u, v \in X$ are adjacent if they are connected by an edge. *Inputs* $E^+(x)$ and *outputs* $E^-(x)$ for the node are denoted by $E(x) = E^+(x) \cup E^-(x)$.

- Degree of the node is defined as $\deg(x) = |E(x)|$. The *indegree* is denoted $\deg^-(x)$ and the *outdegree* is denoted as $\deg^+(x)$. The degree defines $\deg(x) = \deg^-(x) + \deg^+(x)$ ingoing and outgoing edges to / from the node.
- A node $x \in X$ with $\deg^-(x) = 0, \deg^+(x) > 0$ is called a *source* or *start*.
- Similarly, a *goal* node is marked as $\deg^+(x) = 0, \deg^-(x) > 0$.
- Other nodes $x \in X$ with degree $\deg(x) = \deg^-(x) + \deg^+(x)$ are called *internal* vertices of the graph.

Reachability errors verified in a graph are the following:

- *Deadlock* vertex is an internal node $x \in X$ when $\deg^+(x) = 0, \deg^-(x) > 0$ that does not belong to the set of goal nodes.
- *Endless loop* contains a circular sequence of nodes leading to unreachable node.

Figure 3 illustrates an example of a workflow graph. The example graph is defined by the sets presented below.

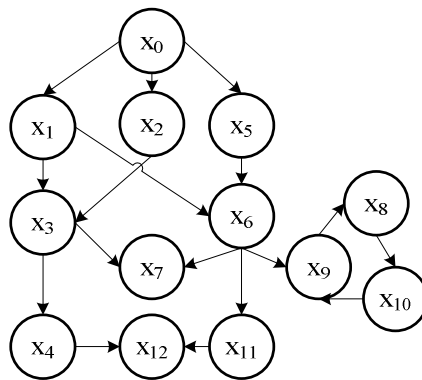


Fig. 3. An example of a workflow graph

- The set of nodes of the given graph is: $X = \{x_0, x_1, x_2, \dots, x_{12}\}$;
- The set of edges of the graph is: $W = \{(x_0, x_1), (x_0, x_2), (x_0, x_5), (x_1, x_3), (x_1, x_6), (x_2, x_3), (x_3, x_4), (x_3, x_7), (x_4, x_{12}), (x_5, x_6), (x_6, x_7), (x_6, x_9), (x_6, x_{11}), (x_8, x_{10}), (x_9, x_8), (x_{10}, x_9), (x_{11}, x_{12})\}$;
- Start node is: $\{x_0\} \subset X$ where $\deg^-(x_0) = 0, \deg^+(x_0) > 0$;
- Goal node is: $\{x_{12}\} \subset X$, where $\deg^+(x_i) = 0, \deg^-(x_i) > 0, i = 12$.

Errors can be checked in the graph:

- *Deadlock* node is an internal node $\{x_7\} \subset X$, where $\deg^+(x_7) = 0, \deg^-(x_7) > 0$ and it does not belong to the set of goal nodes.
- Nodes in *endless loop* are $\{x_9, x_8, x_{10}\} \subset X$.

4 Implementation of Knowledge Based Verification System

Knowledge based verification system (KBVS) is implemented using Amzi! Prolog [2]. Amzi! Prolog is a powerful implementation of the Prolog language. Its integrated development environment includes an editor, interpreter (listener) and debugger for developing Prolog modules. KBVS is showed in Figure 4.

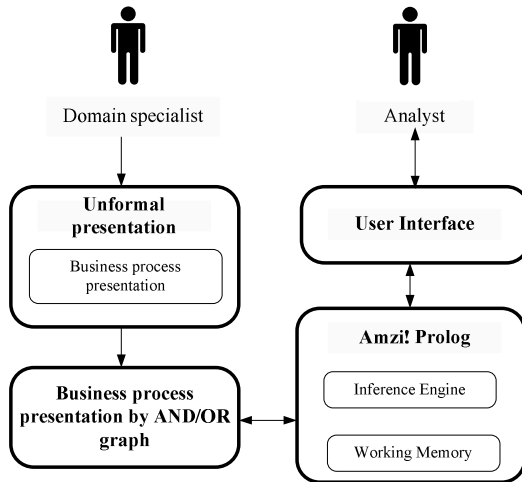


Fig. 4. Architecture of knowledge based verification system

Analysis of business process flows provided by KBVS (Fig. 4) consists of the following steps:

1. Representation of business flows by AND/OR graphs. In graph analysis algorithm (presented in the previous chapter) edge directions are changed to the opposite.
2. Transformation of AND/OR graph into Prolog clauses.
3. Verification of business process flows, using Prolog environment.

Figure 5 illustrates the example workflow graph in bottom-up style. The graph notation of business process can be very simple rewritten in clauses of Prolog. The clauses that examine the workflow graph are presented in the same figure 5. Succeeding figure 6 depicts verification results.

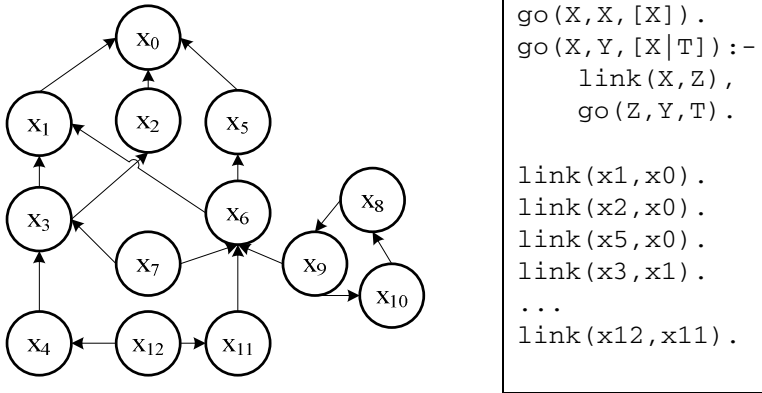


Fig. 5. Bottom-up style graph and its presentation in Prolog clauses

```

-----
Reachability analysis:

Reachable path -->
X= [x12,x4,x3,x1,x0]

Reachable path -->
X= [x12,x4,x3,x2,x0]

Reachable path -->
X= [x12,x11,x1,x0]

Reachable path -->
X= [x12,x11,x6,x5,x0]
Reachability analysis:

Reachable path -->
X= [x12,x4,x3,x1,x0]

Reachable path -->
X= [x12,x4,x3,x2,x0]

Reachable path -->
X= [x12,x11,x1,x0]

Reachable path -->
X= [x12,x11,x6,x5,x0]

Not reachable arcs:

(x3,x7),
(x6,x7),
(x6,x9),
(x8,x10),
(x9,x8),
(x10,x9),

```

Fig. 6. Graph verification results

Explanation of the graph verification results – reachable paths found from *source* node $x_0 \in X$ to *goal* node $x_{12} \in X$. Not reachable edges go:

- through node $x_7 \in X$. It corresponds to a *deadlock* because it does not belong to the set of goal nodes;
- through a set of nodes $x_8; x_9; x_{10}$. It correspond the *endless loop* that contains a circular sequence of nodes leading to unreachability.

5 Approach Demonstrating Example

Figure 7 shows a business process diagram [18] in BPMN notation representing an ordering process. The example introduces the main elements of the language: events, activities, gateways and sequence flow. The process model starts with an event. An ordered set of activities to analyze the order and to check the stock are performed. If the ordered products are in stock, then the lower branch is selected. Otherwise the product has to be manufactured first, so that the lower branch needs to be chosen.

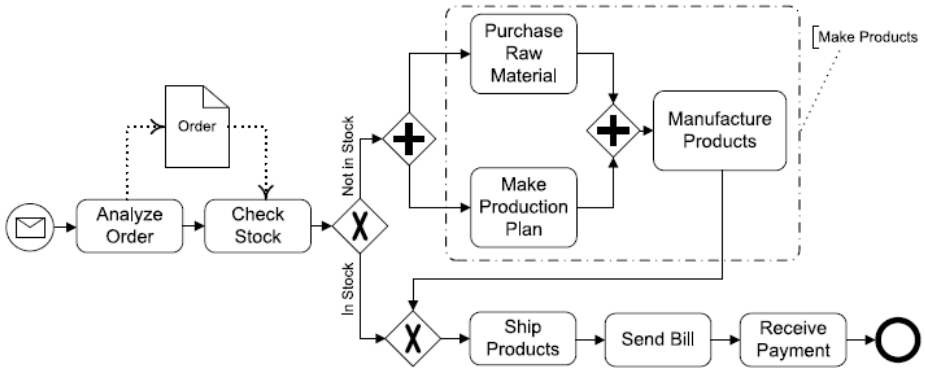
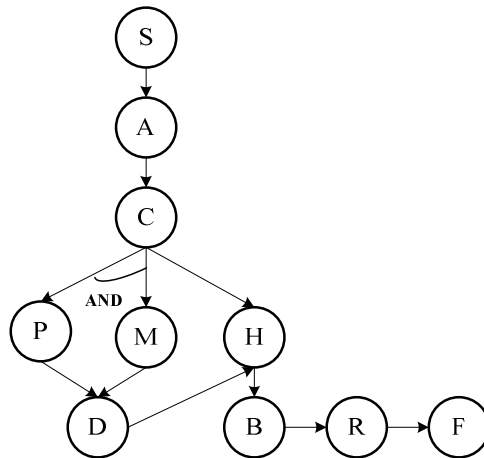


Fig. 7. Ordering business process BPMN diagram (Source: [18, 210 p.])



S – start, A - analyze order, C- check stock, P – purchase raw materials, M – make production plan, D – manufacture products, H- ship products, B - send bill, R- receive payment, F- finish.

Fig. 8. Ordering business process mapping to AND/OR graph



According to authors [14] the BPMN notation can be mapped to a graph representation. In figure 8 we present mapping of the considered example to AND/OR graph.

Figure 9 presents graph in bottom-up style in Prolog. Reachability verification report is pictured in the same figure.

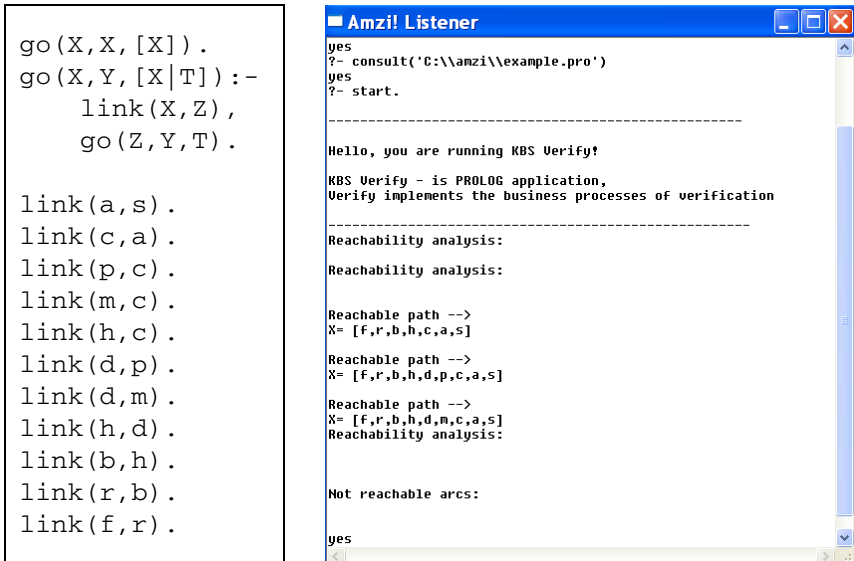


Fig. 9. Graph in Prolog and reachability verification report

6 Conclusions

Modeling of business processes and its succeeding implementation may be erroneous, some mistakes may be left in model specifications.

Unfortunately, the most often used business process models like UML, BPMN, Petri nets are not intended to be directly executed. They need to be transformed into executable languages. This requires a lot of programming code to run the business process verification procedure.

This paper presents a novel approach of business process workflow verification using knowledge based system. The approach combines benefits of graph notation for representation of business workflows and knowledge -based techniques for their verification. Verification is performed using created knowledge based verification system in Prolog. Created KBVS interface allows user involved in problem solution to interfere, find and correct mistakes during verification process.

A main benefit of our approach is that it provides a simple business process representation and verification technique that does not require complex tools and related background knowledge on workflow formalization.

The approach can be seen as a possible alternative solution additionally to the already existing modeling and verification techniques.

Acknowledgements. The work described in this paper has been carried out within the framework the Operational Programme for the Development of Human Resources 2007-2013 of Lithuania “Strengthening of capacities of researchers and scientists” project VP1-3.1-ŠMM-08-K-01-018 “Research and development of Internet technologies and their infrastructure for smart environments of things and services” (2012- 2015), funded by the European Social Fund (ESF).

References

1. Aalst, W.: Trends in business process analysis: From verification to process mining. In: Cardoso, J., Cordeiro, J., Filipe, J. (eds.) Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS 2007), pp. 12–22. Institute for Systems and Technologies of Information, Control and Communication, INSTICC, Medeira (2007)
2. Amzi!, <http://www.amzi.com>
3. Cormen, T.H., Leiserson, C.H., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)
4. Corradini, F., Polzonetti, A., Re, B., Falcioni, D.: An ECLIPSE Plug-in for Formal Verification of BPMN Processes. In: Proceedings of the Third International Conference on Communication Theory, Reliability, and Quality of Service, pp. 144–149 (2010)
5. Davis, R.: Business Process Modeling with ARIS. A Practical Guide. Springer-Verlag, New York, Inc. (2001)
6. Jakstonyte, G., Boguslauskas, V.: Graphic model regulating the application of land site taxation deductions. *Engineering Economics* 21(3), 238–243 (2010)
7. Karami, N., Iijima, J.: A logical approach for implementing dynamic business rules. *Contemporary Management Research* 6, 29–52 (2010)
8. Lohmann, N., Verbeek, E., Dijkman, R.: Petri Net Transformations for Business Processes – A Survey. In: Jensen, K., van der Aalst, W.M.P. (eds.) Transactions on Petri Nets and Other Models of Concurrency II. LNCS, vol. 5460, pp. 46–63. Springer, Heidelberg (2009)
9. Polyvyanyy, A., Weske, M.: Hypergraph-Based Modeling of Ad-Hoc Business Processes. In: Ardagna, D., Mecella, M., Yang, J. (eds.) BPM 2008 Workshops. LNBIP, vol. 17, pp. 278–289. Springer, Heidelberg (2009)
10. Pranevicius, H., Miseviciene, R.: Verification of business rules using logic programming means. In: Proceedings of the International Conference Modeling of Business, Industrial and Transport Systems, pp. 99–106. Transport and Telecommunication Institute, Riga (2008)
11. Pranevicius, H., Miseviciene, R.: Verification of business process workflows. *Technological and Economic Development of Economy* 18(4), 623–635 (2012)
12. Rima, A., Vasilecas, O., Smaizys, A.: Comparative analysis of business rules and business process modeling languages. *Computational Science and Technologies* 1(1), 52–60 (2013), Special issue for research innovations fundamentals
13. Russell, N., Wil, M.P., Hofstede, A., Wohed, P.: On the suitability of UML 2.0 activity diagrams for business process modeling. In: Proceeding APCCM 2006 Proceedings of the 3rd Asia-Pacific Conference on Conceptual Modeling, vol. 53, pp. 95–104. Australian Computer Society, Inc. (2006)
14. Sadiq, W., Orłowska, M.E.: Analyzing process models using graph reduction techniques. *Information Systems* 25(2), 117–134 (2000)

15. Sadiq, S., Orlowska, M., Sadiq, W., Foulger, C.: Data flow and validation in workflow modelling. In: ADC 2004 Dunedin, Conferences in Research and Practice in Information Technology, vol. 27, pp. 1–8 (2004)
16. Schumm, D., Karastoyanova, D., Leymann, F., Nitzsche, J.: On Visualizing and Modelling BPEL with BPMN. In: Grid and Pervasive Computing Conference, GPC 2009, pp. 80–87 (2009)
17. Smaizys, A., Vasilecas, O.: Business rules based agile ERP systems development. *Informatica* 20(3), 439–460 (2009)
18. Weske, M.: *Business Process Management Concepts, Languages, Architectures*. Springer-Verlag New York, Inc. (2012)
19. WfMC. *Workflow Management Coalition Workflow Standard: Workflow Process Definition Interface – XML Process Definition Language (XPDL) (WfMC-TC-1025)*. Technical report, Workflow Management Coalition, Lighthouse Point, Florida, USA (2002)
20. Ye, J., Sun, S., Song, W., Wen, L.: Formal semantics of BPMN process models using YAWL. In: *Second International Symposium on Intelligent Information Technology Application*, pp. 70–74. IEEE (2008)

Web-Based Analytical Information System for Spatial Data Processing^{*}

Viacheslav Paramonov, Roman Fedorov, Gennagy Ruzhnikov,
and Alexandr Shumilov

Institute for System Dynamics and Control Theory of Siberian Branch of Russian
Academy of Sciences (ISDCT SB RAS), Irkutsk, Russia
{slv, fedorov, rugnikov}@icc.ru

Abstract. It is considered an actual task of creation of analytical information system for spatial data processing. Currently research organisations, government agencies and local (municipal) governments have accumulated, update and use large amounts of scientific spatial data. As a rule data presented in different formats, which don't allow effective processing of them. It is proposed a development of analytical information system for spatial data processing which is represented in special web-based resource as GeoPortal. This kind of analytical information resource helps to ensure interoperability between different actors in the information exchange and improve the quality of researches.

Keywords: GIS, geospatial data, geoportal, Web-services, e-map, data storage.

1 Introduction

In recent years, many research organisations, government agencies, local (municipal) governments have accumulated various types of spatial data such as vector layers, remote sensing readings, GPS, GLONASS etc. Generally, organisations gather and store data in the most suitable way for their purposes. It leads to difficulties of data exchange between different organisations. There are many different methods of spatial data processing. Often these methods are used by developers only. Others have many difficulties to use them, for example obtaining program, installation, studying user interface, different formats of input data etc. However, complex research projects require an integration of data and processing methods of various researchers. The problem of data integration between different organisations requires solutions of many tasks such as information transmission, storage allocation, data formats coordination, usage of specific software, copyright protection etc. Therefore, the development of an analytical information system joining various methods and data is important task to integrate researchers and ensure their collaboration.

^{*} The research is partly supported by Russian Foundation for Basic Research, grant 12-07-98005-p_сибирь_a and Russian Academy of Science, grant FNM-49.

This article describes the development of analytic information system for geospatial data processing in Irkutsk Scientific centre. The major goal of this project is to build a shared informational space for spatial data exchange and processing for scientific, educational and government organisations in Irkutsk region. Main tasks of the system development are:

- establishment of a shared information space for research support in Irkutsk region;
- providing scientific spatial and thematic information resources of Irkutsk region;
- providing access to modern scientific methods of spatial data researches in Irkutsk region;
- providing web-services for gathering spatial data;
- implementation and use of international standards for representation of spatial information and modern methods and techniques for information exchanges;
- usage of existing data represented in different databases of spatial and thematic data.

2 Related Projects

The idea of creation and developing of Analytical information systems (AIS) for geospatial data processing is actual. It is possible to find many WEB-based informational resources for storage and handling spatial data which are often named as GeoPortals. For instance, such GeoPortals were developed in United States of America - <http://geo.data.gov/>, Canada - <http://geoconnections.nrcan.gc.ca/>, New Zealand - <http://geodata.govt.nz/>, European Union (EU) - <http://epp.eurostat.ec.europa.eu> etc.

The EU directive INSPIRE [2] defines common standards of interaction in the processes of working with spatial data and development of geospatial data informational system. The INSPIRE principles helps to communicate with EU informational geo-resources.

The whole set of free access resources are divided into two categories: cartographic services for information search and thematic-oriented services. Most of the Russian geospatial data informational systems are cartographic services such as:

- <http://rosreestr.ru> - Federal Service for State Registration, Cadastre and Cartography. It is a federal executive authority subordinated to the Ministry of Economic Development of the Russian Federation. It performs functions in the areas of state registration of rights to real estate and transactions; keeping of the state real estate cadastre; land administration and state land control; surveying and mapping; navigation support for the transport complex; state geodetic supervision; supervision of self-regulated organizations of appraisers; supervision (control) over the activities of self-regulated organizations of arbitration managers.
- <http://maps.dataplus.ru> - manageable Web server for hosting anything on the Web. From media streaming to web applications, IIS's scalable and open architecture is ready to handle the most demanding tasks. Server has API for user interaction with cartographical data.

- <http://geosamara.ru> – resource represents electronic maps and satellite images of Samara city and region. Resource is a part of work on the creation of a universal territorial cartographic system of Samara region inventory.
- <http://maps.yandex.ru> - search and reference service offering users detailed maps of some regions in America, Europa and Asia. The service is able to help to calculate distances, print maps and plan trips.
- <http://www.mirkart.ru> – the resource contains a set of interactive maps of some Russia and world regions. Resource services allow to make/delete user labels and icons. There is no any program API.
- <http://map.2gis.ru> –2GIS is regularly used for personal and business purposes in over 200 cities of Russia, Ukraine, Kazakhstan and Italy. Users can obtain information about geographically objects in cities areas, build routes between objects and calculate distances.
- <http://www.eatlas.ru> - cartographic reference and information website which contains detailed information of some regions of Russia. Cartographic services of website allows to map scale, print maps areas and calculate distances between geographic features.

3 Data Input, Output and Processing

Let’s consider the major functions for data input and processing in AIS which is called as Geoportal:

- data storage providing data getting in suitable for processing way;
- services for spatial data creating in Geoportal;
- execution of distributed methods (services) in Geoportal;
- delivery of data from data storage to service.

The architecture of the Geoportal is shown in Fig. 1.

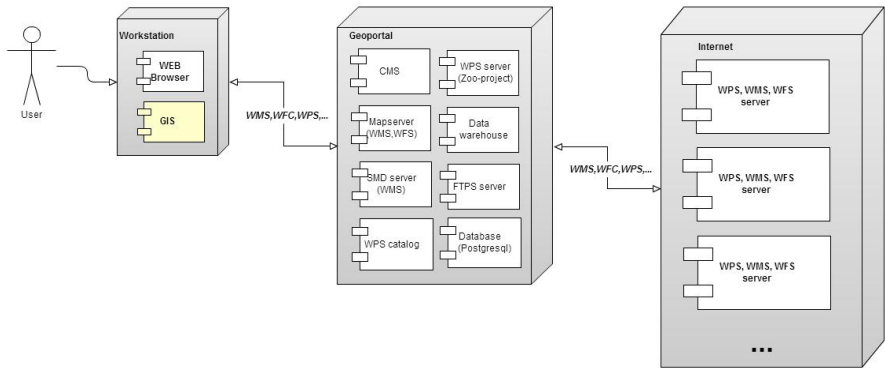


Fig. 1. GeoPortal architecture



The interaction between a user and the GeoPortal is carried out through any web-browser with JavaScript support. The GeoPortal contains the following modules:

- CMS - the core of Geoportal. It provides an interaction of all GeoPortal modules and allows an user to interact with the system. The CMS Calipso (<http://calip.so/>) has been used for this purpose.
- Mapserver (<http://mapserver.org/>) – set of libraries for spatial data visualisation;
- SMD server – it allows to visualize data of large-volume;
- WPS (Web Processing Service) catalog – WSP-service registration;
- WPS server – management of geodata processing services. Currently, Zoo-project has been used [3];
- Data warehouse – it provides storage of user' data;
- FTPS server – it allows to safely upload and download user data. We use a FileZilla [4] server;
- Database – it provides storage and processing tabulated data.

GeoPortal services are enabled to interact with other services in Internet network if they have WPS support.

3.1 Data Storage

Specialized data warehouse is required to collect and process various spatial data. The storage of GeoPortal is required to be reliable, safe and protected. The data storage server and its software must support the handling of large volume data. Each user of GeoPortal has a special file-directory for uploading and storing any data. The creation of directory and determination of its access rights is implemented by GeoPortal CMS modules. The file system management in the user's directory is carried out by the special File manager and FTPS server [4] integrated with GeoPortal.

There are conversation services for documents in Microsoft Excel 97, 2003 and CSV (Comma Separated Values) formats. The conversation is possible for documents which are in data storage. The conversation module has methods of data processing providing recognition the structure of the tables contained in the Excel-documents, determine data types in Excel tables. The result of module processing is a creation and filling tables in GeoPortal database.

The database management system PostgreSQL (with PostGIS extension) is used for user and service data storage and processing. PostGIS extension allows storing spatial data and provides support of OGC standard [6]. The database management is carried out by CMS user interface functions.

3.2 File Manager

File Manager is a special service integrated in the CMS. It is used for all the basic operations over file system through a standard web-browser. These operations include uploading and downloading data from/to an user computer, creating folders, copying data and etc. Also it is possible to download and upload user data in archives (ZIP for

example). There are specialized services that enable compress and decompress user data in the storage system.

File Transfer Protocol (FTP) might be used for transfer of big volume of data. This protocol provides reliable data transmission. However it has serious security issues. Basically, data transmitted via it doesn't have any kind of protection. The user data is transmitted in clear (non protected) view that unacceptable for information such as a component of intellectual property for example. In this case it is suitable to use FTPS (File Transfer Protocol + SSL) transmission. At this protocol all the information is encrypted which provides safety for user data and parameters his account.

To implement GeoPortal access by FTPS Server FileZilla was installed and configured. FileZilla is a open source file server with GNU-license. It supports FTP, SFTP, FTPS data transmission protocols. This server is cross-platform so it can be used even if the software platform, which GeoPortal based, will be change [6].

3.3 Services for Creating Spatial Data

Services for creating spatial data has been developed. They allow user to create (without programming) relational tables in PostgreSQL with spatial attributes. For a table an user interface is generated by the services (fig. 2). The user interface implements following functions:

- data visualization on map and as a table;
- input data, including spatial data (points, polylines, polygons);
- sorting, filtering data.

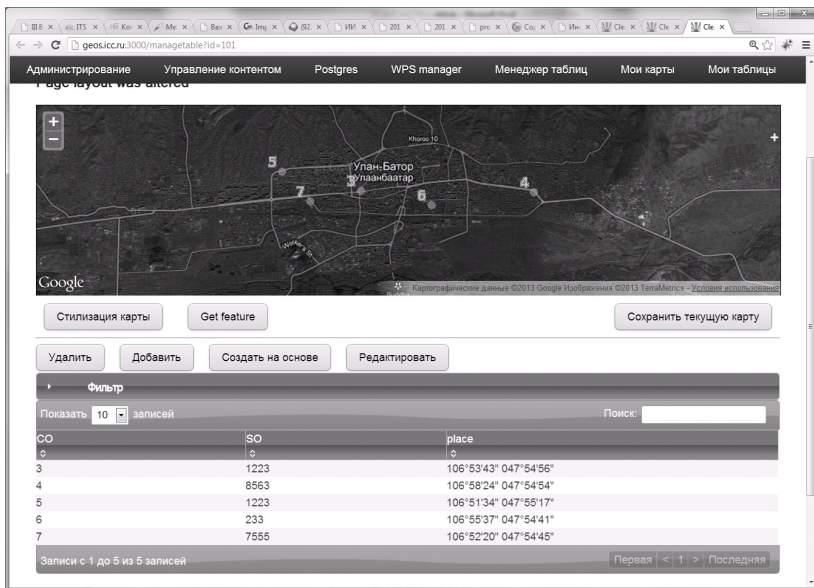


Fig. 2. The user interface for table editing

Data editing and analysing should also be based on the classifiers. Currently, the services for creating spatial data can use the following classifications:

- International Classification of Diseases (ICD) -10 [7];
- The International Plant Names Index (IPNI) [8];
- Taxonomic Classification of the biological diversity of animals in Russia [9].
- For all classifications there are controls for data editing.

3.4 Metadata

To store and process user's data it is useful to use metadata. Metadata, for example, helps to describe user data and organize a possibility for their share usage.

Metadata and their catalogues are used for control of search, creation, registration, storage, updating, processing and statistical analysis of processing spatial data. Catalogues of spatial metadata are data base (DB) containing indexed metadata about spatial data. Selection a profile set of metadata depends on what kind of information objects contained in this AIS service. It should be noted that currently metadata are used for search only. With the development of data processing services it is necessary to expand the use of metadata for process automation. Therefore, the development of tools for working with metadata (including metadata sets, languages, metadata descriptions, etc.) are essential to ensure global access to information and its productive use for automated obtaining of the result. The multiformity of existing AIS don't allow to create a single universal profile of metadata for widespread use. The profile which accumulates all usable standards will be very complex, unwieldy and hard to understand and imply the high cost of its implementation, support, maintenance, personnel training etc. Therefore the methodology metadata development and use should comply with the following principles:

- support for multiple profiles metadata sets that obviously cover the future needs of AIS users;
- software for displaying all the profiles in one which adopted as the standard of data exchange;
- storage of all metadata in the centralized database catalogue (repository);
- the original set of metadata must be possible to convert into one of the approved exchange standards.

The common metadata database storage can have such records as:

- at the common, content, purpose, protocols, formats, sets the conditions for access to spatial data which are data exchange format (Dublin core or ISO 19115 for example);
- at the level of concrete data object - in one of AIS metadata formats;
- normalized records of specific objects - original elements (objects) set mapping to some exchange format.

A relation database contains an extensible set of metadata about attributes. The database metadata stored can be translated into any given metadata format.

4 Services

Open GIS Consortium has created standard Web Processing Services (WPS) for spatial methods interaction through Internet. Using this standard it is possible to execute distributed methods through Internet. Within AIS it has been developed following components:

- WPS catalogue, it needs for registration of WPS services;
- WPS server, it needs for executing local methods;
- a module for service executing.

4.1 WPS Catalogue

WPS catalogue implements registration of WPS services. A user must input address of service during the registration. Then the catalogue gets metadata about existing services. The user must choose a service. Then the catalogue gets a metadata about the service parameters. Then the user must define for every parameter a widget which will help input data, including data from the data storage. For every registered service the catalogue creates JavaScript methods which implements executing.

4.2 WPS Server

Besides executing local methods WPS server needs for supporting WPS standard. Zoo-Project [8] was chosen as a WPS server. It allows executing methods designed on various programming languages. It was modified for JavaScript language support and compilation possibility in Windows operation system.

4.3 Services Executing

The module for service executing has been designed for working on browser. On base of the registration information the module generates a form for parameters input. For every parameters is used a defined widget. Then the service is executed by calling JavaScript method. The services must use library GDAL [9] for reading and writing data.

4.4 Examples of Some Services

Service for Calculating the Density of Point Objects in the Cells of a Regular Grid

A service for calculation density of points in the cells of a regular grid has been developed. Layer of vector objects is a service input. As output we get the number of objects in the cells of a regular grid, in GeoTIFF [9] format. The user can specify the size of the cell and the treatment area.

Service for Calculating the Density of Linear Features in the Cells of a Regular Grid

The service has been developed for calculation the density of linear objects in the cells of a regular grid was developed. The input data of the service is the vector objects layer. The output is the total length of the linear objects in the cells of a regular grid, represented in GeoTIFF format.

Service for Statical Analysis

The service for statically analysis of geospatial data was created. The service represented as a compiled dynamic library written on C++ language for WPS method of ZOO server. The mathematical methods into the library use methods of Alglib library [10]. Alglib allows to implement many statistical and algebraic operations over scalar and vector data.

In our case input data represented two thematic layers in GeoTIFF format. Spatial data in these layers are converted into one-dimensional arrays. After that regression and correlation coefficients are determined. This kind of analysis helps to understand and estimate data dependence and influence. The output data is a conclusion about analyzed data dependence. The practical example of service usage is an establish of relationship between air pollution from forest fires and respiratory diseases in Bratsk city of Irkutsk region.

5 Data Visualization and Displaying

The module for visualisation spatial data has been created. Graphical data can be represented in different formats supported by GDAL library. Format examples are SHP and

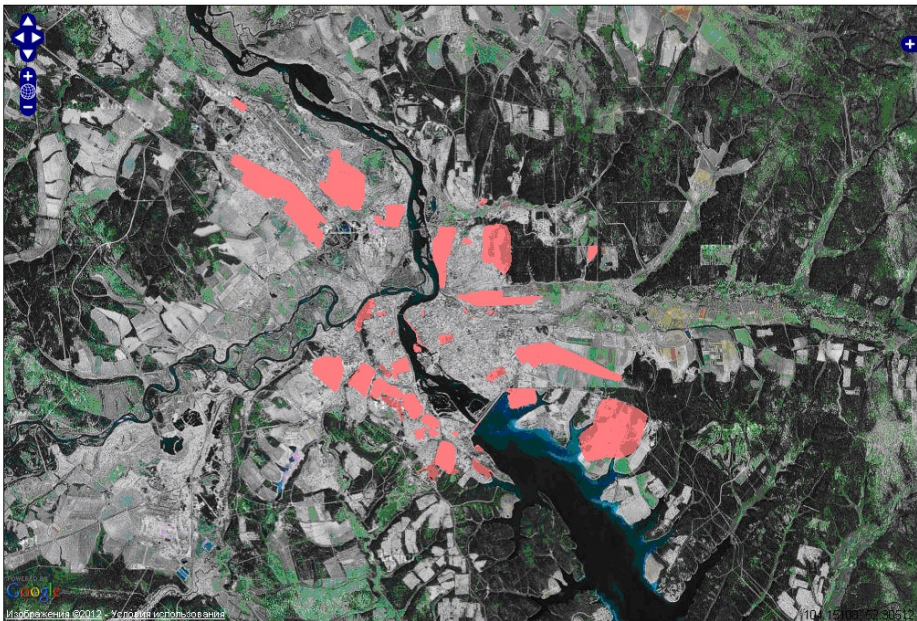


Fig. 3. Ixodes ticks activity in Irkutsk city area

GeoTIFF. The module is based on OpenLayers (<http://openlayers.org/>) and MapServer (<http://mapserver.org/>) libraries. It is enable to specify data displaying (gradient filling of cells or objects in depends of them value). It is possible to build a level curves. Google maps and Yandex maps are used for data displaying. Thus the created module provides thematic map creation and access to them via Internet.

Generated thematic map indicates a Ixodes ticks activity in the area of Irkutsk city [11] shown on fig. 3.

6 Conclusions

In this paper, we have presented the approaches and methods of creation of Analytical information system for geospatial data processing. The main idea of this project is to develop the system which would allow to organize interaction and shared access to various spatial data from different scientific and government organisations. The combination of the services for creating spatial data and WPS-services allows simplifying usage services with user's data. The user is not required to post additional data created on Web-servers. The wide set of controls allows to generate a user-friendly interface to call services. Temporary GeoPortal is accessible in Intranet network of Irkutsk scientific-educational complex. GeoPortal architecture allows you to extend the functionality by WPS-services accessible via the Internet.

References

1. Belussi, A., Catania, B., Clementini, E., Ferrari, E.: Spatial Data on the Web, 326 p. Springer (2007)
2. Infrastructure for Spatial Information in the European Ccommunity, <http://inspire.jrc.ec.europa.eu/>
3. Bychkov, I.V., Ruginov, G.M., Hmelnov, A.E., Fedorov, R.K., Gachenko, A.S., Shigarov, A.O., Paramonov, V.V.: Creation of the spatial data infrastructure for the territorial development management // Kemerovo State university herald (Вестник Кемеровского государственного университета), 52(4), 30–37 (2012) (in Russian)
4. Kresse, W., Fadaie, K.: ISO Standards for Geographic Information, 322 p. Springer (2004)
5. ICD 10 International Statistical Classification of Diseases And Related Health Problems / World Health Organization, 1200 p. (2004)
6. The International Plant Names Index (IPNI), <http://www.ipni.org/>
7. Information system "Biodiversity in Russia", <http://www.zin.ru/BioDiv/index.html>
8. ZOO open WPS-platofrm, <http://www.zoo-project.org/>
9. Shekhar, S., Xiong, H.: Encyclopedia of GIS, 1370 p. Springer (2008)
10. Bochkanov, S.: ALGLIB, <http://alglib.net>
11. Ruginov, G.M., Danchinova, G.A., Fedorov, R.K., Khasnationov, M.A., Paramonov, V.V., Lyapunov, A.V.: Modern technologies for informational and analytical evaluation of activity and forecast of spatial distribution of ixodid ticks by the example of Irkutsk City.// Bulletin of the Siberian Branch of the Russian Academy of Medical Sciences (Бюллетень Сибирского отделения ПАМН), 32(6), 55–59 (2012) (in Russian)

System Architecture Model Based on Service-Oriented Architecture Technology

Tarkan Gurbuz¹, Daina Gudoniene², and Danguole Rutkauskiene²

¹ Middle East Technical University, Ankara, Turkey
tarkan@metu.edu.tr

² Kaunas University of Technology, Kaunas, Lithuania
{daina.gudoniene, danguole.rutkauskiene}@ktu.lt

Abstract. This paper involves the research of the tools that have a positive effect on the quality, effectiveness and value of eLearning. Due the complexity of actual software systems, including web portals and networks, it is becoming more and more difficult to develop software systems suitable for their intended usage. To tackle this problem, we can develop an integrated system for curriculum planning and delivery by using new technologies with a range of individualized constructive learning strategies and social skills acquired through constant communication, active sharing of knowledge and experience, joint activities in various groups, teamwork and training (learning) environments and social networks, with development and evaluation of the work performance.

Keywords: service-oriented architecture, information technologies, applications, eLearning.

1 Introduction

There are different categories of the tools to be used for different aims and possibilities for curriculum implementation (realization of learning events: imitate, receive information, exercise, explore, experiment, create, self-reflect, debate); technological properties (e.g. synchronous, asynchronous, web based, PC application, mobile app, open source, free service); similar or related tools; application domain (language learning, intercultural competences, ICT skills, time management skills, study habits skills, etc.). This paper presents a system architecture model designed and based on service-oriented architecture technology. This architecture allows flexible system-oriented service delivery with the ability to implement effective services and systems integration solutions.

The system is structured into the following logical levels: user level and system interfaces. At this level, users are realized via information systems interface with their necessary functions according to the system requirements. Users will have all the necessary functions. The model is based on the integrated system to be implemented by tools for online communication and collaboration to assure successful study process.

Aim of the research is to select system's architecture model based on service-oriented architecture (SOA) technology which can be integrated to an educational platform to implement eLearning design principles. The objectives are

- to select the tools which assure needs for successful communication, research, collaboration and
- to demonstrate the importance of social tools for eLearning design

The research methods used in this study are literature review, document analysis, content analysis.

2 Review on Tools Based on Service-Oriented Architecture Technology

Any eLearning strategy must include methods for designing and deploying learning solutions, change management, communication planning, performance support solutions, knowledge management services and technologies [1]. Many of actual educational software systems are adapted and used in Lithuania. Some of them are localized and used in national or institutional levels. Systems such as Moodle – eLearning course delivery system, VIPS - a video lecture system, additional integrated social networking tools are integrated into internal institutional systems and could be used by academic society to assure successful ICT based study process. We may not have an agreed set of characteristic forms of effective e-learning, but it is possible for the educational community to identify some effective existing SAM models. These would embody good design practice in a way that might impose requirements on the underlying eLearning architecture [2]. By analyzing the essential characteristics of a range of proven learning activities, we can generate a set of requirements for the architecture to support. For example, a proven existing learning activity might enable students to work simultaneously across a network on a design tool, such as a graphics program and share the results in separate windows. This learning activity therefore generates the computational requirement, which should be possible for 'any' shareable application to be used in this way.

Today, eLearning mainly takes the form of online courses. From the open resources distributed online to the design of learning to the offerings found from colleges and universities everywhere, the course is the basic unit of organization. As a consequence, the dominant learning technology employed today is a type of system that organizes and delivers online courses - a learning management system (LMS). This piece of software has become almost ubiquitous in the learning environment: companies such as WebCT, Blackboard, and Desire2Learn have installed products at thousands of universities and colleges. These products are used by tens of thousands of instructors and students [3]. The learning management system takes learning content and organizes it in a standard way as a course divided into modules and lessons supported with quizzes, tests, discussions and, in most cases, integrated into the college or university's student information system.

Today Moodle (PDF, Modular Object Oriented Dynamic Learning Environment) an open source, web-based virtual learning environment designed to help teachers organize

online learning courses is used at Kaunas University of Technology (KTU), Lithuania. Because Moodle is an open system, it is distributed for free and can be adapted to your needs, without any license agreement. Furthermore, the language you want to translate can be used without restriction. The system is successfully used not only in higher education but also in vocational and general education as the main features of Moodle are simple and convenient: courses can be sorted by different categories, search can be performed, training courses in the listed course descriptions, easy installation, already installed Moodle system can be supplemented with new modules, integrated data security measures.

Another tool selected for recording and online translations of the lectures is VIPS, a system for broadcasting conferences or lectures. Its purpose can be divided into two parts: firstly, it allows teachers to live broadcast their courses given by students; secondly it allows other people to broadcast the conference live via the internet. Streaming video is currently saved and stored on the server, ready for later viewing. VIPS system assigned a space for record creation. Those records are always stored in one place, so they are easy to reach to other downloaders users. This is very convenient because the space name can be the same as the teacher taught module name. All users have the ability to keep track of the selected areas. They can create a list of spaces and then be aware that the new record has not been reviewed there. When a remote area of building service is used data about the area goes into the system, used by the service. Also, the wider potential of information might be: consumer video preview, review date; related video to review; related spaces suggestions; the most popular or newest video information; current areas of information about the new record.

Social network Elgg is intended to be used for communication and collaboration. Drupal CMS, Moodle, Elgg, VIPS and simulation gaming system will be integrated with each other in both functional and user level: Drupal CMS central authentication service (CAS) allows the Drupal CMS to act as a central user authentication and log-storage base and maintain SSO (Single Sign-On) protocol with Moodle, Elgg, and other systems (Fig1).

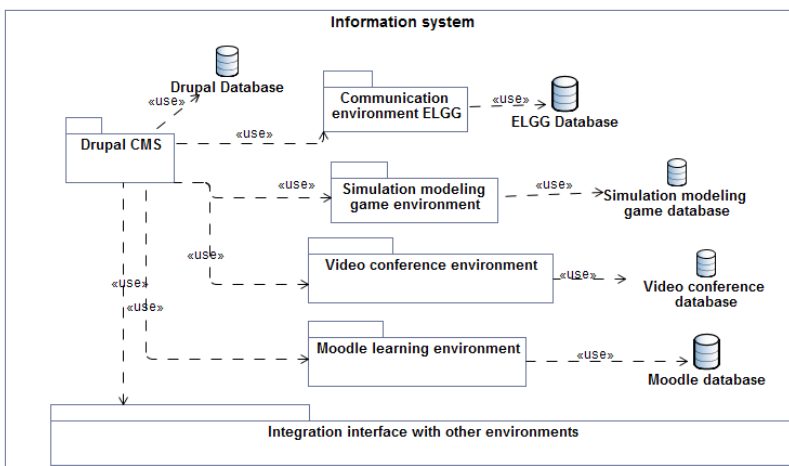


Fig. 1. System components integration

For effective use of digital learning content, it is essential to have the material as learning objects and organize them in a proper way, so that it would not get lost in big amounts of data and be reused again in creating new courses or complementing the old ones. Content can be reused in different contexts, not only in original, in which it was authored because of the learning object technology [4]. Reuse is important due to the fact that authoring of high quality e-learning material is expensive and attempts to lower costs of e-courses design and deployment while keeping high quality standards is desirable [7]. The best way to stimulate the reuse of learning objects is to provide convenient tools for e-courses authoring, which would aid the idea of simple yet effective learning object reuse. The best tools are those which are easily embedded into the process of e-learning [17] and accepted by e-learning participants, so not only e-course authoring is important, but also the process of acquiring and evaluating knowledge, which was received by students from learning objects. Integration between learning management system and learning object repository could allow easy access for institution users to e-learning material search, creation, annotation and various modifications not requiring any special knowledge of web technologies [13]. In so far as users are acquainted with learning management system in their e-learning courses, it would be much easier to introduce learning object repository as an extension to LMS, but not as separate tool.

3 System Architecture Model Based on Service-Oriented Architecture Technology

There is a common KTU login system that provides information about a person trying to connect (for example, name, address, email address, user code), to any of the "login.ktu.lt" harmonized system. User identity shall only be required to confirm username and password (connecting for the first time requires to register). Lithuanian science and studies computer network Litnet is the provider of this system. The transfer of information between a single login system and the system to which you connect is used in SAML (Security Assertion Markup Language) service. SAML is supported in all "mano.ktu.lt" system components. User operation of the system diagram is shown in the example below. It is a unified login system that connects to KTU used systems - thus, there are no integration problems between different e-learning and social networking systems. When connected to one of the service, there is no need to log in to other systems for the second time because the system automatically checks online users and authorize them.

Designed system architecture model is based on service-oriented architecture technology (SOA). This architecture provides the ability to create a flexible system and focuses on the provision of services with the ability to implement effective services and systems integration solutions.

The system will be structured in the following logic levels such as user level and system interfaces. At this level, users are realized via information systems interface with their necessary functions according to the requirements. External systems (called adapters) are provided to them for servicing web services. In addition, the component system is required for external workshops call. The data layer implements all data management, sanctioning, monitoring, archiving and storage components.

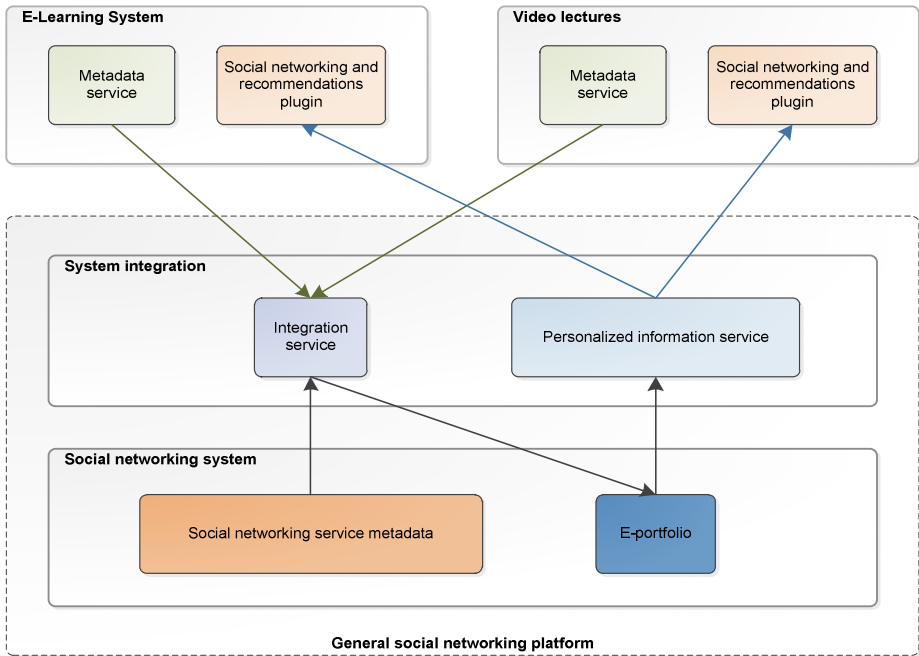


Fig. 2. System architecture SOA model

Overall information system (IS) architecture implements the data flow between the logic levels, omitting the middle layer. It means that the presentation and interface-level data management is based on the operating logic-level modules. Presentations and interface levels are user and system interfaces. At this level user and system interface are realized with their necessary functions according to the requirements.

Modular architecture of SOA concepts enables the following: ensures the flexibility of the system infrastructure, provides a simple system configuration, maintenance, development and support; provides a simple upgrade of the system, and the introduction of new features; logically and physically separates part of the system, reducing software development costs and ensuring any changes to the data transfer module for all the system components and modules; reduces the response time to change business requirements, operating environment and operation regulations.

3.1 User Interface Layer

From an external version of the module, and subsystems of the user interface will be realized under the thin client approach (called "thin-client"), hereinafter also referred to as "Web Interface".

The thin client (web-interface) in the context of this proposal is seen as a program that runs an online browser. "Thin client - client is a computer program that depends entirely on the central server, its main purpose is to display the data to the user and the data exchange programs for clients with a central server system, all data processing tasks moving server.

Although there are many possible variations, most web applications are multi-layer structure (Fig.3). The first layer is a user interface through which the user performs the desired commands. The second acts as the logical layer, coordinating processes, performing calculations and processing and transferring data between the other two layers. The third layer functions as a database management system, which contains all the necessary information. Client program to the user's query to the logical layer modules, the logical layer modules database finds all the required information, which can then be sent back to the second layer, which is processed and presented to the user in the first layer.

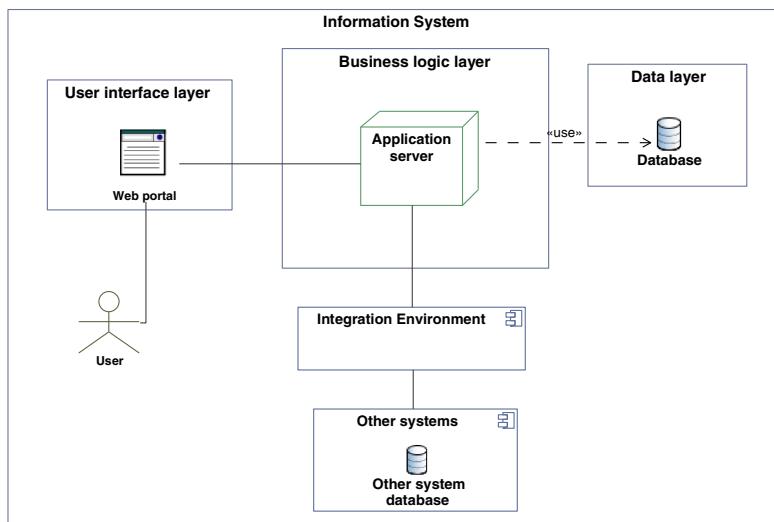


Fig. 3. Principle SOA layer structure

Version of the user interface information system will be developed by using Hyper-text Markup Language (HTML) format, Asynchronous JavaScript and XML (AJAX), meet Web Consortium (W3C) standards and recommendations, have a clear, functional, systematic and attractive design. In addition, Web Content Accessibility Guidelines 2.0 (WCAG) will be applied; they describe and standardize the user interface development with the purpose of making web content more accessible (especially for people with disabilities).

Version will be used for the realization of not less than 1.0 W3C XHTML and level 2 CSS2 (Cascading Style Sheets Language 2 w3.org/Style/CSS/) version. Furthermore, the realization meets the following principles: availability, security, privacy, and the use of open standards.

The main proposed thin client benefits are: simplified system administration. Internet browser installed on the client computer or a program installed or upgraded to the central server is enough. Requirements on the client computer hardware (enough to operate an online browser) are not high and the hardware is easy to reach. The user can operate with the system from any computer's web browser as client. The information can be

accessed from anywhere. Each user of the system is usually your computer which has a web browser. No additional configuration is required. There is a centralized way to organize data stored on reliable and easier to do backups. Updates do not require additional installation of the guest computer, all software is automatically updated and user instructions are easy distributed via the web interface.

Basic thin client weaknesses are: complete reliance on computer networks and the central server system, disruption of communications or inoperative central server, users work stops, print weaknesses. Since the program operates in the online browser, it can only use the browser print options granted, which varies depending on the browser and its faults (such as the use of headers and margins, poor paging). This aspect makes it more difficult to use peripheral devices (scanners, faxes, etc.). Web browser has very limited rights of the client operating system sources and without installing additional software some peripheral devices will fail. Compared to the "thick" client, "thin" client has less user interface elements to choose, some of them are slower and less user-friendly to work (hot keys, etc.).

In order to eliminate the above-mentioned shortcomings, systems will be developed not only in HTML but also in AJAX technology. This ensures some modal windows, more dynamic, compared with only HTML-based solutions, screen format auditability, without sending all the information to the server and not redrawing screen form again. The client from the server side only exchange information and use other AJAX opportunities.

The system architecture of the data will not be realized in any software components, such as data quality and integrity checking, object state changes in work sequence, and so on. These components will be realized at the level of logic.

IS system architecture will be based on the principles of SOA and operational functions will be provided as a service (PDF, etc.). Internal service processes that streamline the provision of direct procedure calls may be disposed directly in the application server procedures. What kind of activities are to be performed as a service (services), or directly, will be agreed upon for the analysis and design.

In the Web services (PDF, workshop) system components will benefit from the IS system functionality that provides a specific service, for example: initiate a business process to record data, to get a report, and so on. In addition, the service component of the system will be used to work in the sequence (called - workflow). Workflow can be used both for system and external systems services.

3.2 Functions Assured According to the User and System-Level Interface

Tailored interface is realized in the Republic of Lithuania the time, date, and currency standards. The interface will use UTF-8 encoding standard for providing and processing data to external system. In general, all the interface data management needs for realizing business logic layer components are applications and middleware servers. This decision is consistent with the service-oriented and multi-level systems architecture principles.

The external version of the graphical interface for external users will be realized by thin-client approach, with minimal need for additional software outside the user's workplace.

If necessary, the user's workplace additional software will be free and will include direct software download of the latest version from the official software distributor. Software installation instructions will be illustrated.

IS will be focused on the user and the graphical interface will be realized based on the best usability. System usability and fitness profile can be found in the system comfort and usability of the user interface.

All of system users will have a single integrated working environment. Interfaces and dialogue with the user is realized in accordance with the standard Lithuanian language rules. All of the error messages of the information system will be presented in the Lithuanian language, according to the character properties.

Error messages will be formulated and appear in Lithuanian language so that the user of the system clearly knows what happened and what actions he should do so in order to continue working.

Data sorting and search rules will be presented in the Lithuanian language and the data consisting of the Lithuanian character sorting will be carried out according to the Lithuanian alphabet.

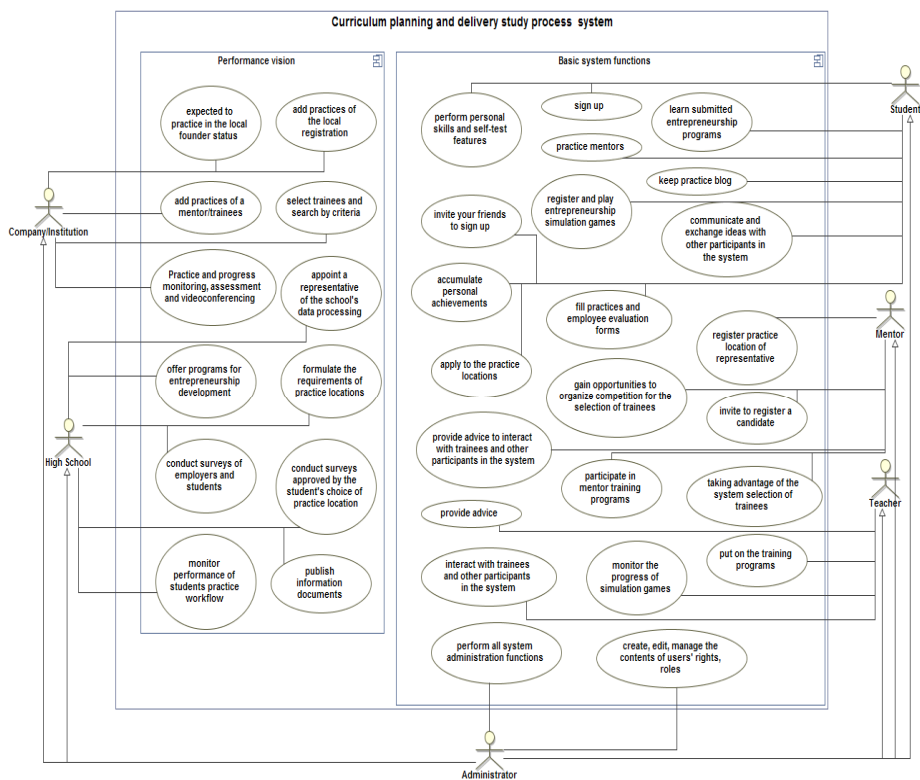


Fig. 4. System functions

Context sensitive help to the user will be created in the system. This will help perform business functions intuitively. Portal System will provide the system user interfaces with existing contextual aid user and portal map display functions:

- shallow browsing of information within a structured content
- deliver information from application-context, which was summoned as a help function
- tracking mode and the opportunity to seek information from the context
- information printing

The system user interface will be constructed on the basis of no lower than version 1.0 W3C XHTML specification, or equivalent and meet modern ergonomic requirements specified in ISO 9241-110:2006 ISO 9241-151:2008 or other equivalent standards and ensure convenient access to the system functions implemented.

It also provides the system administrator's role is to manage all the other roles and content. IS users roles and rights will be reviewed and agreed with the customer service at the stage of analysis tools.

3.3 ICT Model: Theoretical Background

Taking all of the above into consideration, it becomes apparent that we are promoting a constructivist model of education enhanced by the use of ICT. Although innovative ideas on teaching and learning have been progressively introduced over the past few decades, traditional views have been difficult to change. Such views often consider students as “empty vessels” waiting to be filled with knowledge. Students are now learners who come to the classroom with their unique backgrounds, experience, conceptual understanding, learning styles and personal circumstances. Teachers are now becoming learning facilitators rather than reservoirs of knowledge. Psychology of learning has shifted from behaviorism to cognitivism and constructivism.¹

Constructivism states that learning is an active, contextualized process of constructing knowledge rather than acquiring it. However, this does not mean that teachers become redundant – instruction is not “outlawed”- it is another one of a series of ways that knowledge is gained. Constructivism assumes that all knowledge is constructed from the learner’s previous knowledge, regardless of how one is taught. Thus, even listening to a lecture involves active attempts to construct new knowledge.²

Further, in order to use technologies optimally, teachers must be comfortable with a constructivist or project-based, problem solving approach to learning: they must be willing to tolerate students progressing independently and at widely varying paces, trust students to sometimes know more than they do, be flexible enough to change directions when technical glitches occur.³

¹ Chan, D. The Role of ICT in a Constructivist Approach To the Teaching of Thinking Skills
<http://www.learnerstogether.net/PDF/ICT-in-Constructivist-Teaching-of-Thinking-Skills.pdf>

² <http://www.learning-theories.com/constructivism.html>

³ Foa, L., Schwab, R.L. and Johnson, M. (1996) Upgrading school technology. Education Week, 52.

Our constructivist, ICT enabled, education environment will see students participating in knowledge construction, filtering and dissemination through collaborative, technology enhanced activities. These may be separated by time and space but will exist within a meaningful context i.e. real life scenarios and direction where learning will be reviewed, criticized and embedded through peer-exchange and teacher/guide reinforcement.

The graphic in Fig.5 demonstrates the outline of our model which is explained in detail below. The model should be practical for any curriculum area or scenario.

It is also important to recognize that while we refer, here, to the “session” this need not necessarily occur in a single classroom session – it may be extended over time and space to encompass remote work, individual and group work outside the classroom and asynchronous work.

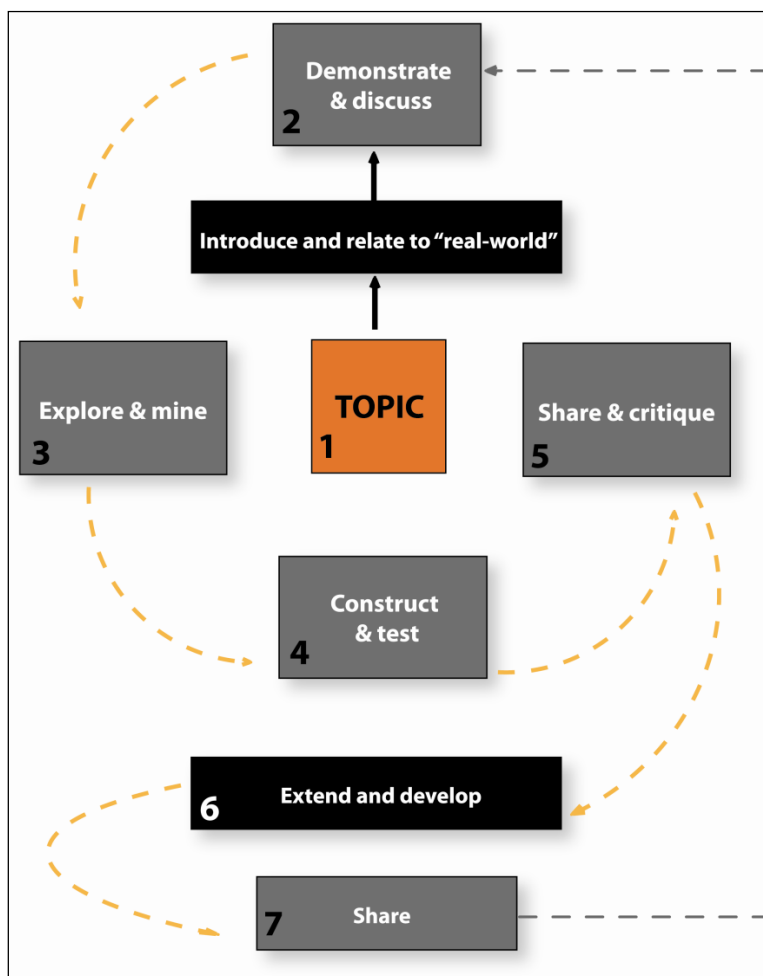


Fig. 5. ICT Model for Education

Information communication technologies are the whole of digital approaches and tools which allow creating, collecting, storing, transforming and disseminating the information. It is needed to emphasize that the purpose of information that is transferred using these technologies is to define, communicate, collaborate, share information, and all of this is ensured by various ICT tools.

4 Challenges in ICT

ICT has an important place the field of education – it is widely used in educational process. Different learning technologies are taking their place in the classrooms. More innovative solutions and learning environments are providing us the possibility to develop the content to be interactive and interesting to any age group. Transferring textbooks into interactive books and creating complementary interactive lessons for both mobile and computer use, may create this opportunity.

ICT has an impact on nearly every aspect of our lives - from working to socializing, learning to playing. With regard to educational context, the purpose of ICT could be comprehended widely as general learning, cooperative learning, reflection, etc. The digital age has transformed the way young people communicate, network, seek help, access information and learn. Teachers must identify that young people are now an online population and access it through a variety of means such as computers, TV and mobile phones.

The learners need to be involved in the online learning process and to benefit from social interaction while learning, consequently, collaborative online learning can be quite productive.

5 Conclusions

ICT is very important in today's education – it adds value to an educational process. The users today tend to choose traditional textbooks that are complemented by their digital equivalents or supplements. Different tools and environments give us the possibility to develop the content to be interactive and interesting to any age group by transferring textbooks into interactive books and create complementary interactive lessons for both mobile and computer use, together with content assessment and social annotation tools.

The system architecture model designed based on SOA technology allows flexible system oriented service delivery with the ability to implement effective services and system integration solutions.

The information system is structured into the following logical levels: user level and system interfaces. At this level, users are realized via IS interface with their necessary functions according to the system requirements. Users will have all the necessary functions. The model is based on the integrated system to be implemented by tools for online communication and collaboration to assure successful study process.

Integration of social networking platform with learning management, academic information and other systems allows to exchange information among them and to

collect user's artifacts into personal learner's e-portfolio. Accumulation of other data about learner's activities and communication patterns allows creating a pool of meta-information that can be used in development of intelligent recommendation widgets.

References

1. Rosberg, M.: eLearning Strategy, The eLearning Guild, <http://www.elearningguild.com/publications/index.cfm?id=7> (accessed May 2013)
2. Laurillard, D.: Design tools for eLearning, http://cms.ascilite.org.au/conferences/auckland02/proceedings/papers/key_laurillard.pdf (accessed May 2013)
3. Downes, S.: eLearning 2.0, <http://elearnmag.acm.org/featured.cfm?aid=1104968> (accessed May 2013)
4. Redecker, C.: Learning2.0: The Impact of Web2.0 Innovation on Education and Training in Europe, Spain, 122 p. (2009)
5. Burneikaitė, N., Jarienė, R., Jašinauskas, L., Motiejūnienė, E., Neseckienė, I., Vingelienė, S.: Informacinių komunikacinių technologijų taikymo ugdymo procese galimybės, Vilnius, 231 p. (2009)
6. Gray, D.E., Ryan, M., Coulon, A.: The Training of Teachers and Trainers: Innovative Practices, Skills and Competencies in the use of eLearning. European Journal of Open, Distance and E-learning (accessed May 2013)
7. <http://www.eurodl.org/?p=archives&year=2004&halfyear=2&article=159>
8. Rutkauskienė, D., Gudonienė, D.: E. švietimas: tendencijos ir iššūkiai. Konferencijos pranešimų medžiaga: Web 2.0 saitynas, Vilnius, 110 p. (2005)
9. Mockus, J.: Investigation of Examples of E-education Environment for Scientific Collaboration and Distance Graduate Studies. Part 2, Informatica 19(1), 45–62 (2008)
10. Kuzucuoglu, A.E., Gokhan, E.: Development of a Web-Based Control and Robotic Applications Laboratory for Control Engineering Education. Information Technology and Control 40(4) (2011)
11. Bajec, M.: A Framework and Tool-Support for Reengineering Software Development Methods. Informatica 19(3), 321–344 (2008)
12. Bersin, J.: Social Networking and Corporate Learning, Certification Magazine, vol. (10), p. 14. MediaTec Publishing Inc (2008)
13. Besson, J., Lupeikiene, A., Medvedev, V.: Comparing Real and Intended System Usages: A Case for Web Portal. Informatica 23(2), 191–201 (2012)
14. Fertalj, K., Hoic-Bozic, N., Jerković, H.: The Integration of Learning Object Repositories and Learning Management Systems. Computer Science and Information Systems (7), 387–407 (2010)
15. Grodecka, K., Wild, F., Kieslinger, B.: How to Use Social Software in Higher Education, Poland (2009)
16. Jucevičienė, P., Valinevičienė, G.: A Conceptual Model of Social Networking in Higher Education. Electronics and Electrical Engineering 6(102), 55–58 (2010)
17. Peters, K.: M-Learning: Positioning educators for a mobile, connected future. IRR ODL 8(2), 66–75 (2007)
18. Targamadze, A., Petrauskiene, R.: Impact of Information Technologies on Modern Learning. Information Technology and Control 39(3), 169–175 (2010)

Towards the Combination of BPMN Process Models with SBVR Business Vocabularies and Rules

Eglė Mickevičiūtė and Rimantas Butleris

Kaunas University of Technology, Department of Information Systems, Studentu 50-315a
Kaunas University of Technology, Centre of Information Systems Design Technologies,
Studentu 50-313a

{egle.mickeviciute, rimantas.butleris}@ktu.lt

Abstract. Combination capabilities of BPMN and SBVR are analyzed in this paper. In order to combine these two standards we have to analyze current proposals. Process modeling focuses on visualization of process with specific notation. However, today there is a need to have business rules separately from process in order to reduce the size of the process model and avoid misunderstandings and miscommunications between analysts and domain experts or between organizations. Therefore, there is a need to combine business process management and business rule management in one user-friendly environment.

Keywords: BPMN, SBVR, business process, business rule, integration, transformation.

1 Introduction

Current business process management (BPM) consists not only of management of business process modeling, it also includes management of business vocabularies and business rules. Nowadays, many people come to the agreement that BPM and business rule management (BRM) should stand together, because these two approaches complement each other. But as the situation shows, these two approaches have their own problems when relating them. Management of business process modeling includes graphical models of a case study, however management of business vocabularies and rules has to deal with limited natural language. In order to have complete business process representation, the combination of business processes (BP) and business rules (BR) is necessary. Business vocabulary is an essential element that links the dynamic and static process aspects. From the business perspective, the use of business rules helps to simplify complex decisions and computations [7]. In this case, rules reduce the process diagram.

Because of existing gap between process modeling and specification of business vocabularies and rules new methods and theories are needed to help these two domains work as complementary rather than competing methods during system development. Later researches showed that neither approach is sufficient to express all required details [12, 8]. In this paper we analyzed Business Process Model and Notation (BPMN) and Semantics of Business Vocabularies and Rules (SBVR) OMG

standards and analyzed some proposals how these two standards can be combined dealing with integration and transformation. Analyzed papers showed that not all was fully implemented in this area and there are a lot of to be done in integrating business rules and business process.

Further in this paper, related work is introduced in section 2. The basics of BPMN standard is presented in section 3. Section 4 deals with another OMG standard – SBVR. In section 5 BPMN and SBVR integration is presented and section 6 describes BPMN and SBVR transformation. Conclusions are presented in Section 7.

2 Related Work

With increasing needs for business agility and cost pressures on IT, BPM is asked to move towards “Dynamic BPM” and “Intelligent Case Management” instead of freezing process flow in hard-to-change IT solutions [9]. Gartner published a report and introduced seven scenarios of how processes and rules can be defined. These seven scenarios differ from the most static to the most dynamic scenario. The reality is a continuum of how much activity is put into rules and how much activity is put into process [14]. Later, in Koehler report [9], these scenarios were critically reviewed and argued that they can be reduced to four key patterns of rule usage. These patterns differ in their usage of domain- vs. meta-level rules and structural vs. operative rules.

Analytic principles were used to identify the overlapping of BR standards (SRML, SBVR) with Petri Net, IDEF3, EPC and BPMN [16]. The best representation power of business processes with minimum overlapping is characterized by combinations: BPMN with SWRL and BPMN with SBVR. Other researchers group suggested three types of mappings of BPMN and SBVR meta-models and chosen one the most appropriate way to map these two meta-models [15]. One of the proposals was to divide two process modeling categories to procedural and declarative modeling [1]. The main aspect is to separate process navigation from restrictions of process elements. The other researchers group suggested an approach for integration of process and rule to complement design approaches for integrated modeling of processes and rules [4]. They studied differences between two notations: business and rule modeling.

There are proposed ways of BPMN transformation to natural language. The presented approach employs SBVR as an intermediate representation to generate natural language due to its easy comprehension [10]. The other article presents their automated approach to translate BPMN based business process models to SBVR based natural language representation [11]. Authors present their tool, which was implemented in Java, with a simple running example.

Some mentioned methods and proposals will be analyzed in section 5 to have a deeper insight in BP and BR combination.

3 Business Process Model and Notation (BPMN)

BPMN is a platform independent process modeling notation [2]. The first aim of BPMN was to develop commonly understandable and usable notation from the business analyst

which provides primary design of process, technical developer who is responsible for realization and finally to business staff, who is responsible for installation and maintenance of a system [5]. The newest version is BPMN 2.0 and it significantly differs from the older version of BPMN: new elements, new diagrams and other new features were added.

BPMN is designed to model the dynamics of the business process. The main elements of BPMN are four types of diagrams and five types of modeling elements. BPMN diagrams are process, choreography, collaboration and conversation diagrams. BPMN modeling elements are divided into groups by their purpose: flow objects, connecting objects, swimlanes, data and artifacts.

BPMN concepts are divided into two groups: main BPMN modeling elements and extended BPMN modeling elements. Extended BPMN modeling elements include variations of elements from the main group.

BPMN meta-model is BPDMM which predecessor is meta-model MOF, and thus ensures full compatibility with other OMG group standards such as UML, SBVR, OSM, BMM. Because of the existing gap between process modeling and specification of business vocabularies and rules we decided to choose BPMN. The fact that BPMN is commonly understandable and usable notation that is compatible with other OMG standards such as SBVR influenced the decision.

4 Semantics of Business Vocabulary and Rules (BPMN)

SBVR is an OMG standard that specifies business knowledge in language understandable for business actors and enables them to use notation that does not require specific IT knowledge. The representation of SBVR is based on Concept Diagram Graphic Notation that refers to the RuleSpeak language [3]. Organizations or systems use XMI schema to exchange business vocabularies and rules. SBVR enables specification of business knowledge in limited natural language that can be easily understandable for business people.

SBVR business vocabulary consists of concepts, fact types. There are two types of business rules in SBVR: operative and structural rules that can be used to restrict business process. Operative rules depend on the actions performed by people and can be violated, while structural rules cannot be violated.

SBVR standard consists of four elements [13]: Meaning and Representation Vocabulary (MRV), Logical Formulation of Semantics Vocabulary (LFSV), Vocabulary for Describing Business Vocabularies (VDBV) and Vocabulary for Describing Business Rules (VDBR). Figure 1 illustrates how these four elements are associated with each other.

There are two different interpretation aspects of SBVR meta-model and vocabulary [13]: Meaning and Representation. Business vocabulary and business rules have to be defined at CIM level in Model Driven Architecture (MDA) as Business Process Modeling.

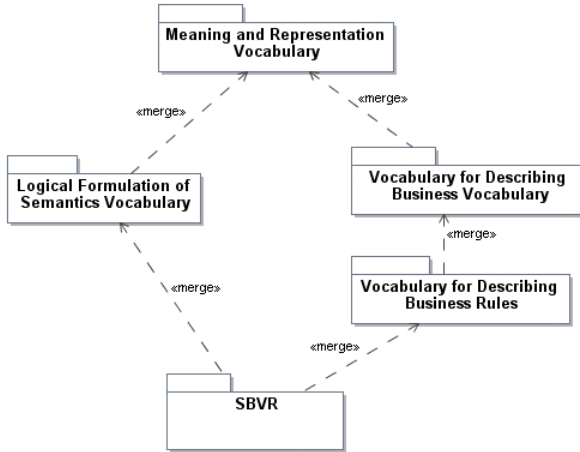


Fig. 1. Four elements of SBVR [13]

Since SBVR appearance, developers tend to express business rules in natural language statements using a list of commonly agreed terms and definitions. Later, OMG and some BPM and BR communities released SBVR standard which was welcomed by business and IT areas. Because the fact that SBVR is OMG standard as well as BPMN, and they are compatible, we have chosen SBVR.

5 Combining BPMN and SBVR

In order to properly combine these two OMG standards some researchers proposed to split SBVR into separate parts [1, 6], because business process model is outside the scope of SBVR. As Agrawal proposed [1], SBVR could be categorized into two types: procedural modeling that defines a process as a sequence of process elements and declarative modeling that defines a process as a set of process elements and declarative statements representing constraints over them. These two types of categorization were named Semantics of Business Process Vocabulary and Process Rules (SBPVR) and SBVR. After this separation the standards could be combined properly.

5.1 Integration of BPMN and SBVR

One researcher group proposed three possible ways to implement the integration between two different meta-models [15]. The best way they presented in the paper was to develop supplementary mapping data structure. Linking of BPMN elements with SBVR specification was implemented through element Text Annotation, by attaching specific stereotype << SBVR >> to it. Table 1 illustrates mapping pairs of elements of BPMN and SBVR. In the presented table, symbol “x” means that a certain element type from BPMN meta-model has its correspondence with some element type from SBVR meta-model.

Table 1. Mapping pairs of elements of SBVR and BPMN meta-models [15]

BPMN Category	BPMN Element	SBVR Noun Concept	SBVR Fact Type	SBVR Rule
Flow Objects	Event	x	x	x
	Activity	x	x	x
	Gateway	x	x	x
Connecting Objects	Sequence Flow	x	x	x
	Message Flow	-	-	-
	Association	-	-	-
Swim-lanes	Data Association	-	-	-
	Lane	x	x	-
	Pool	x	x	-
Data	Data Object	x	x	x
	Data Input	x	x	x
	Data Output	x	x	x
	Data Store	x	x	-
Artifacts	Group	x	-	-
	Text Annotation	-	-	-

Integration of BPMN and SBVR require mapping elements of each standard. All attempts to do that are with limited number of elements of BPMN. In order to have full integration there is a need to have most of elements mapped with SBVR elements, including those who are the most commonly used. Different types of gateways, events, tasks, flows and data objects should be used.

In order to make these two standards work together, differences and similarities should be analyzed. This process involves construct-by-construct analysis of each of the two notations to determine where equivalent exists [4]. However, there are numbers of constructs that cannot be mapped easily because expressiveness could be lost.

There is no proper implementation that would combine BPMN and SBVR in one place. It means that there is no tool where specialist could use SBVR and BPMN at the same time (for example, while creating process model a specialist could use SBVR with graphical notation). In order to implement this solution we need to have aspects (for integration and transformation) that must be evaluated:

1. How to represent SBVR in graphical notation;
2. How to provide business vocabulary and rule management for the user;
3. Should SBVR separation (to separate business rules from business process rules) be used;
4. Should linguistic techniques (transformation problem) be used;
5. Should a possibility for bidirectional transformation exist (related with aspect 3).

5.2 Transformation of BPMN and SBVR

An automated approach to translate BPMN based business process models to SBVR based natural language representation was presented by Malik and Jajwa Sarwan [11]. Operation of their developed BR-Generator tool is based on mentioned approaches and was presented by the running example. It showed how this approach works and it is presented in Figure 2 and Table 2. In this example two actors are involved, seller and auctioning service, to create an auction.

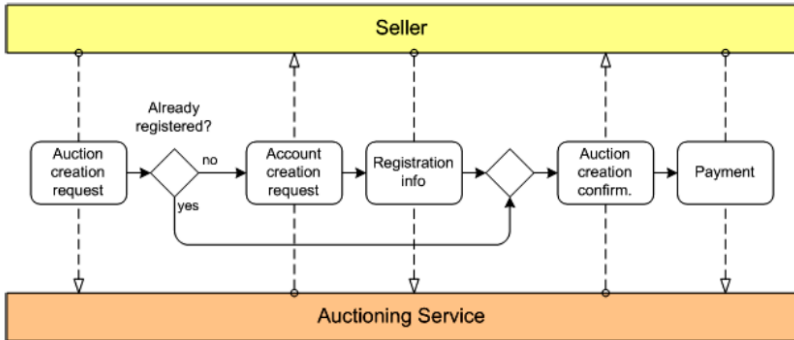


Fig. 2. Auctioning Service Model [11]

To translate BPMN model to natural language representation of SBVR, the BPMN model was given as input to their BR-Generator tool. The tool initially parses the XML representation and extracts the SBVR vocabulary by performing BPMN to SBVR mapping.

Table 2. SBVR vocabulary generated from BPMN [11]

<i>Details</i>
The Seller sends an auction creation request to Auctioning Service .
If Seller is already registered , it is necessary that the auction creation is confirmed .
If Seller is not already registered , it is necessary that the account creation request is sent to Seller .
It is necessary that the Seller sends the Registration info to Auctioning Service and then the auction creation is confirmed .
The Seller sends a Payment to Auctioning Service .

Mappings of the elements are limited, because this approach uses limited number of business process model elements. Elements which are used are: start event, activity, sequence flow, conditional flow, message flow and association, swimlanes, data objects, group, annotation. As example shows in Figure 2, business process modeling is performed without the use of a good modeling practice: names of activity is not in a good format, must be used verbal rather the noun form of activity. Clearly, this is just a small step for BPMN and SBVR transformation.



In order to have integral and fluent text from business process diagram we have to think over linguistic techniques. In general, there are different techniques to translate process models into natural language text: non-linguistic and linguistic. Non-linguistic approach uses templates. Linguistic approach uses intermediate structures to obtain a deeper representation of the text [10]. Because the non-linguistic approach is simple (it is based on canned text) and clearly it is not impossible to have templates for all process model cases it is considerable to use linguistic approach. We also need to distinguish the parts of speech, which meets certain rules concepts. According to one researchers group there are four main challenges of automatic generation of text: text planning, sentence planning (it includes mappings of elements, what we need for integration), surface realization and flexibility [10].

SBVR separation and vocabulary extension with process concepts is needed for bidirectional transformation of BPMN and SBVR. In order to have complete transformation from SBVR to BPMN there is a need to have process concepts, which describe, for example, the depth of process. Otherwise, some information would be lost.

6 Conclusions

In this paper, we presented analysis of BPMN and SBVR combination approaches which will serve as a beginning for the further work. Combination of these two standards is essential in order to avoid misunderstandings and miscommunication issues while reading and interpreting business models among specialists and organizations.

As papers analysis showed, integration and transformation in this case are two inseparable approaches. In order to have two standards combined we have to have most of element mappings of BPMN and SBVR, use linguistic approaches to separate part of speech to create fluent and correct rules and provide all possibilities of two standards in one environment. In case we want full and bidirectional transformation it should be considered to have also rules for process. In further work we are planning to develop an approach for relating BPMN and SBVR. This approach involves combination of two modeling standards in one environment in order to synchronize the two methods of modeling. It will be an attempt for developing an approach which could be improved later with all needed aspects by demand.

Acknowledgements. The work described in this paper has been carried out within the project VP1-3.1-ŠMM-10-V-02-008 “Integration of Business Processes and Business Rules on the Base of Business Semantics”.

References

1. Agrawal, A.: Semantics of Business Process Vocabulary and Process Rules. In: ISEC 2011 Proceedings of the 4th India Software Engineering Conference, pp. 61–68 (2011)
2. Business Process Model and Notation (BPMN), v.2.0, OMG Document Number: formal/2011-01-03. OMG group, <http://www.omg.org/spec/BPMN/2.0>
3. Business Rule Speak Solutions. Rule Speak, LLC

4. Cheng, R., Sadiq, S., Indulska, M.: Framework for business process and rule integration: A case of BPMN and SBVR. In: Abramowicz, W. (ed.) BIS 2011. LNBIP, vol. 87, pp. 13–24. Springer, Heidelberg (2011)
5. Chinosi, M., Trombetta, A.: BPMN: An introduction to the standard. *Computer Standards & Interfaces* 34(1), 124–134 (2012)
6. Graml, T., Brachth, R., Spies, M.: Patterns of Business Rules to Enable Agile Business Processes. In: Proceedings of the 11th IEEE International Enterprise Distributed Object Computing Conference, pp. 365–375. IEEE Computer Society (2007)
7. Hohwiller, J., Schlegel, D., Grieser, G., Hoekstra, Y.: Integration of BPM and BRM. In: Dijkman, R., Hofstetter, J., Koehler, J. (eds.) BPMN 2011. LNBIP, vol. 95, pp. 136–141. Springer, Heidelberg (2011)
8. Indulska, M., Recker, J., Rosemann, M., Green, P.: Business process modeling: Current issues and future challenges. In: van Eck, P., Gordijn, J., Wieringa, R. (eds.) CAiSE 2009. LNCS, vol. 5565, pp. 501–514. Springer, Heidelberg (2009)
9. Koehler, J.: The Process-Rule Continuum – How can the BPMN and SBVR Standards in-teplay? Lucerne University of Applied Sciences and Arts, Swicerland (2010)
10. Leopold, H., Mendling, J., Polyvyanyy, A.: Generating Natural Language Texts from Business Process Models. In: Ralyté, J., Franch, X., Brinkkemper, S., Wrycza, S. (eds.) CAiSE 2012. LNCS, vol. 7328, pp. 64–79. Springer, Heidelberg (2012)
11. Malik, S., Bajwa, I.S.: Back to Origin: Transformation of Business Process Models to Business Rules. In: La Rosa, M., Soffer, P. (eds.) BPM Workshops 2012. LNBIP, vol. 132, pp. 611–622. Springer, Heidelberg (2013)
12. Recker, J., Indulska, M., Rosemann, M., Green, P.: How good is BPMN really? Insights from Theory and Practice (2006)
13. Semantics of Business Vocabulary and Business Rules (SBVR) specification.v1.0. Object Management Group (OMG) (January 2, 2008)
14. Sinur, J.: The art and science of rules vs. Process flows. Research Report G00166408, Gartner (2009)
15. Skersys, T., Tutkutė, L., Butleris, R., Butkienė, R.: Extending BPMN Business Process Model with SBVR Business Vocabulary and Rules. *Information Technology and Control* 41(4) (2012) ISSN 1392-124X
16. Zur Muehlen, M., Indulska, M.: Modeling Languages for Business Processes and Business Rules: A representational Analysis. *Information Systems Journal* 35(4), 379–390

Process for Applying Derived Property Based Traceability Framework in Software and Systems Development Life Cycle

Saulius Pavalkis and Lina Nemuraite

Department of Information Systems, Kaunas University of Technology,
Studentu 50-313a, Kaunas, Lithuania
saulius.pavalkis@nomagic.com, lina.nemuraite@ktu.lt

Abstract. For implementing the idea of applying derived properties for tracing project artifacts, the Derived Property Based Traceability Framework was created that consists of Model-Driven Domain Specific Language (DSL) engine for extending UML with derived property specifications, traceability schemas, and traceability analysis means. Traceability schemas may be generic, suitable for every purpose, but they often are characteristic to a development method, modeling language or a particular project. The paper presents a process for applying the Derived Property Based Traceability Framework consisting of three parts: process for adapting Derived Property Based Traceability solution for development method or Domain Specific Language; process for applying the solution in a development process, and process for automating the maintenance of traceability relations. Process is illustrated with examples from several case studies.

Keywords: traceability, derived properties, model-driven development, traceability framework.

1 Introduction

Traceability of software and systems models is an important aspect of Model Driven Development. Current state of traceability implementations in CASE tools often lacks flexibility, customizability and other qualities, analyzed by many authors and our previous works [1]. Usually, traceability solutions cause significant overhead and require routine efforts what often discourages from using traceability means at all.

We have proposed the traceability solution [1], based on derived properties, which is directed for solving frequent traceability problems. In particular, traceability solutions lack for automation; they pollute models with traceability information that can be redundant, burdening specification and analysis; additional relationships introduce dependencies and tight coupling among project stages that are incompatible with principles of good architectural design; traceability schemas are hardly customizable and maintainable.

The Derived Property Based Traceability Approach helps to avoid these problems as traceability relations are automatically calculated by a CASE tool when they are needed for analysis or validation of models. Derived attributes and relations of model elements are accessible for developers and analysts in specifications, dialog windows, visualization and analysis means in the same way as primary ones; so they do not require additional skills or specific attention.

The proposed traceability solution involves a traceability metamodel, profile, and the overall framework for implementing the solution [1], which is independent from a particular CASE tool. However, developers may wish to create specific traceability schemas for their chosen development methodologies and/or modeling languages as traceability schemas depend on types of modeling concepts and relationships, which are intended to trace.

Therefore, the goal of the paper is to present a complete process for ensuring traceability including adaption of the framework for different cases and automation of maintaining traceability relations. The overall process for using Derived Property Based Traceability approach consists of three parts: a process for adapting the solution for a particular methodology or language; process for applying the adapted solution in development projects; and a process for automating maintenance of traceability relations. We do not present here the traceability metamodel, profile, framework etc., as such information is available in [1] and [2]; instead, we illustrate the process with traceability schemas, validation rules etc., when needed.

The rest of the paper is structured as follows. Sections 2 – 4 present the process for adapting, applying and automating derived property based traceability means in CASE tools. Section 5 provides overview of experimental approval. Section 6 analyses related work and gives a comparison of the approach with existing capabilities of similar tools. Section 7 presents conclusions and future works.

2 Process for Adapting Derived Property Based Traceability Solution

Process for adapting Derived Property Based Traceability solution is shown in Fig. 1. During creation of a traceability schema for a chosen modeling language, development process or a problem, one has to identify traceable artifacts and create derived properties for traceable links among these artifacts.

Choose Traceability Target. Any modeling language or development methodology can be selected as a traceability target. E.g. it could be the SysML [3] for specifying requirements, BPMN [4] for business analysis, and UML for software design. Standard or custom development processes (e.g. UP or SYSMOD [5]) can be used.

Identify Traceable Artifacts. In this step artifacts, whose evolution through project stages should be traced, are identified. Unless this is a mission critical system or different requirements are specified by standard regulations, usually only main artifacts, which influence stage or project completeness, are in focus. Too many artifacts will introduce overhead for managing traceability. Traceability rules are created for each

relation between main artifacts, which relations are decided to be tracked. In order to achieve two-way traceability, traceability rules are created for deriving properties of both ends of traceability relations. Examples of such artifacts are BPMN Process, UML Use Case, SysML Requirement.

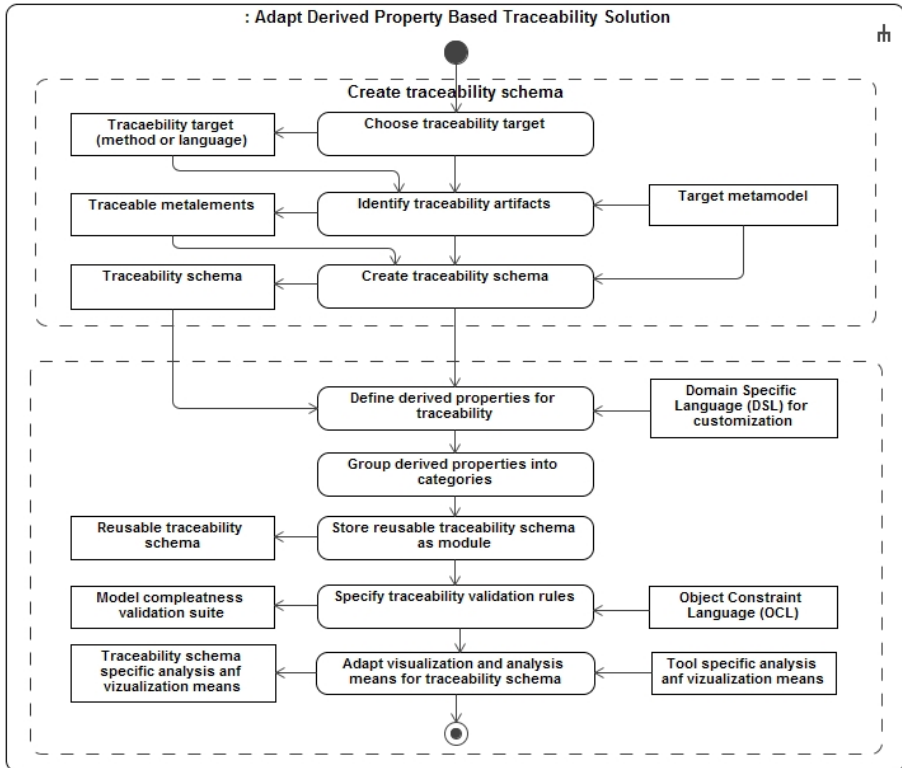


Fig. 1. Process for adapting Derived Property Based Traceability solution

Create Traceability Schema. In order to create traceability schema, metaclasses of artifacts identified in the previous step are associated with tracing relations. Properties reflecting these associations will be owned by associations itself and will make no influence on standard modeling language or process metamodels. Example of traceability schema is presented in Fig. 2.

Define Derived Properties for Traceability. Simple expressions can be used to specify derived properties based on direct relationships, e.g. “Use Case → Satisfy → Requirement”. The advanced Metachain expression should be used for transitive relations, e.g. “Business process → Abstraction → Use case → Abstraction → Requirement → Satisfy → Component”. OCL expressions and scripting languages should be used in more complex cases, e.g. for specifying recursive relations. An example of OCL expression for derived relation between component and use case (Fig. 2) is presented in Fig. 3.

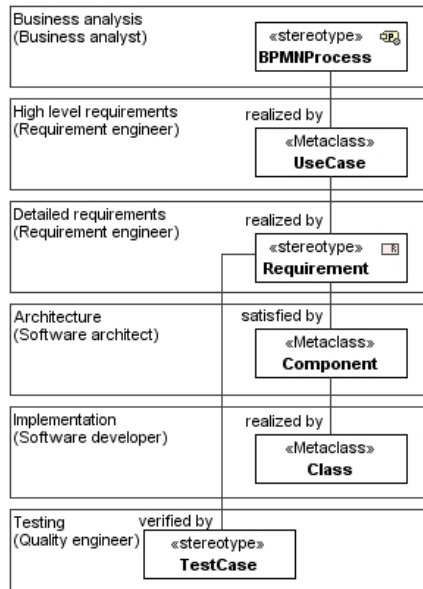


Fig. 2. Traceability schema for software development process

```

context BPMNProcess::RealizedInArchitecture:Component
derive: self.supplierDependency->select(a|a.ocIsKindOf( Abstraction ) ).
client->exists(b|b.ocIsKindOf( UseCase ) ).supplierDependency->
select(c|c.ocIsKindOf( Abstraction ) ).client->exists(d|d.ocIsKindOf
( SysML::Requirement ) ).supplierDependency->select(e|e.ocIsKindOf
( SysML::Satisfy ) ).client->exists(f|f.ocIsKindOf( Component ) )
    
```

Fig. 3. OCL expression for derived property

Group Derived Properties into Categories, e.g. specification and realization groups (if traceability relation is established between artifacts of business process and its implementation, we treat traceability rules as realization ones; if we are going from implementation to business process, we consider them as specification rules).

Store Reusable Traceability Schema as a Module. Traceability schemas (sets of traceability relations) are dependent on traceability context – e.g. modeling language such as BPMN or software engineering process. It is desirable to keep traceability schemas in UML profiles and implement them as separate modules that could be loaded and reused in various projects. Derived properties defined in the loaded module are added to elements of considered models.

Specify Traceability Validation Rules. On the base of traceability schemas we can create validation rules and automate model analysis for checking model completeness (finding model elements not covered with their realizing artifacts, or identifying redundant artifacts); ensuring absence of cyclic traceability relations (i.e. such relations when e.g. one element is involved in both realizing and specifying traceability relations with another element).



OCL allows not only to specify traceability rules, but also to execute them. Having predefined traceability validation rules and using validation engine it is possible to check project for model completeness and cyclic traceability relations. Completeness validation rules are created for checking completeness of traceability (coverage of artifacts), e.g. each Use Case should be traced by at least one Requirement (Fig. 4).

```
Context: UseCase::realizedBy:Requirement
(not self.ownedElement→exists (e|e.ocIsKindOf (UseCase)))
implies self.realizedBy→size()>0
```

Fig. 4. OCL expression for artifact completeness validation rule

Adapt Visualization and Analysis Means for a Particular Traceability Schema.

There are multiple types of UML relationships, properties and custom tags that can be used for traceability visualization. In order to help to quickly visualize traceability, custom (derived) properties are treated in the same way as regular element properties and can be represented on diagrams, validated with validation engine, and inserted into generated documents. Traceability property groups are visible in Element Specifications, Quick Properties, Go To, Reports, etc. Traceability information is available in Relation Maps for multi-level graph type traceability analysis; Dependency Matrix may be used for visualizing single level traceability and analyzing gaps. In order to be able to efficiently create and use traceability visualization means they can be predefined and distributed together with traceability schema.

3 Process for Applying Derived Property Based Traceability Solution

Process for applying Derived Property Based Traceability solution is shown in Fig. 5.

Apply Traceability Schema for Project. If traceability schema is held in a separate module (i.e. reusable project part) it can be loaded in a project and used starting from the beginning of the project or at any moment of already going project. If reusable traceability schema comes together with validation suites and visualization means, tree main immediate changes are observed on traceability module used in the project: 1) traceability properties appear in element specifications, context menu, and other places, and are immediately evaluated; 2) validation suites (if automatic) check model for completeness; incomplete and redundant artifacts are shown; 3) traceability visualization and analysis means (Dependency Matrix, Relation Map Dedicated reports, Generic tables are available and ready to be used).

Perform Coverage Analysis. The Coverage analysis gives coverage information at immediate higher (e.g., Specification) or lower (e.g., Realization) levels having the objective is to visualize and verify that artifacts of different stages, e.g., analysis, design, and implementation, are covered. It allows finding areas of not covered parts and to evaluate coverage metrics, to improve an understanding of the system and acceptance of the system accordingly.

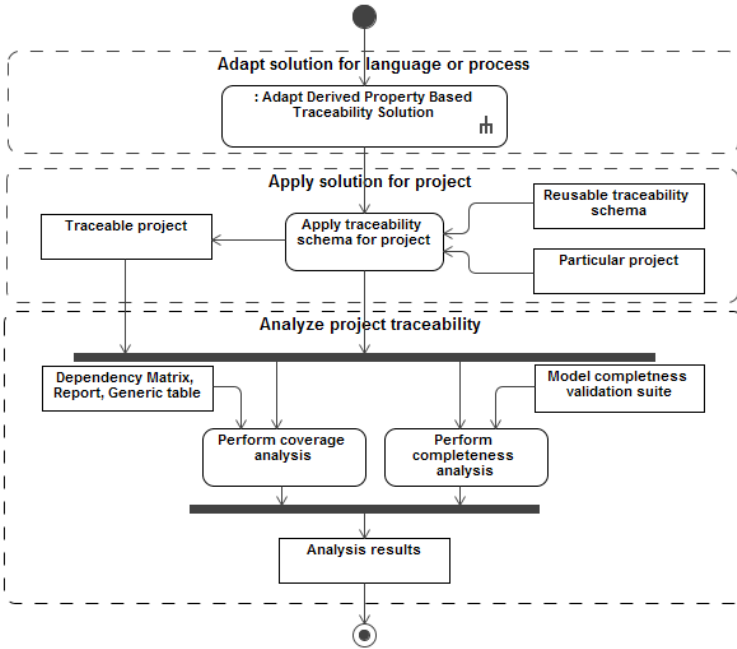


Fig. 5. Process for applying Derived Property Based Traceability solution

Calculate Traceability and Coverage Metrics. Examples of coverage metrics, which could be calculated for requirements of the overall system or level n :

1. The percent F_r of requirements in level n derived from requirements at level $n+1$:

$$F_r = \frac{R'_{n,n+1}}{R_n} 100\%$$

Here $R'_{n,n+1}$ is a number of requirements derived from requirements in level $n+1$; R_n – a number of all requirements in level n .

2. The percent O_r of requirements in level n excluding orphans derived from requirements at level $n+1$:

$$O_r = \frac{R'_{n,n+1}}{(R - O)_n} 100\%$$

Here $R_{n,n+1}$ is a number of requirements derived from requirements in level $n+1$; $(R-O)_n$ – a number of all requirements in level n excluding orphans.

3. The percent V_r of requirements in level n that are not verified (have no incoming *verify* relations):

$$V_r = \frac{V_n}{R_n} 100\%$$

Here V_n is a number of requirements in level n that are not verified.

4. The percent S_r of requirements in level n that are not satisfied by functions (used only at Sub-system level):

$$S_r = \frac{S_n}{R_n} 100\%$$

Here S_n is a number of requirements in level n that are not satisfied by functions (i.e. having no incoming *Satisfy* relationships from *PrincipleSet* or *Activity*).

5. The percent S_r of requirements in level n that are not satisfied by structural elements (i.e. having no incoming *Satisfy* relationships from *System*, *Subsystem*, *Product*, etc.):

$$SE_r = \frac{SE_n}{R_n} 100\%$$

Here SE_n is a number of requirements in level n that are not satisfied by structural elements.

6. The percent ST_r of requirements in level n that are not covered with Safety and Tests (i.e., have no outgoing trace relationships to requirements in level $n+1$):

$$ST_r = \frac{ST_n}{R_n} 100\%$$

Here ST_n is a number of requirements in level n that are not are not covered with Safety and Tests.

Perform Completeness Analysis. It is possible to evaluate model against validation rules, which are checked automatically in all model or in a certain scope on demand. Results of validation rules evaluation show model elements, properties and diagrams, which does not satisfy validation constraints. One can see areas not yet covered with artifacts – incomplete ones, and redundant artifacts.

4 Automating Traceability Solution

Using derived property approach, traceability relations are automatically evaluated by derived property engine via calculating derived property values. However, without automation means for creating and updating derived properties, the approach would have a significant overhead, which would greatly discourage its usage in projects. Process for adapting and applying the Framework for Creating Custom Wizards (FCCW) [6] for automation of creating traceability relations and updating traceability information is shown in Fig. 6.

Choose Development Process for Automation. It is the first step in automating traceability. In the paper [6] two examples are presented about applying the proposed method for RUP-based workflow for use case modeling and capturing robustness analysis classes.

Create Process Workflow. Workflow, which will be automated, should be specified using Software Process Engineering Metamodel v2.0. In particular, Process diagram needs to be used.

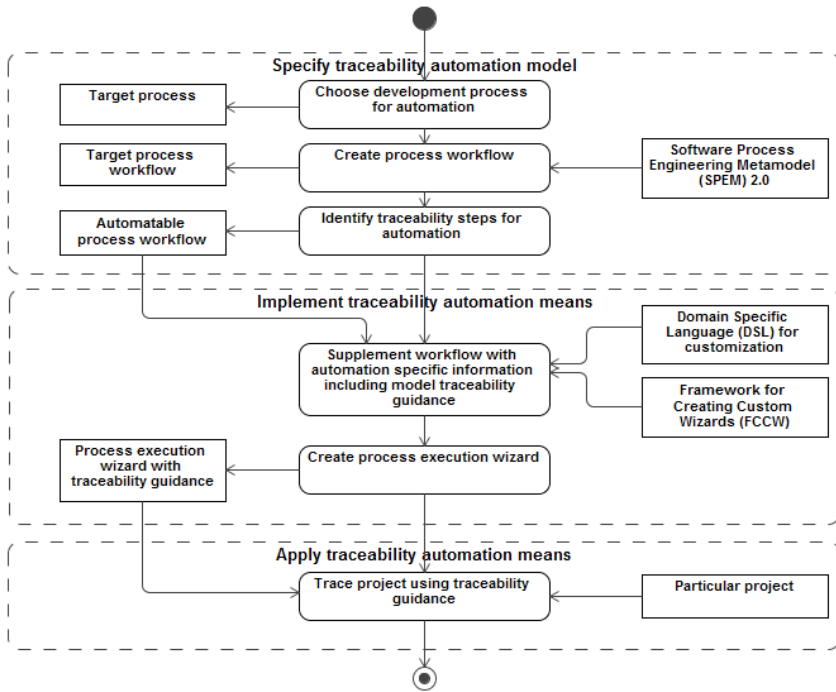


Fig. 6. Process for applying FCCW for automating creation and updating of traceability relations

Identify Traceability Steps for Automation. Further, process steps, which will be automated, are identified. FCCW allows having 4 types of automation: creating an element which symbolizes the target model and defines its name; capturing elements and defining their properties, and listing elements existing in a model; joining elements with editable matrix like table to represent element relations; informing, navigating and invoking other features.

Supplement Workflow with Automation Specific Information. The specified process steps, which will be automated, are stereotyped with FCCW specific stereotypes identifying required automation type. Execution dedicated properties of each step are specified.

Create Process Execution Wizard. On the base of the specified workflow, FCCW stereotypes and their property information, executable wizard specification is created. To be generated wizard will guide through the workflow of traceability creation, analysis and update according to the chosen methodology, providing step-by-step dialog for tracing and creating elements. The wizard output is a model, from which further artifacts can be created: views, documentation, coverage analysis reports, etc.

Trace Project Using Traceability Guidance. The specified wizard can be included into a reusable traceability module together with traceability schema, validation rules and predefined view information. The wizard provides automation for gathering data according rules of modeling language and visualizing, creating and maintaining traceability.

5 Experimental Approval

Three experiments were conducted for evaluating the suitability of the Derived Property Based Traceability approach for implementing traceability solutions for software and system development processes, and modeling language BPMN. The experiments have shown that the approach is capable ensuring consistency of project artifacts, to analyze change impact, and to avoid typical traceability problems for software development [1] and for systems development [7] processes. It is capable to solve traceability problems of BPMN 2 models [8]: lack of traceability between BPMN processes and resource roles, BPMN processes and business concepts, participants and messages, thus allowing validating BPMN 2 models for correctness and completeness, and performing change impact analysis;

Discussion of Threats to Validity. The major threat of validity of the approach is an overhead raised by applying any traceability approach. This thread is eliminated in mission critical projects (e.g. health care, military, nuclear engineering, aerospace) in which traceability is of the great importance. The threat could be minimized in regular projects if only major artifacts are traced, and traceability automation means are used such as editable matrix for traceability relations, traceability validation suites, etc.

Another threat is the reliability of traceability validation results. Even approach is straightforward its results depend on how well it is followed. Also, if we validate coverage of major artifacts, we would not validate a quality of covering artifacts. To do so, validation constraints need to be extended to validate the content of covering artifacts.

6 Related Works

Early empirical studies showing importance of traceability for validating completeness of software or system projects have been published by Gotel and Finkelstein [9], Watkins and Neal [10], Ramesh and Edwards [11]. Aizenbud-Reshef et al. [12] emphasized the importance of automating traceability. We noticed three research directions for automatic creation and maintenance of traceability links: 1) Text mining and information retrieval techniques for recovering traceability links between software artifacts (e.g., [13]–[14]); 2) Establishing traceability links by monitoring users' modifications and analyzing change history; 3) Deriving traceability links from existing ones. The latter principle as less time consuming was used in our Derived Property based approach. We have supplemented it with two additional possibilities for reducing a manual input in creation and maintenance of traceability relations and obtaining a higher usability:

- Creating traceability information during model transformations. Automatic creation of traceability relations during transformation is analyzed in [13], [14]–[17]. As transformations are especially popular in Model Driven Engineering [18]–[25], we treat relations created during transformations as traceability ones.
- Analysis of existing relationships to obtain implied relations [26]. In our approach, a part of traceability information is based on transitive relations.

Comparison of existing traceability solutions in CASE tools with implementation of Derived Property Based Traceability approach in MagicDraw is presented in the Table 1:

Table 1. Comparison of existing traceability solutions in CASE tools

Criteria / CASE tool	Rational Software Architect	Visual Paradigm	Enterprise Architect	Modelio	Reqtify	MagicDraw
1. Traceability schema and rules are easy customizable and model driven	-	-	-	-	+/-	+
2. Capabilities of modeling tool are reusable for traceability analysis and visualization	+	+	+	+	-	+
3. Model is not polluted by traceability information	+	+/-	-	-	+	+
4. Model is loosely coupled	+	+/-	-	-	+	+
5. Creation and maintenance of traceability relations is automatic and flexible	+/-	+/-	-	+/-	+	+/-
6. Suggests traceability schemas	+	+/-	-	-	+	+
7. Coverage/completeness/change management analysis.	+/+/-	-/-/-	-/-/-	+/+/+	+/+/+	+/+/+

Analysis of existing traceability based solutions in CASE tools has shown that the presented solution provides advantages against other currently existing ones. The only equal solution with a similar number of steps to adapt to a custom development method is supported by the non-modeling tool – Geensoft Reqtify. Unfortunately, it requires programmatic integration with a modeling tool and adoption to a custom development method, what is not easy to achieve.

7 Conclusions and Future Works

The use of the proposed process by the real life examples for systems and software modeling projects and BPMN language has shown that the presented process provides the complete, development method independent methodology for adapting, using and automating the proposed traceability solution based on derived properties making it available for every model driven CASE tool.

Implementation of the approach in UML CASE tool MagicDraw has approved the expected quality criteria and was favorably met by MagicDraw users. It may be accomplished much faster and easier in comparison with traceability solutions of other CASE tools, which analysis revealed the advantages of the proposed process. The only equal solution with a similar number of steps to adapt to custom development method is supported by non-modeling tool – Geensoft Reqtify but it requires programmatic integration with a modeling tool and adaption to a custom development method, what is not easy to achieve.



Derived Property Based Traceability Approach already has been successfully adapted by companies including large aerospace and telecommunication corporations and academic institutions.

In our future work, we are planning to deepen our approach on the base of acquired practical experience: to automate transition from traceability metamodel to derived properties as this step could be fully automated; to help creating required traceability solutions by validating non-traced elements and automatically suggesting required relations to be created by using validation engine; to develop more powerful, comprehensive traceability schemas for modeling databases, business processes and enterprise architectures, which would be reusable across a large variety of software projects.

Acknowledgements. The work is supported by the project VP1-3.1-ŠMM-10-V-02-008 “Integration of Business Processes and Business Rules on the Basis of Business Semantics” (2013-2015), which is funded by the European Social Fund (ESF).

The authors would like to thank No Magic, Inc, especially the MagicDraw UML product team for the comprehensive support.

References

1. Pavalkis, S., Nemuraite, L., Butkiene, R.: Derived Properties: A User Friendly Approach to Model Traceability. *Information Technology and Control* 42(1), 48–60 (2013)
2. No Magic, Inc. UML Profiling and DSL (2011), <https://secure.nomagic.com/files/manuals/UML%20Profiling%20and%20DSL%20UserGuide.pdf>
3. OMG. OMG Systems Modeling Language (OMG SysML), Version 1.2. OMG, OMG Document Number: formal/2010-06-01 (2010)
4. OMG. Business Process Model and Notation (BPMN), Version 2.0. OMG, OMG Document Number: formal/2011-01-03 (2010)
5. SYSMOD, The Systems Modeling Process (2011), <http://sysmod.system-modeling.com/>
6. Silingas, D., Pavalkis, S., Morkevicius, A.: MD Wizard - a model-driven framework for wizard-based modeling guidance in UML tools. In: *Proceedings of the International Multi-conference on Computer Science and Information Technology*, pp. 609–615. IEEE Computer Society Press, Los Alamitos (2009)
7. Pavalkis, S., Nemuraite, L.: Lightweight Model Driven Process to Ensure Model Traceability and a Case for SYSMOD. In: *2013 2nd International Conference on Advances in Computer Science and Engineering (CSE 2013)*, pp. 2019–2223. Atlantis Press (2013)
8. Pavalkis, S., Nemuraite, L., Milevičienė, E.: Towards Traceability Metamodel for Business Process Modeling Notation. In: Skersys, T., Butleris, R., Nemuraite, L., Suomi, R. (eds.) *Building the e-World Ecosystem*. IFIP AICT, vol. 353, pp. 177–188. Springer, Heidelberg (2011)
9. Gotel, O.C.Z., Finkelstein, A.C.W.: An analysis of the requirements traceability problem. In: *Proceedings of the 1st IEEE International Requirements Engineering Conference (RE 1994)*, pp. 94–101. IEEE Computer Society, New York (1994)
10. Watkins, R., Neal, M.: Why and how of requirements tracing. *IEEE Softw.* 11(4), 104–106 (1994)

11. Ramesh, B., Edwards, M.: Issues in the development of a requirements traceability model. In: Proceedings of the IEEE International Symposium on Requirements Engineering, pp. 256–259. IEEE Computer Society, New York (1993)
12. Aizenbud-Reshef, N., Nolan, B.T., Rubin, J., Shaham-Gafni, Y.: Model traceability. *IBM Systems Journal* 45(3), 515–526 (2006)
13. Antoniol, G., Canfora, G., Casazza, G., De Lucia, A., Merlo, E.: Recovering traceability links between code and documentation. *IEEE Transactions on Software Engineering* 28(10), 970–983 (2002)
14. Hayes, J.H., Dekhtyar, A., Osborne, J.: Improving requirements tracing via information retrieval. In: Proceedings of the 11th IEEE International Requirements Engineering Conference, pp. 138–147 (2003)
15. Mens, T., Van Gorp, P.: A taxonomy of model transformation. In: Proceedings of the International Workshop on Graph and Model Transformation (GraMoT 2005), March 27. *Electronic Notes in Theoretical Computer Science*, vol. 152, pp. 125–142 (2005)
16. Porres, I.: Rule-based update transformations and their application to model refactorings. *Software and Systems Modeling* 4(2), 368–385 (2005)
17. Van Gorp, P., Janssens, D., Gardner, T.: Write once, deploy N: A performance oriented MDA case study. In: Proceedings of the IEEE International Conference on Enterprise Distributed Object Computing, pp. 123–134 (2004)
18. Schmidt, C.: Model-Driven Engineering. *IEEE Computer* 39(2), 25–31 (2006)
19. Briand, L.C., Labiche, Y., Yue, T.: Automated traceability analysis for UML model refinements. *Information and Software Technology* 51(2), 512–527 (2009)
20. Anastasakis, K., Bordbar, B., Georg, G., Ray, I.: On Challenges of Model Transformation from UML to Alloy. *Journal on Software & System Modeling* 9, 69–86 (2010)
21. Meijler, T.D., Nyttun, J.P., Prinz, A., Wortmann, H.: Supporting fine-grained generative model-driven evolution. *Software and Systems Modeling* 9(3), 403–424 (2010)
22. Bryant, B.R., Gray, J., Mernik, M., Clarke, P.J., France, R.B., Karsai, G.: Challenges and Directions in Formalizing the Semantics of Modeling Languages. *Computer Science and Information Systems* 8(2), 225–253 (2011)
23. Recker, J.: Modeling with tools is easier, believe me – The effects of tool functionality on modeling grammar usage beliefs. *Information Systems* 37, 213–226 (2012)
24. Ablonskis, L., Nemuraitė, L.: Discovery of complex model implementation patterns in source code. *Information Technology and Control* 39(4), 291–300 (2010)
25. Lopata, A., Ambraziunas, M., Gudas, S., Butleris, R.: The main principles of knowledge-based information systems engineering. *Elektronika ir Elektrotechnika* 4(120), 99–102 (2012)
26. Sherba, S.A., Anderson, K.M., Faisal, M.: A Framework for Mapping Traceability Relationships. In: Proceedings of the 2nd International Workshop on Traceability in Emerging Forms of Software Engineering, Montreal, Canada (September 2003)

Developing SBVR Vocabularies and Business Rules from OWL2 Ontologies

Gintare Bernotaityte, Lina Nemuraite, Rita Butkiene, and Bronius Paradauskas

Kaunas University of Technology, Department of Information Systems,
Studentu 50, Kaunas, Lithuania
{gintare.bernotaityte, lina.nemuraite, rita.butkiene,
bronius.paradauskas}@ktu.lt

Abstract. Semantics of Business Vocabulary and Business Rules (SBVR) is OMG adopted metamodel allowing defining noun concepts, verb concepts and business rules of a problem domain in structured natural language based on formal logics. SBVR business vocabulary and business rules are capable of representing ontologies. There are some research works devoted to transforming SBVR into Web Ontology Language OWL2. The reverse way of representing ontology concepts with SBVR structured language was not investigated though there are much more ontologies than SBVR vocabularies. Our research is concentrated on methodology for creating SBVR vocabularies and rules from OWL2 ontologies without a loss of the expressive power, characteristic for ontologies, as some ontology-specific concepts have no direct representation in SBVR. The particular attention is devoted to applying SBVR vocabulary in semantic search.

Keywords: SBVR, OWL 2, business vocabulary, business rules, domain ontology, lexical ontology.

1 Introduction

The goal of the paper is to present principles for developing SBVR [1] business vocabulary and business rules from OWL 2 [2] ontologies. There are two purposes for doing so. The first one is to provide a representation of ontology in a structured language understandable by business participants as ontologies are often used for representing domain knowledge in business applications [3]. In this case, it is desirable to transform all (or as much as possible) ontology concepts into SBVR vocabulary and rules. The second purpose is to allow formulating SBVR questions for performing semantic search in which SBVR questions are transformed into SPARQL queries executable in the source ontology [4], [5].

Semantic search is one of the aims directed towards better facilitation of people and organizations to use natural Lithuanian language in the virtual space in their professional and personal activities. Natural language technologies require sophisticated processing algorithms and vast amounts of resources (dictionaries, corpuses, thesauruses, ontologies) whose development requires involving a lot of professionals with

different skills. Within this project, ontology representation in SBVR would allow developing and investigating methods of semantic analysis and search using structured natural language, and relate them with linguistic analysis and annotation of unstructured texts. For using SBVR business vocabulary and business rules in semantic search, it is not necessary to transform the overall ontology. For formulating SBVR questions, only part of ontology concepts is important. Therefore, we first concentrate on selection of semantic search related concepts though in the wider scope the SBVR based representation of all OWL 2 concepts makes sense.

Current research has to deal with several challenges as similar problems previously have not been solved in Lithuania. One challenge is about relating lexical and domain ontologies. In SBVR, meaning and representation are separate concepts: every meaning may have several representations, and each expression may have several meanings. The representation part of SBVR is similar to what is encompassed by WordNet [6], [7], VerbNet [8] and FrameNet [9], [10] lexical ontologies where various syntactic forms are related to meaning. We think about considering a lexical ontology based on SBVR representations, which is related to SBVR based domain ontology and may considerably improve semantic search and Natural Language Processing techniques. Other important problems are about relating semantic representations with linguistic information, applying SBVR for Lithuanian language, and integration with existing Semantic Web ontologies presented in English language. The aim of this paper is not to give a solution to all these problems but rather to present initial principles for doing this.

The rest of the paper is structured as follows. Section 2 analyses related work and draws a layout for our research. Section 3 presents a domain ontology example. Section 4 describes principles for transforming OWL 2 into SBVR. Section 5 summarizes conclusions and envisages future research.

2 Related Work and Outline of Research

Currently, there is a lot of research aiming at transforming SBVR business vocabulary and business rules into ontologies. Some prototypes [4], [11] and commercial implementations [12], [13] are developed for describing the problem domain in the SBVR Structured English (SSE) language understandable for human and computer programs, and for transforming this description into Web Ontology Language OWL 2 for using it in Semantic Web applications.

The result of SBVRToOWL2 transformation [14] is domain ontology based on preferred representations of SBVR concepts. This transformation does not involve synonyms and synonymous forms except synonymous forms of verb concept wordings corresponding to OWL 2 inverse object properties. SBVR metamodel does not give possibility for specifying preferred inverse relations. Therefore, specific SBVR business rules are specified for desirable OWL 2 inverse object properties.

Kendall and Linehan [11] define reversible transformation without loss of semantic information but the result of reverse transformation does not guarantee an identical representation. This is caused by their solution to transform SBVR synonyms and

synonymous forms into OWL 2 annotations; the reverse transformation from OWL 2 into SBVR is not capable to recover the original representations though they remain semantically equivalent. We hope to solve this problem by separating SBVR synonyms and synonymous forms from domain ontology and creating lexical ontology based on SBVR representations.

Our research has somewhat different requirements than analysed approaches. For creating SBVR vocabularies suitable for semantic search it is desirable to build them upon existing ontologies. However, as far as we can tell, no-one has studied reverse transformation allowing representation of existing ontologies in SBVR structured language, ensuring no loss of semantic information. It is worth to mention that none of the analysed works for transforming SBVR into OWL has considered transformation of SBVR representations. We argue that such consideration has sense in semantic search because there are several lexical ontologies as FrameNet, VerbNet, and WordNet with capabilities similar to SBVR representations.

The Berkeley FrameNet [9], [10] is an on-line lexical resource for English language, based on frame semantics and supported by corpus evidence. The FrameNet consists of lexical units, annotated by hierarchically-related semantic frames, exemplified in annotated sentences. Each frame is related with a meaning.

VerbNet [8] is the largest on-line hierarchical, domain independent verb lexicon for English with mappings to other lexical resources such as WordNet and FrameNet. VerbNet is organized into verb classes, described by thematic roles, argument restrictions, and frames consisting of syntactic descriptions and semantic predicates with a temporal function and semantics of event decomposition.

The other large lexical database of English is WordNet [7]. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept, including antonym, hyponym, hypernym, and other relations with other concepts of WordNet.

Our research aims to create SBVR based Lexical Ontology for Lithuanian language (SLOL) that would have capabilities similar to WordNet, Verbnet and FrameNet. The difference from analysed lexical ontologies is, in semantic sense, that meanings in SBVR business vocabulary and business rules are organized around a conceptual model of a specific problem domain; in lexical sense – that SBVR terms, names and verb concept wordings include not only single words but may be compound, composed of several words. SLOL would not take adjectives and adverbs (as WordNet does) into account, but it would concentrate on representation of SBVR noun concepts and verb concepts. Such representations may be single words and compound ones, comprised of noun and verb phrases. Moreover, SLOL would use only SBVR synonyms and synonymous forms; other WordNet relations (hypernyms, hyponyms, etc.) would be covered in SBVR based domain ontology and would be based on SBVR verb concept relations (associations, property associations, partitive verb concepts, classifications, characterizations, and specializations). Also, our work does not take into account antonyms as they are non-ontological relations.

Lexical ontologies are related with morphological and syntactic information, which is outside of SBVR representations but is indispensable in Semantic Web applications. Linguistic analysis is beyond the scope of our research but SBVR based lexical

ontology should be related with linguistic information as well. We use morphological analysis means (in the form of Web Services) in our SBVR business vocabulary and business rules. In SBVR vocabulary, each case of noun in singular or plural form is treated as different term, i.e. “students” and “student’s” would be synonyms of their primary (preferred) representation. This is not an adequate approach, especially for the Lithuanian language, in which nouns have seven singular and seven plural cases, verbs have various tenses, etc. The stemming Web service used in our SBVR editor [15] allows avoiding this problem by entering nouns, verbs and their phrases once and recognizing them in various cases in verb concept wordings and business rule expressions.

Existing lexical, foundational or standard ontologies as DOLCE, OntoWordNet, SKOS, FOAF are developed in English language. In the multilingual environment, it is desirable to relate SBVR based lexical ontology for Lithuanian language and existing ontologies for English. For multilingual processing and aligning the same meaning in different representations (in one language or in multiple ones), it is desirable to define domain ontologies in a single (e.g. English) language and to use annotation property “label” for defining various representations of that meaning. However, SBVR representations have their own structure, so a label, even if it is multivalued and has a tag for indicating a language, is not sufficient for expressing representations (such an issue applies in a general case for ontologies [16], not only for SBVR based ones).

For ontology based semantic search in a particular domain, three kinds of ontologies have to be provided: linguistic, lexical and domain ontologies. Existing methodologies for developing ontologies, e.g., [17] emphasize modular development allowing construction of new ontologies by reusing existing ones. According to these principles, domain ontology should be constructed from a set of smaller ontologies conceptualizing local, modular domains. Each domain concepts would have their lexical and linguistic information. In [18], authors propose modularisation of domain, terminological and linguistic knowledge. In our research, such modularization aspects also should be taken into account.

3 Running Example

As an example, the domain of political events was chosen. Essential concepts of political event domain are persons, organizations, locations, time, events (meetings, conversations, communications, etc.), and their relations. One of sources for creating ontology for political events was the hierarchy of Extended Named Entities [19], which describes aforementioned entities and their classifications. The Extended Named Entity hierarchy does not claim to be complete but it covers main concepts incident to many domains and is based on practise of information extraction so is useful as such. Fig. 1 presents the ontology class hierarchy based on the Extended Named Entities supplemented with the hierarchy of some political events and object occurrences in the Web documents relevant for semantic search.

For defining the ontology of persons, there are many sources available. OntoLife ontology [20] is devoted to structuring, semantic annotation, analysis and searching unstructured information about persons. The backbone of OntoLife is the entity Person for

modelling the Person’s demographics, biological and legal descriptors, contact methods and online accounts as well as other CV-related information. FOAF ontology [21] describes Person’s social relationships on the Web: Person’s existence in the virtual world, contacts, friends, participation in virtual organizations, etc. Genealogy Tree ontology [22] defines family related properties, relations and events: gender, parent-child and spouse relations, marriage, birth and death events, together with Semantic Web (SWRL) rules for deriving all ancestors and descendants, and other kinship relations. Location ontology could be based on the GeoNames ontology [23], which allows recognizing, annotating and searching various geographical entities (countries, cities, rivers, etc.). Organization and time concepts are included in various foundational ontologies. In the current example, we take restricted person, organization, location and time concepts from Extended Named Entities having in mind that these models should be analyzed more thoroughly in the future.

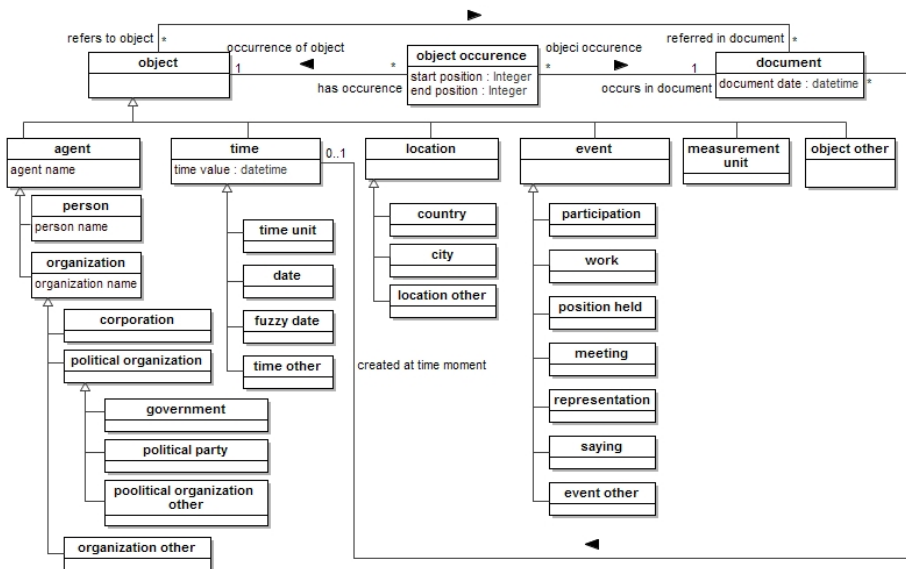


Fig. 1. Ontology class hierarchy based on the Extended Named Entities [19] supplemented with object occurrences and hierarchy of some political events

Event is the most complex and problematic concept. Events in natural language are usually expressed by verbs; majority of the events are represented as n-ary relations. SBVR allows describing n-ary relations; however, the Web Ontology Language is limited to binary ones. For defining n-ary relations in OWL 2, we are forced to objectify them and represent as classes having object or data properties corresponding to roles of the original n-ary relation. This approach has its drawbacks and advantages. The advantage is that roles of n-ary relation are unordered; therefore, the order of roles in an SBVR question often becomes unimportant, and SBVR questions may be more freely formulated. This advantage is relevant for the Lithuanian language, whose grammar does not prescribe strict order of sentence words. The drawback is

that we are forced to change natural language expressions into unnatural constructions, which are difficult to reverse into their original forms. Also, identification of a subject (agent) of an event sometimes becomes complicated.

There are many ontologies and research works devoted to event models, e.g. [24], [25], which describe temporal, spatial, instance, participation, causality, mereology, correlation, documentation, interpretation and other event relations. Currently, we relate event with agent, object (corresponding to SBVR or OWL 2 Thing), time, location (Fig. 2) and occurrence concepts (Fig. 1) thus allowing representing the event as a relation, whose number of roles may vary depending on a type of the event and on completeness of information we have. Occurrences of an event (as well as of other objects) may be identified in many Web documents.

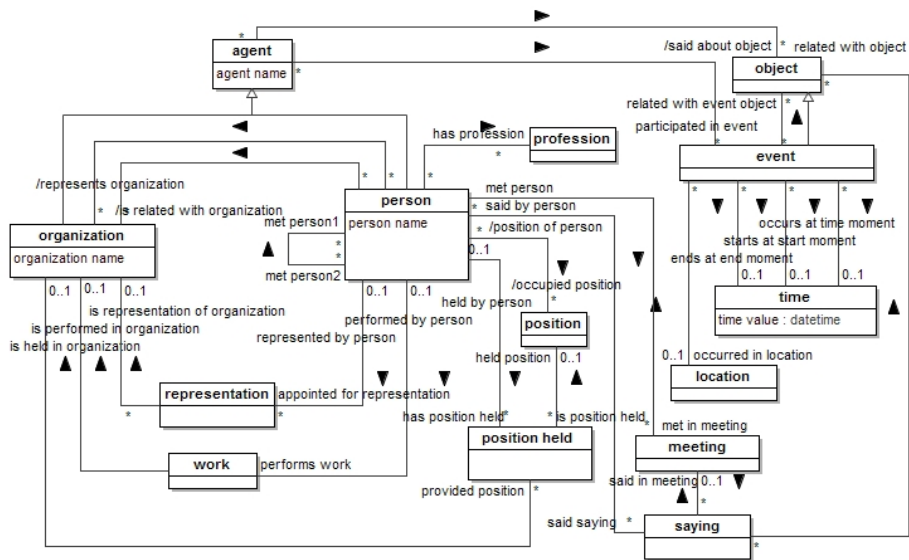


Fig. 2. Initial ontology of some types of political events

4 The Principles for Developing SBVR Vocabularies and Business Rules from OWL 2 Ontologies

Rules of previously developed SBVR into OWL 2 transformation solutions [4, 14] may be adapted to achieve reverse transformation from OWL 2 into SBVR. However, there are some questions requiring specific attention. For developing SBVR vocabularies in Lithuanian on the basis of existing ontologies in English, it is necessary to add labels for Lithuanian language. In order to facilitate the OWL2 to SBVR transformation, it is also desirable to add labels in SBVR style, i.e. connect the words representing single noun concepts or verb concepts with low dashes “_” (e.g. Lithuanian label “political event” will appear also as SBVR label “political event” (for convenience, we extend OWL 2 annotation properties by introducing `label_lt` and

label_sbvr as in such a case it is possible to view OWL 2 ontologies in Lithuanian or SBVR style in the ontology editor Protégé. For representing SBVR concepts we use the SBVR style for terms (political_event), verbs (*participated_in*), Names (Lithuania) and keywords (*that, and, etc.*) [1]. For representing OWL 2 concepts, the abstract syntax of OWL 2 will be used [26].

Generally, for transforming OWL 2 ontology into any language, a manual step would be required to specify or review labels of ontology to be transformed. In transformation, annotation properties “label_sbvr” values will be used instead of class names.

The second question to be considered is the scope of transformation. For applying SBVR vocabulary in semantic search, only a subset of all OWL 2 to SBVR transformation rules would be required because OWL 2 axioms, restrictions and Semantic Web rules mostly are used in the inference, which is always made in ontology before executing SPARQL queries. In the following, the transformation rules will be indicated as semantic Search relevant Rules (SR) and semantic search Irrelevant Rules (IR) that are required for representing the overall ontology in SBVR.

The third question is the modality of rules obtained from ontologies. Ontology axioms and restrictions correspond to SBVR alethic rules, therefore, SBVR vocabulary, obtained from ontology, will have no deontic rules. Decisions for choosing among necessities, possibilities and impossibilities, which are implied in OWL 2, require attention as well.

In the following, the generalized rules for transforming OWL 2 into SBVR are informally presented and illustrated (as far as possible of space limits) by examples of Political events ontology.

Rule 1 (SR). Mapping OWL 2 ontology to SBVR vocabulary:

<pre>OWL2: Ontology (Political_events) → SBVR: <u>Political_events</u> General concept: <u>vocabulary</u></pre>

Rule 2 (SR). Mapping OWL 2 datatypes to SBVR primitive concepts. OWL 2 has a large set of datatypes, including XSD ones, while SBVR has just integers, numbers, and text. SBVR metamodel may be extended by other primitive concepts, e.g. currently datetime and boolean primitive concepts are introduced into SBVR editor [15] but it may not be enough.

Rule 3 (SR). OWL 2 data type values may have restrictions expressed by facet spaces that define data property ranges; these restrictions are mapped to SBVR rules:

<pre>OWL2: DataPropertyRange(:political_rating DatatypeRestriction(xsd:int xsd:maxInclusive "100"^^xsd:integer xsd:minInclusive "0"^^xsd:integer)) → SBVR: It is necessary that <u>political_rating</u> is not greater than <u>100</u> and not less than <u>0</u>.</pre>
--

Rule 4 (SR). Transforming OWL 2 classes into SBVR general concepts:

<pre>OWL2: Declaration (Class (event)) → SBVR: <u>event</u></pre>

Rule 5 (SR). Transforming OWL 2 object properties and data properties into SBVR binary verb concepts. Object properties are mapped to SBVR associations, except for the

object properties having a meaning of “part-whole” relations; these should be mapped to partitive fact types. Automatic transformation cannot recognize that meaning. We can only define a few rules that some set of verbs (consists, is part of, participates, comprise, etc.) have meaning of partitive fact types, but it would not be possible in all cases of “part-whole”. OWL 2 data properties are mapped to SBVR property associations.

```

OWL2: Declaration (ObjectProperty (occupied__position))
ObjectPropertyDomain(occupied__position person)
ObjectPropertyRange(occupied__position position) →
SBVR: person occupied position
    
```

Rule 6 (SR). OWL 2 named individuals along with their assertions are transformed into SBVR individual concepts.

```

OWL2: Declaration(NamedIndividual(Dalia_Grybauskaite))
ClassAssertion(Person Dalia_Grybauskaite)
Declaration(NameIndividual(Angela_Merkel))
ClassAssertion(Person Angela_Merkel)
ObjectPropertyAssertion(met Dalia_Grybauskaite Angela_Merkel)→
SBVR: Dalia_Grybauskaite General concept: person
Angela_Merkel General concept: person
Dalia_Grybauskaite met Angela_Merkel
    
```

Rule 7 (SR). SubClassOf (SubObjectPropertyOf, SubDataPropertyOf) axiom between OWL 2 classes (object properties, data properties) is transformed into specialization between a pair of SBVR concepts (noun concepts or verb concepts):

```

OWL2: SubClassOf(city location) →
SBVR: city General concept: location
    
```

```

OWL2: SubObjectPropertyOf(participated_in__event,
met_in__meeting) →
SBVR: person met_in meeting
General concept: agent participated_in event
    
```

Rule 8 (IR). Transforming OWL2 functional properties into SBVR business rules having “at most one” quantification over SBVR fact type (see an example below). Similarly, OWL 2 existential class expressions are transformed into SBVR business rules having “at least one” quantification over associative or partitive fact type. Similarly, OWL 2 cardinality restrictions are transformed into SBVR business rules having “at most n”, “at least n”, “exactly”, “numerical range” quantifications.

```

OWL2: ObjectProperty(occured_in_location)
FunctionalObjectProperty(a: occured_in__location) →
SBVR: It is necessary that event occured_in at most one locati-
tion.
    
```



Rule 9 (SR). Transforming OWL 2 inverse object properties into synonymous forms of verb concepts and SBVR business rules. It is a special case of synonymous form extraction from domain ontology. In other cases, synonymous forms are included in SLOL.

OWL2: InverseObjectProperty(sayd_saying said_by__person) →
 SBVR: person said saying
 Synonymous form: saying is_said_by person
 It is necessary that person said saying if saying is_said_by person.

Rule 10 (IR). Transforming OWL 2 symmetric, asymmetric, reflexive, irreflexive and transitive properties and object property chains into SBVR business rules using implications.

Rule 11 (SR). Transforming OWL 2 complements into SBVR impossibilities or negations.

OWL2: SubClassOf(corporation ObjectComplementOf(government))
 → SBVR: It is impossible that corporation is government.
 or: It is necessary that not corporation is government.

Rule 12 (SR). OWL 2 Equivalent Classes axiom is used for SBVR formal definitions and categorization schemes or segmentations (Rule 13) except when this axiom is defined between single classes. Such classes as well as object properties or data properties are transformed into SBVR verb concepts using the pattern concept1 is_coextensive_with concept2:

OWL2:
 EquivalentClasses(agent_participated_in_event_at_location participation) →
 SBVR: agent participated in event at location
is_coextensive_with participation

Rule 13 (SR). OWL 2 class hierarchies allowing classifying individuals according different criteria [14] should be transformed into SBVR categorization schemes and segmentations. However, OWL 2 does not explicitly define categorization criteria (SBVR categorization types) for such schemes. Sometimes it would be possible to identify these criteria by analysing subclass definitions but mainly it would be a problem. Therefore, this task is left for analyst who should specify the annotation property “categorization_type” added to OWL 2 annotation properties. Different specializations of the same primary class in ontology are defined by different hierarchies, e.g., a primary class “Person” may have specializations by gender and by occupation (Fig. 3). Each specialization has a top class equivalent to its primary class; subclasses in these specializations have definitions for classifying individuals according the assigned criterion, e.g. politician with high rating would be defined as politician with rating greater than 30%.

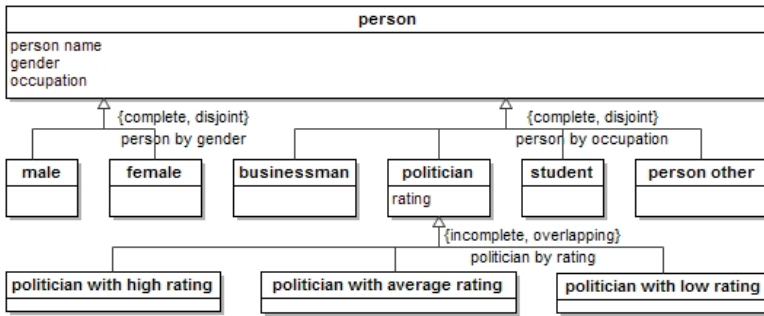


Fig. 3. Categorization schemes and segmentations

```

OWL2: EquivalentClasses (:person :person_by_occupation)
ClassAssertion(:occupation :politics)
EquivalentClasses(:politician ObjectIntersectionOf
(ObjectHasValue(:occupation politics):person_by_occupation))
Declaration(AnnotationProperty(:categorization_type))
AnnotationAssertion(:categorization_type
:person_by_occupation :occupation) →
SBVR: Person by occupation
    Necessity: segmentation for general concept person
        that subdivides person by occupation.
    politician General concept: person
    Necessity: is_included_in Person by occupation
    
```

Rule 14 (SR). All OWL 2 classes, object and data properties, which are disjoint, do not require additional rules in SBVR. But according the open world assumption, if there is no explicit statement about disjointness of some concepts, it is possible that such concepts are the same. Therefore, they require additional rules in SBVR, e.g.:

```

SBVR: It is possible that corporation is_coextensive_with government.
    
```

Similarly, all individuals that are not defined as different or are the same individuals require specifying a possibility about their equivalence, e.g.:

```

SBVR: It is possible that Grybauskaite is Dalia Grybauskaite.
    
```

5 Conclusion and Future Works

The paper presents the principles for creating SBVR vocabularies from ontologies for improving ontology-based semantic search. SBVR vocabularies would allow using the SBVR structured natural language for formulating questions, which could be transformed into SPARQL queries and executed on OWL 2 ontologies. Analysis of existing solutions allows formulating hypotheses about possibility to use separate

domain ontologies and domain specific lexical ontologies based on SBVR representations for semantic search. Moreover, we argue that domain ontology presented in e.g. English language may be used for semantic search in any language having SBVR based lexical ontologies for each language.

Current paper was focused on transforming OWL 2 domain ontology concepts into SBVR. Such transformation requires preparation made by a human. Domain ontology should be manually annotated with SBVR preferred representations in the desired language and some additional concepts as synonymous forms for inverse object properties and categorization types.

The presented transformation principles reflect the key solutions proposed for transforming domain ontologies into SBVR. The main attention was paid to transformations relevant for semantic search as a part of ontology axioms or Semantic web rules is important for inference, which is executed in ontology before executing SPARQL queries and is not used for formulating SBVR questions. The set of OWL2 to SBVR transformation rules was tested on examples of OWL 2 ontologies and SBVR vocabularies and requires further research. The complete solution for semantic search requires more efforts, especially for creating lexical ontologies and rules for relating them with SBVR meaning.

Acknowledgements. The work is supported by the project VP1-3.1-ŠMM-10-V-02-008 “Integration of Business Processes and Business Rules on the Basis of Business Semantics” (2013-2015), which is funded by the European Social Fund (ESF).

References

1. OMG. Semantics of Business Vocabulary and Business Rules (SBVR). SBVR 1.1 RTF Convenience document. OMG Document Number: dtc/2012-06-10, pp. 1–436 (2012)
2. W3C. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax. In: Motik, B., Patel-Schneider, P.F., Parsia, B. (eds.) W3C Recommendation, October 27, vol. 134, pp. 1–134 (2009)
3. El Ghali, A., Chniti, A., Citeau, H.: Bringing OWL ontologies to the Business Rules Users. In: Bikakis, A., Giurca, A. (eds.) RuleML 2012. LNCS, vol. 7438, pp. 62–76. Springer, Heidelberg (2012)
4. Sukys, A., Nemuraite, L., Paradauskas, B., Sinkevičius, E.: Transformation framework for SBVR based semantic queries in business information systems. In: Bustech 2012: the Second International Conference on Business Intelligence and Technology, Nice, France, July 22-27, pp. 1–6. IARIA (2012)
5. Sukys, A., Nemuraite, L., Paradauskas, B.: Representing and transforming SBVR question patterns into SPARQL. In: Skersys, T., Butleris, R., Butkiene, R. (eds.) ICIST 2012. CCIS, vol. 319, pp. 436–451. Springer, Heidelberg (2012)
6. Lin, F., Sandkuhl, K.: A Survey of Exploiting WordNet in Ontology Matching. In: Bramer, M. (ed.) Artificial Intelligence in Theory and Practice II. IFIP, vol. 276, pp. 341–350. Springer, Boston (2012)
7. WordNet: A lexical database for English, <http://wordnet.princeton.edu/>
8. VerbNet: A Class-Based Verb Lexicon, <http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>
9. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet Project. In: COLING 1998 Proceedings of the 17th International Conference on Computational Linguistics, vol. 1, pp. 86–90 (1998)

10. Dzikovska, M.O., Swifty, M.D., Allen, J.F.: Building a computational lexicon and ontology with FrameNet. In: Fillmore, C.J., et al. (eds.) LREC, Lisbon, pp. 53–60 (2004)
11. Kendall, E., Linehan, M.H.: Mapping SBVR to OWL2. IBM Research Report, RC25363, WAT1303-040 (2013)
12. Collibra Data Governance Software, <http://www.collibra.com/>
13. ONTORule Project: ONTOlogies meet Business RULES, <http://ontorule-project.eu/>
14. Karpovic, J., Nemuraite, L., Stankeviciene, M.: Requirements for Semantic Business Vocabularies and Rules for Transforming Them into Consistent OWL2 Ontologies. In: Skersys, T., Butleris, R., Butkiene, R. (eds.) ICIST 2012. CCIS, vol. 319, pp. 420–435. Springer, Heidelberg (2012)
15. Nemuraite, L., Skersys, T., Sukys, A., Sinkevičius, E., Ablonskis, L.: VETIS tool for editing and transforming SBVR business vocabularies and business rules into UML&OCL models. In: Targamadze, A., Butleris, R., Butkiene, R. (eds.) Information Technologies 2010: Proceedings of the 16th International Conference on Information and Software Technologies, IT 2010, Kaunas, Lithuania, April 21–23, vol. 384, pp. 377–384 (2010)
16. Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M.: LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Journal Web Semantics: Science, Services and Agents on the World Wide Web* 9(1), 29–51 (2011)
17. Rector, A.L.: Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL. In: K-CAP 2003, pp. 121–128. ACM, New York (2003)
18. Declerck, T., Gromann, D.: Combining three Ways of Conveying Knowledge: Modularization of Domain, Terminological, and Linguistic Knowledge in Ontologies. In: Proceedings of the 6th International Workshop on Modular Ontologies, Graz, Austria, CEUR-WS, Aachen. CEUR Workshop Proceedings, vol. 875, pp. 28–40 (2012)
19. The Definition of Sekinefs Extended Named Entities, http://nlp.cs.nyu.edu/ene/version7_1_0Beng.html
20. Kargioti, E., Konopoulos, E., Bassiliades, N.: OntoLife: An Ontology for Semantically Managing Personal Information. In: Iliadis, L., Maglogiannis, I., Tsoumakas, G., Vlahavas, I., Bramer, M. (eds.) Artificial Intelligence Applications and Innovations III. IFIP AICT, vol. 296, pp. 127–133. Springer, Heidelberg (2009)
21. Brickley, D., Miller, L.: FOAF Vocabulary Specification 0.98 (2010), <http://xmlns.com/foaf/spec/>
22. Nemuraite, L., Paradauskas, B.: A methodology for engineering OWL 2 ontologies in practise considering their semantic normalisation and completeness. *Electronics and Electrical Engineering* 4(120), 89–94 (2012)
23. GeoNames, <http://www.geonames.org/>
24. Kaneiwa, K., Iwazume, M., Fukuda, K.: An upper ontology for event classifications and relations. In: Orgun, M.A., Thornton, J. (eds.) AI 2007. LNCS (LNAI), vol. 4830, pp. 394–403. Springer, Heidelberg (2007)
25. Scherp, A., Franz, T., Saathoff, C., Staab, S.: F-A Model of Events based on the Foundational Ontology DOLCE+DnS Ultralite. In: Proceedings of International Conference on Knowledge Capture (K-CAP), California, pp. 137–144 (2009)
26. Motik, B., Patel-Schneider, P.F., Parsia, B.: OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax. W3C Proposed Recommendation (September 22, 2009)

Exploring Key Factors of Pilot Projects in Agile Transformation Process Using a Grounded Theory Study

Taghi Javdani Gandomani, Hazura Zulzalil, Abdul Azim Abdul Ghani,
Abu Bakar Md. Sultan, and Khaironi Yatim Sharif

Faculty of Computer Science and Information Technology,
University Putra Malaysia (UPM), Serdang, Malaysia
tjavidani@yahoo.com,
{hazura, azim, abakar, khaironi}@upm.edu.my

Abstract. Changing development approach from disciplined to agile methods is an organizational mutation that requires many issues to be considered to increase its chance of success. Selecting an appropriate pilot project as initial project that is going to be done through an Agile method is a critical task. Due to the impact of such a pilot project on successful Agile transformation, understanding its critical factors helps organizations choose the most suitable project to start Agile transition. Conducting a Grounded Theory, showed that organization should considered some key factors of a pilot: Criticality, Duration, Size and Required resources. Besides these factors, the results showed that organization should be aware of the risk of successful pilot project in their next Agile projects. The study also showed that pilot selection mostly is done by Agile coaches or is forced by customer.

Keywords: gile Software Development, Agile Transformation, Agile Pilot Project, Grounded Theory.

1 Introduction

Agile software development by offering different values comparing to disciplined methodologies, is tempting software companies. Although formal introduction of agile methods was done by creating agile manifesto at 2001[1], their prevalence has increased in recent years.

Altering development method from disciplined methods to agile methods, called Agile Transformation Process (ATP), is not an easy and ordinary change and needs considerable time and effort [2]. Furthermore, companies face with several challenges and issues along the transformation [3]. Changing development approach is a socio-technical change and affects all aspects of organization. Altering people roles and responsibilities lead to changing their behaviors and mindsets [4]. In sum up, while using agile methods are so attractive, fully adoption is so hard.

Between multiple factors that may affect ATP, including effective training, coaching, team set up, transition framework and so forth, selecting pilot project is a critical issue [5]. Pilot project is the initial project which during it, a company tries to adaptation to

agile methods or practices. There is no common and popular approach about characteristics of pilot projects in ATP. While some of the Agile transformation drivers and prerequisites have studied in several researches, less attention has been paid to pilot project and its key characteristics.

During a Grounded Theory study on ATP, different approaches and key characteristics of agile pilot projects were emerged. This study intends to focus on this issue and help software companies by elucidating real facts about Agile pilot projects.

The rest of this article organized as following: Section 2 provides a concise background about agile transformation process; Section 3 presents research methodology; Section 4 provides results of the study; Section 5 provides discussion on findings and future works; Section 6 presents limitations of the study and finally Section 7 provides a conclusion.

2 Agile Transformation Process

As previously mentioned, migrating to agile needs essential cultural and organizational changes in organization. Explaining all aspects of ATP is beyond the scope of this paper, but in this section some important issues are provided briefly.

ATP as an infrastructural project needs a detail action plan which encompasses all related issues. So far, many studies have been done on almost all aspects of ATP. Some of these studies have focused on highlighting challenges and problems that companies may face with during ATP [6, 7]. Organization and management, people, process and tools are the areas that most of the challenges can be seen in them [3].

In some others, a few transformation and adoption frameworks have been proposed [8, 9]. However, none of the proposed frameworks are not enough popular or easy to use and contain a huge organizational overhead.

At the same time, in some other studies human aspects as critical factors are studied [10-13]. In these studies, people mindsets and cultural issues are emphasized. Since agile methods focus on people, paying enough attention to their related issues is so critical and necessary. Furthermore, in some others, drivers and prerequisites of ATP are investigated. Addressing these drivers including training, customer collaboration, people and management commitment and buy-in, champions, practice selection and so forth are just some of the results of these studies.

On the other hand, many articles explained journey of companies and organizations to agile. Lessons learned and strengths and weaknesses of companies during ATP are highlighted in these studies [7, 14, 15]. The findings of these studies give other companies better vision to avoid colliding with obstacles in their migration to Agile.

Choosing a pilot project is also one of the significant items which can affect ATP strongly. For instance, in a company that intends to apply Scrum for agile adoption, choosing a pilot project that its customer does not any sense about agile values and collaboration in development process, causes adaptation to be so hard.

In sum up, in ATP many factors should be considered together and covering all of them in a single article is impossible. Following the selected research methodology, only a mini literature review was needed before conducting the study [16]. After developing theory, the findings will be discussed in light of literature review. In the next sections, this study focuses on pilot project in real agile transformation and on the basis of real data proposes possible considerations about choosing pilot project.

3 Research Methodology

We used Grounded Theory (GT) as research methodology to performing this study. This methodology was created by Glaser and Strauss [17] and by defining a systematic approach tries to discover the grounded theory on the basis of the substantive and grounded data [18, 19]. Strauss explained that GT uses a "systematic set of procedures to develop an inductively derived grounded theory about a phenomenon" [20]. This method is so helpful to answering questions like, "what is going on in an area?" by generating formal or substantive theory [20]. However this method usually have been used in social studies, it is useful for a wide range of topics in software engineering, especially those which are related to people behaviors and human aspects [21].

GT was chosen because of multiple reasons. 1) Agile methodologies are people oriented and their values and practices rely on people and on the other hand, GT assists researchers to do studies on people interactions and their behaviors. 2) For phenomena that are not studied in deep or in global perspective, GT is completely suitable; meanwhile, agile transformation process as a separate empirical study in real environment was not studied yet during ATP and in most of the times, it was studied from a specific rather than global perspective [22]. 3) There were enough successful evidences of using GT in agile software development in the recent years [23-27].

This methodology is useful in studies that researches cannot define upfront hypotheses and they are looking for main concerns of participants in real environments [28]. In this case, research questions should emphasize on a wide area rather than specific topic [27]. This study followed such a strategy and by coding substantive data and abstracting them in a multi-level analysis process, core concern and its related categories and properties were emerged, as Glaser explained in his instructions [29].

3.1 Data Collection

Since GT starts with data collection [17], it was performed as the first step of this study. By publishing an invitation for expert agile practitioners in on-line communities, enough volunteers registered for participating in this study. The main requirement for participants was having at least one transformation experience. Then, several semi-structured and on-line interviews were conducted with them using open ended questions. Table 1 shows the participants whom their point of views were used in this study. Since all of the participants were from different countries, face-to-face interview was not possible. Selected candidates were 32 agile experts from 13 different countries and in this article they are referred by their numbers, P1 to P32. They were

using combination of agile methods, mainly Scrum, XP and Kanban as the most popular methods these days [30].

Interviews started with a few general questions about the respondents' background and experience. Afterwards, some general questions about their experience in dealing with issues and challenges during ATP. Consistent with GT, questions were not addressed specific items or issues and were focused on general concepts [16]. Data collection was continued up to reaching a saturation level which means emerging no new code related to specific concept or category [17]. At the time of writing this report, data collection in other categories is still ongoing.

3.2 Data Analysis

Data analysis was done in a multi level process. Using Nvivo as powerful software package for handling data and facilitating data analysis was so helpful. It decreased substantially risk of human errors in all levels of the data analysis. In the first level, all transcripts were reviewed line-by-line, and *key points* of them were found [17]. Each key point was assigned with an *Open code*, called *Free node* in Nvivo. Using *Constant comparison*, comparing newly emerged open code with previous codes in the same and pervious transcripts, helped to grouping the codes and finding out *Concepts*, a higher level of abstraction in data analysis [17]. Iterative using this technique on emerged concepts, lead to emerging *Categories*, which were a higher abstraction level comparing to concepts [17].

Axial coding, making connection between emerged categories [20], was used for putting back fractured data together. This technique tries to showing potential relationship between categories or their subcategories and properties [20]. Importance and key factors of pilot projects were consequence of applying this technique. The next step was Selective coding; used for integrating and refining the theory [20]. Following the GT instructions, this step of data analysis focused on only variables that related to pilot projects[18]. The emerged core category of this study was "*Iterative Agile Transition Process*" and "*pilot Projects*" was one of the core category related categories.

Using *memoing* for adding several memo in collected data or after each interview gave a good chance to improve quality of collected data [29]. When almost all codes were reached to an accepted saturation level, conceptual sorting helped the authors to depict emerged theory outline [20]. Figure 1 depicts the levels of coding in this study.

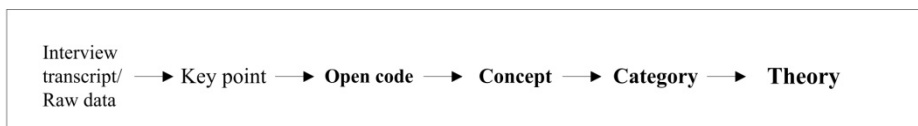


Fig. 1. The levels of data analysis based on the Grounded Theory method

Table 1. Demography of the participants (SM=Scrum Master, AC=Agile Coach, CON= Agile Consultant, DEV= Agile Developer, PM= Project Manager, QA= Quality Assurance, HDD=Head of Development Department)

No.	SD exp. (Yrs)	Agile exp. (Yrs)	Country	Position	Methods	Company size	Transition Duration (month)
P1	14	8	Finland	HDD	Scrum, XP, Kanban	70+	12+, Ongoing
P2	25	15	USA	AC	Scrum, Kanban	65+	12+
P3	7	7	USA	PM	Scrum, XP, Kanban	500+	6+, Ongoing
P4	10	2	Bulgaria	PM	Scrum, XP, Kanban	200+	6+, Ongoing
P5	10	2	Iran	PM	Scrum, Kanban	150+	12+, Ongoing
P6	11	8	Australia	CON	Scrum, Kanban, FDD	1000+	12-15
P7	6	2	Greek	DEV	Scrum	20+	12+, Ongoing
P8	10	5	Germany	PM	Scrum, Kanban	50+	8+, Ongoing
P9	20	10	Spain	HDD	Scrum	200+	24
P10	20	3	Spain	SM	Scrum, Kanban	200+	24+, Ongoing
P11	10	4	India	AC	Scrum, XP, Kanban	50+	+6, Ongoing
P12	16	2	USA	HDD	Scrum, Kanban	1600+	6+, Ongoing
P13	14	6	Finland	AC	Scrum, Kanban	20+	3-30
P14	15	3	Iran	MGT	Scrum, Kanban	50+	12
P15	10	2	Indonesia	CON	Scrum	200+	3+, Ongoing
P16	21	10	USA	PM	Kanban	65+	12
P17	19	5	Sweden	PM	Scrum, Kanban	50+	24+, Ongoing
P18	8	2	Sweden	DEV	Scrum	40+	24
P19	13	6	India	PM	Scrum	200+	USA:18, India:24
P20	11	3	USA	HDD	Scrum, Kanban	1200+	6+, Ongoing
P21	16	7	USA	SM	Scrum, XP	250	18
P22	11	5	France	AC	Scrum, Kanban	2000+	12+, Ongoing
P23	16	8	USA	AC	Scrum, XP, Kanban	200+	6-24
P24	15	7	USA	SM	Scrum, XP	40+	6+, Ongoing
P25	8	4	USA	DEV	Scrum, XP	300+	15+
P26	13	6	India	AC	Scrum, XP	50+	12+
P27	14	5	USA	SM	Scrum, Kanban	40+	6+, Ongoing
P28	15	6	Germany	AC	Scrum, Kanban	50+	15+
P29	10	1	Norway	PM	Scrum	40+	12+
P30	35	1	USA	DEV	Scrum	100+	6+, Ongoing
P31	17	4	USA	QA, PM	Scrum	50+	12
P32	25	2	USA	AC	Scrum, Scrumban	200+	12, Ongoing

3.3 Theory Building

Theory building, called *theoretical coding* in GT, was the last step of the study. Several approaches were proposed about theoretical coding. Induction or theory emergence was stressed by Glaser [16], while Strauss believed on a systematic approach and validation criteria [20] and Charmaz emphasizes on role and effect of researchers on theory building [31]. Following Glaser's view and using *Process Family* [18] or *Temporal Family* [19], formed the general theory of this study.

AS mentioned in previous section, "Iterative Agile Transition Process" was the primary theory and involved several related categories such as "Continuous Assessment", "Transition Facilitators", "Iterative Transition Framework", "Pilot Projects", "Prerequisites of Transition" and so forth.

Pilot project, as one of the related categories of the core category will be discussed in this article. Of course, due to the limited space, all of the quotes and codes, concepts and properties cannot be provided and only those that help to clarify the emerged concepts will be proposed. Figure 2 depicts emergence of Agile Pilot Project theory on the basis of GT data analysis.

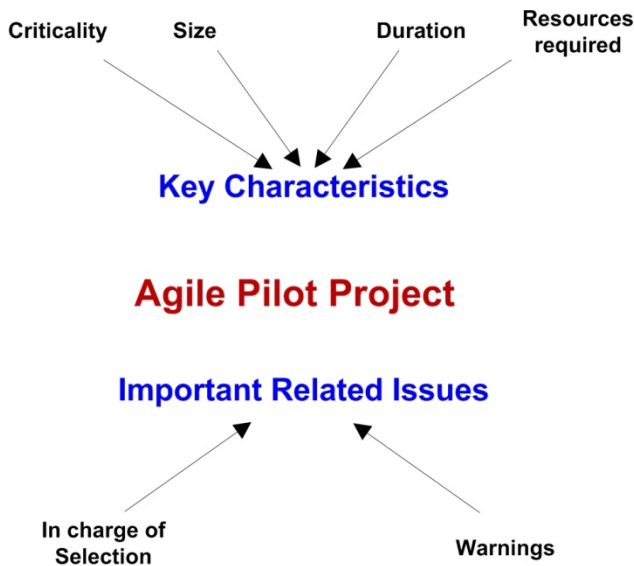


Fig. 2. Emergence of Agile Pilot Project Key Characteristics and Related Issues

4 Pilot Projects in Agile Transition

Attending the experienced participants made an opportunity to add their valuable experiences to this study. The results of this study discovered key characteristics of pilot projects in Agile transition. Furthermore, the results showed that some other factors should be considered in project selection process. This section explains the results for both items.

4.1 Key Characteristics of Agile Pilot Project

Although several issues were explained about the main characteristics of a candidate pilot project, they could be grouped in four main categories: Criticality, Duration, Size and Required resources.

Criticality and Importance

There was no consensus on degree of criticality and importance degree of a pilot project. Some of the participants believed that it should be only a training project or a “pet project”. (P1)

“Pilot should be a training project; you pick it up for transformation to agile only, not for doing an agile project, since you are not agile yet” P18, Agile Developer

“I would say that define your pilot project as learning project not as yes, we do a very very effective agile project because it won’t be very very agile project, because your company is not agile yet.” P13, Agile Coach.

Some other of the participants believed that pilot should be a real project or at least one part of a real project. They believed that training projects are useful but not effective for Agile adoption.

“I think that agile transformation should be started with a real project. This is the only way for real migration. Facing with real challenges and issues will help people to learn more and get more powerful, otherwise, they may increase their knowledge in agile, but they cannot use it in a real environment.” P4, Project Manager.

“Without a real project, teams cannot learn the real challenges...” P28, Agile Coach.

Nonetheless, they emphasized that real project should be low risk. This feature leads to paying more attention to transition rather than project risks and challenges.

“It is better that you start with some specific and non-critical projects. This strategy helps you to pay attention more to people problems and after successful migration...” P4, Project Manager.

The participants referred to their experiences about choosing training or real projects. At the same time, most of them believed that pilot should be managed carefully, regardless of its nature, training or real. This point was more emphasized by some of them. (P9, P23)

Duration

Despite of previous characteristic, regarding to duration of agile pilot project, there was a considerable consensus among the participants. They believed that pilot project should be small. They declared that large scale projects have lots of challenges and teams’ focus would be on these challenges and not on transformation to Agile.

“Let it be small project. It cannot show your agility but can help you being agile.” P16, Project Manager.

Choosing small projects leads to more rapid transition and adoption. It helps teams to prove that agile works well.

“Small junks i.e., begin with a small project in order to prove that agile is successful.” P12, Head of Department.

Although no one addressed fix duration for pilot projects, duration about three or four months, was mostly agreed by the participants.

Size

Selecting projects that need minimum number of teams was emphasized by most of the participants.

“Pilot project should be chosen carefully. I suggest a project with no more than 12 people. There are two advantages in starting small: minimizing disruption and leveraging what the pilot group learns about doing Agile in environment when transitioning other groups.” P23, Agile Coach.

Starting with only one team and then adding more teams, if necessary, was also suggested by some of them. This strategy lets coaches to focus on specific team(s) and help them to change their development approach easily and quickly.

“Pilot should consist of one or two teams. Lots of challenges will not occur if fewer teams get involve.” P26, Agile Coach.

“I always start with maximum 10 persons in 2 teams. Even with such a few number of people, I have faced with many challenges.” P22, Agile Coach.

Fewer numbers of teams participating in Agile transition, generally decrease risk of Agile transition. Indeed, increasing number of teams leads to increasing potential challenges.

“One way for handling challenges during agile migration is picking a project that needs the minimum number of teams.” P24, Scrum Master.

Small projects consisting no more than two or three teams were recommended by many of the participants. Furthermore, *“Co-located teams”* was strongly stressed for running selected pilot. (P2, P26).

Required Resources

The last important factor that emerged in this study was attention to the resources required for running a pilot project.

“For choosing a pilot project, attention to team capabilities is a critical factor. It is the best to choose a pilot that all required resources are available. For instance, if Product Owner is not available in a pilot, expecting agile adoption in PO related practices is wrong.” P19, Project Manager.

“We did a pilot but did not select well. We selected a pilot with a group of “virtual” team members that were not dedicated. Did not have a real Product Owner named and Scrum Master was a Project manager.” P20, Head of Department.

Two of the participants mentioned to *“pilot team”* (P2, P13); they believed that pilot project should be selected based on capabilities of *“pilot teams”*.

“If business people or customers have not enough engagement, choosing their project as a pilot for agile transformation would be an irrational decision.” P6, Agile Consultant.

The participants believed that, because of the critical role of pilot project in Agile transition, required resources, technical and business resources should be available for it. In the other words, pilot should be picked up based on the available resources and their capabilities and characteristics.

4.2 Pilot Project Selection Related Issues

The study showed that besides of the characteristics of pilot project, several other important issues also needed to be considered. These issues are related to pilot project selection and its impact on ATP. These results are explained in this section.

In Charge of Pilot Selection

There was no common agreement on “in charge of pilot selection” among the participants. Some of the participants declared that Agile coaches are responsible for choosing pilot projects. They believed that coaches have enough knowledge about all related potential challenges, thus, they are responsible for pilot project selection.

“A coach is only one who is eligible for selecting the right pilot project. Experienced Agile coaches with enough knowledge can predict future challenges and by choosing best pilot try to decrease potential challenges” P3, Project Manager.

“Selecting best case for pilot is one the duties of an Agile coach.” P32, Agile Coach.

On the other hand, some others believed that in real environments, companies are limited to their customers’ decisions.

“It is always difficult because generally the project has already been chosen by the customer who fantasizes about the agile...” P19, Scrum Master.

“It is right that a coach or a manager choose pilot project, but it’s finally its customer that accept to participant in transformation or no. Therefore, I believe that role of customer is more important than others.” P27, Scrum Master.

Finally, some of the participants addressed a committee involving coach, management, customer and senior team members, as in charge of pilot selection.

“Most often I ask top manager to nominate a committee for handling transition issues such selecting pilot, required training...” P13, Agile coach.

Warnings

The participants warned about some facts of pilot projects and their impacts on whole transition process. Some of the participants explained how in some cases after successfully doing a pilot, most often teams have a wrong perception of their capability in Agile adoption.

“The risk of a pilot project is that it may not be representative for the rest of the organization...” P2, Agile Coach.

Running pilot under enough control and almost best conditions, is only an organizational try to being Agile.

“Pilot projects are problematic because they are usually staffed with the best and most motivated individuals and have management backing.” P1, Head of Department.

While choosing best pilot “*is an art*” (P21) and facilitates Agile adoption in other teams, wrong selection makes a critical challenge for Agile migration.

“Wrong selection of pilot project is harmful and causes a lot of cost and failures in transformation process. Such a pilot jeopardizes transition process.” P28, Agile Coach.

5 Discussion and Future Works

While several studies have been done on different aspects of ATP, a few studies focused on pilot project and its important factors. However, choosing pilot project as one of the critical factors of ATP, regardless of its factors and characteristics, was emphasized by many studies.

Many studies explained their journey to Agile without focusing on how they selected pilot projects. At the same time, a few studies explained that their selection was based on some simple criteria such as customer request, business limitation or management request [7, 32, 33]. However, in a few studies, Agile adoption and transformation process considered to be performed with carefully considering pilot project [8, 9]

Boehm et al. believed that specifications of pilot project should be considered before applying any agile methods [34]. In their risk-based framework they addressed some factors such as size, criticality, dynamism, culture and personnel that affect on decision of using any Agile methods. Although their main purpose was not addressing factors of a pilot project, their innovative idea was interesting.

Highsmith discussed about importance of pilot project. As an early agilist, he believed that if a pilot is not important enough, many people don't do their best in running it [2]. He believed that in this case team members may discard difficult Agile practices. Cohn by describing some important factors of Agile pilot projects, addressed some critical factors for each pilot project: importance, business sponsor engagement, size, duration and people [5]. He also suggested that team set up should be considered before pilot selection.

Honious et al. stressed that pilot selection is the most critical and challenge task for migrating to agile [35]. They explained how their initial pilot had some appropriate features that helped them in Agile transition.

Absence of a pilot project was addressed as a critical challenge in some studies [36]. Finally, in some other studies, researchers explained how features and characteristics of pilot projects affect migration process [14, 37, 38]. Furthermore, adopting those Agile practices which need specific roles, if those roles are not available or not supported by pilot, are serious challenges in moving to Agile [3].

Due to weak academic background about Agile pilot project and its related issues, this subject is fertilize for conducting more researches. Proposing a model for pilot selection, based on the critical characteristics and organization capabilities and limitations can be a potential future work. Furthermore, tracking the challenges that cause by wrong or weak selection of pilot projects and strategies that can be used to deal with this issue can be studied by application of a multiple-case study.

6 Limitation

All the emerged key points, codes, concepts and categories of this study directly came from the data that was collected from real environments; therefore, the findings of the study are grounded in substantive environments [29]. Since access to resources was limited to the above participants, this study cannot claim that its findings are universal. However, it claims that its findings have characterized and described the context studied [39].

7 Conclusion

Performing a Grounded Theory study involving 32 agile practitioners about Agile transformation process in real environments, showed that besides of many critical factors, pilot project has substantial effect on Agile transition. The study explained that Agile experts addressed several key characteristics of pilot projects. The key feature that software companies and organizations should consider them for choosing their pilot projects are: Criticality, Duration, Size and Required resources. However proposing the best features of a pilot is not easy, some of the recommended features are: small size with no more than two or three teams, short time project with around three or four months duration, low risk project and supporting by right personnel.

The study also showed that some newly-Agile teams may have a wrong perception of their Agility capability. Successful pilot do not mean that they are enough expert in Agile. Also, the study showed that pilot selection most often is forced by Agile coach or customer.

Acknowledgements. This study was financially supported by the University Putra Malaysia (UPM) under the International Graduate research Fellowship (IGRF). The authors also are thankful to the participants of the study who shared their valuable experiences in this study.

References

1. <http://www.agilemanifesto.org>
2. Highsmith, J.A.: Agile Software Development Ecosystems. Addison-Wesley Professional, Boston (2002)
3. Gandomani, T.J., Zulzali, H., Ghani, A.A.A., Sultan, A.M., Nafchi, M.Z.: Obstacles to moving to agile software development; at a glance. *Journal of Computer Science* 9, 620–625 (2013)
4. Conboy, K., Coyle, S., Wang, X., Pikkarainen, M.: People over process: Key challenges in agile development. *IEEE Software* 28, 48–57 (2011)
5. Cohn, M.: *Succeeding with Agile: Software Development Using Scrum*. Addison-Wesley Professional, Boston (2009)
6. Pikkarainen, M., Salo, O., Kuusela, R., Abrahamsson, P.: Strengths and barriers behind the successful agile deployment-insights from the three software intensive companies in Finland. *Empirical Software Engineering* 17, 675–702 (2012)
7. Ganesh, N., Thangasamy, S.: Lessons learned in transforming from traditional to agile development. *Journal of Computer Science* 8, 389–392 (2012)
8. Sidky, A., Arthur, J., Bohner, S.: A disciplined approach to adopting agile practices: The agile adoption framework. *Innovations in Systems and Software Engineering* 3, 203–216 (2007)
9. Qumer, A., Henderson-Sellers, B.: A framework to support the evaluation, adoption and improvement of agile methods in practice. *Journal of Systems and Software* 81, 1899–1919 (2008)

10. Tolfo, C., Wazlawick, R.S., Ferreira, M.G.G., Forcellini, F.A.: Agile methods and organizational culture: Reflections about cultural levels. *Journal of Software Maintenance and Evolution* 23, 423–441 (2011)
11. Vijayasarathy, L., Turk, D.: Drivers of agile software development use: Dialectic interplay between benefits and hindrances. *Information and Software Technology* 54, 137–148 (2012)
12. Sutharshan, A.: Enhancing Agile methods for multi-cultural software project teams. *Modern Applied Science* 5, 12–22 (2011)
13. Sohaib, O., Khan, K.: Integrating usability engineering and agile software development: A literature review, pp. V232–V238 (2010)
14. Srinivasan, J., Lundqvist, K.: Agile in India: Challenges and lessons learned. In: 3rd India Software Engineering Conference, ISEC 2010, pp. 125–130. ACM, New York (2010)
15. Gandomani, T.J., Zulzalil, H., Ghani, A.A.A., Sultan, A.B.M.: Towards comprehensive and disciplined change management strategy in agile transformation process. *Research Journal of Applied Sciences, Engineering and Technology* 6, 2345–2351 (2013)
16. Glaser, B.: *Basics of Grounded Theory Analysis: Emergence Vs. Forcing*. Sociology Press, Mill Valley (1992)
17. Glaser, B., Strauss, A.: *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction Chicago (1967)
18. Glaser, B.G.: *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*. The Sociology Press, Mill Valley (1978)
19. Glaser, B.G.: *The Grounded Theory Perspective III: Theoretical Coding*. Sociology Press, Mill Valley (2005)
20. Corbin, J.M., Strauss, A.C.: *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory (3e)*. SAGE Publications, Inc., Thousand Oaks (2008)
21. Hoda, R., Noble, J., Marshall, S.: Using grounded theory to study the human aspects of software engineering. In: *Human Aspects of Software Engineering*, pp. 1–2. ACM, USA (2010)
22. Dingsøyr, T., Nerur, S., Balijepally, V., Moe, N.B.: A decade of agile methodologies: Towards explaining agile software development. *Journal of Systems and Software* 85, 1213–1221 (2012)
23. Ghanam, Y., Maurer, F., Abrahamsson, P.: Making the leap to a software platform strategy: Issues and challenges. *Information and Software Technology* 54, 968–984 (2012)
24. Hoda, R., Noble, J., Marshall, S.: The impact of inadequate customer collaboration on self-organizing Agile teams. *Information and Software Technology* 53, 521–534 (2011)
25. Baskerville, R., Pries-Heje, J., Madsen, S.: Post-agility: What follows a decade of agility? *Information and Software Technology* 53, 543–555 (2011)
26. Hoda, R., Noble, J., Marshall, S.: Developing a grounded theory to explain the practices of self-organizing Agile teams. *Empirical Software Engineering* 17, 609–639 (2011)
27. Coleman, G., O'Connor, R.: Using grounded theory to understand software process improvement: A study of Irish software product companies. *Information and Software Technology* 49, 654–667 (2007)
28. Parry, K.W.: Grounded theory and social process: A new direction for leadership research. *Leadership Quarterly* 9, 85–105 (1998)
29. Glaser, B.: *Doing Grounded Theory: Issues and Discussions*. Sociology Press, Mill Valley (1998)
30. <http://www.versionone.com/state-of-agile-survey-results/>

31. Charmaz, K.: *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE Publications Ltd., London (2006)
32. Drobka, J., Noftz, D., Raghu, R.: Piloting XP on four mission-critical projects. *IEEE Software* 21, 70–75 (2004)
33. Cohen, D., Lindvall, M., Costa, P.: An Introduction to Agile Methods. *Advances in Computers* 62, 1–66 (2004)
34. Boehm, B., Turner, R.: Using risk to balance agile and plan-driven methods. *Computer* 36, 57–66 (2003)
35. Honious, J., Clark, J.: Something to believe in. In: *AGILE Conference*, pp. 203–210 (2006)
36. Hajjdiab, H., Taleb, A.S.: Agile adoption experience: A case study in the U.A.E. In: *IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS 2011)*, pp. 31–34. IEEE Computer Society, Washington, DC (2011)
37. Fulgham, C., Johnson, J., Crandall, M., Jackson, L., Burrows, N.: The FBI gets agile. *IT Professional* 13, 57–59 (2011)
38. Heidenberg, J., Matinlassi, M., Pikkarainen, M., Hirkman, P., Partanen, J.: Systematic piloting of agile methods in the large: Two cases in embedded systems development. In: Ali Babar, M., Vierimaa, M., Oivo, M. (eds.) *PROFES 2010*. LNCS, vol. 6156, pp. 47–61. Springer, Heidelberg (2010)
39. Adolph, S., Hall, W., Kruchten, P.: A methodological leg to stand on: Lessons learned using grounded theory to study software development. In: *2008 Conference of the Center for Advanced Studies on Collaborative Research: Meeting of Minds*, pp. 166–178. ACM, Ontario (2008)

Incompleteness in Conceptual Data Modelling

Peter Thanisch¹, Tapio Niemi², Jyrki Nummenmaa¹, Zheyong Zhang¹,
Marko Niinimäki², and Pertti Saariluoma³

¹ School of Information Sciences, University of Tampere, Tampere, Finland

² Helsinki Institute of Physics, University of Helsinki, Finland

³ University of Jyväskylä, Finland

{peter.thanisch,jyrki}@sis.uta.fi, {tapio.niemi,man}@cern.ch,
zheyong.zhang@uta.fi, pertti.saariluoma@jyu.fi

Abstract. Although conceptual data modelers can "get creative" when designing entities and relationships to meet business requirements, they are highly constrained by the business rules which determine the details of how the entities and relationships combine. Typically, there is a delay in realising which business rules might be relevant and a further delay in obtaining an authoritative statement of these rules. We identify circumstances under which viable database designs can be constructed from conceptual data models which are incomplete in the sense that they lack this "infrastructural" detail normally obtained from the business rules. As such detail becomes available, our approach allows the conceptual model to be incrementally refined so that each refinements can be associated with standard database refactorings, minimising the impact on database operations. Our incremental approach facilitates the implementation of the database earlier in the development cycle.

Keywords: Conceptual data modeling, entity-relationship, database refactoring.

1 Introduction

In his general principles of conceptual modeling, Bernhard Thalheim states that engineering a conceptual model must address the problem of "incompleteness both in specification and in coverage of the application domain" ([9], p. 77). In the present paper, we explore a way to represent some of that incompleteness in the conceptual model. Our research emphasizes Thalheim's observations that, viewed as artifacts, "models represent the state of knowledge of the user" ([9], p. 78) and that models "are changing artifacts due to changes imposed by . . . development rules for partial delivery of models, partial usage and deployment" ([9], p. 86).

Contrary to the impression given by many textbook accounts of conceptual data modeling, it is not always the case that the conceptual modeler is provided with a complete and consistent description of the world-to-be-modeled. In particular, incremental development methodologies [5] such as Agile, demand that analysis, modeling and design activities occur in short, time-boxed increments,

along with all of the other development activities, none of which is necessarily completed in the initial increments. In the context of such development methodologies, an early, working version of the database must be delivered and must be capable of continuing to work, despite refactorings in subsequent increments.

This creates twin challenges for the conceptual data modeling community, since not only is it necessary to take into account the consequences of incompleteness in the input to conceptual modeling, it is also necessary to view the output from conceptual modeling as an incomplete model, followed by a sequence of incremental refinements which are progressively more complete. This latter consideration is a challenge because it is in the nature of incremental development methodologies that once the database is in place, it will be used by the development team. The project's progress will be greatly hampered if the applications which access the database cease to work in the following increment as a result of a change to the database design. Consequently, in the present paper we emphasize a limited notion of incompleteness which, for the most part, means that only refactorings, rather than arbitrary re-designs, are needed as incompleteness in the conceptual model is incrementally eliminated.

In a typical medium or large organization, the information needed by the conceptual modeler is provided by various forms of analysis. Here we mention just three, which collectively cover all of the scenarios and examples in Sections 4 and 5; see Table 1.

Table 1. Examples of Information-Gathering Delays

Information gatherer	Source of Information	Typical causes of delay
Requirements Analyst	Stakeholders	Ambiguity in requirements and disagreements about scope
Business Analyst	Subject Matter Experts (SMEs)	Identifying and interviewing the appropriate SME
Data Analyst	Information Systems Specialists	Resolving issues about data definitions, availability, refresh rates, etc.

In practice, the information required is rarely written down in a form where it is immediately useable by the project. It is very often the case that the subject matter expert must be consulted in person. Typically, the experts are busy and elusive, so it can take days before they can be interviewed. Thus the delay between the point in time at which the conceptual modeler realizes that an item of information is needed and the point in time at which the analyst is able to extract the relevant information from the appropriate expert could straddle successive time-boxed increments of the development methodology. Hence the

modeler may have to proceed with producing the conceptual data model, despite being aware that the information on which the diagram will be based is incomplete. We define a notion of incompleteness which can be incorporated into an enhanced entity-relationship (“EER”) diagram and which has the property that the diagram can be used as the basis for a viable database design. This approach to conceptual modeling goes some way to meeting the aforementioned twin challenges of incremental development.

The content of the rest of this paper is as follows. Section 2 discusses related research. Our notion of incompleteness emphasizes particular aspects of the EER model, which we discuss in Section 3. In Sections 4 and 5, we present the categories of incompleteness which we have identified in the context of EER modeling, showing how each such category can be represented diagrammatically and incorporated into the database design. Our concluding remarks are in Section 6.

2 Related Research

Our perspective on requirements and modeling is closely related to the work of Salay *et al.* [7], except that we have a more restricted notion of uncertainty. Salay *et al.*'s modeling framework simply requires that there exists at least one way of resolving all of the uncertainties that results in a consistent model. However, our more restricted notion of uncertainty means that we are able to proceed with the design and implementation without resolving the uncertainty. The purpose of Salay *et al.*'s approach is to provide the modeler with a framework for reasoning about uncertainty so that contradictions and refinements can be identified.

Thalheim and Wang [10] describe a technique for data migration which has some resemblance to our work, though for an entirely different purpose. They specify how to manage schema refinement in order to facilitate data migration. This can involve a transition from a more generic to a more specific schema.

“Fuzzy” conceptual modeling, a field recently reviewed by Ma and Yan [6], models applications in which the data, and possibly also the entities and the relationships, are subject to inconsistency, imprecision, vagueness, uncertainty or ambiguity. Our methodology incorporates a restricted form of “fuzziness”, but it is purely transient since it derives from the conceptual modeler's incomplete knowledge of the world-to-be-modeled and we assume that world to be ultimately know-able. The area of overlap between our work and fuzzy conceptual modeling is that we need to generate a provisional database design based on a conceptual model which incorporates forms of incompleteness and this is similar to generating a database design from a fuzzy conceptual model. Another area of commonality is that we also need a graphical representation of incompleteness. The graphical style which we have adopted is based on the style in fuzzy conceptual modeling [6].

In the last decade, iterative and incremental development methodologies [5], such as Agile, have come into widespread use in industry. Such methodologies are characterized by time-boxed design and development activities which operate according to strict deadlines. The emphasis is on the momentum of the

increments, even when this means that analysis and design activities are incomplete. The database design community has responded to such methodologies by investigating techniques whereby physical database designs can be re-factored as a consequence of changes in successive development increments [1]. As yet, however, the preliminary phase of database development, namely conceptual modeling, has not been adapted for use in incremental development methodologies. For data warehouse design Golfarelli et al. [4] have devised a technique whereby development tasks can be aligned with a project's "sprints" (iterations). The techniques which we report in the present paper can be used in conjunction with Golfarelli et al.'s approach to prioritizing work so as to deliver high-value aspects of the data warehouse earlier in the project.

Our work uses results in the field of database refactoring [1]. In Sections 4 and 5, we describe in outline how the database design can be changed when an incompleteness is resolved. The database administrator responsible for such changes would have to follow the precepts on how to effect database changes, as outlined by Ambler and Sadalage[1]. However, an important difference is that refactoring involves changing the database design without changing the semantics, whereas our results involve changing the database as a consequence of changes in our understanding of the semantics. Furthermore, the database designer knows in advance what refactorings might be required in the future and can plan accordingly.

Our work also uses the standard techniques for generating a database design from a conceptual model; see for example, Teorey et al. [8].

3 Structure and Infrastructure in the Enhanced Entity-Relationship Model

In the context of the EER model, we distinguish between two aspects of an EER diagram, which we refer to as the *structure* and the *infrastructure*. The structure comprises the shapes and lines in the ER diagram, whereas the infrastructure comprises the information concerning the various constructs which make up the structure. In terms of Fidalgo et al.'s EER metamodel [2], the structural items in an EER diagram correspond to Fidalgo et al.'s *meta-entities*, whereas the infrastructure items in an EER diagram correspond to their *meta-attributes*. For example, two entity classes and a relationship between them are part of the structure, whereas the information concerning the connectivity and cardinality of the relationship and whether participation of an entity instance in the relationship is mandatory or optional are all a part of the infrastructure. Similarly, in a generalization hierarchy, the parent entity class and the child entity classes are part of the diagram's structure, whereas information on whether the child entity classes overlap or are disjoint is a part of the diagram's infrastructure. See Table 2 for a summary.

When constructing an ER diagram, the conceptual modeler, working from the requirements, has considerable latitude in the design of the diagram's structure. Having creatively designed the diagram's structure, however, the modeler

Table 2. Structure and Infrastructure

Structural Feature	Examples of Infrastructural Features
Entity	Weakness
Relationship	Participation, Connectivity, Cardinality, Degree,
Generalization hierarchy	Disjoint or overlapping subtypes
Aggregation hierarchy	Completeness
Attribute	Purpose (e.g. descriptor, identifier, etc.)

is highly constrained by the business rules with regard to the diagram's infrastructure. The diagram's structure depends largely on the requirements and the modeler's design, whereas the diagram's infrastructure is determined largely by the business rules. One consequence of this is that a delay in discovering relevant business rules will tend to delay completion of the diagram's infrastructure, but not necessarily the diagram's structure. In the following sections, we describe how the modeling process can cope with this differential delay.

4 Categories of Infrastructure Incompleteness

As we shall see in Section 4, for most of the infrastructure elements there is a small and discrete list of values that the element can take. For example, the relationship connectivity infrastructure element can take on the values "many-to-many", "one-to-many" and "one-to-one". Wherever the element is used in an EER diagram, it will take one of those values. Our approach to conceptual modeling leverages a particular characteristic of these lists of values, namely that one of the values is more generic than the others, in the sense that if you use the generic value in your diagram, the resulting database will be able to manage the data even if the actual value of the infrastructure element in the world-to-be-modeled turns out to be one of the others in the list. Continuing with our example, if we do not know the connectivity for a relationship, we could use the many-to-many option. The resulting database can manage the data even if the relationship turns out to be one-to-many or one-to-one. Although that database will be viable, it will have several undesirable characteristics: it will be unnecessarily complex because it will use a bridge table, query and other transaction processing will be unnecessarily inefficient and integrity maintenance will be harder.

For each category of incompleteness, we provide a general scenario and a specific example in order to explain how the specific kind of incompleteness can arise in practice, how it can be incorporated into the EER diagram convention, how the database designer can use the resulting diagram and, where appropriate, how the database can be refactored if it turns out that one of the less generic alternatives can be used. The scenarios and examples which we provide have

been deliberately chosen to reflect the wide range of circumstances which can lead to uncertainty and incompleteness in the conceptual modeling process.

For the database refactoring operations, wherever possible we refer to the catalog of refactoring operations published by Ambler and Sadalage [1].

In the following sections, we indicate that a model element is potentially redundant (depending on which business rules are eventually found to hold) by depicting the element with dashed lines [6]. As business rules are discovered and incompleteness in information is diminished, those modeling elements will either be changed to regular modeling elements with solid lines or, in the case that they are found to be redundant, eliminated from the diagram. Although the resulting ER diagram can be more complicated than a diagram which ignores incompleteness, our intention is still to use the diagram as the common medium which can be understood by all concerned parties (stakeholders, analysts, subject-matter experts, conceptual modelers and database designers).

4.1 One Entity Class or Two Entity Classes?

Scenario. It is not possible to determine whether a part of the world-to-be modeled should be represented as (a) one entity class or (b) two entity classes linked by a relationship.

Example 1. The world-to-be-modeled includes employees, but the conceptual modeler needs to know more about the business rules. Can a person have more than one employment with the company? If so, would such a person have more than one employee ID, more than one manager, etc.? If a person with multiple employments is to be treated as several different employees then the conceptual model might include a separate entity class for Person with a one-to-many relationship to an Employee entity class.



Fig. 1. One Entity Class or Two Entity Classes?



Fig. 2. Identifier or part of a Composite Identifier?

In Figure 1, the attribute shape is surrounded by a dashed-line entity class shape in order to indicate that this entity class is potentially redundant, depending on which business rules hold.

Database Design Method. The database design implements a separate table corresponding to the potentially-redundant entity class, along with the relationship. Clearly, this design will be correct (though inefficient) even if the second entity class is redundant.

Refactoring Technique. If the table and the corresponding relationship are eventually discovered to be redundant, the following refactoring technique can be used to eliminate the redundant structures from the design.

Merge Tables [1].

4.2 Attribute Status: Identifier or Part of a Composite Identifier?

Scenario. A particular attribute is known to be an identifier (rather than a descriptor), but it may not be possible to determine whether it is actually a part of a composite identifier.

Example 2. The requirements analyst considers using a person's social security number as the identifier for the Employee entity class. The project scope, which has not yet been fully determined, must include details about local employees, but might also include details about foreign employees. A social security number might be an identifier for an employee, but there are complications if the scope is extended to include foreign employees. Different countries have their own social security number systems, a foreign employee might, by coincidence, have the same number as a local employee. It might be necessary to combine the social security number with some other attribute, e.g. nationality, as a composite identifier.

In Figure 2, the attribute name is underlined with a dashed line to indicate that the attribute might not be an identifier by itself, though it is believed to be a part of a composite identifier.

Database Design Method. Use a surrogate key as the primary key. Use all of the attributes which might collectively form the composite identifier as an alternate key.

Refactoring Technique. Drop Column [1].

4.3 Attribute Value Multiplicity

Scenario. It may not be possible to determine whether a particular attribute should be modeled as a) Single-valued or b) Multi-valued

Example 3. The business stakeholders would like details of all degrees held by employees. However, it seems that the only information available might be the highest degree held by each employee. If it is possible to obtain all degrees then the attribute is multi-valued. If, however, it is only possible to obtain information on the highest degree then the attribute is single-valued.

Database Design Method. The potentially multi-valued attribute is modeled as a separate entity class and a relationship established from the parent class.



Fig. 3. Attribute value multiplicity

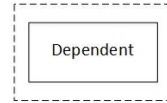


Fig. 4. Weakness of an entity class

Refactoring Techniques. Add Column, Drop Foreign Key Constraint, Drop Table [1].

4.4 Weakness of an Entity Class

Scenario. It may not be possible to determine whether or not an entity class is weak. I.e. does an instance of the entity class in question derive its identity from the identifying attributes of its parent attributes, or does it have an internal identifier that uniquely determines the existence of the entity instance?

Example 4. The Personnel entity class comprises entity instances for employees and the Dependent entity class comprises entity instances for dependents of employees. When a person leaves the company, that person's Personnel entity instance ceases to belong to the Personnel entity class and any corresponding instances in the Dependent entity class are also removed. However, the requirements analyst has been told that there may exist a business rule that when an employee dies in the service of the company, the dependents continue to be supported by the company. There are no current examples of this in the company. The requirements analyst is unable to obtain a definitive statement that there really is such a business rule.

Database Design Method. The table representing the potentially-weak entity class is given a surrogate key.

Refactoring Technique. Introduce Cascading Delete [1]. When the incompleteness is resolved, if this entity class is not weak then the identifier can become an alternate key.

4.5 Degree of a Relationship

Scenario. Although it may appear that there is a ternary relationship between entity classes E_1 , E_2 and E_3 , it be possible to model the relationship as a set of binary relationships. Whether or not this is feasible can be by identifying set of business rules has been discovered which relate E_1 , E_2 and E_3 . The business analyst must establish which of the following functional dependencies correspond to the business rules: $E_1 \rightarrow E_2$, $E_1 \rightarrow E_3$, $E_2 \rightarrow E_1$, $E_2 \rightarrow E_3$, $E_3 \rightarrow E_1$, $E_3 \rightarrow E_2$, $\{E_1, E_2\} \rightarrow E_3$, $\{E_1, E_3\} \rightarrow E_2$, $\{E_2, E_3\} \rightarrow E_1$.

Example 5. Suppose that an application includes the entity classes Engineer, Project and Laptop. The database must model the fact that an engineer uses a laptop on a project. But are there any relevant business rules? For example: Is it permissible that an engineer use more than one laptop? If so, can the engineer use more than one laptop on the same project? Is it permissible for two engineers to use the same laptop? Is it permissible to use a laptop on more than one project? For example $Engineer \rightarrow Laptop$ means that any given engineer cannot use more than one laptop and $\{Project, Engineer\} \rightarrow Laptop$ means that a given engineer on a given project cannot use more than one laptop.

If a version of the conceptual model must be produced before the business analyst is able to ascertain which subset of the above functional dependencies actually corresponds to the business rules, then the ternary relationship is marked as a incompleteness in the EER diagram.

Database Design Method. Use standard methods to implement the ternary relationship; see [8]. Use a view for querying and stored procedures for updates in order to hide a possible subsequent refactoring which would transform the ternary relationship into a sequence of binary relationships.

Refactoring Technique. Add ForeignKey Constraint, Drop Table [1].

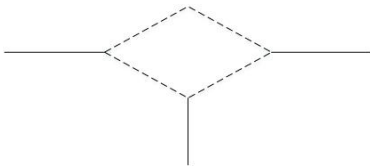


Fig. 5. Degree of a Relationship



Fig. 6. Connectivity of a Relationship

4.6 Relationship Connectivity

Scenario. Connectivity is a constraint on the connection of a given entity instance in the relationship. If the business rule concerning whether participation is optional or obligatory has not yet been ascertained then this element is marked as incomplete in the EER diagram; see Figure 6.

Example 6. In our running example, if there is no business rule that for each department there must exist an employee who manages the department then connectivity is optional.

Database Design Method. Use a bridge table to model the potential many-to-many relationship.

Refactoring Technique. Add ForeignKey Constraint, Drop Table [1].

4.7 Optionality of the Occurrence of an Entity Instance in a Relationship

Scenario. It may not be possible to determine whether the occurrence of an entity instance in a relationship is (a) mandatory or (b) optional. If the involvement of an entity class in the relationship is mandatory then each entity instance in the entity class must be involved in at least one occurrence of the relationship. If, however, there need not be such an occurrence, then it is considered optional then there can exist an entity instance which is not involved in the relationship.

Example 7. Using our earlier example, there may or may not be a business rule that a Department must have a Manager.



Fig. 7. Optionality of the Occurrence of an Entity Instance in a Relationship

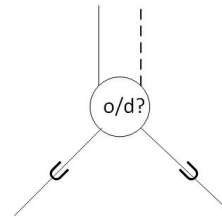


Fig. 8. Generalization: Subtypes Overlapping or Disjoint

Database Design Method. The occurrence of an entity instance in this relationship is modeled as if it were optional.

Refactoring Technique. Make Column Non-Nullable, Add Foreign Key Constraint [1].

4.8 Generalization: Subtypes Overlapping or Disjoint

Scenario. In a generalization hierarchy, it is not known whether the subtype entity classes overlap or are disjoint.

Example 8. In a vehicle generalization hierarchy, the subtype entity classes include different kinds of vehicle, e.g. trucks and boats. Analysis has not yet determined whether an amphibious vehicle is to be classified as a truck or a boat or both or neither (i.e. a separate entity class). Thus it is not known whether the subtypes overlap or are disjoint.

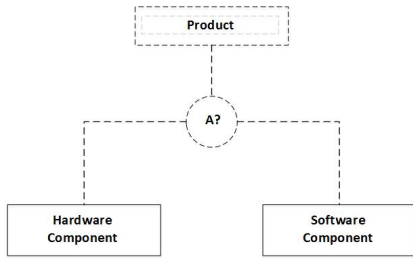


Fig. 9. Completeness or Otherwise of an Aggregation Hierarchy

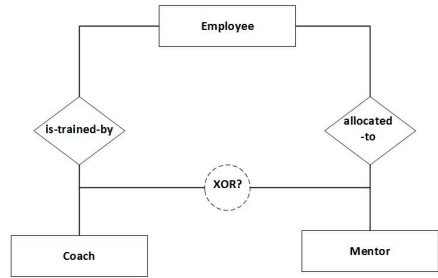


Fig. 10. Exclusion Constraint: Exclusive OR or Inclusive OR?

Database Design Method. Create the additional entity class and treat the generalization subtypes as if they overlap.

4.9 Redundancy of Aggregation Hierarchy

Scenario. A potential change to the business affects whether it is appropriate to model some entity classes as being a part of an aggregation hierarchy.

Example 9. In the past, the company has sold a range of products, each of which comprises a hardware component and a software component. However, a pending court case may force the company to un-bundle these two items so that, in the future, the hardware component and the software component may have to be sold separately as different products. See Figure 9.

Database Design Method. The aggregation hierarchy is implemented in the usual way. Views and stored procedures are required as the external interface in order to minimize change if the aggregation hierarchy is removed.

4.10 Multiple Relationships: Inclusive or Exclusive OR

Scenario. In the model, one entity class, E_1 is related, via a pair of relationships, R_1 and R_2 , to a pair of entity classes, E_2 and E_3 such that each entity instance of E_1 is related to an entity instance in either E_2 or E_3 . The conceptual modeler does not know whether a business rule exists which prohibits a given entity instance in E_1 from being related via both R_1 and R_2 to entity instances in E_2 and E_3 , respectively.

Example 10. Each employee either gets trained by a coach or is allocated to a mentor. However, there may be a business rule defining circumstances where an employee has both a coach and a mentor. See Figure 10.

Database Design Method. The database design allows for the possibility that an entity instance can participate in both of the relationships.

Refactoring Techniques. If a business rule prohibiting this is discovered then a check constraint can be used to prohibit the presence of both foreign key references in a given Employee record.

5 Structural Incompleteness

In this section, we present two modeling scenarios in which incompleteness at the structural level can be incorporated into the conceptual modeling process in much the same way as incompleteness at the infrastructural level.

5.1 Relationship Redundancy

Scenario. It may be the case that an entire relationship might be redundant in the sense that it can be removed from the model as a result of a transitive anti-closure.

Example 11. Suppose that analysis has identified a many-to-one relationship `Manages_Emp` from the `Employee` entity class to itself and the relationships "Manages" and "Works In" between the `Employee` and `Department` entity classes. Then the `Manages_Emp` relationship will be redundant so long as (i) the other relationships are many-to-one and (ii) the direct relationship from `r` has the same meaning as the relationship from `Employee` to `Department` to `Manager`. If analysis is unable to establish that both of these conditions are true, then the potential redundancy of the `Employee-Manager` relationship is marked on the diagram; see Fig. 11.

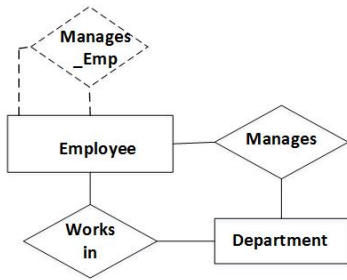


Fig. 11. Relationship Redundancy

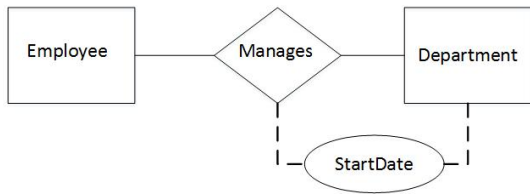


Fig. 12. Attribute of a Relationship or an Entity Class

Database Design Method. Include the foreign key from `Employee` to `Manager` in the table design, but create a view definition for the association. All reports, stored procedures, etc. must use the view definition. If, eventually, the relationship is removed from the model then only the foreign key definition and the view definition need to be changed.

Refactoring Techniques. Drop Foreign Key Constraint and Drop Column [1].

5.2 Assign an Attribute to a Relationship or to an Entity Class

Scenario. In some situations, it is necessary to assign an attribute to a relationship, rather than to an entity class. In particular, many-to-many relationships or ternary relationships typically require one or more attributes, such as the start date of each instance of the relationship. This could be incomplete if knowledge of the connectivity of a relationship is incomplete. This is because the attribute is only needed in the relationship if it turns out that the relationship is many-to-many. We note that this is a further example of where the resolution of one instance of incompleteness is dependent on the resolution of another instance of incompleteness.

Example 12. Suppose that the modeler is already aware of incompleteness of information about the degree or connectivity of a relationship. If the Manages relationship between Employee and Department turns out to be one-to-one or many-to-one then the Start-date attribute, which indicates when the Employee started to be the manager of the Department, is most conveniently a part of the Employee entity class. If, on the other hand, the relationship is one-to-many or many-to-many then the attribute should be added to the Manages relationship.

Database Design Method. Add the attribute to the relationship.

Refactoring Technique. Introduce New Column and Drop Column [1].

6 Concluding Remarks

In practice, conceptual data modeling is a process driven by a sequence of interactions between the conceptual modeler and those who provide information to, or consume information from, the conceptual modeling process. Unfortunately, expositions on conceptual data modeling tacitly assume a “waterfall” development methodology: a stage only starts once the previous stage completes and the conceptual modeller has all the information required to complete the modeling exercise from the beginning. We have gone some way towards reconciling contemporary incremental development methodologies with traditional conceptual data modeling by emphasizing the latter’s collaborative nature. The technique which we have proposed is complementary to other techniques with a similar goal [4].

In iterative development methodologies, a time-boxed increment might have a duration of anywhere in the range of two days up to two weeks or more. Our approach demonstrates that conceptual data modeling can be more closely aligned with incremental methodologies, such as Agile, since it allows the analysis team to pass on an incomplete requirements document to the conceptual modeler, who can represent that incompleteness in the conceptual model diagram. This diagram can be scrutinized by the stakeholders and subject-matter experts in

the usual way. Furthermore, it can be used by the database designer to produce a viable database earlier in the project life cycle.

We have also distinguished between the *requirements* which partly determine the structure of the conceptual model and the *business rules* which largely determine what we call the “infrastructure” the model.

Our techniques for handling incompleteness are not restricted to use with incremental development methodologies. They are equally applicable wherever a model must be produced despite the incompleteness of the preliminary analysis. It should also be apparent that our techniques are applicable to conceptual modeling languages other than those based on the EER model.

In particular, when the UML Class Diagram is used for conceptual data modeling [8], a practically identical collection of infrastructural features can be identified and can be associated with similar incompleteness categories.

References

1. Ambler, S., Sadalage, P.J.: *Refactoring Databases: Evolutionary Database Design*. Addison Wesley (2006)
2. Do Nascimento Fidalgo, R., De Souza, E.M., España, S., De Castro, J.B., Pastor, O.: EERMM: A Metamodel for the Enhanced Entity-Relationship Model. In: Atzeni, P., Cheung, D., Ram, S. (eds.) *ER 2012 Main Conference 2012*. LNCS, vol. 7532, pp. 515–524. Springer, Heidelberg (2012)
3. Galindo, J., Urrutia, A., Carrasco, R.A., Piattini, M.: Relaxing constraints in enhanced entity-relationship models using fuzzy quantifiers. *IEEE T. Fuzzy Systems* 12, 780–796 (2004)
4. Golfarelli, M., Rizzi, S., Turricchia, E.: Sprint planning optimization in agile data warehouse design. In: Cuzzocrea, A., Dayal, U. (eds.) *DaWaK 2012*. LNCS, vol. 7448, pp. 30–41. Springer, Heidelberg (2012)
5. Larman, C., Basili, V.R.: *Iterative and Incremental Development A Brief History*. *Computer* 36, 47–56 (2003)
6. Ma, Z.M., Yan, L.: A Literature overview of fuzzy conceptual data modeling. *Journal of Information Science And Engineering* 26, 427–441 (2010)
7. Salay, R., Chechik, M., Horkoff, J.: Managing Requirements Uncertainty with Partial Models. In: *Proc. of Requirements Engineering*, pp. 1–10 (2012)
8. Teorey, T., Lightstone, S., Nadeau, T.: *Database Modeling and Design: Logical Design*, 4th edn. Morgan Kaufmann, San Francisco (2006)
9. Thalheim, B.: The science and art of conceptual modelling. In: Hameurlain, A., Küng, J., Wagner, R., Liddle, S.W., Schewe, K.-D., Zhou, X. (eds.) *Transactions on Large-Scale Data- and Knowledge-Centered Systems VI*. LNCS, vol. 7600, pp. 76–105. Springer, Heidelberg (2012)
10. Thalheim, B., Wang, Q.: Towards a theory of refinement for data migration. In: Jeusfeld, M., Delcambre, L., Ling, T.-W. (eds.) *ER 2011*. LNCS, vol. 6998, pp. 318–331. Springer, Heidelberg (2011)

Semi-supervised Learning of Action Ontology from Domain-Specific Corpora

Irena Markievicz¹, Daiva Vitkute-Adzgauskiene¹, and Miniija Tamosiunaite²

¹ Faculty of Informatics, Vytautas Magnus University, Kaunas, Lithuania
{i.markievicz,d.vitkute}@if.vdu.lt

² Bernstein Center for Computational Neuroscience, University of Gottingen
m.tamosiunaite@if.vdu.lt

Abstract. The paper presents research results, showing how unsupervised and supervised ontology learning methods can be combined in an action ontology building approach. A framework for action ontology building from domain-specific corpus texts is suggested, using different natural language processing techniques, such as collocation extraction, frequency lists, word space model, etc. The suggested framework employs additional knowledge sources of WordNet and VerbNet with structured linguistic and semantic information. Results from experiments with crawled chemical laboratory corpus texts are given.

Keywords: action ontology, semi-supervised ontology learning, natural language processing, corpus linguistics, domain-specific corpus.

1 Introduction

Design and use of intelligent, knowledge-based systems requires an adequate domain model which is normally designed in the form of an ontology presenting main concepts and their associations necessary for reasoning purposes. For example, such an ontology applied in robotics activity scenarios allows to define the knowledge field of a robot aimed at carrying out tasks in a specific-domain, e.g. in a kitchen specific, or chemistry lab specific domains. Task-oriented ontologies are usually designed as action ontologies, with action verbs being their main concepts and, also, different elements describing the action environment (e.g. action objects, tools, location, time, etc.).

This paper deals, specifically, with construction issues of action ontologies, concentrating on automated ontology building methods, i.e. on so-called ontology learning methods.

The main classifying points for ontology learning approaches are: a) a priori knowledge at the input (texts, preprocessed texts, dictionaries, other ontologies, etc.); b) learning methods (statistic vs. logical, etc.) [1]. Based on the scope of a priori knowledge, unsupervised and supervised ontology learning methods are defined. Unsupervised ontology learning is based on concept and association extraction from domain-specific texts, often containing some basic linguistic annotations (e.g. morphological annotations, dependency parses). Supervised ontology learning

assumes the use of supplementary labeled information (e.g. specifically annotated training corpora), structured semantic information (e.g. taxonomies, ontologies) and regular-expression based lexical patterns used for concept and association extraction.

This paper presents a semi-supervised method for action ontology building, using both unsupervised information extraction from domain-specific corpora, and, also, input from other ontologies or external databases with structured semantic information as well as corresponding lexical patterns for information extraction.

Experimental investigation is based on building of an action ontology for a robotics scenario, using a domain-specific corpus with crawled online material on chemistry laboratory processes. The corpus texts describe chemistry laboratory experiments, basic rules, instruments and techniques. The overall size of the experimental chemistry lab corpus (further referred to as the CHEMLAB corpus) is 1,971,415 running words. Collected texts were morphological annotated and lemmatized using Stanford University NLP tools for English language (<http://nlp.stanford.edu/software/>).

2 Related Works

Related works can be grouped into those dealing with action ontology construction, and those dealing with automation of general domain ontology building processes.

Research works on action ontologies are in most cases oriented towards the development of domain-specific ontology models (knowledge structure) and reasoning mechanisms. Research domains are usually related either to natural language interfaces to agent systems [2,3], or structures for organizing work in robot-based systems [4,5]. However, little or no attention in these cases is paid to the automation of ontology creation process, with manual procedures prevailing, e.g. using ethnographic methods, study of human behavior and work practice [4]. Individual attempts of automated design of knowledge bases for understanding user situations and actions are usually rather limited to a priori knowledge structure, e.g. using semi-structured instruction texts [6].

References on automation of general domain ontology building process cover different design methods are much more, mainly based on transformations and merging of other existing ontologies, on domain text mining and use of external knowledge resources. [7] and [8] give a good summary of available automatic and semi-automatic ontology extraction techniques. Approaches using external knowledge resources, mainly WordNet [9] and those making use of different Natural Language Processing (NLP) methods [10] are prevailing. Semi-supervised methods, combining concept mining in domain texts and relationship extraction from WordNet are also presented in some works [11].

Our difference is in offering a semi-supervised ontology building method, specifically tailored for a domain-based action ontology design. It is based on text mining and NLP methods, combined with automated information extraction from several external knowledge bases – WordNet and VerbNet.

3 Action Ontology Learning Model

General methodology for ontology building from texts can be described using the following meta-model [1]:

$$M = \{D, LA, T, S, C, TR\}, \quad (1)$$

where D is document collection (text corpus), LA are linguistic annotations for corpus texts, T is terminology collection, S is synonym collection, C is ontology concept collection, TR are ontology relations (associations).

Domain corpus texts, possibly with linguistic annotations, are used as the input to different NLP tools, resulting in terminology collection, further grouped into synonym collection (synsets). These are further used as building blocks for ontology concept collection, and the latter is finally enriched with corresponding associations between concepts.

There are different ontology development methodologies available. However, for building ontology from scratch using domain-specific corpus texts and integrating other knowledge sources, ontology engineering methodology named *Methontology* [12] is the most appropriate, as it suggests a framework for cyclical, multi-step ontology building, i.e. “ontology growing” based on the use of evolving ontology prototypes. For each prototype, *Methontology* suggests to start from planning, i.e. determining the time and resources necessary for each ontology building task. Ontology building starts from ontology specification, giving the domain, purpose, scope, knowledge source information. Further ontology development tasks include conceptualization (building a conceptual model), formalization (specifying techniques and tools) and implementation. Ontology development is accompanied by parallel activities of knowledge acquisition – extraction, integration, evaluation. Integration with other ontologies or knowledge databases should be described before implementation starts. Also, the evaluation of outcomes is foreseen by planning of control and quality assurance processes.

Further, the application of this model to the automated design of an action ontology for a robotics scenario is presented.

Each robot has limited number of actions, which it can execute. The action ontology should be based on those actions and it should add related actions and action environment information in the process of ontology *growing*. Knowledge sources, that are relevant in this case, consist of a domain-specific text corpus (chemistry laboratory domain is considered) and other related ontologies and other sources with structured semantic information. Linguistic database of English language WordNet (<http://wordnet.princeton.edu/>) and domain-independent verb lexicon for English language VerbNet (<http://verbs.colorado.edu/verb-index/>) were selected as the most appropriate external knowledge sources for action ontology building.

A conceptual model of the action ontology for a robotics scenario is given in Fig.1.

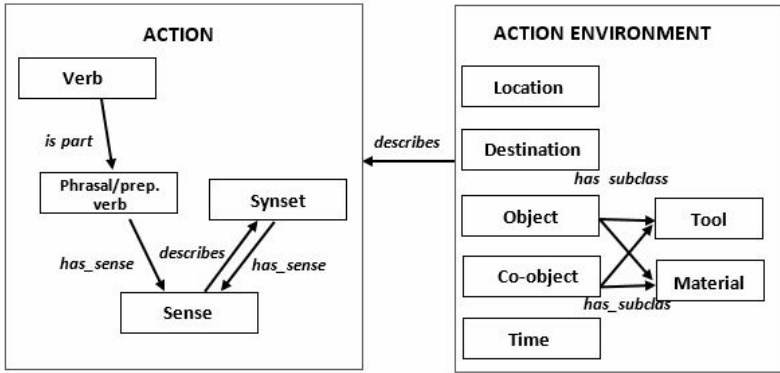


Fig. 1. Conceptual model of an action ontology

The presented action ontology conceptual model assigns appropriate action *synset* for each action and, also, all action details required for action execution (action environment). Action *synset* contains verbs, prepositional verbs and phrasal verbs, having the same sense. Environment description includes all the necessary elements for robot activity: time, location, destination, involved tools, involved material, etc.

Fig.2 presents the general process-structure of the semi-supervised action ontology building approach. Action verbs, extracted from morphological annotated corpus, are grouped into action synsets. In this process, external lexical data sources of WordNet and VerbNet databases are involved. These databases are also employed in action environment building. The elements in action environment synsets are grouped by the semantic roles, indicated by VerbNet frames and WordNet relations. Relations and axioms between ontology elements include semantic event chains and manually-built rules.

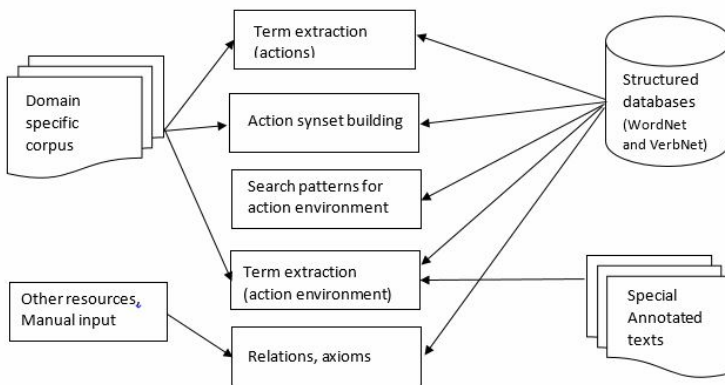


Fig. 2. General process-structure of the semi-supervised action ontology building



The following sections give a step-by-step presentation of the action ontology building process, illustrated by experimental examples from the chemistry laboratory domain.

4 Term Extraction (Actions)

Extraction of terms denoting actions is the first step in the action ontology implementation process. With domain-specific corpus available, the most reasonable way to start building of the glossary of the most common actions is by building a verb-frequency list and filtering out the most frequent actions. In order to have a complete action representation, term-specific linguistic patterns, which include verbs, prepositional verbs (verb + preposition) and phrasal verbs (verb + [direct object] + adverb) and other multiword verbs (verb + direct object, verb + modifier) are used. Results of an experiment with the chemistry laboratory corpus by applying the above mentioned patterns on a morphologically annotated corpus is presented in Table 1.

Table 1. Actions by frequency - examples on MIX and PUT action groups

PUT	Frequency	MIX	Frequency
put in	353	mix with	342
put on	149	mix of	189
put into	111	mix to	178
put of	94	mix together	99
put away	43	mix for	53
put back	32	mix up	45
put to	29	mix into	30
put together	18	mix at	24
put not (n't)	17	mix as	23
put off	16	mix until	21
put out	15	mix not (n't)	17
put at	12	mix under	16
put down	11	mix by	13

Larger frequency values point to the importance of an action verbs. When planning the ontology building process, these actions should be taken care first of all. Also, the experiment results point to the need of sorting out action verbs into synonymic groups, synsets, as actions linked to the same main verb can have entirely different meanings (e.g. “*put on*” and “*put off*”).

5 Building Synsets of Similar Actions

A verb usually has more than one sense and its' sense can change in collocation with other words, e.g. a direct object name, a preposition or a certain modifier (e.g. *don't*).

Data from Table 1 contains examples of verbs with similar meaning, which can be marked as synonyms: *put in* = *put into*, *put out* = *put away*. It also contains verbs with opposite meaning: *put in* ≠ *put off*, *put in* ≠ *put out*, *put* ≠ *put not* (*n't*), *mix* ≠ *mix not* (*n't*).

Grouping actions into the synsets with the same sense is the next step of action ontology learning. This process involves external domain-independent lexical databases – *WordNet* and *VerbNet*. *WordNet* contains English language nouns, verbs, adjectives and adverbs. It describes the following relations between words: for nouns – hypernyms, hyponyms, holonyms and meronyms, for verbs – hypernyms, troponyms, for adjectives – relativeness, similarity, participation, for adverbs – common adjectival core. *VerbNet* groups English language verbs into conceptual classes. Each verb is described by roles and restrictions, its semantic group, frames with common examples and syntactic structure.

Synset is a synonym ring, which groups semantically equivalent data elements. Fig.3 presents an excerpt from synsets of verb “*remove*”, as given by WordNet. Not all of them are adequate to the domain-specific action ontology – for example, *Sense-7* verbs, describing murder, are not adequate to the CHEMLAB domain.

Sense 6
absent, remove -- (go away or leave; "He absented himself")
=> disappear, vanish, go away -- (get lost, as without warning or explanation; "He disappeared without a trace")

Sense 7
murder, slay, hit, dispatch, bump off, off, polish off, remove -- (kill intentionally and with premeditation; "The mafia boss ordered his enemies murdered")
=> kill -- (cause to die; put to death, usually intentionally or knowingly; "This man killed several people when he tried to rob a bank"; "The farmer killed a pig for the holidays")

Fig. 3. Synsets for verb “*remove*” (Source: WordNet)

Similar situation can be observed in *VerbNet* – verb “*remove*” is assigned to a semantic group, containing not just common synonym verbs (*extract*, *delete*, *dismiss*, *separate*, *etc.*), but also more specialized ones (*excommunicate*, *ostracize*) with their meaning dependent on the domain context.

Therefore, the task is to filter out inadequate verb senses and to grow synsets by adding suitable verbs with the same sense, coming from different sources. Word Space Model (WSM), which is based on the hypothesis that words with similar meanings will occur with similar neighbors, if enough text material is available [13], is used for testing *semantic similarity* of verbs. WSM is implemented by calculating feature vectors (frequency of co-occurrence with other words) for each word and measuring the distance between corresponding vectors. The feature vector of a certain verb is calculated, taking every occurrence of this verb in corpus texts, identifying

meaningful words in the sentence-wise neighborhood of each occurrence, and building a vector with calculated measures of association between the verb and each of its neighborhood words. *Pointwise mutual information (PMI)* coefficient, describing relationship between the probability of the co-occurrence of two words and their individual distributions, is normally used as a probabilistic association measure in building such feature vectors:

$$PMI(A, B) = \log \frac{p(A, B)}{p(A)p(B)} = \log \frac{p(A|B)}{p(A)} = \log \frac{p(B|A)}{p(B)}, \quad (2)$$

where $p(A, B)$ is the probability of A and B occurring together in the same context and $p(A)$, $p(B)$ – probabilities of their individual occurrence.

Feature vectors are then compared between each other using the cosine similarity method:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}, \quad (3)$$

where A and B are feature vectors of verbs that are being compared.

Cosine similarity ranges from -1 to 1, where -1 means exactly opposite sense, 0 means independence, and 1 shows strong synonyms.

Table 2 presents an excerpt of feature vectors for verbs “wash” and “rinse” built using the CHEMLAB corpus as a reference.

Table 2. Excerpt of feature vectors for „wash“ and „rinse“

WORD	PMI (wash)	PMI (rinse)	WORD	PMI (wash)	PMI (rinse)
acetone	6,11	7,329	NaOH	4,99	5,756
Acid	6,15	4,108	Precipitant	7,32	6,071
careful	7,204	7,374	Product	4,19	4,276
Dilute	6,55	5,636	Residue	5,81	7,182
discard	11,749	8,58	Sodium	5,45	5,705
distilled	6,981	6,213	Solvent	4,96	3,289
Fume	8,387	8,156	Buret	0,00	8,077
funnel	5,74	4,66	Cake	9,32	0,00
addition	0,00	4,053	Color	0,00	4,386

After applying the cosine similarity method we get:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = 0,772615 \quad (4)$$

As the obtained cosine similarity value is close to 1, we can state, that verbs “wash” and “rinse” are similar and can be included in the same synset.

By applying WSM consequently to verbs in WordNet (WN) sense descriptions and VerbNet (VN) class descriptions, we observe an ontology learning process, named as *synset growing*. Fig.4 illustrates the synset growing process for the verb “*add*”.

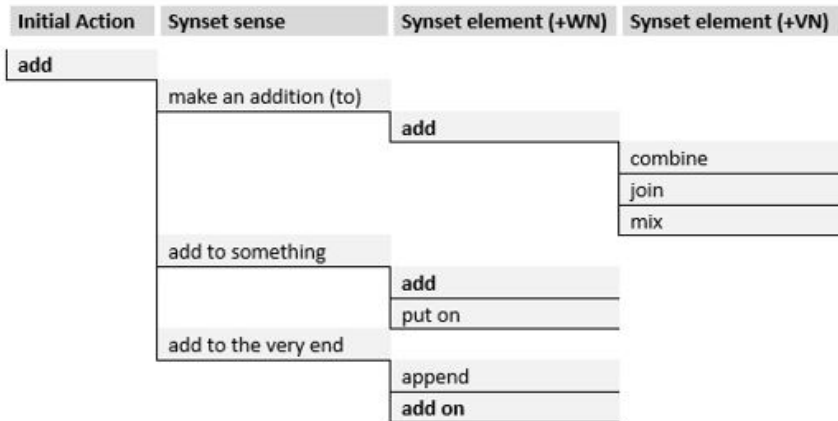


Fig. 4. Synset growing process – example for “*add*”

In this case, inadequate senses (e.g. “*add*” and “*add up*” in the meaning of “*summing up*”) have been filtered as inadequate to the CHEMLAB domain.

6 Action Environment Learning

Each action synset should be described by a certain action environment, containing time, location, duration, destination, actor, tool, material, etc. elements. This process can be organized in 3 steps: 1) text preprocessing and building the glossary of possible environment elements; 2) obtaining rules (search patterns) for action environment element classification; 3) classifying the action environment elements by their roles.

Text preprocessing, leading to building of a glossary of possible action environment elements, involves collocation extraction methods. Collocation is a sequence of words that co-occur more often than it would be by chance (e.g. room temperature).

There are different statistical methods for extracting collocations from text, such Mutual Information, chi-squared test, Log-likelihood ratio, Fisher exact test, Dice coefficient, gravity counts [14], etc. Experiments showed, that for the purpose of identifying action environment elements, log Dice coefficient is adequate [14]:

$$\log Dice(A, B) = 14 + \log \frac{2|A \cap B|}{|A| + |B|}, \quad (5)$$

where $|A \cap B|$ is the frequency of A and B words co-occurrence in text, $|A|$, $|B|$ - frequency of A and B words occurring separately.

Table 3 presents most frequent collocations obtained from the CHEMLAB corpus. Extracted glossary of CHEMLAB environment elements contains not just domain terms (e.g. *periodic table*), but also named entities, such as chemical elements (e.g. *carbon dioxide*, etc.), measurement data (e.g. *room temperature*) and names of tools (e.g. *water bath*).

Table 3. Most frequent CHEMLAB corpus collocations

Collocation	logDice	Freq.	Collocation	logDice	Freq.
reductive amination	13,740	287	aqueous layer	11,851	290
baking soda	13,709	206	science fair project	11,79	213
science fair	13,319	459	diethyl ether	11,784	232
carbon dioxide	13,098	361	reflux condenser	11,715	188
essential oil	12,891	296	acetic acid	11,696	426
periodic table	12,838	244	hydrochloric acid	11,579	359
copper sulfate	12,798	220	organic layer	11,482	202
hydrogen peroxide	12,705	270	small amount	11,404	207
methylene chloride	12,639	487	reaction mixture	11,337	787
sodium hydroxide	12,474	661	sassafras oil	11,222	284
reduced pressure	12,371	239	Chemical Abstracts	11,085	205
room temperature	12,359	551	sodium borohydride	10,872	196
alkali metal	12,285	200	formic acid	10,783	202
ammonium chloride	12,018	309	sodium acetate	10,766	213
sulfuric acid	11,856	504	sodium chloride	10,664	219

With action environment element glossary in place, classification of environment elements according to their action-specific roles must be done. This can be done by applying certain rules or search patterns. Possible sources for such rules may be the VerbNet lexicon with structured description of the syntactic behavior of verbs [15], or, alternatively, syntactic parse trees can be used. Our approach is based on automated extraction of rules from VerbNet lexicon database, mapping VerbNet thematic roles to the elements of the action environment conceptual model. Rules are extracted from VerbNet syntactic and semantic frames for corresponding verbs (Table 4).

In the example with “wash” verb, we obtain 5 possible search patterns, which are then used in action environment classification: *NP-Agent VB NP-Object; NP V; NP V NP PP.instrument; NP V NP PP.location; NP V NP PP.duration*. These patterns are then applied to morphologically annotated CHEMLAB corpus for filling the action ontology with classified action environment elements.

Table 4. VerbNet syntactic and semantic frames for verb „wash“ (Source: VerbNet)

Description	Syntax	Semantics	Example
NP V NP	NP- Agent VB NP- Object	TAKE CARE OF: ThemeRole = (?)Agent Event = during(E) ThemeRole = Object	<i>He washed the solvent layer, dried it and concentrated.</i>
NP V	NP- Agent VB	TAKE CARE OF: ThemeRole = Agent Event = during(E) ThemeRole = (?)Object	<i>Wash the aqueous layer twice.</i>
NP V NP PP.instrument	NP- Agent VB NP- Object PREP- With NP- Instrument	TAKE CARE OF: ThemeRole = (?)Agent Event = during(E) ThemeRole = Object USE: ThemeRole = Agent Event = during(E) ThemeRole = Instrument	<i>The filter cake is washed thoroughly with methanol.</i>
NP V NP PP.location	NP- Agent VB NP- Object PREP- In NP- Location	TAKE CARE OF: ThemeRole = (?)Agent Event = during(E) ThemeRole = (?)Object USE: ThemeRole = Agent Event = during(E) ThemeRole = Location	<i>The top aqueous layer was washed in the funnel.</i>
NP V NP PP.duration	NP- Agent VB NP- Object PREP- During NP- Duration	TAKE CARE OF: ThemeRole = Agent Event = during(E) ThemeRole = (?)Object USE: ThemeRole = Agent Event = during(E) ThemeRole = Duration	<i>The successive washes during the work up.</i>

7 Experimental Results

Experimental research with CHEMLAB domain corpus resulted in developing of a prototype action ontology containing 528 named classes, 3457 axioms (including 1070 logical axioms) and 1855 annotation assertion axioms. The following main classes were used for the action environment elements: ACTIVITY, OBJECT, CO-OBJECT, DESTINATION, TOOL, LOCATION and MATERIAL.

Ontology building process is illustrated for most common action verb from CHEMLAB domain corpus: add, apply, make, mix, pour, put, remove, transfer and wash. DL Expressivity is used for action ontology evaluation [16]. Developed ontology can be described with ALU metrics – allows atomic negation of concepts, that do not appear on the left hand side of axioms, concept intersection, concept union, universal restrictions [17].

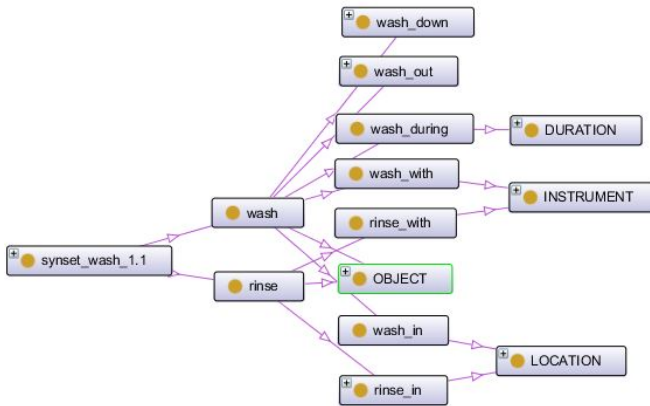


Fig. 5. “Wash” synset with its environment classes

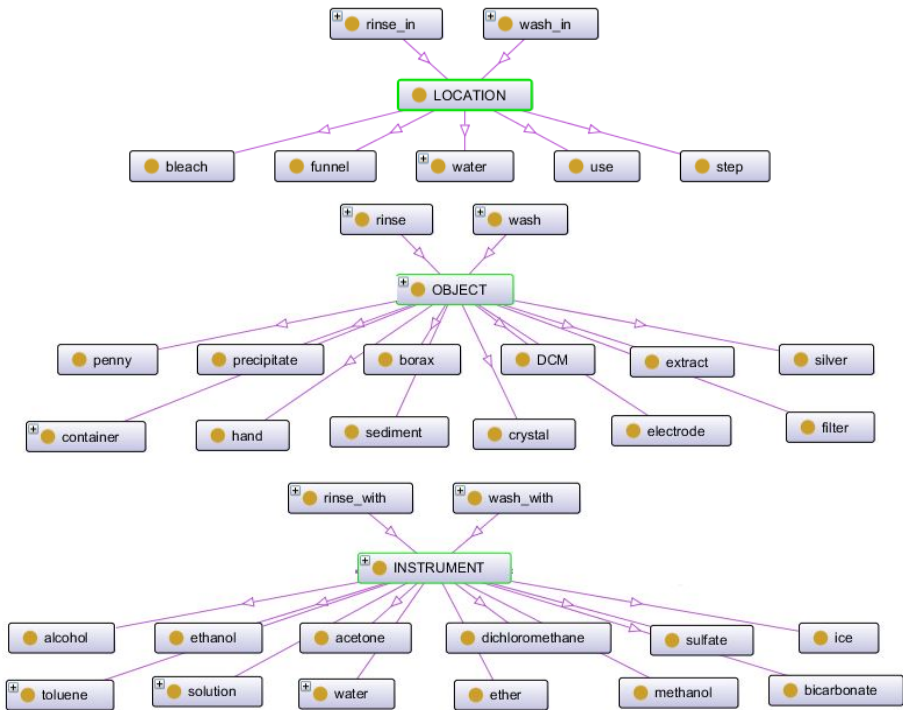


Fig. 6. Action environment examples for location, instrument and object in wash synset

Fig.5 presents the visualization of “wash” action and its environment. The “wash” synset in this case contains two synonyms: *wash* and *rinse*. Both actions can be directly connected with some objects: *filter*, *electrode*, *dish*, etc. Also, these actions are associated with other action environment elements: *duration*, *instrument* and *location*.

Some segments of action environment are presented in Fig.6. The results of the experiments show, that the same element of environment can be defined as *location*, *object* or *instrument* depending on which preposition verb is used. E.g. *water* can be interpreted as *instrument* or *location*, depending on context, as shown in Fig 6.

The elements of actions environment presented above were classified using VerbNet semantic frames. However, this method does not ensure, that all elements of action environment are classified. Different semantic roles of objects depend on action context. The results of the experiment show, that chemistry laboratory domain-corpus contains a lot of data, with multiple meanings and thus raises challenges for future work.

8 Conclusions

The proposed action ontology building approach uses employs NLP techniques: morphological analysis, POS tagging, collocation extraction, word space model for word sense disambiguation, concept classification and semantic tagging.

The results of this study show that structured information from existing knowledge bases (WordNet, VerbNet, etc.) can be of use in designing automated procedures both for ontology concept and relation learning.

A combination of unsupervised and supervised ontology learning methods is efficient for integrating different input data in action ontology building. This integration is specific to the each step of the proposed approach.

The designed prototype action ontology is still missing role hierarchy, inverse and functional properties. Adding cardinality restrictions would be helpful with chemical element measurement data.

More complex environment classification, recognition of hierarchical relations and building restrictions are the main tasks for future research work.

Acknowledgement. The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 (Programme and Theme: ICT-2011.2.1, Cognitive Systems and Robotics) under grant agreement no. 600578, ACAT.

References

1. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, pp. 10–17. Springer, Karlsruhe (2006)
2. Kemke, C.: An Action Representation Formalism for Natural Language Interfaces to Agent Systems. *International Journal of Convergence Information Technology* 2(2), 30–36 (2007)

3. Morgenstern, L., Riecken, D.: SNAP: An Action-Based Ontology for E-commerce Reasoning. In: Formal Ontologies Meet Industry, Proceedings of the 1st Workshop, FOMI 2005 (2005)
4. Wales, R.C., Shalin, V.L., Bass, D.S.: Requesting Distant Robotic Action: An Ontology of Work, Naming and Action Identification for Planning on the Mars Exploration Rover Mission. *Journal of the Association for Information Systems* 8(2), art. 6 (2007)
5. Chatterjee, R., Matsuno, F.: Robot description ontology and disaster scene description ontology: analysis of necessity and scope in rescue infrastructure context. *Advanced Robotics* 19(8), 839–859 (2005); VSP and Robotics Society of Japan
6. Jung, Y., Ryu, J., Kim, K., Myaeng, S.: Automatic construction of large-scale situation ontology by mining how-to instructions from the web. *Web Semantics: Science, Services and Agents on the World Wide Web* 8(2), 110–124 (2010)
7. Bedini, I., Nguyen, B.: Automatic Ontology Generation: State of the Art. PRISM Laboratory Technical Report. University of Versailles (2007)
8. Grigonyte, G.: Building and Evaluating Domain Ontologies: NLP Contributions. Logos Verlag Berlin GmbH (2010)
9. Moldovan, D.I., Girju, R.: Domain-specific knowledge acquisition and classification using WordNet. In: Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference, pp. 224–228. AAAI Press (2000)
10. Nobécourt, J.: A method to build formal ontologies from texts. In: Workshop on Ontologies and Text. Juan-Les-Pins, France (2000)
11. Hu, H., Lium, D.Y.: Learning OWL ontologies from free texts. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics, vol. 2, pp. 1233–1237. IEEE (2004)
12. Mariano, F., Gomez-Perez, A., Juristo, N.: Methontology: From Ontological Art Towards Ontological Engineering. In: Proceedings of AAAI-97 Spring Symposium. Series on Ontological Engineering, pp. 33–40. AAAI Press, Stanford (2004)
13. Shutze, H., Pedersen, J.: A co-occurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management* 33(3), 307–318 (1997)
14. Daudaravicius, V., Marcinkeviciene, R.: Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics* 9(2), 321–348 (2004)
15. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: Extending VerbNet with Novel Verb Classes. In: Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa (2006)
16. Corcho, O., Fernández-López, M., Gómez-Pérez, A.: Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering* 46(1), 41–64 (2003)
17. Baader, F.: Appendix: description logic terminology. *The Description logic handbook: Theory, implementation, and applications*, pp. 485–495. Cambridge University Press (2003)

Speech Keyword Spotting with Rule Based Segmentation

Mindaugas Greibus and Laimutis Telksnys

Vilnius University Institute of Mathematics and Informatics,
Akademijos str., 4, LT-08663 Vilnius, Lithuania
mindaugas.greibus@exigenservices.com, telksnys@mii.vu.lt

Abstract. Speech keyword spotting is a retrieval of all instances of a given keyword in utterances. This paper presents improved template based keyword spotting algorithm. It solves speaker dependent speech segment detection in continuous speech with small vocabulary. The rules based segmentation algorithm allows to extract quasi-syllables. We evaluated the algorithm by experimental with synthetic signals. The algorithm results outperform classical keyword spotting algorithm with experimental data.

Keywords: Speech processing, Speech segmentation, Keyword spotting.

1 Introduction

Natural communication qualities simulation of human-machine get attention by authors in number publications during past few years[1,2,3,4,5]. Veveo Inc.[2] states that conversational machine-human interfaces are the ultimate and natural user interface that will shape the usability of mobile devices.

Keyword spotting(KWS) is a technologically relevant problem, playing an important role in audio indexing and speech data mining applications [6]. KWS is also used for locating occurrences of keyword in speech signal [7]. This problem is similar to speech recognition, although ignoring the additional signal information around the words of interest[8].

In 70s of 20th century Christiansen and Rushforth [9] propose algorithm that uses Sliding Window and based on Dynamic Time Warping(DTW) with Linear Predictive Coding. The proposed algorithm calculates word unit keyword for each frame. Authors point that training for multi speakers is not trivial task and deserves a closer look. Myers et al [10] suggest using a local minimum dynamic time warping algorithm to find the words in speech signal. The main idea is moving test sample center if the reference sample matches.

Rohlicek et al [11] in their paper use Hidden Markov Model(HMM) with different speech signal features. They found that cepstra model shows the best performance which attempts to represent non-keyword speech. Authors mention that changes in model can lead to performance degradation.

Wilpon et al [12] state that duration constraints are the major problem in DTW based template matching recognition system, since each template has

physical duration and this forces algorithm follow this time constraint. They propose use HMM with models as such behaviour is statistically modelled as part of training procedure. They reported that for HMM training should use same number of almost the same number of training and test utterances. The most occurred words of their speech corpus selected as keywords for they experiment.

Weintraub [13] is using Viterbi algorithm to search keyword in a large vocabulary continuous speech with N-best scoring technique. The main idea it is if same keyword shows up few time in N-best list, then matched keyword is the best hypothesis. As author mentioned this algorithm can cause high False-Positive rates.

Szoke et al in the paper [14] compare word and subword level segment together with HMM word recognition using large vocabulary approach. The experiment results show that word level segment vocabulary is more accurate and requiring more computational power. Authors propose use subword recognition for a fast preselection of candidates. In the paper they uses training data twice as much as testing. Keywords selected as them most frequently occurred words.

In another paper Szoke et al [15] describe algorithm based on Gaussian mixture model(GMM) together with HMM. They optimize algorithm for real-time keywords search. They noted that using phoneme model is less accurate than triphonomes, but triphones require more calculation power.

Shao et al [7] propose algorithm that uses dynamical programming algorithm to do fuzzy search on this specially designed index structure. Algorithm has 3 stages: recognising speech into speech syllables, construct index, search keyword. This paper algorithm depends on how accurate segmented to syllables and how well those recognized. Compared with English, Mandarin is a monosyllabic and tonal language so syllable is more efficient for KWS than phone.

Jansen and Niyogi [6] use syllable-based GMM model and Vowel Landmark Detector. They algorithm requires relatively small number of training speakers. Vowel Landmark Detector was found as effective as Sliding Window search. The syllabary of English consists of approximately 12,000 syllables, so collecting enough training examples of each to build a set of detectors for an arbitrary word is practically infeasible. However, it is interesting to note that the most frequent 324 syllables in a typical speech corpus can cover two-thirds of it. Laurinciukaite and Lipeika was using 227 syllables for Lithuanian speech recognition experiments [16].

Grangier et al [5] say that Viterbi decoding is fragile with respect to local model mismatch. In the context of HMM-based KWS, an algorithm miss a keyword, if only its first phoneme suffers such a mismatch, for instance. They proposed use discriminative algorithm based on Support Vector Machine. This research was done together with Google Inc.

Zhang and Glass [17] propose unsupervised learning framework. The algorithm requires training GMM to represent each speech frame with a Gaussian posteriorgram[18]. For each keyword they are using a segmental dynamic time warping [19] technique to compare the keyword examples and test data.

Algorithm requires that speech would be segmented into speech, non-speech segments and requires a big number of training data for GMM.

Zhang et al [20] use HMM with confusion garbage model. The main goal was reduce False-Positive for similar pronunciation with predefined keywords, but with different grapheme.

A supervised approach requires training models such as HMMs or SVMs in a preprocessing step. However, in several applications such kind of training is not possible as no training material is available a priori [4]. The full signal recognition could be easier from point of implementation if there is a high recognition ratio Automatic Speech Recognition system. This approach could require high calculation power. A word level keyword spotting based on DTW is very sensitive to start and end of keyword in a test utterance. A new word reference pattern requires new sample.

The KWS can be defined as speaker dependent word recognition in continuous speech with small vocabulary. In this paper we comparing two algorithms: Multi Feature Extremum and Rule Based (MEaRBS) [21] and Sliding Window [10]. In Sect. 2 the classical KWS algorithm presented. In Sect. 3 proposed algorithmic described. The experimental data defined in Sect. 4 and in Sect. 5 the results analysed. The conclusions can be found in Sect. 6.

2 Keyword Spotting

Many KWS approaches use word, syllable or phone level information. The word level spotting always requires additional training for unseen word [14]. The phoneme provides ability to define new keyword easy, but such segment automatic boundary detection and classification in speech is not trivial [22]. The syllable is trade-off as it is easier to add new keywords if such syllables already exists. Also A syllable should be easier to find than phoneme as in most cases a syllable has longer duration than a phoneme.

The KWS complex problem can be divided into two tasks: change point detection and segment labelling. Keyword or key subword segment boundaries needed to define when a segment occurred in an utterance and what meaning this segment has from language perspective.

Figure 1 shows given continuous speech utterance (top wave form): "Nuaidejus Lietuvos himno paskutiniams akordams". The goal is to find if keyword segment *Lietuvos* appeared in the signal and at what moment. Each segment should have boundaries and a label. The sequence *Lie, tuvos* of quasi-syllables can describe same keyword.

2.1 Signal Processing

Both algorithms process audio signal using overlapping frames: sequence of samples \bar{x}_t (1).

$$\bar{x}_{t+I} = (x_t, x_{t+1}, \dots, x_{t+N}) \quad (1)$$

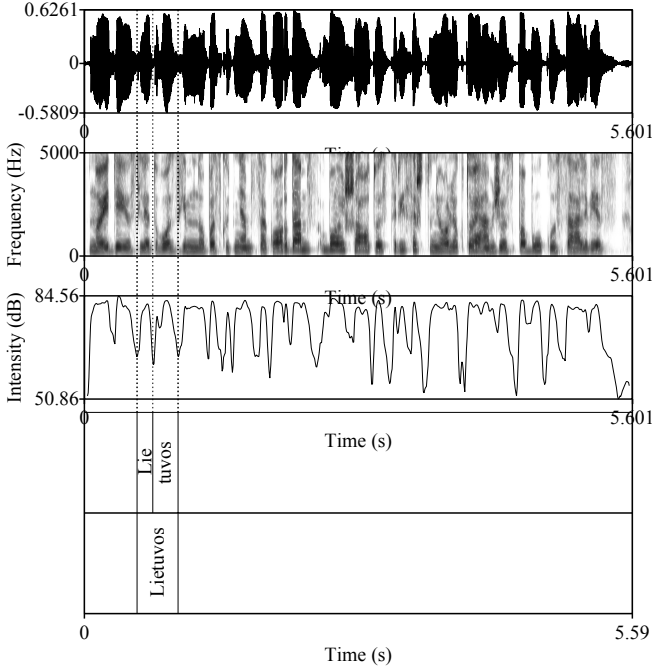


Fig. 1. Phrase wave form, spectrogram, intensity and key segments boundaries(quasi-Syllable and word level)

where x_t - signal sample at t time coordinate, N - Frame length, I - frame shift by number of samples.

Algorithms use 12 Mel-Frequency Cepstral Coefficients(MFCC) \mathbf{y}_t . The values are calculated for each frame(2)

$$\mathbf{y}_{t+I} = MFCC_{12}(\bar{x}_t) \tag{2}$$

2.2 Sliding Window Algorithm

The algorithm [9,10] is based on shifting window through signal. Each time the reference template is compared with test utterance. The DTW algorithm helps measure degree of similarity D_t (6) between reference $\bar{\mathbf{y}}_c^R$ (4) of class c and test $\bar{\mathbf{y}}_t^T$ patterns (3).

Equations (3) define window for test pattern.

$$\bar{\mathbf{y}}_{t+I}^T = [\mathbf{y}_t, \mathbf{y}_{t+I}, \mathbf{y}_{t+2I}, \dots, \mathbf{y}_{t+IK}] \tag{3}$$

where K number of frames in window.



The Reference pattern (4) of class c has length E . This pattern is defined by a training algorithm.

$$\bar{\mathbf{y}}_c^R = [\mathbf{y}_0^R, \mathbf{y}_I^R, \mathbf{y}_{2I}^R, \dots, \mathbf{y}_{IE}^R] \quad (4)$$

For Sliding Window algorithm training and reference utterance duration should be the same (5).

$$E = K \quad (5)$$

For each frame is calculated warping distance between references on the time coordinate $t + I$ (6). Figure 2 reveals this distance on time coordinates. We say that segment found if local minimum of warping distance is in range d smaller than warping distance threshold T_c .

$$D_{t+I} = DTW(\bar{\mathbf{y}}_c^R, \bar{\mathbf{y}}_t^T) \quad (6)$$

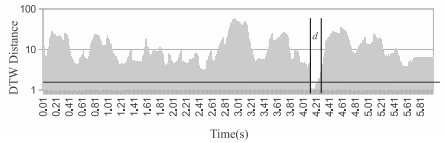


Fig. 2. Sliding Window DTW distances to reference pattern and threshold T_c

Lets say we have a set of reference templates for each key segment of class c , then the threshold T_c is calculated as maximum DTW distance between reference patterns (7) in same class c set.

$$T_c = \max_{\bar{\mathbf{y}}_c^R \in D_c} DTW(\bar{\mathbf{y}}_{c1}^R, \bar{\mathbf{y}}_c^R), \quad (7)$$

where $\bar{\mathbf{y}}_{c1}^R$ - first key segment reference template in set D_c .

3 Proposed Algorithm Based on MEaRBS

Sliding Window $\bar{\mathbf{y}}_t$ requires evaluate similarity distance for each frame. The classical approach requires same reference and test template duration. Also such approach requires high computational power as it is needed calculate for each shift. The same duration of the patterns in most cases does not match speech variability nature. We propose use MEaRBS rules segmentation to improve efficiency calculation.

Horak [23] uses phoneme alignment with synthetic signals. We propose use similar approach for KWS. The main idea is that text to speech synthesizer generates synthetic keyword. It should not generate full word at once, but sequence of quasi-syllables: consonant-vowel-consonant, consonant-vowel or vowel-consonant style. In this way reference sequence of quasi-syllables defines full keyword or keyphrase. The next step is automatically segment test utterance into quasi-syllables. After

the algorithm goes through auto-segments and search for the first quasi-syllable reference in the KWS sequence. If the first matches then proceed next quasi-syllable in sequence till all segments in sequence are found.

The proposed algorithm uses MEaRBS [24] for quasi-syllable segment detection and same as above DTW pattern matching algorithm for segment classification (8).

$$D_{t+MEaRBS}(\bar{x}_{t+i}) = \min_{\bar{y}^R \in R} DTW(\bar{y}^R, \bar{y}_t^T) \tag{8}$$

Figure 1 depicts the algorithm approach. In short the algorithm marks quasi-syllable energetic areas of a signal. Next each quasi-syllable is measured similarity (6) to the reference pattern \bar{y}^R in quasi-syllable speech corpus R .

The algorithm flow chart presented in Fig. 3. It contains phases: framing, quasi-syllable segment boundaries detection, quasi-syllable segment labelling, word segment boundaries detection, word segment labelling.

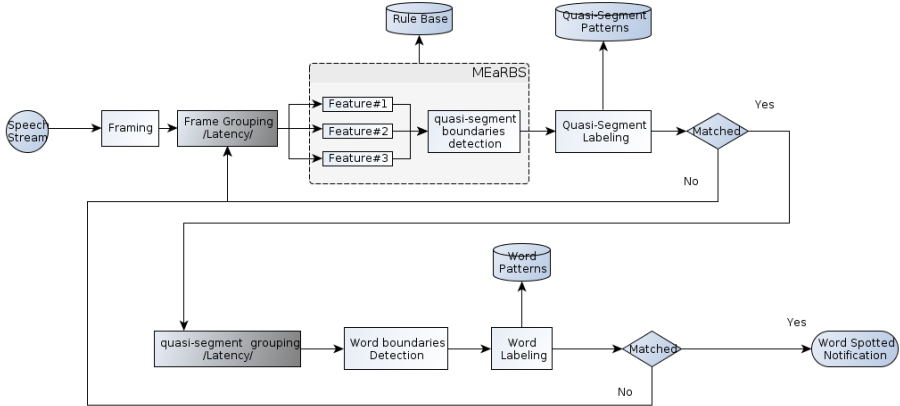


Fig. 3. Algorithm flow chart

As previous experiments showed the segment classification DTW algorithm is very sensitive to the correct boundary detection. This is even more noticeable for short syllable level segments. The segment boundary adjustment could improve matching results. The detected quasi-segment boundaries were adjusted through time coordinate that would be optimised the defined criteria. For this research we used two different criteria: minimum DTW path and maximum delta between two closed samples. Each auto-segment was shifted from -50ms to 100ms from the original position using 10ms step. The numbers were identified empirically. Each quasi-segment classification result was compared with threshold T_c that is related with class c . If the value exceeded threshold, such segment was rejected.



4 Experiment

We have reviewed several aspect of the two algorithms. Which algorithm gives better performance results when applied to KWS experiments? How different speech nature variables effects spotting? What is performance improvement using MEaRBS.

The experiment goal was compare KWS Algorithms based on Sliding Window and MEaRBS.

4.1 Experiment Data Synthesis

In the experiment for simplicity we use one keyword with quasi-syllables Fig. 4: *Lietuvos*, *Lie* and *tuvos*.

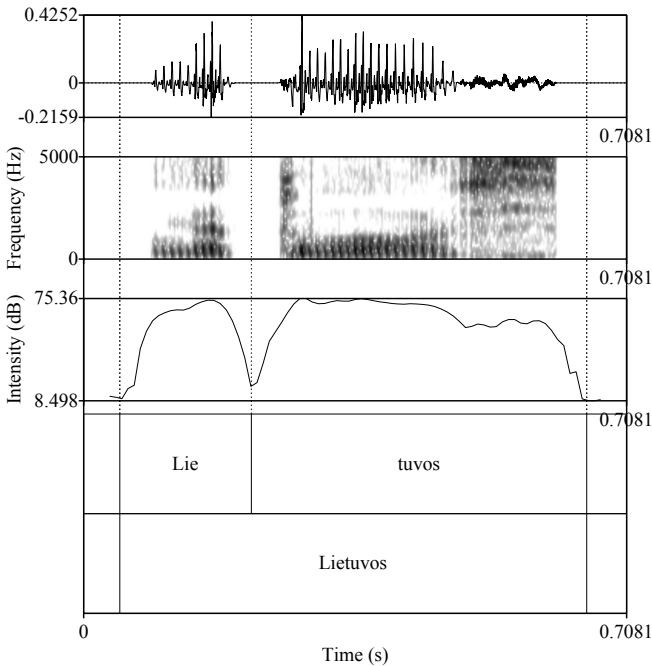


Fig. 4. Synthetic keyword wave form, spectrogram, intensity and segments boundaries(quasi-syllable and word levels)

The Speech Synthesizer could generate signals using 4 variables Fig. 5: pronounceable phrases, speed of speech by defining phone duration, tone variance by defining pitch and white noise level.

Speech phrases for the experiment were generated using defined algorithm that:

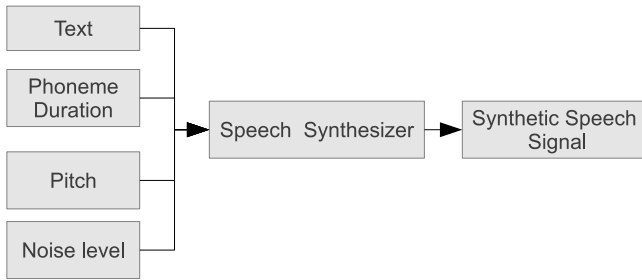


Fig. 5. Speech synthesis

1. each phrase contains 2 keywords,
2. before and after each keyword exists at least one word,
3. in each signal at least one word should be different.

In this was generated 400+1 unique phrases. Examples: *kazkur Lietuvos jeigu Lietuvos nariu skaitys, kadangi Lietuvos apie vidudieni Lietuvos visuomenes aplink, aplink Lietuvos bet Lietuvos sviesuomenes nekomentavo.*

A phone duration could be chosen as a random or constant number. Variation represents a natural speech model. A static phone duration reduces complexity of the model that allows easier interpreter experiment results.

The pitch variation defines how natural speech sounds. Pitch variation depends on previous word and intonation as defined by the Speech Synthesizer algorithm.

The influence of different noise level estimation of was out of scope for this experiment. We are using 30dB noise to power of noiseless synthetic signal.

1 natural and 3 synthetic speech corpora are used as experimental data:

1. G_1 - (wopitch) pitch and average phoneme duration not varying. Audio files is total 32 minutes length.
2. G_2 - (dynlen) only average phoneme duration is varying. Audio files is total 48 minutes length.
3. G_3 - (wpitch) only pitch is varying. Audio files is total 32 minutes length.
4. G_4 - (natural) natural speech. Audio files is total 2.8 minutes length. 40 multi-speaker phrases were taken from LRN0.1 corpora.

In this way it was generated $400 * 2 * 3 = 2400$ keywords or 4800 - quasi-syllables that is synthesized. Total synthetic audio length is 1h 57min.

For each speech corpora was calculated different threshold T_c (7)

4.2 Sliding Window and MEArBS Algorithm Evaluation

The experiments with 4 speech corpora evaluated Sliding Window and MEArBS algorithms. For each experiment detection error was calculated as word error rate: WER (9) and keyword detection error ε (10)

$$WER = \frac{S + I + D}{S + I + D + C}, \quad (9)$$

where S - substitution (when A class segment detected, but B class was true), I - insertion (when no key segment found, but a segment existed), D - deletion (when the key segment exists, but none detected), C - correct (when the key segment exists and the same detected).

$$\varepsilon = \frac{S + I + D}{N}, \quad (10)$$

where N is amount of decision points (amount of windows).

The confidence interval calculated by (11).

$$\varepsilon \pm z_{\alpha/2} \sqrt{\frac{\varepsilon(1-\varepsilon)}{N}}, \quad (11)$$

where $z_{\alpha/2}$ - the 95% confidence level would imply the 97.5th percentile of the normal distribution at the upper tail 1.959964.

The execution time is important for some of algorithm applications. An online implementation requires computation time shorter than audio signal itself: audio signal and processing time ratio expected less than < 1 .

By comparing word error rate, the proposed algorithm does around 1.5-4 times less error than Sliding Window for synthetic and natural corpus in Fig. 6 and Table 1.

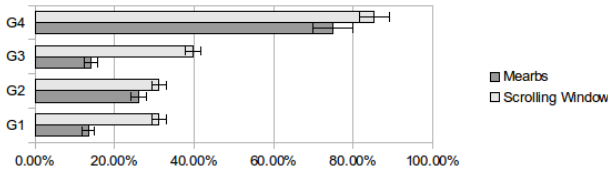


Fig. 6. Word error rate $WER(9)$

Table 1. Quasi-syllable spotting results

Algorithm	Corpus	False-Negative	False-Positive	Correct
MeaRBS	G1	14.39%	11.62%	74.00%
Sliding Window	G1	0.17%	31.00%	68.82%
MeaRBS	G2	0.38%	13.01%	86.61%
Sliding Window	G2	0.00%	31.21%	68.79%
MeaRBS	G3	7.83%	6.20%	85.97%
Sliding Window	G3	4.99%	34.65%	60.36%
MeaRBS	G4	16.43%	58.39%	25.17%
Sliding Window	G4	8.50%	76.77%	14.73%

The proposed algorithm does around 5-25 % less error than Sliding Window Fig. 7 comparing keyword detection error.

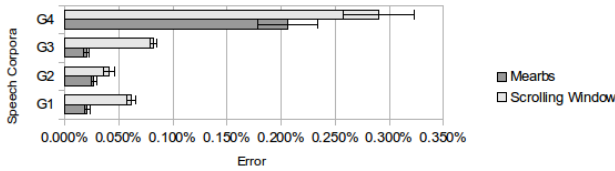


Fig. 7. Keyword Detection Error ε (10)

Current algorithm realisations of proposed algorithm took about 3-5 less time than Sliding Window realisation with same hardware (figure 8).

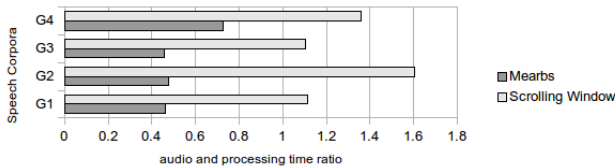


Fig. 8. Audio signal and processing time ratio

The algorithm accuracy could be improved by improving False-Positive rejection methods.

5 Conclusions

In this paper improved keyword spotting algorithm was proposed:

1. The main novelty is keyword spotting is based on quasi-syllables that is extracted with rule base segmentator
2. Experiments was done with controlled environment with synthetic signals.
3. The experiments showed that proposed algorithm performs better 5%-25% from than classical and require in average twice less computational power.
4. The algorithms were evaluated with natural multi speaker signals also.

The proposed algorithm should be improved to use in real-life applications. It is planned to improve False-Positive error stability and multi-speaker quasi-segment classification.

References

1. Maskeliunas, R., Ratkevicius, K., Rudzionis, V.: Some aspects of voice user interfaces development for internet and computer control applications. Electronics and Electrical Engineering 19(2), 53-56 (2013)

2. Veveo: Conversational interfaces whitepaper. Technical report, Veveo (2012)
3. Frinken, V., Fischer, A., Manmatha, R., Bunke, H.: A novel word spotting method based on recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 211–224 (2011)
4. Von Zeddelmann, D., Kurth, F., Müller, M.: Perceptual audio features for unsupervised key-phrase detection. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP, pp. 257–260 (2010)
5. Keshet, J., Grangier, D., Bengio, S.: Discriminative keyword spotting. *Speech Communication*, 317–329 (2009)
6. Jansen, A., Niyogi, P.: Point process models for spotting keywords in continuous speech. *IEEE Transactions on Audio, Speech, and Language Processing* 17(8), 1457–1470 (2009)
7. Shao, J., Zhao, Q., Zhang, P., Liu, Z., Yan, Y.: A fast fuzzy keyword spotting algorithm based on syllable confusion network. In: Eighth Annual Conference of the International Speech Communication Association, pp. 2405–2408 (2007)
8. Ramachandran, R.P., Mammone, R.J.: *Modern methods of speech processing*, vol. 327. Springer (1995)
9. Christiansen, R., n, C.: Detecting and locating key words in continuous speech using linear predictive coding. *IEEE Transactions on Acoustics, Speech and Signal Processing* 25(5), 361–367 (1977)
10. Myers, C., Rabiner, L., Rosenberg, A.: An investigation of the use of dynamic time warping for word spotting and connected speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1980, pp. 173–177. IEEE (1980)
11. Rohlicek, J.R., Russell, W., Roukos, S., Gish, H.: Continuous hidden markov modeling for speaker-independent word spotting. In: *Acoustics, Speech, and Signal Processing* 1989, pp. 627–630. IEEE (1989)
12. Wilpon, J., Rabiner, L., Lee, C., Goldman, E.: Automatic recognition of keywords in unconstrained speech using hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing* 38(11), 1870–1878 (1990)
13. Weintraub, M.: Lvcsr log-likelihood ratio scoring for keyword spotting. In: 1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1995, pp. 297–300 (1995)
14. Szoke, I., Schwarz, P., Matejka, P., Burget, L., Karafiát, M., Fapso, M., Cernocky, J.: Comparison of keyword spotting approaches for informal continuous speech. In: *Proc. of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms* (2005)
15. Szöke, I., Schwarz, P., Matějka, P., Burget, L., Karafiát, M., Černocký, J.: Phoneme based acoustics keyword spotting in informal continuous speech. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) *TSD 2005. LNCS (LNAI)*, vol. 3658, pp. 302–309. Springer, Heidelberg (2005)
16. Laurinciukaite, S., Lipeika, A.: Syllable–phoneme based continuous speech recognition. *Electronics and Electrical Engineering* 6, 70 (2006)
17. Zhang, Y., Glass, J.R.: Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriors. In: *Automatic Speech Recognition & Understanding*, pp. 398–403. IEEE (2009)
18. Aradilla, G., Vepa, J., Boulard, H.: Using posterior-based features in template matching for speech recognition. In: *Int. Conf. on Spoken Language Processing* (2006)
19. Park, A.S., Glass, J.R.: *Unsupervised pattern discovery in speech*, vol. 16, pp. 186–197. IEEE (2008)

20. Zhang, S., Shuang, Z., Shi, Q., Qin, Y.: Improved mandarin keyword spotting using confusion garbage model. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 3700–3703. IEEE (2010)
21. Greibus, M., Telksnys, L.: Speech segmentation analysis using synthetic signals. In: Electronics and Electrical Engineering (2012)
22. Skripkauskas, M.: Lietuviu snekos signalu segmentavimas kvazifonemomis. In: Informacins Technologijos, pp. 76–81 (2006)
23. Horák, P.: Automatic speech segmentation based on alignment with a text-to-speech system. Improvements in Speech Synthesis, pp. 328–338. Wiley Online Library (2002)
24. Greibus, M., Telksnys, L.: Rule based speech signal segmentation. Journal of Telecommunications and Information Technology (JTIT) 1, 37–44 (2011)

Business Intelligence Maturity Models: Information Management Perspective

Alaskar Thamir and Babis Theodoulidis

Manchester Business School, University of Manchester, Manchester M15 6PB, UK
thamir.alaskar@postgrad.mbs.ac.uk,
b.theodoulidis@manchester.ac.uk

Abstract. While Business Intelligence (BI) plays a critical role for businesses in terms of organizational development and creating competitive advantages, many BI projects fail to fully deliver the features and benefits that could help organizations in their decision-making. Rather than depending on software, BI success relies on the capabilities of sensing for appropriate information, data collection, extraction, organization, analysis, and retention of information due to the large volume of information that exists.

Therefore, this paper presents a comprehensive review of existing BI maturity models and elaborates their methodical and conceptual characteristics to determine their gaps in addressing the information life-cycle concept in terms of sensing, collecting, organizing, processing, and maintaining activities. As a result, a conceptual framework is proposed from the literature analysis. The intentions are to build a BI maturity model that can be used to increase the success of BI implementation by basing it on Information Management Practice (IMP), which a model built on the information life-cycle concept.

Keywords: Business Intelligence, Maturity Model, Information Life-Cycle, Information Management Practice, Literature Review.

1 Introduction

Nowadays, Chief Information Officers (CIOs) consider BI to be the most important technological area (Raber et al., 2013; Gartner, 2013), helping them to improve performance and create competitive advantage (Chen, 2012; Vitt et al., 2002). However, Wells (2008) sees BI as the capability of an organization to predict, plan, and solve problems to help in establishing and achieving business goals; and not as being about tools, applications, data and databases.

The role of BI has changed from concentrating on technical capabilities (Wells, 2008) to contributing to strategic decision-making by focusing on the sensing activity to monitor market change in the external environment and explain early threatening signals of risk from unpredicted sources (Gilad, 2004; Frates and Sharp, 2005: 20). Based on information needs, it also contributes to deciding which information is to be exploited in order to maximize opportunities, and avoid problems before they occur (Rouibah and Ould, 2002; Grof, 1999). Furthermore, it also assists in deciding how much they want

information sources; be they external, such as information on competitors and customers, or internal, such as operational databases (Myllarniemi et al., 2009).

In addition, while most organizations deal with the large volume of information that exists within an organizational environment, causing a big data issue, the BI role becomes important in addressing only information that is critical and accurate (Rouibah and Ould, 2002; Gromm and David, 2001). Cackett et al. (2013) state that while information management focuses on organizing the large volumes of semi-structured and unstructured data that are stored in organizations, big data capabilities have to fit with information management design in order to leverage big data in a successful way. For example, a Telecommunications Company can interact with its customers by triggering a customer's location with real data instead of putting fixed campaigns against defined target segments. However, this use of big data should be justified in terms of what new opportunities could be used regarding Price Management, Product and Offering Design, Acquisition and Retention Management, and Loyalty Management (Cackett et al., 2013). Therefore, it is important to address the organization within sensing activities in an appropriate way during BI implementation.

Nevertheless, to make BI more effective, it is important to link sensing, collecting, organizing, and maintaining information activities with organizational success. Despite the complexities in implementing BI systems in terms of sensing and other information life-cycle activities, as discussed above, there has been little empirical research into BI maturity models regarding how to identify the concepts of information life-cycle and business intelligence that can impact on the successful implementation of BI systems, and this gap in the literature is reflected in the low level of contributions on this issue to international conferences and journals. Therefore, this paper identifies gaps in existing BI maturity models (MMs) by analyzing the existing BI maturity models to highlight their shortcomings in addressing BI benchmarking variables. The analysis will also be done from an Information Management Practice (IMP) perspective to show the weaknesses of these models in terms of addressing critical information life-cycle phases.

2 Life Cycle View of Information Management

This part will discuss the information life cycle concept as well as giving a description of the IMP model and its phases and why it is used in BI as a measurement base.

2.1 Information Life Cycle Model

Information management has been defined as a set of activities that transfers through a desired sequence of phases, as each phase is dependent on the other (Kettinger and Marchand, 2011; Alavi and Leidner, 2001; Butler and Murphy, 2007). The life cycle phases have been changed with time in the literature, as most of them are inconsistent in terms of concepts and in including only four phases: collecting, organising,

processing and maintaining information (Kettinger and Marchand, 2011; Ashby, 1956; Taylor, 1968). However, the sensing phase was later included in the information management life cycle by Choo (1998) to address the activities that related to the scanning of the external environment (Kettinger and Marchand, 2011). Based on that, Kettinger and Marchand (2011) suggest an IMP model which includes sensing, collecting, organizing, processing and maintaining phases.

2.2 Information Management Practices (IMP)

The IMP model, which was built by William J. Kettinger and Donald A. Marchand in 2011, is based on a general model of information used, proposed by Choo in 1998. According to Kettinger and Marchand (2011), the IMP model is a theoretical model that is built on path dependency theory for the nature of decision-making phases, where each phase is dependent on the previous phase, and keeps independence as a concept. Moreover, both tacit and explicit knowledge concepts were taken into account in the design of the IMP model concept by focusing on the knowledge of people (Kettinger and Marchand, 2011).

The IMP model implements the growth of information life cycle approaches, and it includes five phases which represent the information management cycle of the IMP model, and they are:

- 1- Sensing Phase: used to detect and identify information concerning:
 - A- Social, economic and political variations which could impact organizations;
 - B- Innovations that are created by competitors which might influence the business;
 - C- New products which satisfy customer demands and market changes;
 - D- Recognition of the problems that could happen with the company's partners and suppliers.
- 2- Collecting Phase: used to collect related information, including:
 - A- To make sure that the right information is provided at the right time; outlining the desires of information for employees is required;
 - B- In order to prevent overloading of information, cleaning information is necessary;
 - C- Key information sources should be identified;
 - D- To ensure that there is accurate and complete collecting of information, training and rewarding employees should be identified.
- 3- Organizing Phase: used to organize the information to ensure cost-saving by minimizing efforts in locating useful data and preventing duplication; the focus is on:
 - A- Indexing and classifying information for appropriate availability;
 - B- Linking databases across the business units and functions within an enterprise;
 - C- Training and rewarding employees for accurately and completely organizing the information for which they are responsible.

- 4- Processing Phase: used for analyzing data which have been organized in the previous stage; processing information includes:
- A- Only suitable information is accessed;
 - B- To drive sensible decisions, databases are analyzed;
 - C- People with outstanding analytical skills are hired;
 - D- Making sure of the appropriate use of information to arrive at decisions; training and rewarding of employees is required in this stage;
 - E- Appraisal of employees' performance should be aligned with their use of information.
- 5- Maintaining Phase: used for future organizational use of information, involving the following:
- A- In order to save efforts and cost, existing information which has been previously collected in one part of the organization will be used again;
 - B- Databases should be updated to make sure they remain current;
 - C- Continuous refreshing of data to make sure that people are using the appropriate and up-to-date information.

However, the proposed BI maturity assessment will be based on the IMP model, as it identifies the cycle of information that includes the sensing phase; to help in assessing the capabilities of information processes within the BI environment. According to Choo (2002), environmental sensing and BI reflect the same meaning as they both focus on immediate competitive situations as well as the political, social and economic factors of the external environment. In addition, to increase quality and clarity of information deal with uncertain situations, sensing phase have to be well developed as emphasis by Marchand, Kettinger, and Rollins (2002). Moreover, Rouibah and Ould (2002) put emphasis on the importance of building sensing or scanning strategies in the BI environment as BI depends on various data collection, extraction, and analysis technologies (Chen et al., 2012; Chaudhuri et al., 2011).

3 Existing BI Maturity Models

The maturity model for Business Intelligence gives support to organizations so that they have a clear perspective of their current position and what they need to do in order to reach the next phase. As Rajteric states (2010), BI maturity models (MMs) offer different strategies for development in this rapidly growing field. Bruin et al (2005) argues that the earlier research could be a good resource to get critical success factors which are required in building maturity model. Therefore, in table 1 below, the existing BI maturity models will be explored and compared to understand what key areas have been addressed by such models.

Table 1. Overview of existing BI maturity model

Model Name	Reference	Topic	Description
The BI Maturity Model	(Stock, 2013)	BI	The three main areas of the model are business enablement, information management, and strategy and program management. It uses a five-grade scale for each part (Stock, 2013). This model focuses principally on the alignment and integration by a linkage KPIs within the organization strategies as well as responsive to business environments. In addition, it focuses on data governance and stewardship, measurement of ROI, quality of data and data management (MDM, metadata), BI programme management office (PMO) analytics skills ,sponsorship and C-level role.
Enterprise Business Intelligence Maturity Model	(Chuah and Wong, 2012)	BI	The three main areas of the Enterprise Business Intelligence Maturity Model (EBIMM) are data warehousing, information quality and knowledge process. It uses a five-grade scale for each part (Chuah and Wong, 2012). This model focuses mainly on the technical viewpoints by emphasizing the data warehouse part rather than the business side. In addition, the documentation of this model is not well established. This model focusing on data and metrics alignment between departments, alignment between KM process and department level (individuals, Department, Enterprise, Extended enterprise), data management policy and information quality conditions, technical programme skills, redundancy of data and management of metadata issues.
Impact-Oriented BI MM	Lahrmann et al (2011)	BI	BI capabilities, BI practices, BI IT, organizational support, individual use, organizational use, individual impact and organizational impact are the main areas of the Impact-Oriented BI maturity model, which uses a five-grade scale for each part (Lahrmann et al., 2011). The Impact-Oriented BI maturity model is a theoretical BI model that based on the IS impact measurement model which created by Gable et al. (2008). In addition, the model is based on comparisons between ten existing BI maturity models, data warehousing, information management, and data management. This model focuses on business requirements methodology, data governance, cost effective development and operations, technical and social capabilities, organizational support, technical architecture and analytical tools, data quality, and data integration.

Table 1. (continued)

American SAP User Group	Hawking et al (2010)	BI	<p>Information analytics, governance, standards processes, and application Architecture are the main areas of this Business Intelligence Development Model, which uses a six-grade scale for each part .According to Hawking et al. (2010), this model was published only for SAP customers; as a result, there is no literature which has discussed and analysed this model critically. However, this model focuses on KPIs, and on the importance of building an alignment between business needs and KPIs in order to drive a standardised view of business performance (Hawking et al, 2010). In terms of BI benchmarking indicator, this model focuses on identification and use of KPIs and analytics. Moreover, it focuses on the BI Competency Centre (BICC), standards and processes of BI, architecture needed for BI application.</p>
Business Intelligence Development Model (BIDM)	Sacu and Spruit (2010)	BI	<p>Temporal characteristics, data Characteristics, decision Insights, output Insights, BI-Process Approaches, Semantics, User, Implementation people, process and technology are the main areas of focus of the Business Intelligence Development Model, which uses a six-grade scale for each part (Sacu and Spruit , 2010). Chuah and Wong (2010) criticize the Business Intelligence Development Model as it is not well documented and lacks a well-defined evaluation. In addition, the model focuses on the technical side more than on the business side. However, this model focusing on data and analysis in terms of refreshing period data focus, and action type. In addition, it focuses implementation type, at department level or enterprise-wide, culture and whether it is a closed loop environment, type of analysis tools at each level, data type, data sources, and granularity level.</p>
TERADATA'S BI and DW maturity model	Miller et al (2009)	BI	<p>Business alignment, architecture practices, performance systems management, BI/decision support, business analytics, data management, data acquisition/integration, business continuity, communication/ training, program and project management are the main areas of Teradata's BI MM, which uses a six-grade scale for each part (Miller et al, 2009). TERADATA'S maturity model is considered to be a process-centric model emphasizing mainly the influence of BI on the business processes (Lahrman et al.,2010). Moreover, the model focuses on the as-is situation of BI and DW and the consistency of the model is not documented (Lahrman et al., 2010). However, this model focusing on analytic vision, business alignment, project management methodology and data warehouse agility. In addition, it focuses on data governance and stewardship, measurement of ROI, training on the data model to know how to address data and interpret it, data Acquisition and Integration techniques, quality of data, and data management (MDM, metadata).</p>

Table 1. (continued)

TDWI's Business Intelligence Maturity Model	(Eckerson, 2009)	BI	Scope, Funding, Sponsorship, Data, Value, Architecture, Development and Delivery are the eight main parts that are used for evaluation in this model, with a five-grade scale for each part .Eckerson (2007) also states that a top-down approach is used in TDWI's BI MM. However, this model focuses on the technical viewpoints by putting emphasis on the data warehouse part, and the business viewpoint could be improved with regard to the organizational and cultural vision (Chuah and Wong, 2011). In addition, the model put emphasis on creating standards for developing BI functionality: Cost-benefits; Sponsorship (CFO, CEO, BI Project, etc.), and Culture by addressing the field of analytics, whether by monitoring business events or delivering paper reports, or by addressing the technical infrastructure through emphasis on analytical tools and data architecture.
Hewlett Package Business Intelligence Maturity Model	(HP,2009)	BI	The HP maturity model covers the dimensions of business enablement, information technology, strategy, and programme management, with a five-grade scale for each part (HP, 2009). As this model focuses mainly on project management and alignment of business aspects, the data warehousing and analytical aspects have not been included which, as Chuah and Wong (2011) note, they should have been. In addition, Lahrmann et al. (2010) state that a HP maturity model is not reliable as it is not documented. However, the model put emphasis on business alignment, BI programme management office (PMO) and BICC., governance, analytics skills ,sponsorship and C-level role, technical infrastructure and quality of data.
BTMM Steria Mummert Consulting (SMC)	SMC (2009)	BI	SMC is an IT consulting company in Germany, and their Enterprise Data Management Maturity Model has three main areas of focus: process, organization, and technology, using a five-grade scale for each part (Chamoni & Gluchowski 2004; Schulze et al. 2009; Neumann, 2009). However, Lahrmann et al (2010) state that the model is not reliable as it is not documented. In terms of BI benchmarking variables, this model focusing on strategic alignment, analytical saturation, and business relevance, BI organisational structure (Project, dedicated BI-organizes, etc.),cost-effective strategy ,IT architecture needed for BI, and data management (data marts, data warehouse, etc.)
Gartner Maturity Model for Business Intelligence and	(Rayner and Schlegel ,2008)	BI/PM	People, processes and metrics or technology are the main three areas of Gartner's Maturity Model, which uses a five-grade scale for each part (Rayner et al. 2008). However, Rajteric (2010) notes that the method used to evaluate the maturity level is not well-defined as it is based on an individual maturity level classification rather than on IT employees' or business users' classifications. Nevertheless, these authors point out that this model focuses on the business viewpoints rather than on the technical view (Chuah and Wong, 2011). Moreover, the strategic

Table 1. (continued)

Performance Management			vision and plan for implementing BI projects are filed to be integrated (Hostmann et al., 2006; Rajteric, 2010). However, this model emphasising the alignment between BI and performance management strategies and business goals, BI competency centre, data policies; capabilities to support policy management and data quality; sponsorship whether from the IT or business side, incentives and the creation of opportunities; enterprise architecture, and data consistency.
SAS Information Evolution Model	(SAS,2009)	IM	People, process, culture and infrastructure are the four main areas of the SAS Maturity Model, which uses a five-grade scale for each part. This model is mainly focused on the information management approach, and its reliability is not well documented (Lahrman et al., 2010). It uses the IEM assessment process to move from one level to another by conducting five steps; determining the current IEM level, gap analysis, recommendation, roadmaps and action plan, and presentation of findings. In terms of BI benchmarking variables, this model focusing on the alignment between human capital, internal processes, culture, and infrastructure aspects. In addition, it focusing on BICC implementation, information skills, training, fact-based decisions and sharing information between units, and information architecture
Business Intelligence Maturity Hierarchy	(Deng,2007)	BI	The Business Intelligence Maturity Hierarchy model uses the knowledge management field as its main area, and it uses a four-grade scale for each part (Deng 2007). It focus on knowledge management field mainly and on technical point of view such as efficiency of reporting, analysis and data-warehousing(Rajteric ,2010).However, the evaluation standards of maturity levels are not defined appropriately (Chuah and Wong, 2010) In terms of BI benchmarking variables, this model focus on Return on investment strategy, experience perception, technical and tools infrastructure, data quality, and integration of data.
Analytical Capability Maturity Model	(Davenport and Harries, 2007)	Analytic	The three main areas of the Analytical Capability Maturity Model are organization, human, and technology; and it uses a five-grade scale for each part (Davenport and Harries, 2007). This model is based on competing in analytics strategy as it emphasises managing analytics with IT processes, governance principles, and analytical architecture, with a focus on consistent, good quality data (Aho, 2010). In addition, the model is based on four pillars: unique strategic capability, high level management support, enterprise-wide analytics, and large-scale motivation (Davenport and Harries, 2007). In term of BI benchmarking variables, this model focuses on insight into customers, markets, and competitor. In addition, it focuses on analytical competencies, executive management support, analytical culture weather if it fact-based culture or test and learning culture, hardware and

Table 1. (continued)

Infrastructure Optimization Maturity Model	(Microsoft, 2007)	BI	<p>software architecture and IT infrastructural issues, quality of data, data integration, and data architecture.</p> <p>This model was built by Microsoft, with its main areas of focus being: efficiency of reporting, analysis, and data warehouse; and uses a four-grade scale for each part (Microsoft, 2007; Kašnik, 2008; Rajteric, 2010). However, Rajteric (2010) states that the Infrastructure Optimization Maturity Model is inadequate for the business intelligence field as it focuses mainly on the products and technologies for commercial purposes; in addition, the assessment criteria for individual maturity levels are not well defined.</p> <p>In term of BI benchmarking variables, this model focuses on IT costs and business value, culture by focusing on collaboration between employees and mobility of BI, IT infrastructure such as SQL Server Analysis Services, data mining, data warehousing, data types and integration.</p>
Enterprise Data Management Maturity Model	(Fisher, 2007)	DM	<p>People, process, technology, risk and reward are the three main areas of the Enterprise Data Management Maturity Model, which uses a four-grade scale for each part. In addition; Fisher (2007) mentioned that the Enterprise Data Management Maturity Model focuses on the maturity of an organization with regard to how data is managed. While Lahmann et al. (2010) state that the Enterprise Data Management Maturity Model is good in addresses and assesses the risks of data, as well as considering a cost-benefits strategy for moving to the next level; but the model is not reliable as it represents a practice mode.</p> <p>However, this model focuses on deploying the roles, responsibilities, and policies to the acquisition, maintenance, and dissemination of data. Moreover, it focuses on employees' technical skills, sponsorship, data management tools across the organization, and data quality monitoring.</p>

Table 1. (continued)

Business information maturity model	(William and William, 2007)	IM	<p>The two main areas of model are information focus and return on investment (ROI); and it uses a three-grade scale for each part. (William and William, 2007). In addition, the main success factors have been used in the model are alignment and governance, leverage and delivery, BI strategic position, BI portfolio management, partnership between business units and IT, information and analysis usage culture, process of improving business culture, process of establishing decision culture, and technical readiness of BI/DW (William and William, 2007).</p> <p>According to Rajteric (2010), the model shows a new perspective on maturity that could add value to the business intelligence maturity assessment domain, as it is assessed from the cultural perspective. Moreover, William and William (2007) used the information culture of organizations as an assessment tool for achieving high business efficiency. In addition, the model is considered by Rajteric (2010) to be well-documented as it shows a full description for each level with a list of questions which help in performing a self-evaluation. However, the technical side of TDWI has been used in this model to cover the technical requirements for BI as the authors are TDWI business partners (Rajteric, 2010)</p> <p>In terms of BI benchmarking variables, this model focusing on the way that information requirements are defined, organizational processes that are in place for using information, cost/benefits of changing an organization culture, and fact-based decision processes.</p>
Data Warehousing Process Maturity	(Sen et al.,2006)	DM	<p>The areas of focus in this model are data quality, alignment of architecture, change management, organizational readiness, and data warehouse size with six-grade scale for each part .The model is incomplete and future work has been considered by the author (Lahrman et al., 2010).</p> <p>In term of BI benchmarking variables, this model focuses on data definition and business rules, technical skills of data warehouse, training to improve technical skills, culture by rewarding fact-based decisions and sharing information. In addition, it focusing on BI applications and IT infrastructure aspects (telecommunication, operating system, etc.) in alignment with data warehouse, data quality, and data warehouse size and architecture. However, as this model focus mainly on DM, it addresses issues like reliability of data and data warehouse size and architecture without considering business side.</p>
AMR Research's Business Intelligence/Performance Management	(Hagerty, 2006)	BI/PM	<p>The three main dimensions of the model are technology, people, process; and it uses a four-grade scale for each part (Hagerty, 2006). According to Rajteric (2010), the model focuses more on performance management than BI, as Hagerty (2006) sees Performance Management as a natural growth of Business Intelligence. Additionally, Kasabian (2007) mentions that BI is considered to be a means of transport by enabling more actual information delivery. However, Rajteric (2010) states that the analysis of this model seems to be difficult due to a lack of available documentation</p>

Table 1. (continued)

Maturity Model			as it is produced by Consultant Company. While Chuah and Wong (2010) criticize this model because it is focused on a balanced scorecard methodology rather than BI, they also point out that the criteria of evaluation are not clear as there is no questionnaire to evaluate maturity levels. However, this model focuses on mapping key performance indicators (KPIs) with organizational strategies. In addition, it focuses on project based aspects, whether multi-department, or single consistent views of the enterprise. Moreover it addresses sponsorship, culture by focusing on performance management as a cultural philosophy, incentives, and data source type.
Ladder of business intelligence	(Cates et al., 2005)	BI	Technology, process and people are the three main areas of the Ladder Maturity Model which work in synchronization using a six-grade scale for each part (Cates et al., 2005). According to Cates et al. (2010), the synchronization of work between technology, process and people leads to two main aspects. First of all, it guides intelligent business to be proactive rather than reactive in addressing problems and improving business processes. Secondly, it allows innovation at every level of the organization so that it is in advance of its competitors. However, Chuah and Wong (2010) criticize the Ladder model as it is not well documented and its maturity levels are not well defined. In addition, the model has been built from a technical point of view, and this means that it is incomplete in terms of BI characteristics. However, this model focusing on information analysis, the process needed, data needed; and frequency of information needed. In addition, it emphasising on IT governance charts and PMO roles, IT governance, sponsorship and business roles (CFO, VP, etc.) technical infrastructure and tools, and data quality and the existence of sources.
Data warehousing stages of growth	(Watson et al., 2001)	DW	The nine main dimensions of the DW maturity model are: data, architecture, stability of the production environment, warehouse staff and users, impact on users' skills and jobs, applications, costs and benefits, and organizational impacts. Each part has a three-grade scale of initiation, growth, and maturity (Watson et al., 2001). The stage of growth theory is used to build the data warehousing stages of growth model. However, the model emphasising benefits associated with data warehouse and costs, the experience and specialization of the warehouse staff, the kinds of applications that utilize warehouse data, and structure of marts and warehouses.

As noted in overview of existing BI MM, there is a small number of maturity models that are information management based; for example, Business information maturity model, which was built by William and William (2007), and SAS Information Evolution Model, which was built by SAS (2009). However, neither of them were complete models because they are not addressing whole information life cycle process in terms of sensing, collecting, organizing, processing, and maintaining. Moreover, while the BI Maturity Model which built by Stock (2013) uses information management as key area, the focus was only on organizing and processing phases rather than use whole information life cycle. In addition, most of the existing BI MMs lack empirical tests as they do not deep enough in terms of addressing BI dimensions, or the key process and assessment levels. However, three main socio-technical aspects of business intelligence maturity model have been proposed in this study according to their importance for BI as has been mentioned in some of the relevant literature, and those are: organizational, human, and technical aspects. Furthermore, the assume that no BI MMs concentrate on the information life cycle, an important part of BI implementation, means that there are shortcomings which need to be overcome. Next section will address these issues by completed content analysis of existing BI maturity models.

4 Content Analysis

Content analysis has been defined by Stone et al. (1966) as “(...) any research technique for making inferences by systematically and objectively identifying specified characteristics within text”. Prasad (2008) addressed six main steps for completing content analyses; start with designing of the research objectives or questions, selection of content , developing content themes, completing units of analysis, preparing a pilot testing, and analyzing the collected data.

As mentioned previously, Brooks et al (2013) criticizes key process of existing BI maturity models, which is used in many BI maturity models, because not included technology, people, and organizational processes. However, to build new maturity model levels, top-down approach can be used, by address definitions and dimensions first (Bruin et al ,2005). In same regard, Steenbergen et al (2009) emphasis on the importance of top down method as it is more suitable for new field. Therefore, this part will examine the BI dimensions which have been addressed in exiting BI maturity models, as well as the IMP phases. The main unit analyses of content analysis among current MMs (organizational, human, and technology dimensions) will be examined by completing two main phases. In the first phase, all synonyms of terms of BI dimensions and benchmarking variables of current BI MMs have been addressed with their current definitions (See Appendix A). In the second phase, an alternative expression has been used to change the names of the dimensions and benchmarking variables (See Appendix A). Finally, the content analysis of those BI MMs is carried out to (Table 2).

4.1 BI Dimensions and Their Definitions

Lahrman et al (2010) mentioned that there is homonymy and synonyms of terms in BI maturity models; as example HP maturity model use “IT” term while Cates et al.(2005) use the term “Technology” . By looking at the definition of dimensions and

benchmarking variables, it is clear that there is a different definition for the same construct in BI maturity models. For example, the human dimension has been defined by Curtis et al. (2010) as “the level of knowledge, skills, and process abilities available for performing an organization’s business activities”. Cates et al. (2010) define people without differentiating between knowledge and skills by saying that “an intelligent business employs human intelligence to its fullest capacity”. In addition, Fisher (2007) addresses people generally by focusing on the type of employee and their contribution to business activities in this way: “who is involved and what contributions must they make”. To solve this issue, one definition has been used as an alternative expression, in order to conduct the content analysis of BI maturity models in a successful manner. Therefore, it is important to have alternative expression to the dimensions and benchmarking variables in the existing BI maturity models to carry out comparison between them in appropriate way. The definition of dimensions and variables has been given by used existing BI maturity models authors (See Appendix A), to help us to define the alternative expression. However, many of the existing BI maturity models have not addressed definitions of their variables as most of them practitioner models. Within three basic dimensions (organizational, human, technology), ten matching benchmarking variables of current BI MMs have been founded as shown in table 2 in next section.

4.2 Content Analysis of BI Maturity Models

In this part, content analysis has been carried out for twenty BI MMs in order to examine the BI dimensions which have been addressed (Table 2) in exiting BI maturity models to be used in next step in methodical analysis of IMP model.

Table 2. Analysis Content: BI Maturity Models

<i>BI Maturity Model / BI benchmarking Variables</i>	Organizational				Human				Technical	
	Analytical processes	organization structure	Governance	Cost-benefits	Skills	Training	Sponsorship	Culture	Technical infrastructure/ Tools	Data
The BI Maturity Model (Stock, 2013)	■	■	■	■	■		■		■	■
Enterprise Business Intelligence Maturity Model (Chuah, and Wong,2012)	■	■	■		■					■
Impact-Oriented BI MM Lahrman et al (2011)	■		■	■	■		■		■	
American SAP User Group (ASUG) (Hawking et al ,2010)	■	■	■						■	
Business Intelligence Development Model (BIDM) (Sacu and Spruit ,2010)	■	■						■	■	■
TERADATA’S BI and DW maturity model (Miller et al ,2009)	■		■	■		■			■	■

Table 2. (continued)

TDWI's Business Intelligence Maturity Model (Eckerson,2009)	■	■	■	■			■	■	■	■
Hewlett Package Business Intelligence Maturity Model (HP,2009)	■	■	■		■		■		■	■
BI MM Steria Mummert Consulting (SMC ,2009)	■	■		■					■	■
Gartner Maturity Model for Business Intelligence and Performance Management (Rayner and Schlegel ,2008)	■	■	■			■	■	■	■	■
SAS Information Evolution Model (SAS,2009)	■	■			■	■		■		■
Business Intelligence Maturity Hierarchy (Deng,2007)				■	■				■	■
Analytical Capability Maturity Model (Davenport and Harries,2007)	■				■	■	■	■	■	■
Infrastructure Optimization Maturity Model (Microsoft, 2007)				■				■	■	■
Enterprise Data Management Maturity Model (Fisher,2007)			■		■		■		■	■
Business intelligence maturity model (William and William, 2007)	■		■	■				■		
Data Warehousing Process Maturity (Sen et al.,2006)			■		■	■		■	■	■
AMR Research's Business Intelligence/Performance Management Maturity Model (Hagerty, 2006)	■	■					■	■	■	■
Ladder of business intelligence (Cates et al.,2005)	■	■	■				■		■	■
Data warehousing stages of growth (Watson et al., 2009)				■	■				■	■
Σ	14	10	12	9	10	5	9	9	17	17

As shown in the analysis section, none of the BI maturity models have applied all the dimensions and benchmarking variables of BI. While some of them focus on organizational factors, such as the HP and Gartner maturity models, the others focus mainly on technical factors. Examples of these are TDWI, Data warehousing stages of growth, and the Ladder maturity model. Human factors like skills, training and culture are addressed only by a few of these models, such as the Analytical Capability maturity model and the impact oriented maturity model. However, although some of them address many factors, the way in which they do so does not seem to be

appropriate. An example of this is the governance factor, which is addressed by the TERADATA maturity model in terms of architecture governance, while the HP and Impact oriented maturity models address it in terms of data governance.

Moreover, some BI maturity models address factors by providing in-depth details while others do not. A case in point is the analytical process factor, addressed by some maturity models by refer to internal environment process without addressing the external environment for that, as does the AMR maturity model, which addresses that by mentioning linking KPIs with organizational strategies without addressing benchmarking variables of customers, markets, and competitors as Analytical Capability Maturity Model. However, next section will address methodical analysis of IMP model in order to link those factors which have been used in existing BI MMs with IMP model.

5 Methodical Analysis

In this part, methodical analysis has been carried out for twenty BI MMs in order to examine the BI dimensions within IMP phase's .The classification of analysis is based on phases of IMP phases, and those are, sensing, collecting, organizing, processing, and maintaining. To complete this task, content analysis procedures have been carried out.

Table 3. IMP Analysis of BI Maturity Models

Maturity Model / IMP phases	IMP phases				
	<i>Sensing</i>	<i>Collecting</i>	<i>Organising</i>	<i>Processing</i>	<i>Maintaining</i>
The BI Maturity Model (Stock, 2013)	■		■	■	
Enterprise Business Intelligence Maturity Model (Chuah and Wong, 2012)			■		
Impact-Oriented BI MM (Lahrmann et al ,2011)			■	■	
American SAP User Group (ASUG) (Hawking et al ,2010)				■	
Business Intelligence Development Model (BIDM) Sacu and Spruit (2010)		■		■	
TERADATA'S BI and DW maturity model (Miller et al ,2009)	■		■		■
TDWI's Business Intelligence Maturity Model (Eckerson,2009)	■	■		■	
Hewlett Package Business Intelligence Maturity Model (HP,2009)	■	■	■	■	
BI MM Steria Mummert Consulting (SMC ,(2009)			■	■	
Gartner Maturity Model for Business Intelligence and Performance Management (Rayner and Schlegel ,2008)	■				

Table 3. (continued)

SAS Information Evolution Model (SAS,2009)	■	■	■		
Business Intelligence Maturity Hierarchy (Deng,2007)				■	
Analytical Capability Maturity Model (Davenport and Harries,2007)	■	■			
Infrastructure Optimization Maturity Model (Microsoft, 2007)		■	■	■	
Enterprise Data Management Maturity Model (Fisher,2007)		■	■		
Business information maturity model (William and William, 2007)	■				
Data Warehousing Process Maturity (Sen et al.,2006)			■		
AMR Research's Business Intelligence /Performance Management Maturity Model (Hagerty, 2006)		■		■	
Ladder of business intelligence (Cates et al.,2005)	■	■			
Data warehousing stages of growth (Watson et al., 2001)			■		■
Σ	9	9	11	10	2

By looking at the comparisons between existing BIMMs in terms of IMP phases, as shown in Table 3, it is clear that none of the BI Maturity Models have applied all the IMP phases. A few of them focus on the sensing phase by addressing external environment issues as defined in the IMP Model. For example, the Analytical Capability Maturity Model focuses mainly on the sensing phase in terms of addressing benchmarking variables of customers, markets, and competitors as well as building deep strategic insights, while the AMR Maturity Model addresses the internal environment side, by emphasizing the importance of linking KPIs with organizational strategies. Furthermore, the SAP Maturity Model addresses the internal environment by focusing mainly on performance management and how to build active KPIs that address business needs; however, it does not focus on the external environment as IMP does. However, the BI Maturity Model, which was built by Stock (2013), addresses both environments successfully.

Additionally, while the SAS Maturity Model addresses the sensing phase by focusing on market alignment and efficiency of driving the performance, including the importance of culture and the human aspects in driving organization objectives and understanding the environmental benchmarking, it successfully addresses some of the important variables of the sensing phase. In contrast, the TDWI addresses the sensing phase by emphasizing the importance of managing expected risks and executive-level visions, either by driving the business or monitor processes without a focus on the methodology for the analytical process, or the skills and knowledge that are needed for this phase. In addition, while the Ladder Maturity Model addresses sensing by focusing on the importance of the industry's best practice research, the information needed to answer questions, information analysis in terms of which information, processes and frequencies are required, and the need to be proactive rather than reactive in enhancing business processes, it does not address the required skills and training for that.

In regard to the collecting phase, the Ladder MM focuses on data sources and the quality that is needed to generate information. In contrast, the HP and the TDWI

Maturity Models focus on unstructured content to be integrated with structured data which could help to find new sources of data that can help to provide organizations with their information needs and be used for more influential analysis. Moreover, the Infrastructure Optimization Maturity Model and the Business Intelligence Development Model address the collecting phase by focusing on the data type, be it structured, semi-structured or unstructured; on the data sources, be they files and database, RSS or web based; and on granularity level. In addition, the governance issue in the collecting phase has been addressed by the Enterprise Data Management Maturity Model by focusing on roles, responsibilities, and policies for the data collection phase, while not addressing the training aspect as the IMP Model did.

In the organizing phase, the TDWI addresses the phase by focusing on the management of data architecture, be it data marts, data warehouses, or enterprise data warehouses, whereas the Enterprise Data Management Maturity Model addresses the metadata environment and maintaining metadata for corporate data structures. In addition, the organizing phase has been addressed by Gartner by focusing on data governance and the existence of BICC which emphasizes BI issues such as business metadata and data assurance. Furthermore, the organizing phase has been addressed by the SAS Maturity Model which focuses on information architecture to deliver information consistently.

In addition, as Data Warehousing Process Maturity focuses mainly on the data warehouse aspect from the technical side, the organizing phase has been addressed by focusing on the reliability of data, data warehouse size and architecture. Additionally, it addresses the organizing phase by including the importance of training and rewarding staff. Moreover, the Enterprise Data Management Maturity Model focuses mainly on the organizing phase by focusing on the technology, policies, and rules that are needed for data management. In addition, it includes the reward aspect to be used as a benefit for data management although it does not address the training aspect as the IMP model did. However, this model focuses mainly on data management rather than BI. Therefore, it has addressed the organizing phase successfully but not the sensing and processing phases.

Regarding the processing phase, the HP and TDWI Maturity Models have addressed the phase by putting emphasis on processing data methods, in monthly reports, interactive reports, dashboards, or embedded analytics. The Data Warehousing Process Maturity Model, meanwhile, addresses the processing phase by focusing on the processing of historical and current data. Moreover, it discusses culture issues, such as rewarding for collaboration, sharing information, and fact-based decision making. In addition, the processing phase has been addressed by the Business Intelligence Development Model by focusing on the culture of processing, and on the processing methods used in the organization, whether they are standard reporting, ad-hoc analysis, trends analysis, data mining, or predictive modelling. Furthermore, the processing phase has been addressed by Impact-Oriented BI MM which focuses on analytic purposes, be they forecasting or operational processes. Also, the Business Intelligence Maturity Hierarchy Model has successfully addressed the processing phase by focusing on experience and types of process at each level, whether they are KPIs at the information level, or cause analysis and what-if analysis at the knowledge level. However, as this model focuses mainly on knowledge management, it has successfully addressed the processing phase by focusing on types of process at each level, but does not address the training, and culture that are needed to complete this phase appropriately.

Finally, the DW Maturity Model and the TERADATA Maturity Model are the only models that address the maintaining phase. While the DW Maturity Model addresses the maintaining phase by recognizing the processes for maintaining, the stability of the production environment and increasing the warehouse, the TERADATA Maturity Model addresses it by putting a focus on business continuity, availability, recoverability, and data protection. However, training needs in maintaining or analytics have not been addressed by either of them.

6 Conceptual Framework Development

This paper has presented IMP as a model which addresses information life cycle phases and the BI dimension with benchmarking variables that are commonly used in current BI maturity models. Figure 2 below represents the themes and factors found in the literature analysis to be implicated in the adoption of a BI assessment model.

If you have more than one surname, please make sure that the Volume Editor knows how you are to be listed in the author index.

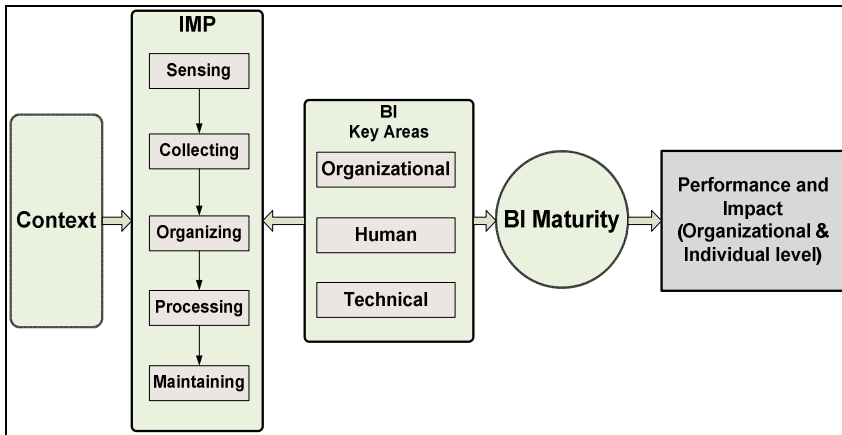


Fig. 1. Conceptual Framework of BI assessment

7 Conclusion and Future Work

In this paper, information life-cycle and Information Management Practice (IMP) have been introduced as new perspectives that are critical for successful BI implementation. Description of an information life-cycle concept and an IMP model has been given. Then, an overview of existing BI MMs has been documented, and compared from a content and IMP model perspective. According to the analysis result, this paper concludes with a conceptual framework link between the information life-cycle, BI capabilities, and organizational performance maturity which will be a base for new BI maturity model future work.

As shown in the analysis section, none of the BI maturity models have applied all the dimensions and benchmarking variables of BI; nor have they addressed all phases of the IMP model. While the existing IMP model addressed only a few BI

benchmarking variables, none of the BI maturity models have applied all the IMP phases. Some of them try to implement the sensing phase in an accurate way, while the others focus mainly on the organizing, processing or maintaining phase the latter only being applied by two models.

Therefore, in order to have a comprehensive BI model that can help in implementing BI, what seems to be important is the maturity assessment which is based on a theoretically derived model of an information life-cycle. This can act as a guide, and help in overcoming the challenges of implementing successful BI by critically determining the impact of the main BI benchmarking factors to be included in any future model.

References

1. Aho, M.: A Capability Maturity Model for Corporate Performance Management, an Empirical Study in Large Finnish Manufacturing Companies. In: Proceedings from the eBRF 2009. Presented in the eBRF 2009 - A Research Forum to Understand Business in Knowledge Society in Jyväskylä, Finland (2009)
2. AlFedaghi, S.: Information Management and Valuation. *International Journal of Engineering Business Management* (2013)
3. Bach, J.: The immaturity of CMM. *American Programmer* 7(9), 13–18 (1994)
4. Biberoglu, E., Haddad, H.: A Survey of Industrial Experiences with CMM and the Teaching of CMM Practices. *Journal of Computing Sciences in Colleges*, S.143–S.152 (2002)
5. Brunelli, M.: BI, ERP top 2007's IT spending list (2006), <http://searchoracle.target.com/originalContent/0,289142,sid41gci1233170,00.html> (accessed May 2013)
6. Bramer, M.: *Artificial Intelligence: An International Perspective*. LNCS, vol. 5640. Springer, Heidelberg (2009)
7. King, W.R., Thompson, T.S.H.: Integration Between Business Systems Planning: Validating a Stage Hypothesis. *Decision Sciences* 28(2), 279–308 (1979)
8. Cackett, D., Bond, A., Gouk, J.: *Information Management and Big Data A Reference Architecture*. Oracle Corporation (2013), <http://www.oracle.com/technetwork/topics/entarch/articles/info-mgmt-big-data-ref-arch-1902853.pdf> (accessed February 2013)
9. Cates, J.E., Gill, S.S., Zeituny, N.: *The Ladder of Business Intelligence*. Happy About Info. CA (2007)
10. Chamoni, P., Gluchowski, P.: Integration trends in business intelligence systems - An empirical study based on the business intelligence maturity model. *Wirtschaftsinformatik* 46(2), 119–128 (2004)
11. Chee, T., Chan, L.-K., Chuah, M.-H., Tan, C.-S., Wong, S.-F., Yeoh, W.: *Business Intelligence Systems: State-of-the-art Review and Contemporary Applications*. Paper presented at the Symposium on Progress in Information and Technology (2009)
12. Chen, H., Chiang, R., Storey, V.: Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36(4), 1165–1188 (2012)
13. Chen, X.: *Impact of Business Intelligence and IT Infrastructure Flexibility on Competitive Advantage: An Organizational Agility Perspective*. Dissertations and Theses from the College of Business Administration. University of Nebraska (2012)
14. Choo, C.W.: *Information management for the intelligent organization: The art of scanning the environment*. Information Today, Medford (1998)

15. Choo, C.W.: *Information Management for the Intelligent Organization: The Art of Scanning the Environment*, 3rd edn. Information Today, Inc., Medford (2002)
16. Chuah, M., Wong, K.: A review of business intelligence and its maturity models. *African Journal of Business Management* 5(9), 3424–3428 (2011)
17. Davenport, T., Prusak, L.: *Information ecology: Mastering the information and knowledge environment*. Oxford University Press, New York (1997)
18. David, R., Felix, W., Robert, W.: Situational Business Intelligence Maturity Models: An Exploratory Analysis. In: *HICSS 2013*, pp. 3797–3806 (2013)
19. Deng, R.: Business Intelligence Maturity Hierarchy: A New Perspective from Knowledge Management. *Information Management* (2007), <http://www.informationmanagement.com/infodirect/20070323/1079089-1.html>
20. Eckerson, W.: *Predictive Analytics. Extending the Value of Your Data Warehousing Investment*. The Data Warehousing Institute (2007), <https://www.tdwi.org/publications/whatworks/display.aspx?id=8452> (retrieved January 2012)
21. Ferris, J.: *How to Compete on Analytics*. The Analytical Center of Excellence. SAS Institute Inc. (2008)
22. Fisher, T.: How Mature Is Your Data Management Environment? *Business Intelligence Journal* 10(3), 20–26 (2005)
23. Fisher, T.: How Mature Is Your Data Management Environment (2007), <http://www.tdan.com/view-articles/5831> (accessed February 2013)
24. Frates, J., Sharp, S.: Using business intelligence to discover new market opportunities. *Journal of Competitive Intelligence and Management* 3, 15–26 (2005)
25. Gable, G., Sedera, D., Chan, T.: Re-conceptualizing Information System Success: The IS-Impact Measurement Model. *Journal of the Association for Information Systems* 9(7), S.377–S.408 (2008)
26. Gartner Press Release. Gartner EXP survey of more than 1,400 CIOs shows CIOs must create leverage to remain relevant to the business (2007), <http://www.gartner.com/itpage.jsp?id=501189/page.jsp?id=501189> (January 25, 2013) (retrieved) (accessed May 2013)
27. Gartner Press Release, *Get Smarter Business Intelligence: Should You Create a BI Competency Center* (2013), <http://www.gartner.com/technology/cio-priorities/>
28. Gilad, B.: *Early Warning: Using Competitive Intelligence to Anticipate Market Shifts, Control Risk, and Create Powerful Strategies*. American Management Association, New York (2004)
29. Grof, A.: *Only the Paranoid Survive How to Exploit the Crisis Points that Challenge Every Company*, 1st edn. Bantam books (1999)
30. Groom, J.R., David, F.R.: Competitive intelligence activity among small firms. *SAM Advanced Management Journal*, 12–20 (Winter 2001)
31. Hagerty, J.: AMR Research's Business Intelligence/ Performance Management Maturity Model, Version 2 (2006), http://www.eurim.org.uk/.../ig/.../AMR_Researchs_Business_Intelligence.p (Accessed February 2013)
32. Hatcher, D., Prentice, B.: *The Evolution of Information Management: A model for enabling companies to get maximum results from existing information* (2004), http://www.ewsolutions.com/resource-center/rwds_folder/rwds-archives/rwds-2004-04/evolution-of-information-mgt (Accessed February 2013)

33. Hawking, P., Jovanovic, R., Sellitto, C.: Business Intelligence Maturity in Australia. Victoria University ERP Research Group (2010)
34. Henschen, D.: 2012 BI and Information Management Trends. Information week report (2011), http://www.ums1.edu/~sauterv/DSS/research-2012-bi-and-informationmanagement_9951311.pdf (Accessed May 2013)
35. Hewlett Packard (HP). "The HP Business Intelligence Maturity Model: De-scribing the BI journey". Hewlett-Packard Development Company, L.P. (2009), <http://www.computerwoche.de/fileserver/idgwpcw/files/1935.pdf> (accessed February 2013)
36. Hostmann, B., et al.: Gartner's Business Intelligence and Performance Management Framework. Gartner Inc. (2006), <http://www.gartner.com> (accessed May 2013)
37. Kasabian, D.: 'I Can See Clearly Now', Business Trends Quarterly (2007), <http://www.btquarterly.com> (viewed on May 16, 2009) (accessed May 2013)
38. Kasnik, A.: 'Model optimization infrastructure', Internal material of ZRSZ, Ljubljana (2008)
39. Kettinger, W.J., Marchand, D.A.: Information Management Practices (IMP) from the Senior Manager's Perspective: An Investigation of the IMP Construct and Its Measurement. *Information Systems Journal* 21(5), 385–406 (2011)
40. Koh, C.E., Watson, H.J.: Data management in executive information systems. *Information and Management* 33, 301–312 (1998)
41. Lahrmann, G., et al.: Business Intelligence Maturity Models: An Overview. In: *itAIS 2010*. Springer, Naples (2010)
42. Lahrmann, G., Marx, F., Winter, R., Wortmann, F.: Business Intelligence Maturity: Development and Evaluation of a Theoretical Model. In: *Proceedings of the 44th Hawaii International Conference on System Sciences* (2011)
43. Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute (2011), http://www.mckinsey.com/mgi/publications/big_data/pdfs/MGI_big_data_full_report.pdf (accessed May 2013)
44. Marchand, D.A., Kettinger, W.J., Rollins, J.O.: Information orientation: The link to business performance. Oxford University Press, Oxford (2002)
45. McGovern, J., Ambler, S.W., Stevens, M.E., Linn, J., Sharan, V., Jo, E.K.: A Practical Guide to Enterprise Architecture. Prentice Hall PTR (2004)
46. Microsoft. Business Productivity Infrastructure Optimization Campaign (2007), http://download.microsoft.com/.../BPIO_Module_25_Summary.ppt (accessed February 2013)
47. Miller, L., Schiller, D., Rhone, M.: DataWarehouse Maturity Assessment Service Lance (2009), <http://www.teradata.com/.../Data-Warehouse-Maturity-Assessment-Service-> (accessed February 2013)
48. Myllarniemi, J., Okkonen, J., Karkkainen, H.: Utilizing Business Intelligence Framework For Leveraging Products Lifecycle Management. In: *The 9th International Conference on Electronic Business, Macau* (2009)
49. Pawar, S.P., Sharda, R.: Obtaining business intelligence on the Internet. *Long Range Planning* 30(1), 110–121 (1997)
50. Raber, D., Wortmann, F., Winter, R.: Situational Business Intelligence Maturity Models: An Exploratory Analysis. In: *46th Hawaii International Conference on System Sciences* (2013)
51. Rajteric, I.: Overview of Business Intelligence Maturity Models. *Int. J. Hum. Sci.* 15(1), 47–67 (2010)

52. Rayner, N., Schlegel, K.: Maturity Model Overview for Business Intelligence and Performance Management, Gartner, Stamford (2008)
53. Riordan, P.: The CIO: MIS Makes its Move into the Executive Suite. *Journal of Information Systems Management* 4(3), 54–56 (1987)
54. Rouibah, K., Ould-Ali, S.: PUZZLE: A concept and prototype for linking business intelligence to business strategy. *Journal of Strategic Information System* 11(2), 111–130 (2002)
55. Sacu, C., Spruit, M.: BIDM: The Business Intelligence development model. Technical report UU-CS-2010-010, Department of Information and Computing Sciences, Utrecht University (2010)
56. SAS. Information Evolution Model (2009), <http://www.sas.com/software/iem> (accessed February 2013)
57. Schulze, K.-D., Besbak, U., Dinter, B., Overmeyer, A., Schulz-Sacharow, C., Stenzel, E.: Business Intelligence-Studie, Steria Mummert Consulting AG, Hamburg (2009)
58. Sen, A., Sinha, A., Ramamurthy, K.: Data Warehousing Process Maturity: An Exploratory Study of Factors Influencing User Perceptions. *IEEE Transactions on Engineering Management* 53(3), S.440–S.455 (2006)
59. Short, J.: Information Lifecycle Management: An Analysis of End User Perspectives (2006)
60. SMC. Steria Mummert Consulting AG (2009), http://www.nomina.de/cognos/pdf/1s017_co.pdf (accessed February 2013)
61. Stock, P.: The Business Intelligence Maturity Model: describing the BI journey. YoungBlood (2013), <http://www.young-blood.co.za/index.php/2013-02-10-10-55-36/mining-and-operations/item/20-bi-maturity-model> (accessed May 2013)
62. Stone, P.J., et al.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge (1966)
63. Swoyer, S.: Come Together: Business Intelligence and Enterprise Content Management Bleed into Each Other. TDWI (2010), <http://tdwi.org/Articles/2010/01/06/Come-Together-BI-and-ECM-Bleed-into-Each-Other.aspx?Page=1> (accessed February 2013)
64. Turban, E., Aronson, J.E., Liang, T.-P., Sharda, R.: Decision Support and Business Intelligence Systems, 8th edn. Pearson Education International, New Jersey (2007)
65. Yeoh, W., Koronios, A.: Critical Success Factors for Business Intelligence Systems. *Journal of Computer Information Systems* 50(3), 23–32 (2010)
66. Yeoh, W., Gao, J., Koronios, A.: Empirical Investigation of CSFs for Implementing Business Intelligence Systems in Multiple Engineering Asset Management Organisations. In: Cater-Steel, A., Al-Hakim, L. (eds.) *Information Systems Research Methods, Epistemology, and Applications*, pp. 247–271. IGI Global, Pennsylvania (2009)
67. Vitt, E., Luckevich, M., Misner, S.: Business Intelligence, Making Better Decisions Faster. Microsoft Press (2002)
68. Wang, R.Y., Lee, Y.W., Pipino, L.L., Strong, D.M.: Manage your Information as a Product. *Sloan Management Review* 39(4), 95–105 (1998)
69. Watson, H.J., Ariyachandra, T., Matyska, R.J.: Data warehousing stages of growth. *Information Systems Management* 18(3), 42–50 (2001)
70. Wells, D.: Business analytics—Getting the point (2008), <http://b-eye-network.com/view/7133> (accessed May 2013) (retrieved)

71. Whitehorn, M., Whitehorn, M.: Business Intelligence: The IBM Solution Data warehousing and OLAP. Springer, NY (1999)
72. William, S., William, N.: The Profit Impact of Business Intelligence. Morgan Kaufmann Publishers, San Francisco (2007)
73. William, N., Thomann, J.: 'BI Maturity and ROI: How Does Your Organization Measure Up?' (2003), http://www.decisionpath.com/docs_downloads/TDWI%20Flash%20%20BI%20Maturity%20and%20ROI%20110703.pdf (accessed January 2013)
74. Wright, S.: The CI marketing interface. Journal of Competitive Intelligence and Management 3(2), 3–7 (2005)
75. Zeid, A.: Driving Innovation – The Information Evolution Model. In: Statistics Canada Information Technology Conference (2009), <http://www.statcan.gc.ca/conferences/it-ti2009/ppt/session15-aiman-fra.ppt> (accessed February 2013)

Appendix A

Dim/ Variables		Business Intelligence(BI) Dimension / Variables definition	
		<i>Resource</i>	<i>Our Definition</i>
Organizational	Organizational	(Ong et al, 2011 :4)	How an organization is structured to support BI related business processes and which activities of coordinating and managing the BI environment are being carried out.
	Analytical processes	(Devonport, 2007: 114); (Ferris, 2008:8); (Cates et al, 2007:9 ; Fisher, 2007:1); Lahrman et al, 2010 :7); (Ong et al, 2011 :4)	Address activities of business processes in how to solve analytical problems or transforming vision into competitive advantages.
	organization structure	(Ong et al, 2011 :4); (Lahrman et al, 2010 :7) Adapted from (Watson, 2001 :45); (Devonport, 1997 :69)	Structure of organization in which units take control and manage the elements of information.
	Cost-benefits	(Watson, 2001 :45); (Hocevar and Jaklic, 2009); (William and William, 2007:201); (William and William, 2007:22)	The costs and benefits of information associated with the BI
	Governance	(Weill, Ross ,2004-b); (William and William, 2007:77); (Ong et al (2011 :5)	Organize approach of principles, practices, and procedures.

Human	Human	(Curtis et al, 2010); (Cates et al (2010); (Fisher, 2007 :1); (Lahrmann et al, 2010:7)	Level of knowledge, intelligence, skills, and process abilities of Who is involved and contributes.
	Culture	(Devonport, 2007: 114);(Ferris ,(2008:8) ;(Lahrmann et al, 2010 :7)	Criteria that are used to address how organizations maintain the BI environment (i.e. fact-based decision- making)
	Training	(Ong et al ,2011 :4)	Criteria that are used to address how an organization acquires the necessary BI skills and competencies to support business goals.
	Analytic Skills and knowledge	(Brink,2003); (Devonport, 2007: 114); (Ferris, 2008:8) ;(Ong et al, 2011); (William and William, 2007:109)	Necessary BI competencies which depend on experience, interests, task complexity, and productivity that ensure that its BI requirements are built and delivered to users, and are effectively identified, validated, prioritized, and managed.
	Sponsorship	(Devonport, 2007: 114); (Ferris, 2008:8) ;(TDWI, 2007:5)	Level of management that engages support, and commits to BI programme.
Technology	Technology	(Ong et al ,2011 :5); (Cates et al, 2007:9); (Fisher, 2007:1); (Devonport, 2007: 114); (Ferris, 2008:8)	Investments in technology and uses of various BI tools and architectures to use the right information to enable effective decision-making, communication and collaboration.
	Tools and Technical infrastructure	(Sen and Sinha, 2005): (William and William, 2007:78) ; (Lahrmann et al, 2010 :7)	Platforms, standard tools, and technologies that will be used to allow BI implementation.
	Data architecture	(Watson, 2001 :45) ;(Lahrmann et al, 2010 :7); (TDWI, 2007:6); (TDWI, 2007:5); (McGovern et al, 2004)	Criteria that are used to address how data are persisted, managed, and utilized within an organization which include structure of marts and warehouses

Modified Stochastic Algorithm for Mining Frequent Subsequences

Loreta Savulioniene and Leonidas Sakalauskas

Institute of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania
l.savulioniene@eif.viko.lt, sakal@ktl.mii.lt

Abstract. The task of market basket analysis is one of the oldest areas of data mining, but still remains very relevant in today's market. Supermarkets have enormous amounts of data about purchases and it is always important to know what items the market basket contains, how it fluctuates, whether it depends on a particular season, etc. In order to solve these tasks various data mining methods and algorithms are applied. One of them is discovering association rules. The article introduces the modified stochastic algorithm for mining frequent subsequences, as well as computer modeling results and conclusions are presented. The essence of the modified stochastic algorithm is to quickly discover frequent subsequences based on the 1-element subsequence discovered by the Apriori algorithm. In the algorithm the database is scanned once, frequent subsequences and association rules are discovered. The confidence of the algorithm is estimated applying probability statistical methods.

Keywords: frequent subsequence, association rule, Apriori algorithm, modified stochastic algorithm for mining frequent subsequences.

1 Introduction

One of the oldest, but even these days very relevant data analysis task is the market basket task [2-4, 8, 12, 17, 19, 22-24]. Every supermarket stores data about customer transactions, analysis and processing of which give new knowledge i.e. answer a lot of relevant questions. It is important for supermarkets to know the items that comprise the main market basket, what items are most frequent and etc. Discovering new knowledge (data analysis task) consists of some steps: data selection; data preparation for analysis; application of algorithms to discover knowledge; presentation of new knowledge.

In order to solve the tasks data mining is used. Data mining is research and analysis of large amounts of data using automated or semi-automated methods in order to find important relation between data, discover models and association rules [16]. Data mining is defined as the method of acquisition, tracking and discovering of new meanings in data [18]. This process is applied in business, medicine and other spheres where large amounts of data should be analyzed and relation between data found i.e. new knowledge from large amounts of data discovered. The aim of data mining is to discover new relevant knowledge from large amounts of data. Data mining helps

analysts to better use available data and make more accurate decisions quicker minimizing the probability of errors. The most important feature of data mining is to select the most important data from dozens of other reducing time [25] and technological resources and optimizing the results obtained [9]. All algorithms used for frequent sequence mining could be classified in two groups: exact algorithms and approximate algorithms. Exact frequent sequence mining algorithms read the whole database many times, and if the database is very large, then frequent sequence mining is not compatible with limited availability of computer resources and real time constraints. In some cases where a precise result is required, then exact methods are irreplaceable by any approximate methods, because the precision is very important. Exact frequent sequence mining algorithms are usually used to solve genetic tasks. In order to solve the market basket task approximate methods are used. Approximate methods are very popular in analyzing data where the computation speed of the algorithm is more important than the accuracy. The approximate sequence mining algorithms were proposed recently which are much faster than exact algorithms, but do not have the theoretical error estimation of algorithm results therefore a number of empirical tests are required to be performed in order to estimate the empirical evidence of errors made identifying frequent sequences. The estimation of errors made identifying frequent sequences gives advantage for data analysis because then it is known what precision might be used when finding frequent sequences in large databases. An effective and efficient algorithm is based on the PAC (Probably Approximate Correct) learning theory to measure and estimate sample error [5]. A popular solution is to apply MRA (Multi - Resolution Analysis) and Shannon's theorem, to quickly obtain acceptably approximate association rules with an appropriate sample size [5]. When solving the problems of discovering association rules, it is important to estimate which subset is frequent, but not to estimate the exact quantity of frequent subsets. The problems of frequent item set mining can be solved using the MDL (Minimum Description Length) principle: the best set of frequent item sets is that set that compresses the database best [17].

One of the most popular algorithms, used to solve this type of problems, is the Apriori algorithm [27] and its modifications AprioriAll, AprioriSome, DynamicSome algorithms [14, 15].

The Apriori algorithm is interactive and it counts a certain amount of sets by scanning the database. The basic idea of sequence search algorithms is that frequent sequences are searched eliminating infrequent subsequences of possible frequent sequences. The principle of the algorithm for discovering association rules is analysis of frequent data sets. First, the search for frequent elements is performed, and then sets-candidates are generated from these elements. In order to shorten association rule search the apriority property is used, i.e. if itemset X is not frequent, then adding any new element A to this itemset does not make the itemset X frequent. If X is not frequent, then $X+A$ is not frequent as well. Frequent one element itemsets are found in the first step of the Apriori algorithm step. Performing this step all the database is scanned and it is estimated how often each element occurs in the database, later only the elements that meet the minimum occurring support are processed. Other steps of the algorithm consist of two parts: generating potentially frequent itemsets and determining the frequent

candidate itemsets [1]. The Apriori algorithm generates the following step candidate itemsets only from the frequent itemsets obtained in the previous step. The main intuition is that any frequent subset of the itemset must be a frequent itemset. Therefore, each candidate itemset containing k elements is generated by joining the frequent itemsets containing $(k-1)$ elements that meet the minimum occurring support. The function of the generation of candidates is significant in this algorithm. Performing the generation of candidates the database is disregarded. In order to get k element itemsets $(k-1)$ element itemsets, which were frequent in the previous step, are used. Each candidate C_k is generated adding frequent $(k-1)$ element itemset to another frequent $(k-1)$ element itemset element. After the generation of itemsets is performed, new candidates that meet the set minimum occurring support are examined.

The article introduces the modified stochastic algorithm for mining frequent subsequences (SMFS1-algorithm), computer modeling results on real market database and conclusions. The SMFS1-algorithm is approximate algorithm. This algorithm is an upgraded version of our earlier created stochastic algorithm [20, 21]. Performing the algorithm the database is scanned once, frequent subsequences are discovered and association rules are designed. The confidence of this algorithm is estimated applying probability statistical methods. The SMFS1-algorithm is suitable for solving the tasks of market basket analysis, service quality, genetics, etc.

2 Description of Association Rules

Data mining is discovery of unknown, nontrivial, practically useful and easy to interpret knowledge in chaotic data. This knowledge is required for decision-making in various fields. The information found in the application of data mining techniques is not known in advance. Knowledge is described by relationships of new features that distinguish one attribute value from other set attributes. The new knowledge set must be applied to new information with some degree of confidence [11]. The new knowledge must be understandable to a consumer. For example, a person easily understands the logical structure “If..., then...”. Moreover, these rules can be used in a variety of database management systems as SQL queries. Association rules allow us to determine the relationship between events [26]. Various data mining methods and algorithms i.e. neural networks, decision trees, clustering, discovering association rules and etc. are used to solve such tasks.

For example, a customer who purchased bread will buy milk with 72% probability.

First time, the discovered association rules were applied to set a typical market basket.

Let $I = \{i_1, i_2, \dots, i_n\}$ be a set of items. Let D be a database of transactions, where each transaction T consists of a set of items such that $T \subseteq I$. Given an itemset $X \subseteq I$, transaction T contains X if and only if $X \subseteq T$ [1, 6, 19].

Definition 1. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$ [6, 28].

Definition 2. The association rule $X \Rightarrow Y$ holds in D with confidence $conf$ if the probability of a transaction in D which contains X also contains Y is $conf$ [6, 23, 28].

Definition 3. The association rule $X \Rightarrow Y$ has support $supp$ in D if the probability of a transaction in D contains both X and Y is $supp$ [1, 5, 6, 23, 29].

Definition 4. Confidence $conf$ of the association rule $X \Rightarrow Y$ is called a value [5, 6, 23] using the following formula (1):

$$conf(X \Rightarrow Y) = \frac{supp|X \cup Y|}{|suppX|}. \quad (1)$$

Confidence of the association rule $X \Rightarrow Y$ value $conf(X \Rightarrow Y)$ indicates the part of items with feature X that also has $X \cup Y$. Often support of an association rule is expressed as a percentage [23]. Discovering of association rules min_supp and min_conf values are defined in [13, 14].

Discovering of association rules consists of two steps [30]:

1. Discovering of frequent itemsets, i.e. discovering of items or features when support is not less than the identified min_supp value.
2. Creation of an association rule according to identified frequent itemsets.

The values of parameters min_supp and min_conf are selected so as to restrict the number of association rules [12]. If these values of parameters are very high, the algorithm will find such association rules that are clear and well known. If these values of parameters are very small, it will generate a large amount of association rules, and this requires considerable technical and time resources.

For example, 75% transactions which include bread also include milk. 3% of the whole transaction database include both items. 75% is confidence and 3% is support of the association rules.

3 Modified Stochastic Algorithm for Mining Frequent Subsequences

3.1 Description of the Algorithm

The SMFS1-algorithm can be used to determine association rules for the analysis of the market basket, genetic tasks, and so on. The goal of the designed SMFS1-algorithm is to determine frequent subsets in large databases and to highlight association rules.

Let us analyze an M -length database D . Namely, randomly selected random length l subsets, containing at least one frequent element, determined by the Apriori algorithm [7], are analyzed. Assume that the subset length analyzed is distributed according to the geometrical distribution with the parameter q , and the spacing between the two subset lengths is also distributed according to the geometrical distribution with the parameter p [9, 20].

It is easy to calculate that the average subset length analyzed is $l=q/(1-q)$, and the average length of the gap between adjacent subsets is equal to $t=p/(1-p)$. Let us

randomly choose N (number of samples) subsets of various lengths for analyzing database D . Subset frequencies c_i of the appropriate length are calculated using the following formula (2):

$$c_i = N_i / N, \quad (2)$$

where $i=1, 2, \dots, n$, N_i is the number of an appropriate length subset, N is the number of all subsets, i is the length of a subset and n is the maximum length of a subset.

3.2 The Pseudo-code of the Modified Stochastic Algorithm

The algorithm consists of the following steps: the transaction database is transferred to the text file; initial values of parameters p and q are input; frequent 1-itemsets found by the Apriori algorithm are input; subsets of each transaction are selected; results are output. The result is subsets ordered by length and support. The developed SMFS1-algorithm is illustrated by the pseudo-code:

```

Lk := {file}; // The data are transferred to the text file
Lk, reading the file
L1 := {frequent 1-itemsets found by the Apriori algorithm}
Input p-the initial value of the parameter p, p ∈ [0; 1]
Input q-the initial value of the parameter q, q ∈ [0; 1]
k:=2; //k represents the pass number
while (Lk-1 <> ∅) do
begin
If p>q then
Ck := Apriori_gen(Lk-1); //The new candidates of size k,
which include at least one 1-itemset found by the Apriori
algorithm
tk := p/(1-p) //The value of omitted elements
For all transactions t ∈ D do
begin
Ct := subset (Ck, t); //For all candidates in t transaction
For each (candidates c ∈ Ct)
pn := assigning a new value p
qn := assigning a new value q
end
else
tk := p/(1-p) //The new value of omitted elements.
pn := assigning a new value p
qn := assigning a new value q
end
Result := ∪k Ck. //Subsets ordered by length and support.

```

3.3 Statistical Characteristics of Modified Stochastic Algorithm

Confidence Probability Range. Let us analyze two independent subset samples, their sizes being n_1 and n_2 , and the most frequent subset occurs k_1 times in the first sample, and in the second sample- k_2 times.

The null hypothesis states that the most frequent subset proportions are identical (3) in the database from which the samples are taken and, in the opposite case the probabilities are unequal (4):

$$H_0: p_1 = p_2 \tag{3}$$

$$H_1: p_1 \neq p_2. \tag{4}$$

No matter we accept or reject the hypothesis H_0 , there are two possible types of errors. They are called the first and second type errors. The first type error is: the hypothesis H_0 is rejected when it is true. The second type error is: the hypothesis H_0 is accepted when it is false. The SMFS1-algorithm is approximate; therefore the first and second type errors are possible. The subset $c_{i1}, c_{i2}, \dots, c_{ik}$ is determined.

The first type error is when the subset is frequent, but the SMFS1-algorithm does not recognize as frequent.

The second type error is when the subset is not frequent, but the SMFS1-algorithm attributes it to frequent subsets.

The interval $[p_1; p_2]$ is called a confidence interval of the parameter p , if $P(p_1 < p < p_2) = \alpha$, the number α is called a confidence level, p_1 and p_2 are called confidence probability ranges. The criterion of accuracy of the SMFS1-algorithm is a probability bound interval.

The bounds of confidence probabilities are estimated according to the formulas (5) and (6) below:

$$p_1 = 1 - \text{BetaInv}\left(\frac{1-\alpha}{2}, n-k, k+1\right); \tag{5}$$

$$p_2 = 1 - \text{BetaInv}\left(1 - \left(\frac{1-\alpha}{2}\right), n-k+1, k\right). \tag{6}$$

N is the number of all fragments, K is the number of appearance of the fragment, BetaInv is a beta quintile of the distribution.

The accuracy of the algorithm was also evaluated comparing the results with that derived by the Apriori algorithm. The first type error of the SMFS1-algorithm is about 2,63 %, the second type error - 6,09 %. The SMFS1-algorithm processing time is shortest, i.e. 140 s, the Apriori – 736 s. The confidence probability bound of the SMFS1-algorithm, criteria statistics and assumption evaluation are presented in earlier works [20], thus, the interval of the algorithm confidence is [0,9537; 0,9993].

Criterion Statistics (labeled by letter u). We have two independent subset samples with their sizes being n_1 and n_2 . In the first sample there occur k_1 and in the second - k_2 elements with necessary attribute value.

Criterion hypothesis ($H_0: p_1 = p_2$) on the attribute detection probability equality statistics in the samples can be estimated in various ways. Criterion statistics u is constructed so that when the hypothesis H_0 is true, it would be distributed according

to the standard normal distribution. Criterion statistics u [10] is estimated according to this formula (7):

$$u = \frac{d_1 - d_2}{\sqrt{\left(\frac{k_1 + k_2}{n_1 + n_2}\right) \cdot \left(1 - \frac{k_1 + k_2}{n_1 + n_2}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \quad (7)$$

If d is labeled $d = (k_1 + k_2)/(n_1 + n_2)$, the formula is as follows (8):

$$u = \frac{d_1 - d_2}{\sqrt{d \cdot (1-d) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \quad (8)$$

Criterion statistics z can be also estimated this way [10], i.e. using the following formula (9):

$$z = (2 \arcsin \sqrt{d_1} - 2 \arcsin \sqrt{d_2}) \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}. \quad (9)$$

Assumption Evaluation. After criterion statistics is estimated, assumption of probability evaluation is performed. When alternative is double ($H_1: p_1 \neq p_2$), the obtained value u , corresponding value P , is calculated as follows (10):

$$P = 2 \cdot (1 - \text{NORMSDIST}(\text{ABS}(u))). \quad (10)$$

p -value means probability risk, when rejecting H_0 first type error is made, therefore H_0 can only be rejected only if obtained p -value is very small, negligible, less than usual standard probability values (0,1; 0,05; 0,01 or 0,001). Hypothesis H_0 probability, that the given statement corresponds reality, is expressed by P -value. Therefore, the bigger P -value, the more confident the null hypothesis is.

In order to detect sequence characteristics of the change point, a model, in which the most frequent fragment (market basket) is determined by the maximum probability method, is designed. The length of the most frequent fragment (market basket) is determined according to the monotonuos rule – the subset of frequent subsequence is frequent subsequence. In the method, that determines sequence characteristics of the change point, the support of the most frequent fragment (market basket) according to its length is calculated. Further, taking advantage of the fact that as long the fragment (market basket) length is less than a certain length, the support of the analyzed fragment (market basket) is almost constant, and when the analyzed fragment (market basket) length is greater than a certain most frequent fragment (market basket), the probability starts to decrease. In this case, the binary process, the value of which is equal to one, is concluded if the statistical criterion is not contrary to the hypothesis about probability conjunction of appearance of two adjacent length fragments (market baskets) and equals zero, if probabilities of appearance of two adjacent length fragments (market baskets) significantly differ. The change point of the binary process characteristics allows us to set the size of a market basket, i.e. the number of items it consists of. Since the number of appearances in the samples of two adjacent length fragments (market baskets) is distributed according to the binomial law, the function of the logarithmic probability of the most frequent fragment (market

basket) to determine the change in probabilities can be designed using the following formula (11):

$$f(k) = \frac{k!}{k_1!(k-k_1)!} \cdot \frac{(N-k)!}{(K-k_1)!(N-k-K+k_1)!} \cdot p_1^{k_1} \cdot (1-p_1)^{k-k_1} \cdot p_2^{K-k_1} \cdot (1-p_2)^{N-k-K+k_1} \quad (11)$$

The function of the logarithmic probability is derived using the following formula (12):

$$\ln(f(k)) = \sum_{i=1}^k \ln i - \sum_{i=1}^{k_1} \ln i - \sum_{i=1}^{k-k_1} \ln i + \sum_{i=1}^{N-k} \ln i - \sum_{i=1}^{K-k_1} \ln i - \sum_{i=1}^{N-k-K+k_1} \ln i + \sum_{i=1}^{k_1} \ln p_1 + \sum_{i=1}^{k-k_1} \ln(1-p_1) + \sum_{i=1}^{K-k_1} \ln p_2 + \sum_{i=1}^{N-k-K+k_1} \ln(1-p_2), \quad (12)$$

k - the change point of the binary process characteristics, k_1 - the number of probability support conjunction till the change point, k_2 - the number of probability support conjunction after the change point, N - the maximum length of the sequence. The length of the most frequent fragment (market basket) corresponds to the minimum of the function of the logarithmic probability. It is convenient to introduce the minimizing function which is the difference of two adjacent values of the function and is equal to (13):

$$\ln(f(k)/f(k-1)) = \ln k - \ln(k-k_1) - \ln(N-k+1) + \ln(N-k-k_1+1) + \ln(1-p_1) + \ln(1-p_2) \quad (13)$$

if k - value of the binary process is equal to 0. If this value equals 1, then the difference of adjacent values of the probability function is equal to (14):

$$\ln(f(k)/f(k-1)) = \ln k - \ln k_1 - \ln(N-k+1) + \ln(K-k_1+1) + \ln p_1 + \ln p_2, \quad (14)$$

where $p_1 = \frac{k_1}{k}$, $p_2 = \frac{k_2}{k}$. The minimum of the probability function coincides with the first value of the variable k , where the difference of two adjacent function values is positive. The calculation starts from the initial value $k=0$. It is noteworthy that the initial probability function value is (15):

$$\ln(f(0)) = \sum_{i=1}^N \ln i - \sum_{i=1}^K \ln i - \sum_{i=1}^{N-K} \ln i + K \cdot \ln p_2 + \ln(1-p_2) \cdot (N-K). \quad (15)$$

For calculation of the values of the logarithmic probability function recursive formulas can be used the following formula (16):

$$k_1(k+1) = k_1(k), \quad k_2(k+1) = k_2(k), \quad (16)$$

if k -value of the binary process is equal to zero and (17):

$$k_1(k+1) = k_1(k)+1, \quad k_2(k+1) = k_2(k)-1, \quad (17)$$

if k -value of the binary process is equal to one.

For example, there is 10 transaction database {ABCDE; AC; ABC; BC; AC; ACD; ACDE; CDE; ABC; AB}. Each transaction includes not more than 5 items. The statistical characteristics of SMFS1-algorithm are calculated using formulas (7-17) and presented in Table 1:

Table 1. The statistical characteristics

<i>Subset length</i>	<i>Frequent item</i>	<i>supp</i>	<i>u</i>	<i>z</i>	<i>P</i>	<i>k</i>
1	C	0,9	-	-	-	-
2	AC	0,7	1,118	1,153	0,264	0
3	ABC	0,3	1,789	1,841	0,073	1
4	ABCD	0,1	1,118	1,153	0,263	0

This algorithm allows us to combine two important criteria i.e. time and accuracy.

4 Computer Modeling

The aim of the experiment is to determine what size the most frequent market basket is and highlight association rules. We have real one-week database of purchase transactions which consists of 10 000 transactions. All analyzed elements in the database are homogeneous (e.g. milk, mineral water, juice and etc.). There are 25 different titles of the items in the database. The fragment of this database is shown in Table 2.

Table 2. The fragment of transaction database

<i>Transaction number</i>	<i>Item title</i>	<i>Quantity</i>
...
1001	I	1
1001	J	1
1001	T	1
...
1002	A	2
1002	C	2
...

The data are transferred to a text file, where each row corresponds to a particular transaction and the items are in alphabetical order. The fragment of the text file is presented in Table 3.

Table 3. The fragment of the text file

<i>ABCDEFGHIJKLMRSTUV</i>
<i>ACEGIKM</i>
<i>ABTUV</i>
.....
<i>ABCDEF</i>
<i>CDEFGHIJKLMRST</i>
.....



This file is processed by the modified stochastic algorithm, when $min_supp=50$, $min_supp=100$, $min_supp=150$; $min_supp=200$, $min_supp=300$, $min_supp=400$, $min_supp=500$, $min_supp=600$. The average processing time of the algorithm is 2 min. 20 s.

During the experiment estimated probability characteristics of the SMFS1-algorithm are $u, z, P(0)$.

Two independent samples of subsequences are analyzed, their sizes are n_1 and n_2 and in the first sample the most frequent subsequence occurred k_1 times, and in the second - k_2 times. Criterion u statistical average variation when subsequence length varies from 3 to 10 is shown in Fig. 1.

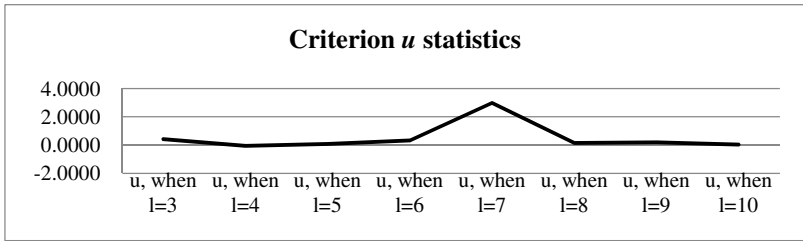


Fig. 1. Criterion u statistical average variation

Criterion statistics u of the support conjunction of two adjacent length subsequences, estimated by the SMFS1-algorithm, showed that the most frequent market basket consists of 6 items.

Criterion z statistical average variation, when subsequence length varies from 3 to 10 is presented in Fig. 2.

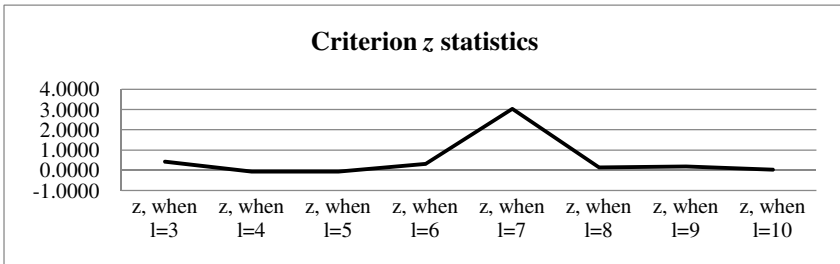


Fig. 2. Criterion z statistical average variation

Criterion statistics z of the support conjunction of two adjacent length subsequences, estimated by the SMFS1-algorithm, showed that the most frequent market basket consists of 6 items.

The average variation of the probability function $P(0)$, when subsequence length varies from 3 to 10 is presented in Fig. 3.



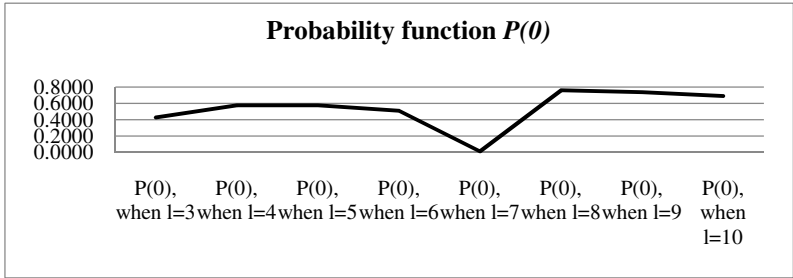


Fig. 3. The average variation of the probability function P(0)

Our algorithm minimizes the probability with $i=7$ as seen in Fig. 3. Values of the probability function $P(0)$, estimated by the SMFS1-algorithm, showed that the most frequent market basket consists of 6 items.

Frequent subsequences were estimated by the experiment. The obtained number of frequent subsequences, defining min_supp values, is presented in Table 4.

Table 4. The number of frequent subsequences

The number of elements contained in the association rule	min_supp value							
	50	100	150	200	300	400	500	600
2-element subsequences	34	30	20	19	18	14	9	5
3-element subsequences	44	31	19	18	17	12	6	0
4-element subsequences	50	31	19	17	15	9	5	0
5-element subsequences	55	33	18	15	11	8	4	0
6-element subsequences	64	29	18	12	11	6	2	0
7-element subsequences	71	28	15	11	10	4	1	0
8-element subsequences	69	23	13	11	7	3	0	0
9-element subsequences	74	24	13	9	8	1	0	0
10-element subsequences	72	23	12	6	6	0	0	0
Total	533	252	147	118	103	57	27	5

Association rules are discovered by estimating min_supp and min_conf values. Association rules when $min_conf \geq 50\%$ are presented in Table 5.



Table 5. Association rules

<i>Association rule</i>	<i>Association rule support</i>	<i>Association rule confidence</i>
M \Rightarrow MP	84 %	55 %
K \Rightarrow KL	78 %	78 %
T \Rightarrow ST	55 %	65 %
ST \Rightarrow RST	70 %	80 %
MP \Rightarrow MPR	63 %	58 %
LM \Rightarrow LMP	63 %	87 %
ST \Rightarrow RSTU	62 %	68 %
ST \Rightarrow STUV	59 %	75 %
MP \Rightarrow KLMP	61 %	85 %
KL \Rightarrow KLMP	64 %	80 %
LM \Rightarrow KLMP	57 %	60 %
PR \Rightarrow MPRS	49 %	69 %
ST \Rightarrow MPRST	54 %	65 %
ST \Rightarrow PRSTU	54 %	60 %
KL \Rightarrow KLMPR	52 %	78 %
PR \Rightarrow MPRST	53 %	55 %

In the database of purchase transactions, which consists of 10 000 transactions, there are 16 association rules, the confidence of which is greater than 50%, estimated.

5 Conclusions

The SMFS1-algorithm is suitable for analyzing large databases. The SMFS1-algorithm, when scanning the database subsequences, which include at least one of the Apriori algorithm with frequent 1-element subset, are taken randomly, makes an initial random sample of the database. Analyzing the random sample statistical conclusions about frequent subsequences and association rules in the initial database are made. During computer modeling the change of probability function and assumption of probabilities are estimated. The results revealed that analyzing a particular database of transactions D, the most frequent market basket consists of 6 items and in the database there are 16 association rules, the confidence of which is greater than 50%, estimated.

In the future, we plan to further extend this work: to estimate interesting criteria as well as to adapt the algorithm scanning the database for several times and to research cases where parameter p is not distributed by geometric law.

Acknowledgments. The authors thank the referees for the valuable comments and suggestions that contributed to significantly improve the quality of the paper.

References

1. Ayres, J., Flannick, J., Gehrke, J., Yiu, T.: Sequential Pattern mining using a bitmap representation. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 429–435. ACM Press, Edmonton (2002)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499. Morgan Kaufmann, Santiago de Chile (1994)
3. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216. ACM Press, Washington, D.C. (1993)
4. Brin, S., Motwani, R., Silverstein, C.: Market Baskets: Generalizing Association Rules to Correlations. In: Proceedings ACM SIGMOD International Conference on Management of Data, pp. 265–276. ACM Press, Tucson (1997)
5. Cai-Yan, J., Xie-Ping, G.: Multi-scaling sampling: An adaptive sampling method for discovering approximate association rules. *Journal of Computer Science and Technology* 20, 309–318 (2005)
6. Cheung, D.W., Han, J., Ng, V.T., Fu, A.W., Fu, Y.: A Fast Distributed Algorithm for Mining Association Rules. In: Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems, pp. 31–43. IEEE Computer Society, Miami Beach (1996)
7. Cho, C.-W., Wu, Y.-H., Chen, A.L.P.: Effective Database Transformation and Efficient Support Computation for Mining Sequential Patterns. In: Zhou, L.-Z., Ooi, B.-C., Meng, X. (eds.) DASFAA 2005. LNCS, vol. 3453, pp. 163–174. Springer, Heidelberg (2005)
8. Coenen, F., Goulbourne, G., Leng, P.: Tree Structures for Mining Association Rules. In: *Data Mining and Knowledge Discovery*, vol. 8, pp. 25–51. Kluwer Academic Publishers (2004)
9. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: *Introduction to Algorithms*, 3rd edn. The MIT Press, Cambridge (2009)
10. Cekanavicius, V., Murauskas, G.: *Statistika ir jos taikymai*. TEV, Vilnius (2000)
11. Gharib, T.F., Nassar, H., Taha, M., Abraham, A.: An efficient algorithm for incremental mining of temporal association rules. In: *Data & Knowledge Engineering*, vol. 69, pp. 737–880. North-Holland (2010)
12. Gyenesei, A., Teuhola, J.: Probabilistic Iterative Expansion of Candidates in Mining Frequent Itemsets. In: Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Melbourne, Florida, USA, pp. 192–195 (2003)
13. Huanyin, Z., Jinsheng, L.: The Research of A-Priori Algorithm Candidates Based on Support Counts. In: *International Conference on Information Technology and Computer Science*, pp. 192–195. TBD, Kiev (2009)
14. Inokuchi, A., Washio, T., Motoda, H.: An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In: Zighed, D.A., Komorowski, J., Zytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 13–23. Springer, Heidelberg (2000)
15. Pallavi, D.: Association Rule Mining on Distributed Data. *International Journal of Scientific & Engineering Research* 3, 1–6 (2012)
16. Rasoulilian, M., Saeed, A.: The Effect of Data Mining Based on Association Rules in Strategic Management. *Journal of Basic and Applied Scientific Research*, 1742–1748 (2012)
17. Raorane, A.A., Kulkarni, R.V., Jitkar, B.D.: Association Rule – Extracting Knowledge Using Market Basket Analysis. *Research Journal of Recent Sciences* 1(2), 19–27 (2012)

18. Sandhu, P.S., Dhaliwal, D.S., Panda, S.N.: Mining utility-oriented association rules: An efficient approach based on profit and quantity. *International Journal of the Physical Sciences* 6(2), 301–307 (2011)
19. Savasere, A., Omiecinski, E., Navathe, S.: An Efficient Algorithm for Mining Association Rules in Large Databases. In: *Proceedings of the 21st International Conference on Very Large Databases*, Zurich, Switzerland, pp. 432–444 (1995)
20. Savulioniene, L., Sakalauskas, L.: Statistical algorithm for mining frequent sequences. *Information Sciences* 58, 126–143 (2011)
21. Savulioniene, L., Sakalauskas, L.: Stochastic algorithm for mining frequent sequences. *Journal of Young Scientists* 4(33), 138–145 (2011)
22. Siebes, A., Vreeken, J., Leeuwen, M.: Item Sets That Compress. In: *Data Mining and Knowledge Discovery*, vol. 23, pp. 169–214 (2011)
23. Srikant, R., Agrewal, R.: Mining generalized Association Rules. In: *Proceeding VLDB 1995 Proceedings of the 21st International Conference on Very large Data Bases*, San Francisco, CA, USA, pp. 407–419 (1995)
24. Toivonen, H.: Sampling Large Databases for Association Rules. In: *Proceedings of the 22nd International Conference on Very Large Databases*, Mumbai, India, pp. 134–145 (1996)
25. Thomas, S., Bodagala, S., Alsabti, K., Ranka, S.: An efficient Algorithm for Incremental Updation of Association Rules in Large Database. In: *Proceedings of 3rd International Conference on KDD and data mining (KDD 1997)*, Newport Beach, California, pp. 263–266 (2007)
26. Umarani, V., Punithavalli, M.: A study on effective mining of Association Rules from huge Databases. *International Journal of Computer Science and Research* 1, 30–34 (2010)
27. Wang, H., Liu, X.: The Research of Improved Association Rules Mining Apriori Algorithm. In: *Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 961–964. IEEE, Shanghai (2011)
28. Yang, J., Zhao, C.: Study on the Data Mining Algorithm Based on Positive and Negative Association Rules. *Computer and Information Science* 2, 103–106 (2009)
29. Zaki, M.J.: SPADE: An Efficient Algorithm for Mining Frequent Sequences. In: *Machine Learning*, vol. 42, pp. 31–60 (2001)
30. Zaki, M.J., Hsiao, C.: CHARM: An Efficient Algorithm for Closed Association Rule Mining. *International Journal of Intelligent Systems Technologies and Applications* 4, 313–326 (2008)

On Two Approaches to Constructing Optimal Algorithms for Multi-objective Optimization

Antanas Žilinskas*

Institute of Mathematics and Informatics,
Vilnius University, Vilnius, Lithuania
antanas.zilinskas@mii.vu.lt

Abstract. Multi-objective optimization problems with expensive, black box objectives are difficult to tackle. For such type of problems in the single objective case the algorithms, which are in some sense optimal, have proved well suitable. Two concepts of optimality substantiate the construction of algorithms: worst case optimality and average case optimality. In the present paper the extension of these concepts to the multi-objective optimization is discussed. Two algorithms representing both concepts are implemented and experimentally compared.

Keywords: Multi-objective optimization, global optimization, optimal algorithms, statistical models.

1 Introduction

Nonlinear multi-objective optimization is very active research area. Depending on the properties of a multi-objective optimization problem, different approaches to its solution can be applied. The best direction developed is optimization of convex problems; for the latter problems the methods that generalize the ideas of classical mathematical programming suit well [9]. For the problems with not so nice objectives, metaheuristic methods are frequently favorable [3], [11]. However, there remains a class of important problems without sufficient attention of researchers: namely, the problems with black-box, multimodal, and expensive objectives.

Generally speaking, the solution to a multi-objective optimization problem

$$\min_{x \in A} f(x),$$
$$f(x) = (f_1(x), \dots, f_m(x))^T, \quad A \subseteq R^d, \quad (1)$$

can be described as a set of objective vectors which well represents either the set of Pareto optimal solutions $P(f)$, or its favorable subset; for a rigorous analysis we refer to [9]. The following two cases can be pointed as extreme: the approximation (discrete representation) of the whole Pareto optimal set, and an

* The support by the Research Council of Lithuania under Grant No. MIP-063/2012 is acknowledged.

objective vector sufficiently close to a desirable one. In the real world applications, the notion of solution can change. Frequently, in the starting optimization phase, a rough approximation of the whole Pareto optimal set is of interest; in the intermediate phase, a subset of the Pareto optimal set of interest is intended to be approximated more precisely; finally, a specific Pareto optimal solution is sought. A similar strategy is also justified in single-objective global optimization: starting from a uniform search over the feasible region, concentrating the search in prospective subregions, and finishing with a local algorithm chosen according to the local properties of the objective function. Therefore the use of similar approaches to the construction of algorithms for global optimization in single- and multi-objective cases is quite natural.

We focus on the problems where objective functions are expensive because of the complexity of the computational model; expensiveness here means long lasting computation of a value of the objective function. The complexity of computations normally implies not only the expensiveness of the objective functions but also the uncertainty in its properties. Such unfavorable from the optimization point of view properties as non-differentiability, non-convexity, and multimodality can not be excluded. The limitation in collecting general information about the function, and particularly about its minima, strongly requires rationality in distribution of the points where to compute the objective function values. Therefore the algorithms, justified by the principles of rational decision theory, are of especial interest. The class of Lipschitz continuous functions is one of the most widely used models for single objective non-convex optimization [6]. The worst-case optimality is the standard concept in the analysis of the algorithms' optimality with respect to a deterministic model of problems/data [1]. The worst-case optimal optimization algorithms for Lipschitz continuous functions have been investigated by Sukharev [16], [17], who has shown that the optimal passive algorithm is coincident with the optimal adaptive algorithm and can be reduced to the covering of the feasible region by the balls of the minimum radius. The worst-case optimal search can be interpreted as an antagonistic game in terms of the game theory where, for the current point selected by the search algorithm, an adversary defines the most inappropriate values of the objective functions. The most inappropriate (not informative) values of the objective function are equal for all the points selected. When an ordinary optimization problem is considered, the assumption about a rational adversary, selecting the most inappropriate function values at all optimization steps, seems not very realistic. Assuming the adversary semi-rational, Sukharev has proposed in [16], [17] a *best sequential algorithm* where the actual information at the current iteration is taken into account but the worst-case information is supposed for the subsequent iterations; unfortunately this theoretically interesting algorithm was too complicated for practical implementation. A one-step worst-case optimal algorithm, frequently referenced as the algorithm by Pijavskij-Shubert, is much simpler [12], [13]. The results concerning worst-case passiv/adaptive algorithm are generalized to the case of multi-objective optimization in [22]. In the present paper we consider a multi-objective analogue of the Pijavskij-Shubert algorithm.

In black box single objective optimization of expensive functions the statistical models of multimodal functions have proven very helpful [10], [14], [15]. Recently several papers have been published which propose multi-objective optimization algorithms generalizing single-objective optimization algorithms based on statistical models of objective functions. The proposed algorithms are straightforward generalizations of the single-objective prototypes, and their theoretical analysis is absent. In the present paper we discuss a new approach to the construction of global optimization algorithms optimal in the sense of the rational decision theory [4] where uncertainty is modeled by statistical models of multimodal functions. Two bi-objective optimization algorithms representing both approaches discussed are implemented and tested. The results of numerical experiments are included to illustrate the performance of the implemented algorithms.

2 A Bi-objective One-Step Worst-Case Optimal Algorithm

A special case of (1) is considered where $m = 2$, $d = 1$, and objective functions belong to the class of Lipschitz functions $\Phi(L)$:

$$f(x) = (f_1(x), f_2(x))^T \in \Phi(L),$$

$$|f_k(x) - f_k(t)| \leq L_k \cdot \|x - t\|, \quad k = 1, 2, \tag{2}$$

for $x \in A$, $t \in A$, $L = (L_1, L_2)^T$, $L_k > 0$, $k = 1, 2$, and A is supposed to be a bounded closed interval. The class of Lipschitz continuous functions is advantageous for constructing global optimization algorithms because of relatively simply computable lower bounds for the function values. The Pijavskij-Shubert algorithm is one step optimal in a sense that the current value of an objective function is computed at the point which is found by the minimization of the gap between the current minimum of computed values, and the Lipschitzian lower bound. The availability of such bounds enables also the assessment of the quality of a discrete representation of the Pareto front for bi-objective Lipschitz optimization. Similarly to the Pijavskij-Shubert algorithm, we propose a one-step optimal multi-objective optimization algorithm where a current point for computing $f(\cdot)$ is the minimizer of the gap between the current simple upper bound for the Pareto front $P(f)$, and Lipschitzian lower bound for $P(f)$.

Definition 1. Let us denote $F^n = (f^1, \dots, f^n)^T$, $f^i = f(x_i)$, $i = 1, \dots, n$. The subset of $\bigcup_{i=1}^n \{z : z \in R^2, z \geq f^i\}$ which consists of weakly Pareto optimal solutions is called a simple upper bound for $P(f)$ and is denoted by $U(F^n) \in R^2$.

As follows from (2) the functions $g_k(x)$, $k = 1, 2$, define the lower bounds for $f_k(x)$, $x \in A = [a, b]$:

$$g_k(x) = \max (f_k^{o_i} - L_k(x - x_{o_i}), f_k^{o_{i+1}} - L_k(x_{o_{i+1}} - x)), \tag{3}$$

$$x_{o_i} \leq x \leq x_{o_{i+1}}, i = 1, \dots, n - 1,$$

where x_{o_i} , $i = 1, \dots, n$, denote the increasingly ordered points x_i , and $f_k^{o_i}$ denote the corresponding values of the objective functions.

Definition 2. *The Pareto front of the bi-objective problem*

$$\min_{x_{o_i} \leq x \leq x_{o_{i+1}}} g(x), \quad g(x) = (g_1(x), g_2(x))^T, \tag{4}$$

is called a local Lipschitz lower bound for $P(f)$, and it is denoted as V_i .

The idea of the algorithm is to tighten the Lipschitz lower bounds for the non-dominated solutions, and to indicate the subintervals of $[a, b]$ with dominated objective vectors. Let us consider the $n + 1$ optimization step where $x_{o_i}, f^{o_i}, i = 1, \dots, n$, are known. The gap ε_n between $V(F^n)$ and $U(F^n)$ where

$$\varepsilon_n = \max_{Y \in V(F^n)} \min_{Z \in U(F^n)} \|Y - Z\|, \tag{5}$$

can be easily computed since $V(F^n)$ consists of the non-dominated solutions of the set $\bigcup_{i=1}^{n-1} V_i$. Let the maximum gap be attained at the current iteration at V_j . Then $[x_{o_j}, x_{o_{j+1}}]$ is subdivided, and the point of subdivision x_{n+1} is the current worst-case optimal decision. In other words, the proposed algorithm computes the vector of objectives at the point x_{n+1} which minimizes the maximum of Lipschitz tolerances. It can be shown that x_{n+1} is defined by simple analytical formulae. The interval $[x_{o_j}, x_{o_{j+1}}]$ is replaced by its subintervals $[x_{o_j}, x_{n+1}]$ and $[x_{n+1}, x_{o_{j+1}}]$, and the current information is updated to prepare the next iteration. The pseudocode of the algorithm is presented in the Appendix.

3 A Generalized Version of the P-algorithm

Let us start from the analysis of a single-objective optimization problem (1) where $d = 1$. By the concept of black box optimization, the uncertainty in properties of $f(x)$ is supposed which is typical, e.g. in applications where the values of $f(x)$ are computed by unfamiliar software. To justify a search strategy in the described situation of uncertainty a "rational optimizer" should define a model of uncertainty, e.g. like a statistical model of uncertainty in the expected utility theory [4]. We focus on the statistical models of uncertainty although the other models such as fuzzy logic and rough sets also would be interesting to investigate.

Let us consider the current minimization step where n function values have been computed at previous steps: $y_i = f(x_i), i = 1, \dots, n$. A rational choice of a point for the next computation of the objective function value can not be performed without assessment of the uncertainty in the result of that computation. The only objective information on $f(\cdot)$ is $x_i, y_i, i = 1, \dots, n$. Besides that objective information normally some subjective information is available, e.g. the experience of solution of similar problems in the past. As shown in [19] the very general assumptions on rational perception of uncertainty imply a random variable model of the supposed for the computation value of the objective function, i.e. those assumptions imply a random variable ξ_x as a model of the unknown value $f(x), x \neq x_i, i = 1, \dots, n$. We refer to [15] for the the bottom-up

constructing of a computational statistical model of objective functions where the mentioned above result of the existence of the statistical model in the form of random variable has been augmented by constructive details. Such a construction of the statistical model is more advantageous than the selection of a known stochastic function for a model of objective functions. The notable exception is Gaussian-Markovian stochastic processes which conditional distribution is defined by simple formulas, and therefore they are attractive models for the construction of one-variable global optimization. The Wiener process was the first stochastic process model successfully used for construction of one-variable global optimization [2], [8], [18]. In the present paper we will also use this statistical model for the construction of the multi-objective P-algorithm. The first global optimization algorithm based on a stochastic function model was proposed in [8]. That one-variable algorithm was constructed using the Wiener process for a model. The current computation of the objective function value is performed by that algorithm at the point of maximum improvement probability:

$$x_{n+1} = \arg \max_{x \in \mathbf{A}} \mathbf{P}\{\xi(x) \leq y^{on} \mid \xi(x_1) = y_1, \dots, \xi(x_n) = y_n\}, \quad (6)$$

where $\xi(x)$ is the Wiener process accepted for a statistical model, and $y^{on} < \min_{1 \leq i \leq n} y_i$, is a value intended to improve. That algorithm was substantiated axiomatically in [20] where it was named the P-algorithm. For the generalization of the P-algorithm to the multidimensional ($d > 1$) case and its theoretical analysis we refer to [15].

Recently the P-algorithm has been extended to the case of multi-objective optimization [21]. The bi-objective P-algorithm is defined by a formula similar to (6)

$$x_{n+1} = \arg \max_{x \in \mathbf{A}} \mathbf{P}\{\Xi(x) \leq Y^{on} \mid \Xi(x_1) = Y_1, \dots, \Xi(x_n) = Y_n\}, \quad (7)$$

where $\Xi(x) = (\xi_1(x), \xi_2(x))^T$ is a vector valued stochastic function the components of which are independent Wiener processes, Y^{on} is a reference vector not dominated by the observed objective vectors $Y_i = (f_1(x_i), f_2(x_i))^T$. The function $P(\cdot)$, which is maximized in (7), means the probability that the Gaussian random vector $\Xi(x)$ dominates Y^{on} . $P(\cdot)$ can be expressed as the product of two Gaussian cumulative distribution functions:

$$P\left(\frac{y_1^{on} - m_1(x)}{s_1(x)}, \frac{y_2^{on} - m_2(x)}{s_2(x)}\right) = \prod_{i=1}^2 \Phi\left(\frac{y_i^{on} - m_i(x)}{s_i(x)}\right),$$

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt. \quad (8)$$

Although the model of a family of Gaussian random variables is proved well suitable for many problems, it is not universal. For example, the Gaussian model based P-algorithm has appeared not efficient for test problem (15) as shown below. The applicability of Gaussian model is also doubtful for the applied problems

similar to that considered in [7]. In the present paper we consider the assumptions concerning rationality of search that are more general than in the previous papers. The family of random vectors $\Xi(x) = (\xi_1(x), \xi_2(x))^T$, $x \in \mathbf{A}$, is accepted as a statistical model of $f(x)$. The location and spread parameters of $\xi_i(x)$, denoted by $m_i(x)$, $s_i(x)$, are essential in characterization of $\xi_i(x)$. For the more specific characterization of $\Xi(\cdot)$, e.g. by a multidimensional distribution of $\Xi(\cdot)$, the available information normally is insufficient. If the information about, e.g. correlation between $\xi_1(\cdot)$ and $\xi_2(\cdot)$ would be available, the covariation matrix could be included into the statistical model. However, in the present paper we assume that the objectives are independent, and the spread parameters are represented by a diagonal matrix $\Sigma(x)$ which diagonal elements are equal to s_1, s_2 . We assume that the utility of choice of the point for the current computation of the vector value $f(x)$ has the following structure

$$u_{n+1}(x) = U(m(x), \Sigma(x), Y^{on}), \tag{9}$$

where $m(x) = (m_1(x), m_2(x))^T$, and Y^{on} denotes a vector desired to improve.

At the current iteration, a point for computing the value of $f(x)$ is sought by an optimization algorithm which maximizes $u_{n+1}(x)$. The rationality of the multi-objective optimization algorithm is formulated as the invariance with respect to the scales of objectives. Such a rationality assumption guarantees that the optimization process runs identically for the objectives presented in different scales, and it can be expressed by the following properties of $U(\cdot)$:

$$\begin{aligned} U(m(x) + c, \Sigma(x), Y^{on} + c) &= U(m(x), \Sigma(x), Y^{on}), \quad c = (c_1, c_2)^T, \\ U(C \cdot m(x), C \cdot \Sigma(x), C \cdot Y^{on}) &= U(m(x), \Sigma(x), Y^{on}), \quad C_i > 0, \\ C &= \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}. \end{aligned} \tag{10}$$

Since the multi-objective minimization problem is considered, the small objectives function values are desirable to obtain at every iteration; therefore we postulate that for $\mu = (\mu_1, \mu_2)^T$, where $\mu_i \geq m_i$, $i = 1, 2$ and at least one inequality is strict, the following inequality is valid

$$U(m, s, y) > U(\mu, s, y). \tag{11}$$

A function which satisfies assumptions (10) has the following structure

$$U(m(x), \Sigma(x), Y^{on}) = \pi \left(\frac{y_1^{on} - m_1(x)}{s_1(x)}, \frac{y_2^{on} - m_2(x)}{s_2(x)} \right). \tag{12}$$

If moreover, assumption (11) is satisfied then $\pi(\cdot)$ is an increasing function of both variables. The conclusion expressed by (12) states that a rational choice of a point for the current computation is reduced to the maximization of aggregated objectives $(y_i^{on} - m_i(x))/s_i(x)$. Such a conclusion is not surprising since the implementation of optimal in some sense single-objective optimization algorithms normally involves optimization of an auxiliary function which formalizes

the concept of optimality; e.g. the multi-objective P-algorithm is reduced to the maximization of the product of two cumulative Gaussian distribution functions (8). The substantiation of rationality of a broader class of scalarizations (12) opens also a broader potentiality of the development of multi-objective optimization algorithms based on statistical models of objective functions. However, the investigation of compatibility of a priori information about the properties of objective functions with particular scalarization methods is needed to realize the mentioned potentiality. As an example, the bi-objective π -algorithm has been implemented. The family of random variables with piecewise linear $m(x)$ and piecewise quadratic $s^2(x)$ is used as a statistical model; for substantiation of this model we refer to [15]. A product of two arctangents is used for $\pi(\cdot)$. Then the $n + 1$ step of the π -algorithm is defined as the follows:

$$x_{n+1} = \arg \max_{x \in \mathbf{A}} \arctan \left(\frac{y_1^{on} - m_1(x)}{s_1(x)} + \frac{\pi}{2} \right) \cdot \arctan \left(\frac{y_2^{on} - m_2(x)}{s_2(x)} + \frac{\pi}{2} \right), \quad (13)$$

where the information collected at previous steps is taken into account in computation of $m_i(x) = m_i(x|x_j, y_j, j = 1, \dots, n)$ and $s_i(x) = s_i(x|x_j, y_j, j = 1, \dots, n)$. The pseudocode of the algorithm is presented in the Appendix.

As mentioned above, a further investigation is needed to find scalarizations best corresponding to the supposed properties of the objective functions. By this implementation we wanted to check if the rather arbitrarily chosen function $\arctan(\cdot) \cdot \arctan(\cdot)$ could be of comparable worth, for the construction of statistical models based multi-objective optimization algorithms, with the Gaussian cumulative distribution function. The experimentation with this version of the algorithm can be helpful also for the further development of the statistical models based on the axiomatic approach proposed in [19].

4 Experimental Results

We were interested to assess the performance of two newly proposed algorithms, i.e. the modified P-algorithm, and the one-step worst case optimal algorithm for Lipschitz continuous functions. The results of the P-algorithm and of the uniform random search have been included for the comparison. The inclusion of the results of uniform random search may seem redundant because of the simplicity of the algorithm. Nevertheless, these results are informative since the uniform search is the worst-case optimal algorithm for Lipschitz continuous functions as shown in [22]; the randomization is applied to exclude the possibilities of bias in the results. A common for all algorithms termination conditions was applied: the algorithms have been stopped after 100 computations of values of the objective functions. Such a number of computations has been selected taking into account the assumption that the algorithms of interest are supposed for expensive problems.

The performance of the proposed algorithms is demonstrated by solving two typical test problems. These test problems have also been used in [21] for the experimentation with the P-algorithm. The first multi-objective test considered

consists of two (slightly modified) Rastrigin functions which are widely used (see, e.g. [15]) for testing single objective global minimization algorithms:

$$\begin{aligned} f_1(x) &= (x + 0.5)^2 - \cos(18(x + 0.5)), \\ f_2(x) &= (x - 0.5)^2 - \cos(18(x - 0.5)), \quad -1 \leq x \leq 1. \end{aligned} \quad (14)$$

The second problem used was proposed in [5]; see also [3], (pp. 339-340). We present below its definition for one-dimensional decision variable

$$\begin{aligned} f_1(x) &= 1 - \exp(-(x - 1)^2), \\ f_2(x) &= 1 - \exp(-(x + 1)^2), \quad -4 \leq x \leq 4. \end{aligned} \quad (15)$$

Both test problems are difficult as it is clear from their feasible objective regions shown in Fig. 1. The set of the Pareto optimal solutions of problem (14) is relatively small, and the multimodality of the objective functions implies discontinuity of the Pareto front. Both objectives of problem (15) are similar to the worst case objectives since they are almost constant over all feasible region, and have a local minimum of the form of a sharp spike; for the discussion on worst-case Lipschitzian objectives we refer to [22]. The P-algorithm and the modified P-algorithm require Y^{on} be defined as an input parameter. Since some information on minima of the objectives usually is available, the selection of the ideal vector as Y^{on} seems reasonable. In the experiments below the vectors $Y^{on} = (-1, -1)^T$ and $Y^{on} = (0, 0)^T$ have been used for in solving the test problems (14) and (15) correspondingly. The one-step optimal algorithm described in Section 2 requires Lipschitz constants of the objective functions be defined as input parameters; the values of Lipschitz constants $L = (21, 21)^T$ and $L = (1, 1)^T$ have been used in solving the test problems (14) and (15) correspondingly.

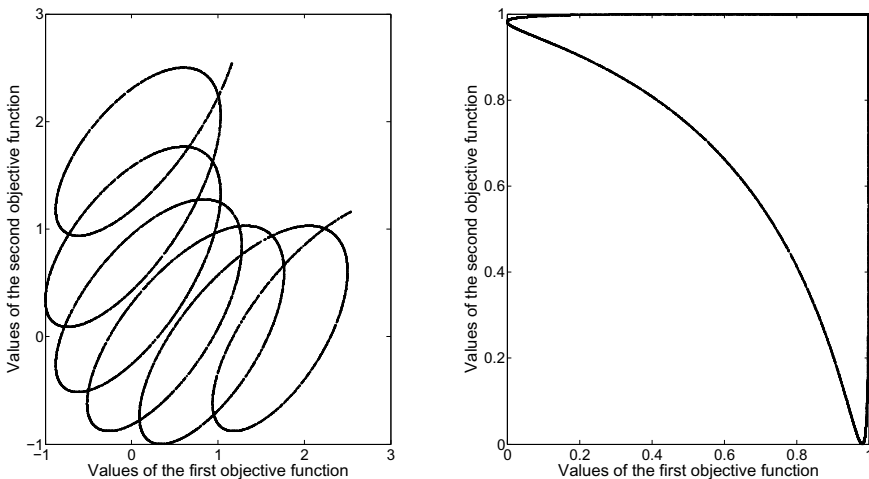


Fig. 1. The feasible objective regions of test functions (14) (left) and (15) (right).

Several metrics have been proposed in recent publications for the comparison of multi-objective algorithms; for the comprehensive list and discussion we refer to [3]. Generally speaking, it is aimed to assess the quality of approximations of the Pareto set and the efficiency of algorithms, used to compute these approximations. In the present paper, we consider only the approximation quality. The performance of random search methods is routinely assessed by statistical methods; the mean values and standard deviations of the considered metrics are assessed and presented below. The performance of the P-algorithm and the modified P-algorithm is assessed similarly; recall that the first 50 observation points by these algorithms are selected randomly with a uniform distribution over \mathbf{A} , therefore each optimization process is different, and the results are random. The performance of the randomized algorithms is evaluated by the mean values and standard deviations (based on the results of 200 runs) of the metrics discussed below. The one-step optimal algorithm is deterministic, and its performance is illustrated by the results of the single run for each test problem. The minimization was stopped after 100 observations, and the following performance criteria were computed: NN -the number of non-dominated points found, NP - the number of points in the true Pareto set, MD - the maximum "hole" length in the Pareto set, and DP - the maximum distance between the found non-dominated solutions outside the Pareto set and the Pareto set; in all cases the distances correspond to the Euclidean norm. MD is computed as the maximum of two maxima: the maximum distance between the nearest neighboring solutions found in the Pareto set, and the maximum distance between the limit points of the Pareto set and their nearest neighbors in the set of the found Pareto solutions. NP shows how many observations were successful in the sense of hitting the Pareto set. The difference between NN and NP gives the estimate of the number of "wrong" solutions, i.e. the number of non-dominated solutions found outside the Pareto set. The so-called error ratio $(NN - NP)/NN$ is a frequently used metric for the assessment of multi-objective algorithms. The metric MD characterizes the spread of solutions over the Pareto set very well. The metric DP characterizes error in evaluating the Pareto set implied by the "wrong" solutions. For the randomized algorithms the mean values and standard deviations of the listed above metrics are presented in Table 1; for the one-step optimal algorithm the metrics have been evaluated from a single run, and therefore the column supposed for the standard deviations consists of zeros. The performance metrics, presented in Table 1, show that the modified P-algorithm outperforms the P-algorithm, especially in the case of test problem (15). The worse performance of the P-algorithm can be explained by the inadequacy of Gaussian model to the test problems considered. Test problem (15) is similar to the worst-case problem defined in [22], and therefore a relatively good value of MD for this problem achieved by the uniform random search is not surprising. However, the results of the uniform random search for test problem (14) are not so good, moreover they would be difficult to interpret correctly as it is clear from Fig. 2. The performance of the newly proposed algorithms, the modified P-algorithm and the one-step optimal algorithm, with respect to metrics MD and DP is similar. The number of found

Table 1. The mean values and standard deviations of the performance metrics for the considered algorithms presented at the left and right columns correspondingly

Algorithm	P-algorithm				Modified P-algorithm			
Problem	Problem (14)		Problem (15)		Problem (14)		Problem (15)	
NP	30.0	2.0	34.0	14.8	47.3	2.3	49.9	2.6
NN	32.9	2.0	36.4	13.7	48.3	2.3	50.8	2.6
MD	0.102	0.014	0.43	0.150	0.065	0.008	0.15	0.023
DP	0.040	0.002	0.007	0.011	0.001	0.001	0.000	0.000

Algorithm	One-step opt.				Uniform			
Problem	Problem (14)		Problem (15)		Problem (14)		Problem (15)	
NP	26.0	0	65.0	0	6.7	2.6	25.0	4.3
NN	34.0	0	66.0	0	9.5	2.2	26.1	4.3
MD	0.064	0	0.027	0	0.29	0.052	0.28	0.094
DP	0.014	0	0.0	0	0.20	0.220	0.004	0.006

non-dominated solutions found by the modified P-algorithm is larger than that found by the one-step optimal algorithm, but the distribution of non-dominated solutions by the one-step optimal algorithm is more rational, as it is seen from Fig. 3. The rationality of the distribution of non-dominated solutions by the one-step optimal algorithm can be explained by the valuable information about the problem consisting in the properly selected Lipschitz constants.

The experimental testing has been performed for a special case of black box multi-objective optimization problems, namely for the bi-objective problems of one variable. These problems are interesting mainly from a theoretical point

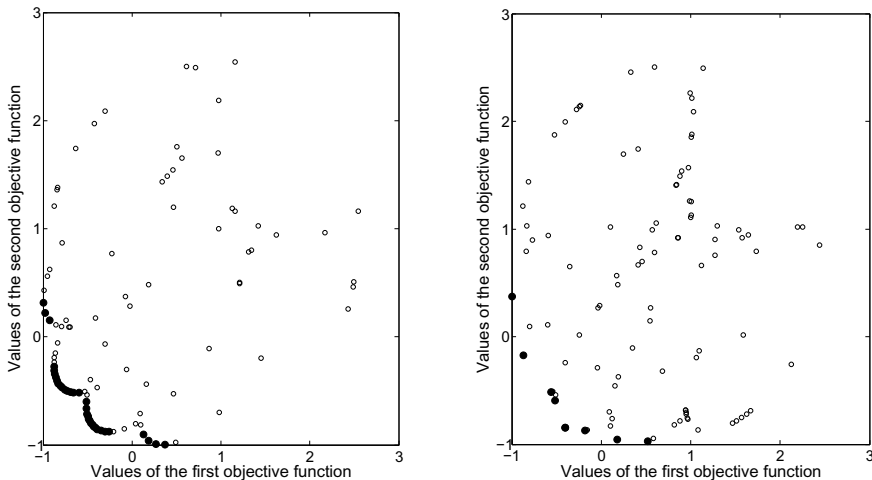


Fig. 2. The points in the feasible objective region of test function (14) generated by the the generalized P-algorithm (left) and the uniform random search (right); the thicker points denote the non-dominated solutions found

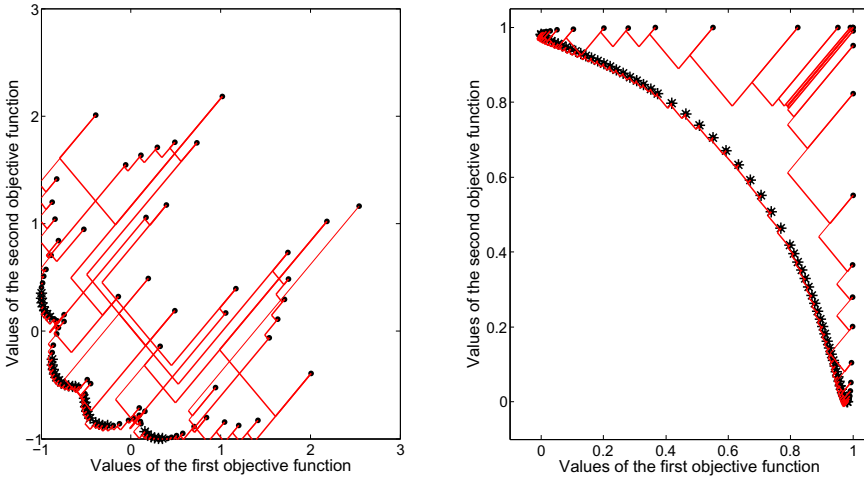


Fig. 3. The points generated by the optimal one-step algorithm in the feasible objective regions of (14) (left) and (15) (right) correspondingly. The non-dominated points are denoted by stars. The lines show the local Lipschitzian lower bound for the Pareto front.

of view. Nevertheless, the results obtained show that the proposed ideas are promising. On these ideas based implementation of algorithms, aimed at real world problems, is not straightforward, and further theoretical investigation will be needed.

5 Conclusions

Two algorithms for black box multi-objective optimization of expensive objectives are proposed generalizing the ideas widely used in single-objective global optimization. The proposed algorithms are one-step optimal either on the average with respect to a statistical model of multimodal functions or worst-case with respect to a class of the Lipschitz continuous functions. A limited testing has shown that the performance of both developed algorithms is similar, and that they outperform previously proposed similar algorithms. Further theoretical investigation is promising to facilitate the development of more general versions of the algorithms based on the proposed ideas.

Appendix

The π -algorithm as well as the one-step worst-case optimal algorithm has been developed in MATLAB. The common input data of both algorithms is: the lower and upper bounds of the feasible interval LB and UB , and the maximum allowable number of iterations N . The specific input data is shown below in the pseudocode of the considered algorithm.

The One-Step Worst-Case Optimal Algorithm

Input: a vector of Lipschitz constants L ,

Initialization: $x_1 \leftarrow LB$, $x_2 \leftarrow UB$, $f^i \leftarrow f(x_i)$,

$x_{oi} \leftarrow x_i$, $f^{oi} \leftarrow f(x_{oi})$, $i = 1, 2$,

initialize: $V(F^2)$, $U(F^2)$, and the set of Pareto optimal solutions P^2 ,

$n = 2$,

while $n < N$

 compute: x_{n+1} , $f(x_{n+1})$,

 update: $\{x_i\}$, $\{f^i\}$, $\{x_{oi}\}$, $\{f^{oi}\}$, $V(F^n)$, $U(F^n)$, P^n ,

$n \leftarrow n + 1$,

end

output: P^n .

The π -algorithm

Input: a reference vector Y^{oN} ,

Initialization: $x_1 \leftarrow LB$, $x_2 \leftarrow UB$, $y_i \leftarrow f(x_i)$,

initialize: $m(x)$, $s(x)$, P^2

$n = 2$,

while $n < N$

 compute x_{n+1} by solving auxiliary problem (13),

 compute $f(x_{n+1})$,

 update: $\{x_i\}$, $\{y_i\}$, $m(x)$, $s(x)$, P^n ,

$n \leftarrow n + 1$,

end

output: P^n .

References

1. Arora, S., Barak, B.: Computational Complexity a Modern Approach. Cambridge University Press (2009)
2. Calvin, J., Žilinskas, A.: A one-dimensional P-algorithm with convergence rate $O(n^{-3+\delta})$. J. Optimization Theory and Applications 106, 297–307 (2000)
3. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. J. Wiley, Chichester (2009)
4. Fishburn, P.: Utility Theory for Decision Making. J. Wiley, Chichester (1970)
5. Fonseca, C., Fleming, P.: On the performance assessment and comparison of multi-objective optimizers. In: Ebeling, W., Rechenberg, I., Voigt, H.-M., Schwefel, H.-P. (eds.) PPSN 1996. LNCS, vol. 1141, pp. 584–593. Springer, Heidelberg (1996)
6. Horst, R., Pardalos, P., Thoai, N.: Introduction to Global Optimization. KAP, Boston (2000)
7. Jančauskas, V., Mackutė-Varoneckienė, A., Varoneckas, A., Žilinskas, A.: Multi-objective optimization aided visualization of graphs related to business process management. Comm. in Comp. and Inform. Sci. 319, 87–100 (2012)
8. Kushner, H.: A versatile stochastic model of a function of unknown and time-varying form. J. Math. Anal. and Appl. 5, 150–167 (1962)

9. Miettinen, K.: *Nonlinear multiobjective optimization*. KAP, Boston (1999)
10. Mockus, J.: *Bayesian approach to global optimization*. KAP, Boston (1988)
11. Nakayama, H., Yun, Y., Yoon, M.: *Sequential Approximate Multiobjective Optimization Using Computational Intelligence*. Springer, Berlin (2009)
12. Pijavskii, S.: An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics* 12, 57–67 (1972)
13. Shubert, B.: A sequential method seeking the global maximum of a function. *SIAM J. Numer. Anal.* 9, 379–388 (1972)
14. Strongin, R., Sergeyev, Y.: *Global Optimization with Non-convex Constraints: Sequential and Parallel Algorithms*. KAP, Boston (2000)
15. Törn, A., Žilinskas, A.: *Global Optimization*. LNCS, vol. 350, pp. 1–225. Springer, Heidelberg (1989)
16. Sukharev, A.: On optimal strategies of search for an extremum. *USSR Comput. Math. and Math. Physics* 11, 910–924 (1971) (in Russian)
17. Sukharev, A.: Best strategies of sequential search for an extremum. *USSR Comput. Math. and Math. Physics* 12, 35–50 (1972) (in Russian)
18. Žilinskas, A.: Optimization of one-dimensional multimodal functions, Algorithm AS-133. *Journal of Royal Statistical Society, ser. C* 23, 367–385 (1978)
19. Žilinskas, A.: Axiomatic approach to statistical models and their use in multimodal optimization theory. *Mathematical Programming* 22, 104–116 (1982)
20. Žilinskas, A.: Axiomatic characterization of a global optimization algorithm and investigation of its search strategy. *Operat. Res. Letters* 4, 35–39 (1985)
21. Žilinskas, A.: A statistical model-based algorithm for black-box multi-objective optimization. *International Journal of Systems Science* (2012) (Published on Internet July 4, 2012), doi:10.1080/00207721.2012.702244
22. Žilinskas, A.: On the worst-case optimal multi-objective global optimization. *Optimization Letters* (2012) (Published on Internet September 14, 2012), doi:10.1007/s11590-012-0547-8

Recognition of Voice Commands Using Hybrid Approach

Vytautas Rudžionis¹, Kastytis Ratkevičius², Algimantas Rudžionis²,
Gailius Raškinis³, and Rytis Maskeliunas²

¹ Vilnius university Kaunas faculty Muitines str. 8, Kaunas, Lithuania
vytautas.rudzionis@vukhf.lt

² Kaunas university of technology, faculty of Informatics,
Studentu str. 65, Kaunas, Lithuania
kastytis.ratkevicius@ktu.lt

³ Vytautas Magnus University Informatics faculty, Vileikos 8, Kaunas, Lithuania

Abstract. Computerized systems with voice user interfaces could save time and ease the work of healthcare practitioners. To achieve this goal voice user interface should be reliable (to recognize the commands with high enough accuracy) and properly designed (to be convenient for the user). The paper deals with hybrid approach implementation issues for the voice commands recognition. By the hybrid approach we assume the combination of several different recognition methods to achieve higher recognition accuracy. The experimental results show that most voice commands are recognized good enough but there is some set of voice commands which recognition is more complicated. In this paper the novel method is proposed for the combination of several recognition methods based on the Ripper algorithm. Experimental evaluation showed that this method allows achieve higher recognition accuracy than application of blind combination rule.

Keywords: Multimodal interface, voice user interface, speech engine adaptation, voice commands, hybrid approach.

1 Introduction

As had been shown in many sources and applications that voice user interfaces and speech processing technology in general are of enormous benefit for the people working in the healthcare industry. The main rationale for the application of voice processing technologies and voice based interfaces in the healthcare industry is the desire to save the time of highly qualified medical personnel which is routinely spent on operations of documentation as well as the desire to speed up and to ease the information search and presentation. In this way, time spent on documentation and other trivial tasks could be allocated for the tasks requiring higher qualification. There exists also many other ways and motivations for implementation of voice user interfaces into the practice of healthcare institutions. Among them we could mention the possibility to ask and to receive necessary information by voice (this often could be performed faster than in more usual keyboard based interface) or the possibility to

use information system with a wide range of modern devices (especially tablet PCs or other mobile devices). It should be noted that if the speech recognition is combined with modern means of communication (computer, internet and telephony), entirely new possibilities to perform medical documentation anytime and anywhere may occur [1] and new types of medical services could be applied.

From the voice user interface developer perspective all these speech recognition using applications could be classified into two big classes. One class is the applications using continuous speech as the basic input while the second class uses voice commands (often even quite long phrases composed from several words or even sentences) as the basic input mode. The first class of applications is more flexible and often provides the user with the higher degree of freedom when pronouncing the utterances fed to the speech engine. The second class of applications requires that the user will utter the phrase from a predefined list of possible phrases (usually called commands) in a strict and predefined way. This doesn't mean that some command should be pronounced in a single way since single command may have several predefined ways of pronunciation but in any case exists only some limited set of such pronunciations. Despite it seems that the second class of voice user interfaces is inconvenient in practice but many valuable and practically useful applications could be designed using such approach. This is based on the fact that in many applications only limited professional vocabulary is necessary. At the same time second class of applications usually possesses higher recognition accuracy and robustness to speaker and environmental variability. From the user point of view the higher recognition accuracy is more important factor than the higher degree of pronunciation development in most of the situations. Since we are convinced that it is still to early develop reliable and practically useful Lithuanian continuous speech recognition system (with high enough recognition accuracy) we concentrated our efforts to develop the prototype Lithuanian speech recognition system for healthcare practitioners using voice commands recognition principle which should be useful for the practitioners in the field [2].

The main research problem is to find the ways to ensure high enough recognition accuracy. In this paper we will deal with the problems implementing hybrid voice user interface design approach. By the term hybrid approach we understand the incorporation of several different recognition algorithms or methods. The basic idea behind the hybrid approach is that different recognition methods are able to extract and to process different kinds of information present in the acoustic signal and if they are used together this could lead to the overall increase of recognition accuracy and robustness. It should be noted that in many of the current state-of-the-art speech recognition systems hybrid recognition principles are implemented in one or another way (e.g. some speech recognizers works using MFCC features while others works in parallel using PLP features, or several HMM based recognizers are used with different training and most likely acoustic states search strategies implemented, etc.) [3], [4], [5]. In the case of Lithuanian voice command recognition hybrid approach is important also because it may potentially enable to use foreign language trained speech recognition engine adapted to recognize Lithuanian commands with the proprietary Lithuanian speech recognizer. Foreign language recognizer should allow

to exploit big amounts of acoustic data used to train these recognizers (for the economy reasons there are no and probably will be no such amounts of Lithuanian acoustic data prepared to train speech recognizers as exist for such languages as English or Spanish). Our experience with the adaptation of foreign language speech engines to recognize Lithuanian voice commands showed that it is possible to achieve very high recognition accuracy for many Lithuanian voice commands using only the appropriate selection of their phonetic transcription. Such approach enabled us to make the development of some limited vocabulary applications easier and more economically viable. Earlier research also showed that not all voice commands may be equally efficiently recognized using adaptation.

At the same time it became clear that not all voice commands that are necessary to realize for some successful voice based service could be recognized equally well using adapted recognition engine. Proprietary Lithuanian speech recognizer may potentially better deal with some acoustic situations that aren't present in other languages and it is necessary to develop specific acoustical models. It is necessary to use a proprietary recognizer to recognize "problematic" voice commands well enough. The need to combine the results provided by two different recognizers requires implementation of hybrid approach.

The problem how to combine different recognizers still remains largely open and needs more research. There were proposed various methods to combine the recognition results obtained from different sources. The most popular method to combine recognition results is the method called heteroscedastic discriminant analysis [6]. There were attempts to apply hybrid recognition principles using SVM and HMM methods together [7]. But before finding the most efficient ways to combine the hypotheses produced by various recognizers still lot of other questions should be solved. Among those problems such issues as the possibilities to get complementary information from different speech recognizers, to find when and in which contexts foreign language recognizer could be used and when it is necessary to use purely Lithuanian acoustic models, to find the limits of adaptation possibilities for foreign language speech engine to recognize Lithuanian voice commands and many other issues.

This paper presents some of our experiments trying to evaluate the possibilities to apply hybrid approach trying to improve overall recognition accuracy of the medical information system. These include evaluation if the different recognizers could provide supplementary information. We are also proposing novel method based on the Ripper logical rules training algorithm to combine the recognition results provided by very different recognizers.

Further paper is organized as follows. In Chapter 2 experimental data used in the study is presented. Chapter 3 provides some results of our experiments trying to find out if two classes of different recognizers could provide supplementary information for making the final decision. In Chapter 4 novel hybrid decision method is presented. Chapter 5 presents proposed scheme for the hybrid recognition system implementation in practice.

2 Data for Experimental Evaluation

One of the key elements in developing a robust and efficient speech recognition system is the employment of proper speech corpus. The main purpose of the corpus is to provide acoustic-phonetic material (recordings) for the training process, i.e. for finding the parameters of acoustical models. Speech corpus should comply with a wide range of requirements. Few of them we will mention explicitly:

- corpus should be as good as possible in representing acoustic-phonetic content which will be used by the system;
- corpus should be as good as possible in covering the variety of speakers which will use the system.

The primary aim of the system under development is to provide services for the people working in the healthcare. It has been decided that voice user interface must be able to deal with the names of the drugs, with the names of the illnesses and diseases as well as with some other words and phrases often met in medical workers practice. The large part of medical terms used in practice by healthcare professionals is contained in the official list of diseases and disorders that is approved by the Ministry of Health. This list contains 14179 diseases and disorders. It is composed of more than 88000 lexical tokens (not all are of medical origin) and has 10955 unique lexical types. It may be surprising that specific medical terms are used much less often than general terms (e.g. switches). 5991 lexical types cover 75 percent of the whole list. This analysis showed that it is unfeasible to develop a full scale medical Lithuanian recognition system in short time and some type of compromise is necessary. But also analysis showed that not all diseases or drug names are used equally often (frequency of medical terms used has been provided by industrial partner “Softdent Ltd.” which has extensive statistical data about the use of medical terms and drug names among healthcare practitioners). A big part of the daily voice requests could be successfully handled using a relatively small number of voice commands. In collaboration with the industrial partners who have the expertise in developing computer systems for medical professionals we selected the 731 diseases names, complaints and drug names contained in 777 lexical tokens. This list represents the most frequently used medical terms in Lithuania.

When the list of the phrases has been selected the speech corpora was collected. Each voice command in the set has been recorded by 7 different male speakers and 5 different female speakers in laboratory conditions. Laboratory condition means that recordings were made in relatively silent environment. Such assumption is quite well grounded when the system would be deployed in medical institution. Then every speaker pronounced each voice command 20 times. This means that there are 240 utterances of each voice command in the corpora. Full semi-automatic and manual validation with respect to the completeness and correctness has been performed. The size of the medical speech corpus is about 100 hours. On the other hand, it was decided to use speech resources already at our possession, i.e. earlier collected speech corpora (containing about 50 hours of speech) to train purely Lithuanian acoustic models. All these data has been used in experiments described below.

3 Investigation of the Possibilities to Apply Hybrid Approach

The first problem that should be evaluated before trying to implement hybrid approach is the possibilities to get benefits from using more than one speech recognizer. Hybrid approach may be useful if at least one of the recognizers provides the correct answer and it is possible to associate the correct answer with a higher degree of confidence. In other words hybrid approach could be useful if the outputs of different recognizers will provide at least partly uncorrelated results (or at least in some situations) and at least one of these outputs is correct. Our some earlier pilot studies showed that in principle hybrid recognizer combined from adapted foreign language recognizer and proprietary Lithuanian recognizer should have potential to lead to the overall increase in recognition accuracy. E.g. in our earlier study we observed that using the names of ten Lithuanian digits of 244 total errors provided by the proprietary Lithuanian recognizer 106 errors were related to the digit “trys”. All these misrecognitions were given to the adapted Spanish recognizer and 93 out of 106 utterances were recognized correctly. In the second case Lithuanian recognizer using prototype of medical speech corpora made 236 errors (out of 44560 utterances). 8 worst recognized commands (135 falsely recognized utterances) were presented to the Spanish recognizer which reduced error rate nearly twice (to 70 errors) [8]. Here we are presenting the more detailed study about the possibilities to implement hybrid approach for the recognition of Lithuanian medical terms.

The first experiment was carried out to evaluate the overall efficiency of the adapted foreign language recognizer and proprietary Lithuanian recognizer to recognize all 731 voice commands from Lithuanian medical corpora. As for the adapted foreign language recognizer Microsoft Spanish speech engine was used. The selection of Spanish recognizer for experiments was caused by two major factors: the availability of engine and its efficiency. There are relatively few commercial speech recognizers available for widely spoken languages. Our earlier experiments showed [9] that Spanish recognizer provides best performance using it for Lithuanian voice command recognition (probably due to the proximity of the acoustic structure of two languages). In these Lithuanian voice commands were transcribed using the methodology developed in our previous studies [9] but without thorough optimization seeking to find best transcriptions. Proprietary Lithuanian recognizer has been trained as the triphone based CD-HMM using the 50 hours Lithuanian transcribed speech corpora [10]. CD-HMM models were obtained using HTK tool [11]. In these corpora recordings of the speakers used in the testing stage weren't used.

One of the first observations drawn from the analysis of results is that recognition accuracy varied greatly among different speakers and both recognizers performed quite different in respect to various speakers. Table 1 shows the recognition accuracy obtained for different speakers both using adapted Spanish and proprietary Lithuanian recognizers. The letter M near the name of the speaker means that speaker is male while the letter F near the name means female speaker. Despite that overall performance was better for purely Lithuanian recognizer the adapted Spanish recognizer enabled to achieve better performance. These results shows that exists bog inter-speaker variability which could be even bigger than inter-language variability.

Table 1. Error rate for each speaker using adapted Spanish and proprietary Lithuanian recognizers

Speaker name	Recognizer	
	Adapted Spanish	Purely Lithuanian
M1	15,49%	5,98%
M2	11,34%	11,28%
M3	11,00%	10,13%
F1	10,33%	7,60%
M4	9,72%	4,59%
M5	8,27%	5,98%
F2	7,19%	9,42%
F3	6,21%	6,29%
M6	6,18%	4,39%
F4	5,45%	4,91%
F6	4,84%	4,87%
M7	3,97%	24,54%

Table 2. Most often incorrectly recognized voice commands using adapted foreign language recognizer and proprietary Lithuanian recognizer

Adapted foreign language recognizer		Proprietary Lithuanian recognizer	
Command name	Number of missrecognitions	Command name	Number of missrecognitions
_zemas_kraujo_spaudimas	238	_vitaminas_cE	90
_roZinE	214	_ketonalis	82
_elokonas	211	_pananginas	80
_plavikšas	208	_renY	80
_influcidas	199	_gUZYs	65
_dikloberlas	194	_mikardis	53
_loperamidas	194	_kelio_sutinimas	51
_nemiga	191	_meningokokinE_infekcija	50
_relaniumas	186	_alkis	49
_xlorheksidinas	185	_esencialis	49
_milgama	185	_rinitas	48
_galUniu_silpnumas	183	_trentalis	44
_lokrenas	183	_diltiazemas	41
_ibufenas	178	_ranigastas	41
_ploni_kaulai	177	_persenas	40
_renY	177	_meningokokinis_meningitas	39
_ventolinas	177	_lineksas	39
_viduriu_laSai	177	_xeminiai_nudegimai	37
_analginas	175	_vEmimas	36
_magvitas	175	_nimezilis	36
_vitaminu_a_ir_dE_tepalas	173	_uvaminas	36
_lopediumas	171	_aponilis	35
_roZE	167	_lomeksinas	34
_flosinas	166	_raumenu_tonuso_praradimas	30
_viduriavimas	164	_isla	30

Another important analysis could be done looking to the most often misrecognized voice commands using different types of recognizers. Table 2 provides the list of 25 most often incorrectly recognized voice commands using adapted foreign language recognizer and proprietary Lithuanian recognizer. The most important observation we seek to draw from the evaluation is to find if and how strongly are the sets of misrecognized words.

Looking to the results in the Table 2 we could easily see that the vocabularies and the structure of incorrectly recognized commands differ significantly. Looking carefully to the list of 25 most often incorrectly recognized medical voice commands using adapted foreign language and Lithuanian speech recognizers we see that there is no even single command met in both lists. In other words commands that were incorrectly recognized with one recognizer did not overlap with the commands incorrectly recognized with another recognizer which means that recognizers are supplementary. We could conclude that in the vocabulary sense these two different recognition approaches also could be used as supplementary ones and potentially lead to the improved recognition accuracy. These results also suggest that incorrect recognitions using adapted foreign language engine are more “concentrated”: 25 commands which had worst performance contained bigger part of errors in this case than 25 worst recognized commands using proprietary Lithuanian recognizer.

Results of these investigations leads to the conclusion that results obtained from two different recognizers may be used as a basis for the hybrid recognition approach: the recognition errors obtained from adapted foreign language recognizer and proprietary Lithuanian recognizer based aren't strongly related. The distribution of errors between different speakers has some common characteristics but at the same time these distributions are very different for some speakers. Even bigger differences are seen when comparing the set of commands that were recognized improperly with different recognizers. Comparing the sets of 25 commands that were recognized with the highest number of errors it could be seen that these sets are completely different. These results show that the error obtained from one recognizer potentially may be corrected using another one.

4 Investigation of the Hybrid Approach Efficiency

The realization of hybrid recognizer is an open and still not investigated question. This is especially true when there is necessary to combine the results obtained from so different recognizers – adapted foreign language engine (SP) and CD-HMM based triphone Lithuanian recognizer (LT). E.g. often used heteroscedastic discriminant analysis principle is difficult to apply when we need to combine hypothesis provided by very different recognition engines and the similarity or likelihood of the hypothesis are expressed in very different parameters. In this paper we propose a novel hybrid recognizer implementation approach which showed promising results.

In this study we used two different independent recognizers. Both recognizers were same ones that were used in previous experiments. Recognition results of 731 voice commands from medical speech corpus were used in construction of hybrid one.

Since each of the voice command has been pronounced by 12 different speakers 20 times there were 175440 commands in the recognition tests. All results obtained from both recognizers could be grouped into several subsets. These subsets are summarized in the Table 3.

The results in the table allows to conclude that the accuracy of LT recognizer was 98.58% (T=T, T-, TF), while the accuracy of SP recognizer was 78.24% (T=T, -T, FT). The goal of the investigation is find out if the results of SP recognizer could be used to improve the LT recognizer's performance.

Table 3. Decisions of recognizers obtained grouped into subsets

Subset	Description	Number of phrases in the subset
T=T	Both recognizers produces same hypotheses and both hypotheses are correct	135898
F=F	Both recognizers produces same hypotheses and both hypotheses are incorrect	178
T-	Recognizer LT produces correct decision while recognizer SP don't produces any decision	3398
F-	Recognizer LT produces incorrect decision while recognizer SP don't produces any decision	48
-T	Recognizer SP produces correct decision while recognizer LT don't produces any decision	7
-F	Recognizer SP produces incorrect decision while recognizer LT don't produces any decision	1
--	Both recognizers don't produce any decision	1
TF	Recognizers produces different hypothesis, LT produces correct decision	33650
FT	Recognizers produces different hypothesis, SP produces correct decision	1357
FF	Recognizers produces different hypothesis, both produces incorrect decision	902
Overall		175440

Principles of machine learning were used to find the decision rule. The aim was to develop the rule to separate two classes TF and FT. Each object in the class was is described by the recognition results of both recognizers. Training set was composed from 35007 objects. But these classes has significant disproportion since TF has 33650 objects while FT only 1357. It means that blind classification rule („if both recognizers produces different hypothesis use the LT hypothesis“) should lead to the 96.12% overall recognition accuracy. Each object in the training set has been described using 70 features Among those features are such parameters as confidence of the result provided by SP recognizer, average log probability of the LT recognizer hypothesis, proportion and likelihood of all sounds present in the hypothesis produced by both recognizers and some other parameters (such as gender probability, silence probability at the start and the end of the utterance, etc.). The Ripper logic rules

learning algorithm [12] has been used for training. The example of the learned rule is presented below:

SP :- *lt_delta_prob* <= 0.73 & *sp_supp* <= 0.41 & *lt_a* >= 18.8 & *sp_prob* >= 495 & *sp_i* >= 9.1 (144/1).

Set of rules found using Ripper algorithm was arranged. This means that rules are applied in a given order: if the first rule can't be applied the second rule should be applied and so on. Applying standard 10 times cross checking procedure to the rules set derived with Ripper algorithm 98.73% ± 0,24 recognition accuracy was obtained. This accuracy exceeds the accuracy that could be achieved using blind rule (96.12%).

Brief analysis of the rules allowed made several observations. Nearly in all rules is present conjunctive *lt_delta_prob* <= *threshold*. It means that the hypothesis of SP recognizer will be used only if LT recognizer produces lower than some level likelihood. In many rules is present conjunctive *sp_supp* <= *threshold*. It means that SP hypothesis will be used only when SP decision is the same as LT decision and this decision isn't much worse than one of the LT alternatives. Conjunctives which are checking the phonetical structure of the voice commands are well adjusted to the type of errors produced by recognizers. It could be supposed that acoustic HMM models in some contexts and for some speakers aren't trained well enough and SP recognizer deals with them better.

5 Computer System Architecture for Hybrid Recognition System

Hybrid recognition approach produces higher requirements for computerized system which will implement this system. Our approach is based on a client – server architecture displayed in Figure 1. User accesses a frontend either via website, web-app, or local implementation (eg. an app in his device). In principle the only thing application front-end should do is to collect the audio data and to return the necessary results and recognition parameters (if preferred). The gathered audio prompt is then sent to our Hybrid speech processing server. We have built a web-service based API, allowing two way communications via any compatible browser or device. On the speech server side, the received input is first parsed to the very fast in performance, commercial closed-source Spanish recognizer which we have adapted to work with Lithuanian transcriptions. This recognizer works with up to 95% recognition accuracy (depending on a Vocabulary used). If a confidence level of a recognized phrase is high enough, or if there are enough highly probably answers to offer an N-best strategy to the user – a text prompt is formulated and returned back to the user. If the recognition is not accurate, not recognized, etc. – the received prompt is then immediately passed to HMM based native Lithuanian recognizer which has a very good recognition accuracy (up to 99% depending on a vocabulary used) but is much slower due to a large data set and the low performance HTK based algorithms, not supporting threading of modern processors. This recognizer then finds a set of best statistically viable variations, checks those in a predefined vocabulary, selects the best answer (most accurate) and then a prompt is formulated and sent back to the user. In

case of a negative recognition (both engines) a user is then asked to pronounce the utterance more slowly and clearly. This proprietary Hybrid approach allows achieving a 99,1 % overall speech recognition accuracy and a very good performance and load on server's CPU, as the CPU intensive HMM Lithuanian recognizer is fired only when necessary.

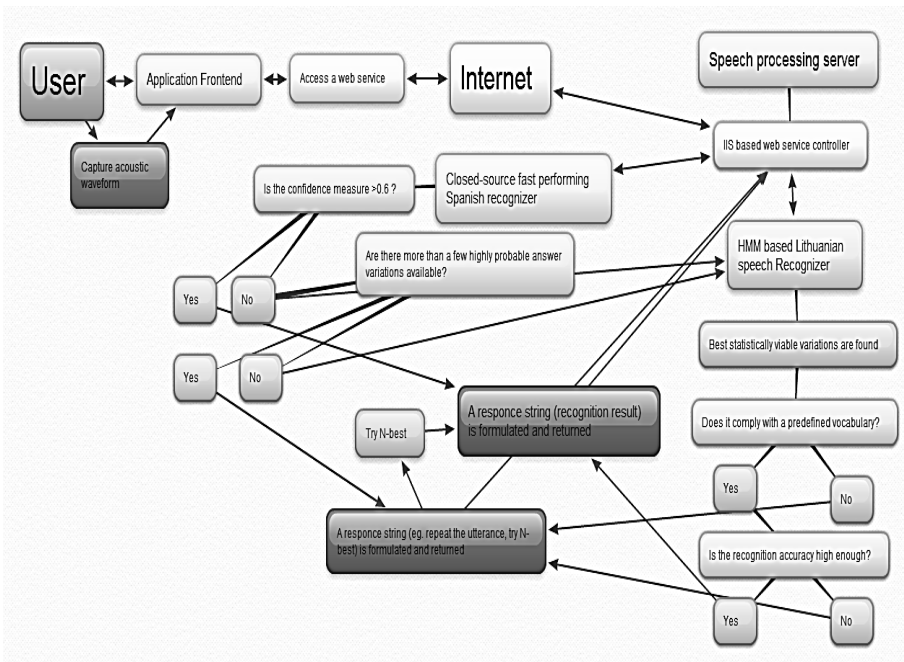


Fig. 1. Hybrid recognition approach implementation

6 Conclusions

The healthcare is one of the most promising areas of human activity where voice user interfaces could be applied. The success of the application depends on various factors. The voice commands recognition accuracy is one of the most important factors influencing acceptance of the application with speech recognition implemented. Lot of useful applications could be developed using voice commands recognition approach.

Hybrid approach is one of the ways to achieve higher recognition accuracy of speech processing system. This implies combination of hypotheses provided by different recognition engines in order to get higher recognition accuracy. This could be achieved if those hypotheses are in some degree uncorrelated. The experimental evaluation showed that recognition of adapted to recognize Lithuanian medical terms



commercial Spanish speech recognition engine and proprietary Lithuanian CD-HMM based speech recognizer has the necessary degree of independence. This level of independence could be seen both on a level of different speakers and both at the voice commands level. The good illustration of the fact is that 25 most often incorrectly recognized with adapted foreign language speech engine and 25 most often incorrectly recognized with proprietary Lithuanian recognizer are completely different ones.

The novel hybrid approach based on Ripper logic rules training algorithm has been proposed. The method allowed achieve the higher recognition accuracy comparing both with the case if “blind” hybridization rule would be used or if only the best single recognizer would be applied.

The results obtained should be seen as the provisional solution. The search for another hybrid training rules and methods should be carried out further.

Acknowledgments. Parts of this work were done under research project No.:31V-34/13 “Hybrid recognition technology for voice interface” (INFOBALSAS) funded by Agency for Science, Innovation and Technology.

References

1. Suendermann, D., Pieraccini, R.: SLU in Commercial and Research Spoken Dialogue Systems. In: Tur, G., De Mori, R. (eds.) *Spoken Language Understanding*, pp. 171–194. John Wiley & Sons, Ltd. (2011)
2. Rudzionis, V., Ratkevicius, K., Rudzionis, A., Maskeliunas, R., Raskinis, G.: Voice Controlled Interface for the Medical-Pharmaceutical Information System. In: Skersys, T., Butleris, R., Butkiene, R. (eds.) *ICIST 2012. CCIS*, vol. 319, pp. 288–296. Springer, Heidelberg (2012)
3. Saon, G., Chien, J.-T.: Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances. *IEEE Signal Processing Magazine* 29(6), 18–33 (2012)
4. Tur, G., Stolcke, A.: The CALO Meeting Speech Recognition and Understanding System. In: *Proc. IEEE Spoken Language Technology Workshop*, pp. 69–72 (2008)
5. Chelba, C., Xu, P., Pereira, F., Richardson, T.: Distributed Acoustic Modeling with Back-off N-grams. In: *Proc. of ICASSP 2012*, pp. 4129–4132. IEEE (2012)
6. Kumar, N., Andreou, A.: Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition. *Speech Communication* 25(4), 283–297 (1998)
7. Ganapathiraju, A., Hamaker, J., Picone, J.: Hybrid SVM/HMM architectures for speech recognition. In: *Proc. of Interspeech*, vol. 11, pp. 504–507 (2000)
8. Rudzionis, V., Raskinis, G., Maskeliunas, R., Rudzionis, A., Ratkevicius, K.: Comparative Analysis of Adapted Foreign Language and Native Lithuanian Speech Recognizers for Voice User Interface. *Elektronika ir Elektrotechnika* 19(7) (in press, 2013)
9. Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Rudžionis, V.: Investigation of foreign languages models for Lithuanian speech recognition. *Elektronika ir Elektrotechnika* 3, 15–20 (2009)

10. Vaičiūnas, A.: Statistical Language Models of Lithuanian and their Application to Very Large Vocabulary Continuous Speech Recognition. Summary of PhD thesis, Vytautas Magnus University, Kaunas (2006)
11. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book, Cambridge (2000)
12. Cohen, W.W.: Fast Effective Rule Induction. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 115–123 (1995)

Estimation of the Environmental Impact on the Accuracy of Signal Recognition

Gintarė Čeidaite¹ and Laimutis Telksnys^{1,2}

¹ Department of System Analysis, Vytautas Magnus University Kaunas, Lithuania
g.ceidaite@if.vdu.lt

² Vilnius University Institute of Mathematics and Informatics Vilnius, Lithuania
telksnys@ktl.mii.lt

Abstract. The problem of the random signals recognition system's adaptation to the variable environmental conditions is discussed. The constructive method that demonstrates possibilities to create recognition systems able to adapt to changing working conditions. The efficiency of the method is demonstrated by experiment analyzing the recognition of random signals in environments with different characteristics. The results demonstrate that the suggested method also can be useful in the development of the speech recognition devices operating in various environments.

Keywords: random signals' recognition, adaptive recognition modelling, recognitions accurate estimation.

1 Introduction

While operating machines and mechanisms that function in variable mediums of atmosphere or hydrosphere it is necessary to solve the problems of random signals recognition in changing environments. Thus it is necessary to have a basic knowledge about the possibilities of recognition devices to adapt to the changing conditions. Similar problems arise when developing devices of automatic speech recognition. Currently the mobility of people grows extremely fast and while people work in different kinds of environments speech recognition device must give accurate results in all of them.

Significantly different characteristics of the environments can generate unacceptable recognition results. It becomes important to adapt recognition systems to those different environments' characteristics, necessary to keep recognition accurate and to eliminate unnecessary environments impact to the accuracy of speech recognition. Therefore, the recognition systems' possibilities to adapt to the changing environments characteristics is under investigation.

The paper is concentrated on the issues of random signal recognition accuracy. A constructive method that gives possibilities to evaluate the accuracy of random signals recognition when recognition conditions change is presented. The efficiency of the method is demonstrated by experiment analyzing recognition of random signals in environments with different characteristics.

2 Statement of the Problem

The situation when dynamic system can be in one of these states $\Omega_l (l=1, \dots, L)$ and function in one of these environments $A_p (p=1, \dots, P)$ is analysed.

Dynamic system states are described by means of following equation :

$$X_i(l, p) = \sum_{j=1}^{m_1} a_j(l, p) X_{i-j}(l, p) + \sum_{k_1=0}^{m_2} b_{k_1}(l, p) V_{i-k_1}, \quad (1)$$

where $i = 1, \dots, N$, N - random value, $a_j(l, p)$ $b_{k_1}(l, p)$ - equation coefficients that describe l states in environments p , $V_i (i=1, \dots, N)$ is Gaussian sequence which average $EV_i = 0$ and variance $EV_i^2 = 1$.

We make the decision that the recognition system is being trained in environment A_l , and later it can be moved into another environment $A_p (p=2, \dots, P)$.

Signal monitoring (1) equation is defined as follows:

$$x_i(l, p, \mu) = \sum_{j=1}^{m_1} a_j(l, p) x_{i-j} + \sum_{k_1=0}^{m_2} b_{k_1}(l, p) v_{i-k_1}(\mu), \quad (2)$$

where $(i=1, \dots, n)$, n - realization of random value N , $l=1, \dots, L$; $p=1, \dots, P$; $\mu=1, 2, \dots$ are the order number of realization of $x_i(l, p, \mu)$ and v_i is the realization of random value V_i .

It needs to be decided to which set of state $\Omega_l (l=1, \dots, L)$ belongs realization $x_i(l, p, \mu)$ observed in environments $A_p (p=1, \dots, P)$.

The dynamic time scale warping method [5] is used to ascertain the distances between observed signals $x_i(l, p, \mu)$ and standards of dynamic systems states' $E_{l_1} (l_1=1, \dots, L_1)$.

The decision made is $d_s : x_i(l, p, \mu) \in \Omega_s$, that the signal observed belongs to the state Ω_s if $\Delta(x_i(l, p, \mu), E_s) = \min_{l=1, \dots, L} \Delta(x_i(l, p, \mu), E_{l_1})$ where $\Delta(x_i(l, p, \mu), E_{l_1})$ is the distance of observed signal $x_i(l, p, \mu)$ till standards $E_{l_1} (l_1=1, \dots, L_1)$ reflecting states of dynamic system.

3 Problem Solving

The signals described in equation (2) are used in analysis of environment impact evaluation to the accuracy of random signals recognition generated by dynamic

systems. Analyze situations when dynamic system from environment A_1 shift to environment $A_p(p=1, \dots, P)$

In this situation we have:

- Detect the moment u when dynamic system moves from environment in which it was trained A_1 into new environment $A_p(p=2, \dots, P)$ and
- Evaluate the regularities /possibilities to adopt to the new environment $A_p(p=2, \dots, P)$.

The designed rule is that the recognition system at the time moment u shifts from environment A_1 to the environment $A_p(p=2, \dots, P)$ if $\alpha_p(p=1) \notin [\xi_{s_1}; \xi_{s_2}]$,

$$\alpha_p(p=1) = \begin{cases} \Delta(x_i(l, p, \mu), E_2) - \Delta(x_i(l, p, \mu), E_1), & x_i(l, p, \mu) \in \Omega_l(l=1) \\ \Delta(x_i(l, p, \mu), E_1) - \Delta(x_i(l, p, \mu), E_2), & x_i(l, p, \mu) \in \Omega_l(l=2) \end{cases} \quad (3)$$

$$\Delta(x_i(l, p, \mu), E_s) = \min_{l=1, \dots, L} \Delta(x_i(l, p, \mu), E_{l_1}(l_1=1, \dots, L_1)) \quad (4)$$

$\Delta(x_i(l, p, \mu), E_{l_1})$ is realized as the distance between observed realization $x_i(i=1, \dots, n)$ and pattern references $E_{l_1}(l_1=1, \dots, L_1)$. Pattern references $E_{l_1}(l_1=1, \dots, L_1)$ of state $\Omega_l(l=1, \dots, L)$ are defined as follows

$$E_{l_1}(l_1=1, \dots, L_1) = \{e_1(l_1), e_2(l_1), \dots, e_h(l_1)\} \quad (5)$$

Patterns $E_{l_1}(l_1=1, \dots, L_1)$ are created from sub-patterns formed of features set $e_t(l_1)$

$$e_t(l_1) = \{Zcr(x_z^k(l, p, \mu), Se(x_z^k(l, p, \mu)))\} \quad (6)$$

$k=1, \dots, dn, t=1, \dots, h$, h is a number of sub-patterns used to create pattern $E_{l_1}(l_1=1, \dots, L_1)$. Zero crossing rate $Zcr(x_z^k(l, p, \mu))$ [10]

$$Zcr(x_z(l, p, \mu)) = \sum_{k=1}^{dn} \frac{|\operatorname{sgn}(x_z^k(l, p, \mu)) - \operatorname{sgn}(x_z^{k-1}(l, p, \mu))|}{2} \quad (7)$$

$$\operatorname{sgn}(x_z^k(l, p, \mu)) = \begin{cases} 1, & x_z^k(l, p, \mu) \geq 0 \\ -1, & x_z^k(l, p, \mu) < 0 \end{cases} \quad (8)$$

and $Se(x_z^k(l, p, \mu))$ signal energy character systems [7]

$$Se(x_z(l, p, \mu)) = \sum_{k=1}^{dn} (x_z^k(l, p, \mu))^2 \tag{9}$$

$\Omega_l (l = 1, \dots, L)$ features are calculated as follows:

realization signal $x_z(l, p, \mu) \ z = 1, \dots, Z$ is processed frame-by-frame in overlapping intervals dn , with frame size denoted dl . $e_t(l_1)$ features are extracting from frame vectors $x_z^k(l, p, \mu) \ k = 1, \dots, dn$ of realization signal $x_z(l, p, \mu)$ [7] These feature systems are very practical to process simulated speech signals [9]. It is also possible to use another features [5].

To find the the distance $\Delta(x_i(l, p, \mu), E_{l_1})$ to the dynamic system state reflected patterns $E_{l_1} (l_1 = 1, \dots, L_1)$ of the observed realization $x_i(l, p, \mu)$ is used Sakoe-Chiba method [2],[8] where $\Delta(x_i(l, p, \mu), E_{l_1})$ is calculated as follows:

- distance constituent $D(j_1, i_1)$ is evaluated, where component i_1 determines k -th characteristics of vector $i_1 = 1, 2, \dots, dn$ of pattern $e_t(l_1)$ and element j_1 of observed realization $x_i(l, p, \mu)$ k -th vectors features $j_1 = 1, 2, \dots, dn$. The values of distances $\delta(i_1, j_1)$ are calculated:

$$\delta(j_1, i_1) = \sqrt{\frac{Se(x_z^{j_1}(l, p, \mu)) - Se(x_z^{i_1}(l, p, \mu)))^2 - (Zcr(x_z^{j_1}(l, p, \mu)) - Zcr(x_z^{i_1}(l, p, \mu)))^2}{}} \tag{10}$$

- all features distances to every single sub-pattern $e_t(l_1)$ of observed realization $x_i(l, p, \mu)$ are measured

$$\Delta(x_i(l, p, \mu), e_t(l_1)) = \min \left(\begin{matrix} D(j_1, i_1 - 1) \\ D(j_1, i_1 - 1) \\ D(j_1 - 1, i_1 - 1) \end{matrix} + \delta(j_1, i_1) \right) \tag{11}$$

- character distance $\Delta(x_i(l, p, \mu), E_{l_1})$ of realization $x_i(l, p, \mu)$ to patterns $E_{l_1} (l_1 = 1, \dots, L_1)$ reflecting dynamic system states is calculated

$$\Delta(x_i(l, p, \mu), E_{l_1}) = \min_{t=1, \dots, h} \Delta(x_i(l, p, \mu), e_t(l_1)) \tag{12}$$

ξ_{S_1} and ξ_{S_2} means min and max values that $\alpha_p (p = 1, \dots, P)$ can obtain for state $\Omega_l (l = 1, 2)$ while working in exact environment $A_p (p = 1, \dots, P)$. Values $\alpha_p (p = 1)$ has a standard normal distribution with probability density function

$$f_p(x) = \frac{1}{\sqrt{2\pi\sigma_p}} e^{-\frac{(x-\bar{x}_p)}{2\sigma_p^2}} \tag{13}$$

So if state $\Omega_l(l=1,2)$ is functioning in primary environment $A_p(p=1)$, parameters $\alpha_p(p=1)$ interval values ξ_{S_1} and ξ_{S_2} are calculated

$$\begin{cases} \xi_{S_1} = \bar{x}_p - \frac{\sigma_p^2}{2} \\ \xi_{S_2} = \bar{x}_p + \frac{\sigma_p^2}{2} \end{cases} \tag{14}$$

Where \bar{x}_p and σ_p^2 are mean and variance of $\alpha_p(p=1)$

$$\bar{x}_p = \frac{1}{k_1} \sum_{i=1}^{k_1} \alpha_p(p=1, \dots, P; i) \tag{15}$$

$$\sigma_p = \sqrt{\frac{1}{k_1 - 1} \sum_{i=1}^{k_1} (\alpha(p=1, \dots, P; i) - \bar{x}_p)^2} \tag{16}$$

where k_1 is selected constant. If state $\Omega_l(l=1,2)$ shifts from the environment A_1 to environment $A_p(p=2, \dots, P)$, interval values ξ_{S_1} and ξ_{S_2} are calculated

$$\begin{cases} \xi_{S_1} = \alpha_p(p=2, \dots, P) - \sigma_p \\ \xi_{S_2} = \alpha_p(p=2, \dots, P) + \sigma_p \end{cases} \tag{17}$$

This method, let get primary values ξ_{S_1} and ξ_{S_2} in new environment. Late these values ξ_{S_1} and ξ_{S_2} must be recalculate using (18)

Realization signals $x_i(l, p, \mu)$ of states $\Omega_l(l=1,2)$ generated in environment $A_p(p=1, \dots, P)$ for modelling testing situation are described as follows:

$$x_i(l, p=1, \dots, P, \mu) = x_i(l, \mu) + A_p(p=1, \dots, P) \tag{18}$$

where $x_i(l, \mu)$ defines realization of state $\Omega_l(l=1,2)$ not functioning in environment $A_p(p=1, \dots, P)$.

Recognition quality $P(x_i(l, p, \mu))$ of realization $x_i(l, p, \mu)$ in specific environment $A_p(p=1, \dots, P)$ defines adaptation possibilities.



$$P(x_i(l, p, \mu)) = \frac{T_a(x_i(l, p, \mu))}{r} \tag{19}$$

Here $T_a(x(l, p, \mu))$ is the number of correctly recognized state's $\Omega_l(l = 1, \dots, L)$ quantity of realizations in environment $A_p(p = 1, \dots, P)$

$$T_a(x_i(l, p, \mu)) = \sum_{i=1}^r d_s(x_i(l, p, \mu)) \tag{20}$$

r is in environment $A_p(p = 1, \dots, P)$ accomplished states $\Omega_l(l = 1, \dots, L)$ quantity of observed realizations, d_s is the correct recognition derivation of realization $x_i(l, p, \mu) \in \Omega_l(l = 1, \dots, L)$. Generated realizations $x_i(l, p, \mu)$ of dynamic system states $\Omega_l(l = 1, \dots, L)$ perform in environment $A_p(p = 1, \dots, P)$ randomly. Lehmer method is used to control the sequence of realization [6].

The following steps are to be done to model the performance of generated realizations $x_i(l, p, \mu)$ of state $\Omega_l(l = 1, 2)$ in sequences environment $A_p(p = 1, \dots, P)$:

firstly the set $F(ii = 1, \dots, \omega)$ of generating state $\Omega_l(l = 1, 2)$ realizations $x_i(l, p, \mu)$ is made

$$F(ii = 1, \dots, \omega) = \{x_i(l = 1, p, \mu = 1); \dots; x_i(l = 1, p, \mu = n_{11}); x_i(l = 2, p, \mu = 1); \dots; x_i(l = 2, p, \mu = n_{12})\} \tag{21}$$

where n_{11} is the quantity of realizations $x_i(l, p, \mu) \in \Omega_l(l = 1)$, n_{12} is the quantity of realizations $x_i(l, p, \mu) \in \Omega_l(l = 2)$ and ω corresponds the size F_{ii}

$$\omega = n_{11} + n_{22} \tag{22}$$

First, the a priori probability $P_a(x_i(l, p, \mu)) = \frac{n_{11}}{\omega}$ is of realizations $x_i(l, p, \mu) \in \Omega_l(l = 1)$ and $x_i(l, p, \mu) \in \Omega_l(l = 2)$ functioning in the environment $A_p(p = 1, \dots, P)$.

Second, the chain $\varphi(i_2 = 1, 2, \dots)$ of realizations $x_i(l, p, \mu)$ of states $\Omega_l(l = 1, 2)$ is processed

$$\varphi(i_2 = 1, 2, \dots) = \{\varphi(1), \varphi(2), \varphi(3), \dots\} \tag{23}$$

$$\varphi(i_2) = F_{ii} \tag{24}$$

$$ii = (\nu \cdot (i_2 - 1) + d) \bmod \omega \tag{25}$$

ν - multiplier which can be $\nu < \omega - 1$, d - constant with various values but usually equal to 0. The adaptation system operating principals are shown in Fig. 1

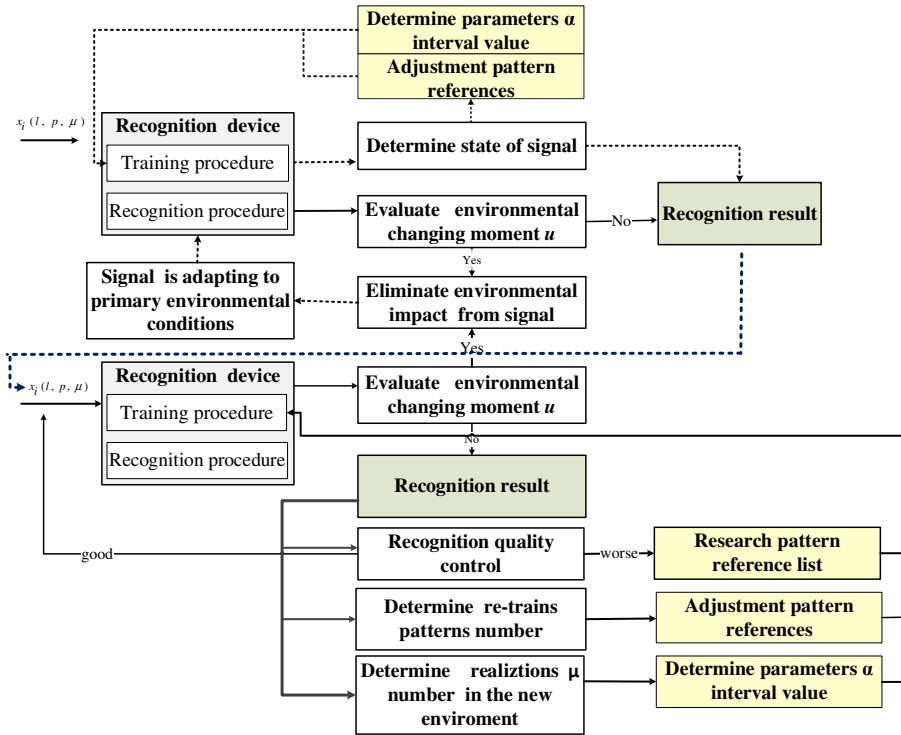


Fig. 1. The model of recognition system adaptation to changing environment conditions

At the recognition system adaption process is very important to detect the event u when dynamic system shifts to new environment $A_p (p = 2, \dots, P)$ and to start re-train procedure. Fig. 2 presents adaptation procedure when the moment u is detected.

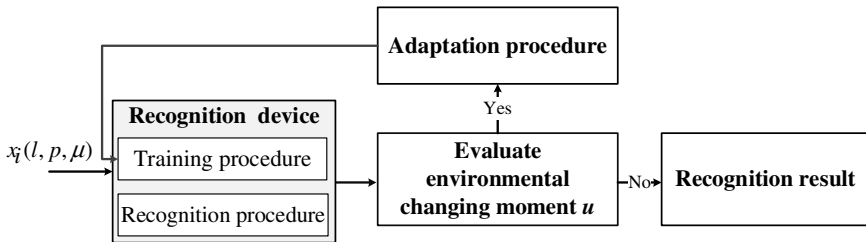


Fig. 2. Procedure of recognition system adaptation to the new environment

Adaptation procedure can make mistakes during re-training procedure, i. e. procedure can use invalid features that have direct influence to the quality $P(x_i(l, p, \mu))$ of state recognition during the formation of state $\Omega_l (l = 1, 2)$ reflecting patterns $E_{l_1} (l_1 = 1, \dots, L_1)$ (16). Regular pattern re-checking procedure helps to detect fallacious patterns and to remove them from the training list.

4 Experimental Investigation

The efficiency of the method is demonstrated by experiment where the recognition precision of two signals set $\Omega_l (l = 1, 2)$ when system recognizing signals at the moment u shifts from environment A_1 to environment A_2 is observed.

To make decisions ds , to decide to which state $\Omega_l (l = 1, 2)$ observed realization $x_i(l, p, \mu)$ depends the recognition device based on the method of dynamic time warping is used [1]. Sakoe-Chiba method is used to find distances $\Delta(x_i(l, p, \mu), E_l)$ [2], [8]. The experimental analysis is done for described situation. $\Omega_l (l = 1, 2)$ states are generated to control signals (2) in the environment A_1 .

$$x_i(1, 1, \mu) = 0.3x_{i-1}(1, 1, \mu) + 0.51x_{i-2}(1, 1, \mu) + v_i (i = 1, \dots, n) \quad (26)$$

$$x_i(2, 1, \mu) = 0.55x_{i-2}(1, 1, \mu) + v_i (i = 1, \dots, n) \quad (27)$$

Ω_1 state signals in environment A_2 are

$$x_i(1, 2, \mu) = x_i(1, 1, \mu) + x_i(\mu) (i = 1, \dots, 20000; \mu = 1, \dots, 200) \quad (28)$$

where

$$\begin{aligned} x_i(\mu) &= -0.75x_{i-1}(\mu) - 0.5x_{i-2}(\mu) + v_i(\mu) \\ (i &= 1, \dots, 20000; \mu = 1, \dots, 200) \end{aligned} \quad (29)$$

Ω_2 state signals in environment A_2 are

$$\begin{aligned} x_i(2, 2, \mu) &= x_{i-1}(2, 1, \mu) + x_i(\mu) \\ (i &= 1, \dots, 20000; \mu = 1, \dots, 200) \end{aligned} \quad (30)$$

At first recognition device is trained to recognize realizations $x_i(l, p, \mu)$ of states $\Omega_l (l = 1, 2)$ that exist in environment $A_p (p = 1)$. States $\Omega_l (l = 1, 2)$ are trained with patterns $E_{l_1} (l_1 = 1, 2)$ (5)

$$E_{l_1} (l_1 = 1) = \{e_1(l_1); e_2(l_1); e_3(l_1); e_4(l_1); e_5(l_1)\} \quad (31)$$

$$E_{l_1} (l_1 = 2) = \{e_1(l_1); e_2(l_1); e_3(l_1); e_4(l_1); e_5(l_1)\} \quad (32)$$

Here the interval range of environments change moment u defining parameter $\alpha_p (p = 1)$ values variation $\alpha_p (p = 1) \in [\xi_{S_1}; \xi_{S_2}]$ is defined when state $\Omega_l (l = 1, 2)$ functions in environment A_1 . $\xi_{S_1} = 0.01; \xi_{S_2} = 0.04$.

During the research dynamic system in environment A_1 and A_2 generates the amount of $\mu = 200$ realizations $x_i(l, p, \mu) \in \Omega_l (l = 1, 2)$.

Performance sequence $\varphi(i_2 = 1, 2, \dots, 400)$ for state realization $x_i(l, p, \mu) \in \Omega_l (l = 1, 2)$ is modeled (20) when $n_{11} = 30, n_{12} = 20$.

Recognition process is presented in Fig. 1 and experiment recognition results are presented in Fig. 3.

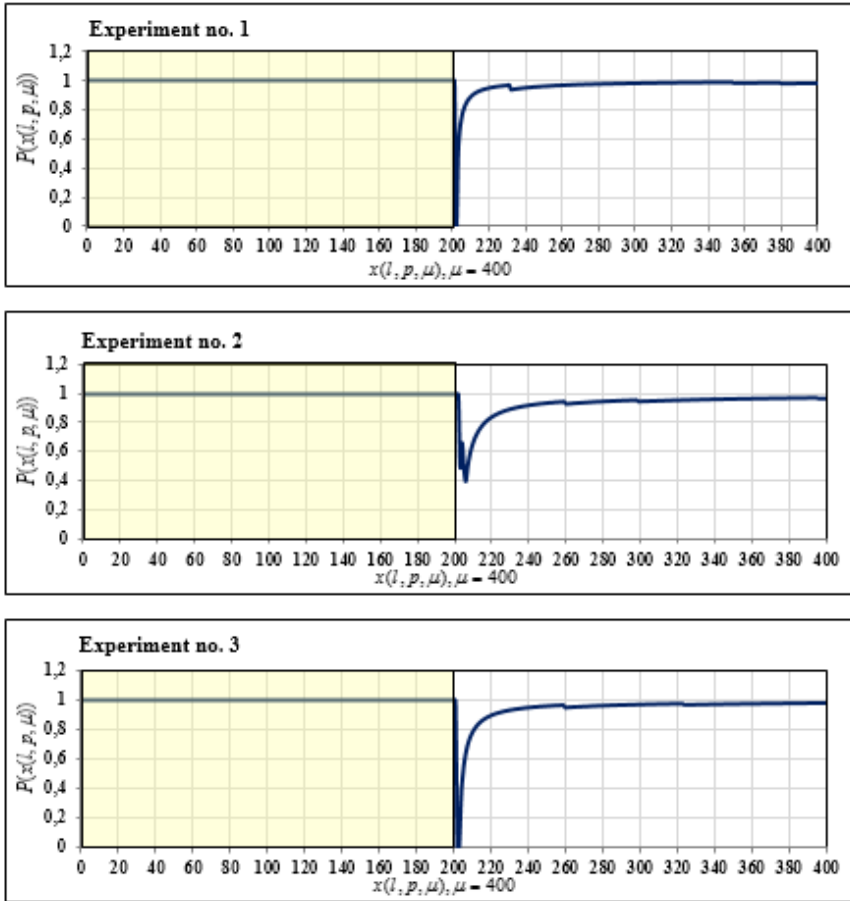


Fig. 3. Experiment result of random signal recognition

In yellow background are presented recognition results when dynamic systems functions in environment A_1 , and in white background are presented recognition results when dynamic systems functions in environment A_2 .

The analysis was made using three different appearance in environment $A_p (p = 1, 2)$ sequences of states $\Omega_l (l = 1, 2)$ realizations $x_i(l, p, \mu)$. The sequence of state realization appearance was different thus received recognition results are also different.

Experiment helped to detect problems that influence training procedure:

States $\Omega_l (l = 1, 2)$ pattern retraining can be complicated if state $\Omega_l (l = 1, 2)$ is generating realization $x_i(l, p, \mu)$ performance on low frequency;

It is especially important to evaluate correctly the points of parameters $\alpha_p (p = 1, \dots, P)$ sharpness interval $[\xi_{S_1}; \xi_{S_2}]$ (9). If points during the basic systems training were defenced incorrectly, the quality of systems recognition $P(x_i(l, p, \mu))$ (16) is received false and the recognition system can fail to adapt to the new environment conditions.

If incorrect features were used to retrain state $\Omega_l (l = 1, 2)$ pattern $E_{l_1} (l_1 = 1, \dots, L_1)$ in new environment, it negatively influences the control recognition quality $P(x_i(l, p, \mu))$

Analyzed adaptive recognition system model (Fig.1) is not dependent on the device that accepts the recognition decisions. The functioning of model is based on received recognition results, thus this model can be adapted to other recognition devices that use various methods.

5 Conclusions

In this paper a constructive method of random signals recognition precision assessment when random signals recognition environment changes is presented.

The method opens possibilities:

- To determine time moments u when random signals recognizing system shifts from one environment to another;
- To adapt recognition systems to new environments in which systems were not trained;
- To evaluate random signals recognition precision when adaptation procedure is in process in new environment.

As a result of this research is a software system allowing to make experiments assessing random signals recognition peculiarities when recognition conditions change.

References

1. Theodoridis, S., Pikrakis, A., Koutroumbas, K., Cavouras, D.: Introduction to Pattern Recognition: A Matlab Approach. Elsevier (2010)
2. Čeidaitė, G., Telksnys, L.: Analysis of factors influencing accuracy of speech recognition. Electronics and Electrical Engineering 9(105), 69–72 (2010)
3. Šalna, B., Kamarauskas, J.: Evaluation of Effectiveness of Different Methods in Speaker Recognition. Electronics and Electrical Engineering 2(98), 67–70 (2010)
4. Rudžionis, A., Ratkevičius, K., Rudžionis, V.: Speech in Call and Web centers Electronics and Electrical Engineering 3(59), 58–63 (2005)
5. Tamuliavičius G.: Pavienių žodžių atpažinimo sistemų kūrimas Ph D. Thesis (2008)

6. Moler, C.: Numerical Computing with Matlab, pp. 257–265 (2004)
7. Lawrence, R.R., Ronald, W.: Schaf: Introduction to Digital Speech Processing Foundations and Trends® in Signal Processing 1(1-2), 1–194 (2007)
8. Müller, M.: Information Retrieval for Music and Motion, pp. 69–70
9. Bachu, R.G., Kopparthi, S., Adapa, B., Barkana, B.D.: Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal. ASEE (2008)
10. Prasad Das, B., Parekh, R.: Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers. International Journal of Modern Engineering Research (IJMER) 2(3), 854–858 (2012)

Automated Method for Software Integration Testing Based on UML Behavioral Models

Dominykas Barisas, Eduardas Bareiša, and Šarūnas Packevičius

Software Engineering Department, Kaunas University of Technology,
Studentų st. 50, LT-51368 Kaunas, Lithuania

{dominykas.barisas, eduardas.bareisa, sarunas.packevicius}@ktu.lt

Abstract. Nowadays, testing is often considered more important than the code itself. Therefore, in order to test large and complex systems test automation methods are needed, which help evaluating whether the software is working properly. The main goal of the research is to improve effectiveness of integration testing by creating an automated method based on UML behavioral models. Test input data generation using symbolic execution was applied and it gave full structural coverage, which increased the quality of integration testing. Testing method allowed automating the testing process and increased the effectiveness of tests in comparison with other methods. Experiments showed that 96% of all mutations were successfully detected, and automated test data generation based on symbolic execution increased the detection of mutants by 6-19% in comparison to other model-based testing methods.

Keywords: software testing, integration testing, model-based testing, symbolic execution, constraint solver.

1 Introduction

Testing is a part of software development, which is used in every phase of the development cycle and takes more than 50% of the software development time [3, 22].

Software bugs can cause system crashes, incorrect behavior and huge losses. Therefore reliability and quality of software is important. Software design and implementation conformance is required for good quality.

Integration testing is the level of test used to check how the different parts of a system work together and whether the components of a system communicate and pass correct information between each other [9]. All objects or modules are integrated to form the complete software package as it is usually indicated by the high level design. The average effectiveness of detected integration errors is 35% [19] compared to other testing phases, which shows the importance of integration testing.

In order to lower costs of software development various techniques of automatic integration testing are used. The aim of this work is to investigate new method of automatic software integration testing based on software models and symbolic execution.

This paper is organized as follows: section 2 gives an overview of existing integration testing methods, then section 3 describes integration testing method, which

makes use of the concepts provided in section 2. Further, section 4 evaluates the created test method practically and provides experimental results. The paper concludes with section 5.

2 Related Work

The use of models has become popular not only in software design and development, but is widely used for testing as well [5]. There is a number of advantages as well as difficulties and shortcomings of various model-based approaches. Many object-oriented methods have been used as solutions to address the increasing demand for assuring software quality. Many different UML models have been used for object integration testing including state machine, sequence and communication diagrams.

Existing model based integration methods and test data generation techniques are discussed in this section.

2.1 Model Based Integration Testing

One of the researches in the area of integration testing has been made for software component testing [25]. The proposed methodology is based on model of component integration testing. However, this kind of test automation is still under investigation and requires further researches.

D. Sokenou proposes test sequence generation from sequence diagrams used together with state diagrams [24]. Tests are generated for integration testing. In this work, the main information is parsed from sequence diagrams, and the object states derived from state diagrams. However, this approach does not completely solve an oracle problem, and it does only propose a technique for test path generation.

Garousi describes a technique based on UML 2.0 sequence diagrams [12]. To analyze control flow Concurrent Control Flow Graph (CCFG) is generated. OCL is used to define a consistency rules between a sequence diagram and a CCFG.

Fraikin defines a method for testable sequence diagrams generation and a test tool that supports automated generation of test stubs, called SeDiTeC [10]. System behavior is specified using sequence diagrams. These diagrams are used as an input by SeDiTeC and it generates test stubs. This method uses only information provided by the sequence diagrams and does not describe how test input data could be generated.

Hartmann describe an approach for generating and executing system tests [13]. System behavior is modeled using UML use cases and activity diagrams. Test coverage is measured by the transition coverage.

Kansomkeat introduced a method that automatically generates and selects test cases from UML state chart diagrams [14]. Initially, an intermediate model is constructed, which is called Testing Flow Graph (TFG) and represents flows from UML state chart diagrams. Secondly, test cases are generated from TFG. Test coverage is defined by the coverage of the state and transition of diagrams.

The technique presented in this work improves integration testing of object-oriented software by taking into account all class states interacting in a sequence diagram. In Table 1 various methods were compared using different aspects, such as the use of UML models, presence of OCL constraints, identification of object states, test sequence generation and path coverage, automated test input data generation and oracle problem.

Table 1. Comparison of model based integration testing methods

Test method	Use of UML models	Use of OCL constraints	State based approach	Test sequences (path coverage)	Automated test data generation	Oracle problem solved
Test generation method using UML collaboration diagrams [1]	x			x		
TOTEM - state based class testing method [7]	x	x	x			x
UML based system testing method [6]	x	x		x		
UML based integration testing method [26]	x	x	x			
Test case generation method using UML sequence diagrams and OCL expressions [17]	x	x		x		
Test sequence generation method based on UML sequence diagrams [23]	x	x		x		
Test sequence generation method based on UML sequence and state diagrams [24]	x	x	x	x		
SeDiTeC – testing method based on sequence diagrams [10]	x					
Test case generation method using UML statechart diagrams [14]	x		x	x		x
Test case generation method based on state diagrams [16]	x		x	x		x
UML based test sequence generation method [15]	x	x		x		x
Statistical Test Cases generation method using UML State Diagrams [8]	x		x		x	x
Testing method for interacting software components [11]	x		x	x		
SCOTEM - a state-based integration testing method based on UML models [2]	x	x	x	x		x

Most of the compared model based methods use UML models and OCL constraints for test case/sequence generation providing path traversal algorithms. However, only a few of them propose a way for automated test input data generation and solve oracle problem including the experimental method evaluation. In this research, we aim to cover all the analyzed aspects. System object interactions in all possible states are modeled using state machine and sequence diagrams. Test paths are generated, which compose a graph and then executed using various coverage criteria.

2.2 Test Data Generation Techniques

Many researchers investigated automatic test data generation in recent years. A lot of attempts to automate the test generation process are limited by the size and complexity of software. Metaheuristic search techniques offer several means to solve these problems [20]. They include such algorithms as Hill Climbing, Simulated Annealing and Evolutionary Algorithms.

Structural test data generation is based on analysis of the internal structure of the program, but it is not required to execute the program. Symbolic Execution is used for test data generation, which assigns expressions to program variables while going through the code structure. Concolic execution is similar to symbolic execution, where the execution is based on symbolic inputs and solving constraints of a program path. Concolic execution includes symbolic execution with concrete random execution of the program, which allows simplifying constraints based on the concrete values. Symbolic execution based test generation is directed and test inputs are generated by taking program paths at the symbolic level, and these inputs are guaranteed to execute the paths determined in advance. Therefore, these test inputs are not redundant and each of them executes a different program path.

Structure oriented approaches represent a more successful strategy and executed a number of program structures, which gives better code coverage because each structure gets attention in the form of an individual search. Symbolic execution helps implementing a number of code coverage metrics, like generating test inputs that invokes all lines of software code at least once and executes all return statements. Data generated in this way allows reaching high code coverage and executes each path of a method at least once.

3 Concept of Testing Process

Proposed method combines the integration testing method based on UML behavioral models and automated test data generation using symbolic execution, and it allows solving a test oracle problem. The presented method generates test cases from UML sequence and state machine diagrams, which allows testing a correct class integration of object-oriented software. Analysis of the source code of system under test allows automatically generating test input data, which provides complete structural method coverage. This leads to automation of all phases of testing: test generation, execution, and result evaluation.

3.1 Main Steps

The purpose of the proposed test approach is to detect faults related with the interactions among objects in a system. The process can be separated into activities illustrated in Fig. 1 [4].

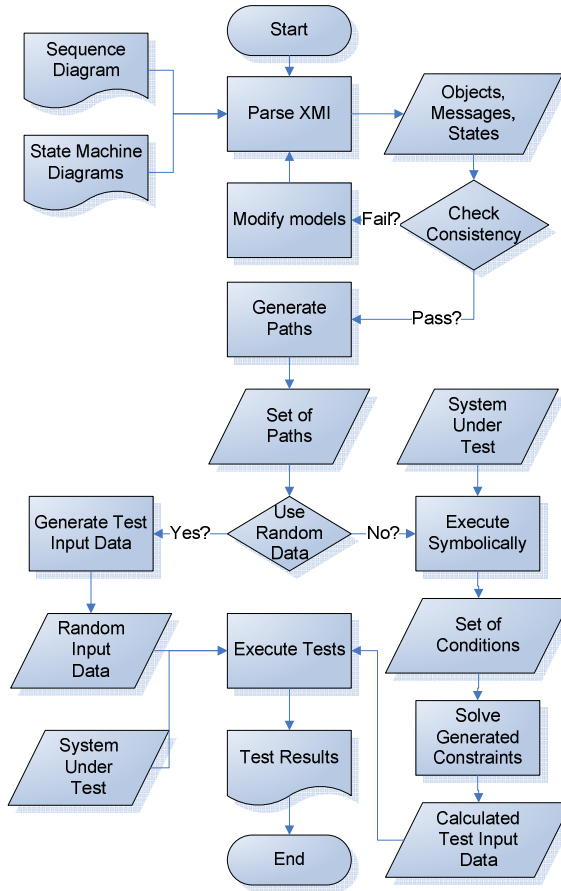


Fig. 1. A flowchart representing a concept of the proposed testing process

The process begins with reading UML diagrams. XMI Parser parses the sequence and state machine diagrams and generated test paths for the integration testing. Generated test paths are executed using the specified path coverage criteria. Test framework initializes each object of the system under test, uses the generated test data and runs the test. Test results are returned by the test framework.

There is a need to specify guard conditions (state invariants). Each object in the sequence diagram corresponds to an instance of a class and should have a corresponding state machine diagram. Connections in the proposed model can emerge between objects in the sequence diagram and between states in the state machine diagram.

Connections in the sequence diagram should have a unique number, operation, receiver and sender objects. Connections between states include a unique number, operation, source state and target state. The algorithm iterates through the messages in the sequence diagram and reads related information from state machine diagrams. All possible states of the source and target objects are collected.

3.2 Test Data Generation

Symbolic execution based test data generation algorithm is provided below:

```

Algorithm GenerateTestData(SC): TESTDATA
Input
SC: source code of one method
Output
TESTDATA: Generated concrete variable values

1. PATHCONSTRAINT := emptyStack
2. for all LINE ∈ SC do
3.   if ( conditionalStatement(LINE) )
4.     CONDITION := extractCondition(LINE, true)
5.     addConstraint(CONDITION, PATHCONSTRAINT);
6.   end if
7. end for
8. index := Length(PATHCONSTRAINT) - 1
9. while not empty(PATHCONSTRAINT) do
10.  c := pop(PATHCONSTRAINT)
11.  if ( PATHCONSTRAINT and not c )
12.    input := solve(PATHCONSTRAINT, not c)
13.    TESTDATA := generateData(PATHCONSTRAINT, input, index)
14.    index := index - 1
15.  end if
16. end while
17. return TESTDATA

```

The constraint programming solver is used to generate test input values by evaluating the conditions. A set of conditions represent a state and one value is found for each state.

4 Evaluation of Test Generation Method

This section describes the implementation details of the testing framework and provides experimental results.

4.1 Implementation Details

Various tools and libraries were used to implement the proposed method and evaluate results practically. First of all, UML2 models were built using IBM Rational Software Architect, Version 7.0.0 and exported as XMI files.

A tool that was used to execute programs symbolically is Java PathFinder (JPF). It takes the source code of the system under test (SUT), executes it using symbolic values and returns a set of conditions for each path.

4.2 Experimental Evaluation

This section evaluates the proposed integration testing method against existing systems. Models of two software applications (Elevator and ATM) were created in order to evaluate the proposed method. Test paths were composed and executed using the generated test data. A graph was constructed from UML models and test paths were composed from the sequences of messages. Finally, the test results were compared with the expected results, where the expected results are the object states before and after the message call, which allows solving an oracle problem.

The goal of experiment is to evaluate the effectiveness of the generated tests, detect faults and compare the test quality having different program coverage with tests and test data generation techniques. A number of mutation operators [18, 21] were used to evaluate the method.

Test results using random test input data generation are provided in Fig. 2.

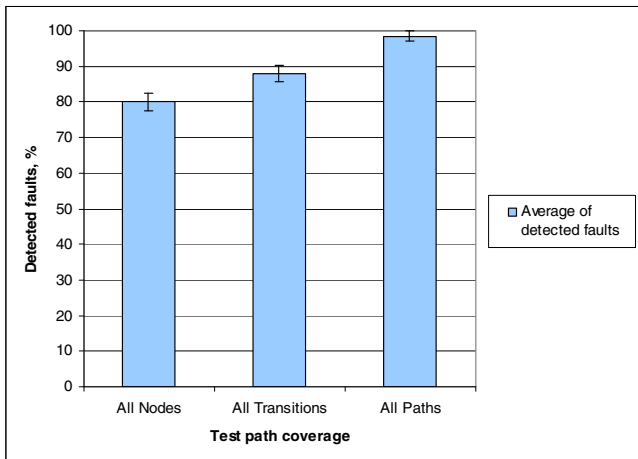


Fig. 2. Number of detected faults for various graph coverage criteria

Different graph coverage criteria gave different results. The more paths are covered, the more mutations detected.

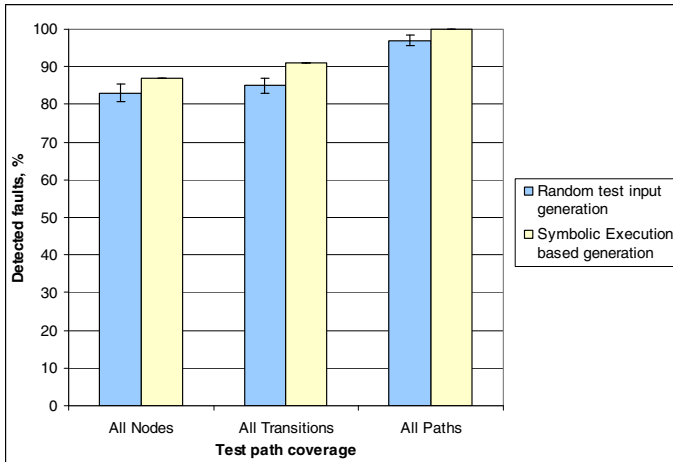


Fig. 3. Random test input generation against symbolic execution based generation

Fig. 3 shows the number of detected faults using random test input generation and symbolic execution based test input generation where a set of test inputs are generated to cover all edges, all transitions and all paths of a message, which trigger state transition.

The result shows that the number of detected faults increased by 9%. Comparing to random test input generation this number increased for those methods, which contain conditional statements and not all of the branches changed the state of a class.

Table 2 provides metrics of the SUT: lines of code (LOC) and cyclomatic complexity (CC) for both applications (Elevator and ATM) that were tested.

Table 2. Metrics of the tested applications

System Under Test	LOC	CC
Elevator	416	6
ATM	3048	12

Number of test cases generated for each application is given in Table 3.

Table 3. Number of generated tests for the tested applications

Test method	Elevator		ATM	
	Test cases	Method invocations	Test cases	Method invocations
Full path coverage	9	177	12	255
Full path coverage + full method coverage	9	398	12	568

It can be noted that although test methods have the same number of generated test cases, the number of methods to be called is different. The second method has a larger

number of generated tests because a greater amount of test inputs leads to a greater amount of method invocations, which are used to cover all branches within a method. Moreover, methods have different parameter types: *boolean* type parameters can take only two values - *true* or *false*, but *integer* and *real* numbers can be selected from a relatively infinite set of values.

Table 4 provides comparison of generated test metrics:

- ✓ The average test efficiency (ATE) - the percentage shows the average number of test cases that detected faults;
- ✓ Number of undetected faults (UFN) - the amount of mutations undetected by generated tests;
- ✓ Percentage of undetected faults (UFP) – percentage of undetected mutations compared to a number of introduced mutations.

Table 4. Comparison of generated test metrics

Metric	Elevator		ATM	
	Full path coverage	Full path coverage + full method coverage	Full path coverage	Full path coverage + full method coverage
ATE	17 %	18,8 %	14,2 %	15,4 %
UFN	5	1	7	0
UFP	23 %	0,05 %	22 %	0 %

The average test efficiency is not high, but more than half of the faults are found in generated test sets. It means that not every test of the entire set of tests detects a specific mutant, but the majority of mutants is still detectable.

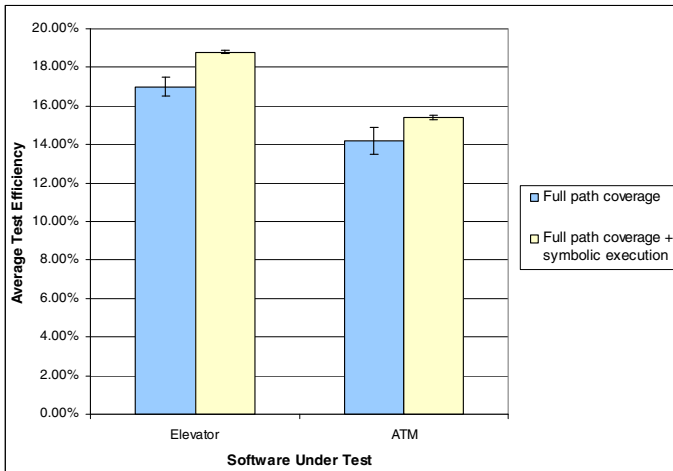


Fig. 4. Average test efficiency

However, the number of generated tests cannot be reduced because it is an optimal set and otherwise it may not detect some of mutations. The average test efficiency is presented in Fig. 4. As can be seen from the results, in some cases a number of undetected mutants can be reduced to 0. This means that automatically generated tests and test data may find all mutants in software.

Method was compared to other state-based integration testing method SCOTEM [2], which uses collaboration and state diagrams and manually provides test input data for test case generation. Fig. 5 depicts the effectiveness of both methods having single and all transition path coverage. Arithmetic tutor application was used as a SUT.

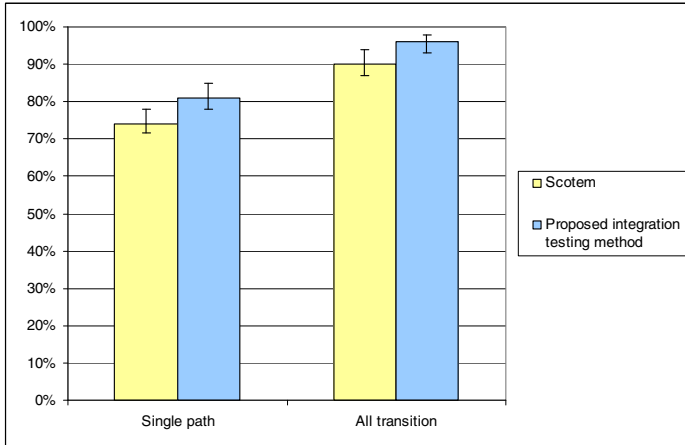


Fig. 5. Proposed method effectiveness against SCOTEM

Furthermore, the proposed method was compared to other methods named State based class testing method [7] and SCOTEM taking the average of test effectiveness of all tested applications. Fig. 6 illustrates the results.

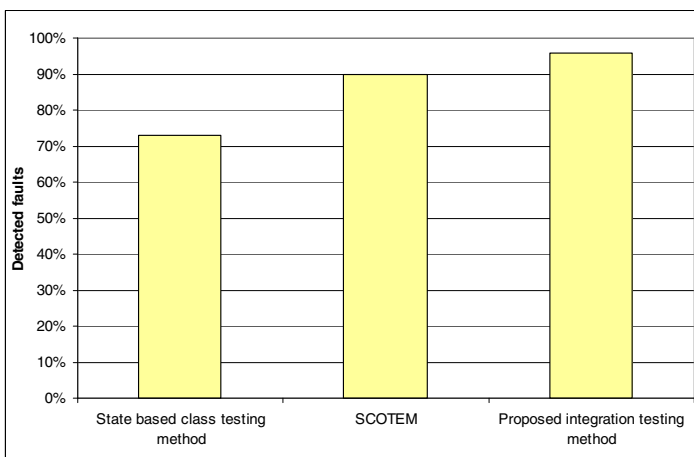


Fig. 6. Proposed method comparison with other methods

The proposed method gave better results approximately by 6-19%. Automated test input generation using symbolic execution and the combination of different UML behavioral models had most of the impact on the better results.

4.3 Experimental Results

Experiments showed that 80-96% of the mutations were detected successfully. Symbolic execution based test input data generation gave better results compared against random test input generation and mutant detection was increased by 6-9%. However, it took more time because symbolic execution looked for symbolic values and had to solve the constraints in order to generate test input values. Time increases when trying to cover more paths. One of the future investigations could be finding a solution to identify a set of paths, which has the highest possibility to detect faults in the system.

Moreover, the method was compared to existing method called SCOTEM. Results showed 6% better fault detection rate using the proposed method which was mostly because of improved test input data generation.

In order to use the proposed testing method, it is needed to be aware of it in advance - in the design phase of the software life cycle, because the architecture models needs to be compliant with the testing requirements. The testing method is rather complex and needs to be simplified in order to use it for integration testing widely. In real software systems, the number of state transitions can grow exponentially, therefore the testing process may become difficult and time consuming.

One of the future investigations could be finding a solution to identify a set of paths, which has the highest possibility to detect faults in the system. This could reduce time needed for test case generation and execution. Another improvement of this method could be an optimization for a better quality test input data generation.

5 Conclusions

This paper presented a method for the testing process based on software behavioral models. Using this method, the object graph was generated and the system was tested by checking communication between objects in different states. The following tasks were discussed and accomplished:

1. The effectiveness of integration testing method was evaluated with benchmark applications (ATM, Elevator), different test path coverage criteria were applied. A number of mutants were introduced in the Elevator and ATM application. Experiments showed that 96% of the mutations were detected successfully;
2. The effectiveness of test input generation based on symbolic execution was evaluated against random test input generation. The need for random test input generation was eliminated and parameter values were calculated to reach higher code coverage. Experiments showed that the introduced method increased mutant detection by 6-9%;
3. The method was compared with existing methods using the implemented testing framework. Results showed that fault detection rate was improved by 6-19% using

the proposed method due to the combination of different UML behavioral models and automated test input generation using symbolic execution. Moreover, it was noticed that test execution time depends on the path coverage criterion, therefore it is advised to use all transition coverage, which still gives 90% and higher error detection rate and reduces test generation time.

References

1. Abdurazik, A., Offutt, J.: Using UML collaboration diagrams for static checking and test generation. In: Evans, A., Caskurlu, B., Selic, B. (eds.) UML 2000. LNCS, vol. 1939, pp. 383–395. Springer, Heidelberg (2000)
2. Ali, S., Briand, L.C., Rehman, M.J.-U., Asghar, H., Iqbal, M.Z.Z., Nadeem, A.: A state-based approach to integration testing based on UML models. *Inf. Softw. Technol.* 49(11–12), 1087–1106 (2007)
3. Ammann, P., Offutt, J.: *Introduction to Software Testing*. Cambridge University Press (2008)
4. Barisas, D.: Automated method for software integration testing based on UML behavioral models. Dissertation, Kaunas University of Technology, p. 110 (2012)
5. Bouquet, F., Grandpierre, C., Legnard, B., Peureux, F., Vacelet, N., Utting, M.: A subset of precise UML for model-based testing. In: *Proceedings of the 3rd International Workshop on Advances in Model-Based Testing*, pp. 95–104. ACM, London (2007)
6. Briand, L.C., Labiche, Y.: A UML-Based Approach to System Testing. In: Gogolla, M., Kobryn, C. (eds.) UML 2001. LNCS, vol. 2185, pp. 194–208. Springer, Heidelberg (2001)
7. Briand, L.C., Penta, M.D., Labiche, Y.: Assessing and Improving State-Based Class Testing: A Series of Experiments. *IEEE Trans. Softw. Eng.* 30(11), 770–793 (2004)
8. Chevalley, P., Th, P.: #233, and venod-Fosse, Automated Generation of Statistical Test Cases from UML State Diagrams. In: *Proceedings of the 25th International Computer Software and Applications Conference on Invigorating Software Development*, pp. 205–214. IEEE Computer Society (2001)
9. Craig, R.D., Jaskiel, S.P.: *Systematic software testing*. Artech House (2002)
10. Fraikin, F., Leonhardt, T.: SeDiTeC-testing based on sequence diagrams. In: *Proceedings of the 17th IEEE International Conference on Automated Software Engineering, ASE 2002*, pp. 261–266 (2002)
11. Gallagher, L., Offutt, J.: Automatically testing interacting software components. In: *Proceedings of the 2006 International Workshop on Automation of Software Test*, pp. 57–63. ACM, Shanghai (2006)
12. Garousi, V., Briand, L.C., Labiche, Y.: Control Flow Analysis of UML 2.0 Sequence Diagrams. In: Hartman, A., Kreische, D. (eds.) *ECMDA-FA 2005*. LNCS, vol. 3748, pp. 160–174. Springer, Heidelberg (2005)
13. Hartmann, J., Imoberdorf, C., Meisinger, M.: UML-Based integration testing. In: *Proceedings of the 2000 ACM SIGSOFT International Symposium on Software Testing and Analysis*, pp. 60–70. ACM, Portland (2000)
14. Kansomkeat, S., Rivepiboon, W.: Automated-generating test case using UML statechart diagrams. In: *Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology*, pp. 296–300. South African Institute for Computer Scientists and Information Technologists (2003)

15. Kim, S.-K., Wildman, L., Duke, R.: A UML Approach to the Generation of Test Sequences for Java-Based Concurrent Systems. In: Proceedings of the 2005 Australian Conference on Software Engineering, pp. 100–109. IEEE Computer Society (2005)
16. Kim, Y.G., Hong, H.S., Bae, D.H., Cha, S.D.: Test cases generation from UML state diagrams. IEE Proceedings - Software 146(4), 187–192 (1999)
17. Li, B.-L., Li, Z.-S., Qing, L., Chen, Y.-H.: Test Case Automate Generation from UML Sequence Diagram and OCL Expression. In: Proceedings of the 2007 International Conference on Computational Intelligence and Security, pp. 1048–1052. IEEE Computer Society (2007)
18. Ma, Y.-S., Kwon, Y.-R., Offutt, J.: Inter-Class Mutation Operators for Java. In: Proceedings of the 13th International Symposium on Software Reliability Engineering, p. 352. IEEE Computer Society (2002)
19. McConnell, S.: Code complete, pp. 463–477. Microsoft Press (2004)
20. McMinn, P.: Search-based software test data generation: a survey: Research Articles. *Softw. Test. Verif. Reliab.* 14(2), 105–156 (2004)
21. Offutt, J., Ma, Y.-S., Kwon, Y.-R.: The class-level mutants of MuJava. In: Proceedings of the 2006 International Workshop on Automation of Software Test, pp. 78–84. ACM, Shanghai (2006)
22. Patton, R.: Software Testing. Sams (2000)
23. Samuel, P., Joseph, A.T.: Test Sequence Generation from UML Sequence Diagrams. In: Proceedings of the 2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, pp. 879–887. IEEE Computer Society (2008)
24. Sokenou, D.: Generating Test Sequences from UML Sequence Diagrams and State Diagrams. *GI Jahrestagung*, 236–240 (2006)
25. Weiqun, Z.: Model-Based Software Component Testing: A UML-Based Approach, 891–899 (2007)
26. Wu, Y., Chen, M.-H., Offutt, J.: UML-Based Integration Testing for Component-Based Software. In: Erdogmus, H., Weng, T. (eds.) ICCBSS 2003. LNCS, vol. 2580, pp. 251–260. Springer, Heidelberg (2003)

Computational Algorithmic Generation of High-Quality Colour Patterns

Alfonsas Misevičius¹, Evaldas Guogis², and Evelina Stanevičienė³

¹ Kaunas University of Technology, Department of Multimedia Engineering,
Studentų st. 50–400/416a, LT–51368 Kaunas, Lithuania
alfonsas.misevicius@ktu.lt

² Singleton Labs, Narucio st. 31/Studentų st. 65–307, LT–51405/51368 Kaunas, Lithuania
evaldas.guogis@singleton-labs.lt

³ Kaunas University of Technology, Department of Multimedia Engineering,
Studentų st. 50–400/416a, LT–51368 Kaunas, Lithuania
evelinastaneviciene@yahoo.com

Abstract. The purpose of this paper is to describe the computational algorithmic generation of high-quality colour patterns (digital halftones). At the beginning, the formal model for generation of the digital halftones, the so-called grey pattern problem (GPP) is introduced. Then, the heuristic algorithm for the solution of the grey pattern problem is discussed. Although the algorithm employed does not guarantee the optimality of the solutions found, still perfect quality, near-optimal (and in some cases probably optimal) solutions can be achieved within reasonable computation time. Further, we provide the preliminary results of the extensive computational experiments with the extra-large instance (data set) of the GPP. As a confirmation of the quality of the analytical solutions produced, we also give the visual representations of fine-looking graphic halftone patterns.

Keywords: combinatorial optimization, heuristic algorithms, grey pattern problem, digital (colour) halftoning, creative multimedia.

1 Introduction

In this paper, we are concerned with the automatic computer-based generation of the superior-quality colour patterns (digital colour halftone textures and images). The computational algorithmic generation of the digital colour halftones is based on the formal mathematical-theoretical model — the so-called grey pattern problem (GPP). In the context of creation of fine grey textures and digital halftoning, the grey pattern problem was firstly formulated mathematically by Taillard in 1995 [11]. Despite the fact that the initial formulation of the GPP was made in the context of generation of grey (black-white) patterns [13], the GPP can also be applicable to colour halftones [4, 5], so that perfect (ideal) halftones for every colour (both monochromatic and non-monochromatic) can be computationally (algorithmically) constructed. This is also true for the digital colours which are encoded by using the well-known colour

coding model RGB (see the internet page: RGB color model - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/RGB_color_model).

The main contribution of this work is that the new, extra-large scale instance of the GPP is created and superior-quality colour patterns are generated using an efficient heuristic optimization algorithm. We think that the described approach might be helpful in domains such as creative multimedia and computer graphic arts.

The paper is organized as follows. In Section 2, the mathematical formulation of the grey pattern problem is considered. In Section 3, a heuristic algorithm for the computational solution of the GPP is discussed. Further, we present both analytical and graphical results of the extensive computational experiments on the extra-large instance of the GPP. Finally, the brief concluding remarks are given.

2 The Grey Pattern Problem — Formal Model of Generation of the Digital Halftones

2.1 Preliminaries: Formal Model of Generation of the Digital Halftones

Suppose that there is given a grid (matrix) consisting of n "nodes" — small identical squares (or simply, discrete points, "pixels") (also see [11]). The squares are evenly positioned on n_1 horizontal rows and n_2 vertical columns, such that $n = n_1 \times n_2$. Without loss of universality, assume that points are either white or black. (White colour can be corresponded to a surface's (image's) background colour, whereas the black colour may serve as a foreground colour.) Let m ($m \leq n$) be the total number of the black squares, then the number of white squares is equal to $n - m$. A pattern of m black points/squares is formed from n points in the grid and, by juxtaposing these patterns, one gets a grey surface (frame) (see Fig. 1). The grey density (intensity) of the frame is equal to $\frac{m}{n} \left(0 \leq \frac{m}{n} \leq 1 \right)$. The less value of $\frac{m}{n}$, the closer

the frame colour is to background colour, and vice versa. If $m = 0$, then one simply gets a background (white) colour; if $m = n$, then a foreground (black) colour is obtained.

The goal is to have a grey pattern where the black points are distributed as uniformly (regularly) as possible. In the other words, we are seeking to obtain the most fine, subtle grey halftone. This applies for every other foreground colour of RGB colour model. For the programming convenience, it may be recommended that the number of halftones (densities), n , be equal to 2 raised to a power λ , i.e., 2^λ (for example, 2^6 , 2^8 , and so on). Let c be the number of available foreground colours different from a background colour. Then, the maximum number of available halftones increases to nc ($= 2^\lambda c$). Furthermore, if there exist c foreground colours and c background colours, then the maximum available number of halftones becomes theoretically equal to $\frac{nc(c-1)}{2}$, which can offer new opportunities for creating of novel combined halftones (textures).

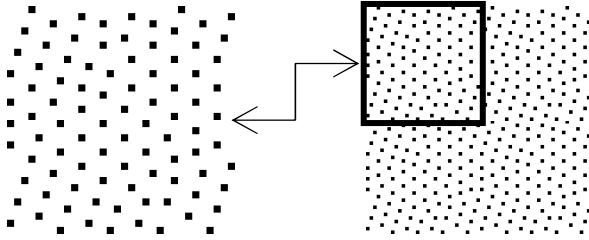


Fig. 1. A graphical illustration of grey surface (frame). The grid consisting of 1024 points (92 black points and 932 white points) (see the left side of the figure) is replicated four times (see the right side of the figure).

2.2 Mathematical Formulation of the Grey Pattern Problem

The grey pattern problem (GPP) is a special case of the well-known combinatorial optimization problem, the quadratic assignment problem [2]. Mathematically, the grey pattern quadratic assignment problem can be formulated as follows [11]. Given two matrices $A = (a_{ij})_{n \times n}$ and $B = (b_{kl})_{n \times n}$ and the set Π_n of permutations of the integers from 1 to n , find a permutation $p \in \Pi_n$ that minimizes

$$z(p) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} \cdot b_{p(i)p(j)}, \tag{1}$$

where the elements of the matrix $(a_{ij})_{n \times n}$ are defined as $a_{ij} = 1$ for $i, j = 1, 2, \dots, m$ and $a_{ij} = 0$ otherwise; the values of the matrix $(b_{kl})_{n \times n}$, i.e., the distances between every two of n points are defined according to the following rule [11]:

$$b_{kl} = b_{(r-1)n_2+s \ (t-1)n_2+u} = \omega_{rstu}, \ \omega_{rstu} = \max_{w_1, w_2 \in \{-1, 0, 1\}} \frac{1}{(r-t+w_1n_1)^2 + (s-u+w_2n_2)^2}, \tag{2}$$

where $k, l = 1, \dots, n, r, t = 1, \dots, n_1, s, u = 1, \dots, n_2, n_1 \times n_2 = n$. The interpretation of the quantity ω_{rstu} is as follows. We may consider m electrons that have to be put on the grid's squares. Then, ω_{rstu} may be thought of as a quantity proportional to repulsion force between two electrons i and j ($i, j = 1, \dots, n$) located in the grid positions $k = p(i)$ and $l = p(j)$ with the coordinates (r, s) and (t, u) ($r, t = 1, \dots, n_1, s, u = 1, \dots, n_2$). (The arrangement of electrons must be done in such a way that the sum of the intensities of the repulsion forces is minimized.)

According to the above formulation, p denotes a permutation and $p(i), p(j)$ denote the corresponding elements of the permutation. Every feasible permutation may be considered as a solution of the GPP. In this way, the objective is to find the best available, optimal solution, i.e., the permutation elements $p(1), p(2), \dots, p(m)$ ($1 \leq p(i) \leq n, i = 1, 2, \dots, m, m \leq n$) such that the sum $\sum_{i=1}^m \sum_{j=1}^m b_{p(i)p(j)}$ (considered as an objective function of the GPP) is as minimal as possible, that is:

$$z(p) = \sum_{i=1}^m \sum_{j=1}^m b_{p(i)p(j)} \rightarrow \text{minimum}. \tag{3}$$

This is the compact formulation of the grey pattern problem which is analogous to that formulated by Drezner in [3]. In formulation (3), only the matrix \mathbf{B} and the values of n , m are necessary; meanwhile, the matrix \mathbf{A} is not needed at all. In our work, we use the formulation (3), rather than the general formulation (1).

The i th ($i = 1, 2, \dots, m$) element of the found analytical solution, i.e., permutation p^* , $p^*(i) = (r - 1)n_2 + s$, gives the location in the grid where the i th black point has to be placed in. (Elements $p^*(m + 1), p^*(m + 2), \dots$ are disregarded.) The coordinates (r, s) of the black point in the grid are derived according to the formulas: $r = \lfloor (p^*(i) - 1)/n_2 \rfloor + 1, s = ((p^*(i) - 1) \bmod n_2) + 1$ (also see Fig. 2).

Note that many analytical solutions (permutations) with the same objective value ($z(p^*)$) may exist: exchanging two black (or two white) points does not change the value of the objective function z .

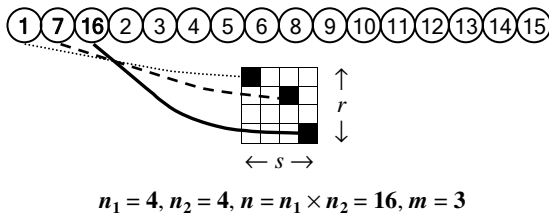


Fig. 2. A graphical illustration of correspondence of the analytical solution (permutation) to the graphical (black) points on grid's positions

2.3 Benchmark Instances of the Grey Pattern Problem

It is possible to create a wide spectrum of many different problems by varying the values of n, n_1, n_2, m and modifying the rule (2) for definition of the distances (values of the matrix \mathbf{B}). For given values of n_1, n_2 (usually, $n_1 = n_2$), the value of m (considered as the particular halftone number) varies between 1 and n ($n = n_1 \times n_2$); here n can be seen as the maximum number of available colour halftones. In this way, n different instances occur, however there is no need to run an optimization algorithm on all n instances: solutions for densities (halftones) $m = n/2 + 1, n/2 + 2, \dots, n$ can be obtained by symmetry from densities $m = n/2 - 1, n/2 - 2, \dots, 1$.

A medium-sized benchmark instance of the GPP is available at the public library of the instances of the QAP — QAPLIB (<http://www.seas.upenn.edu/qaplib>, also see [1]). The instance is called Tai64c (the numeral in the instance name indicates the size of the problem). Tai64c is the benchmark GPP instance for a square grid of 8 by 8 points ($n_1 = n_2 = 8, n = 64$) and $m = 13$ black points. In QAPLIB, there is also available other, larger problem instance denoted by Tai256c. Tai256c is the grey pattern problem for grid of dimensions 16 by 16 ($n_1 = n_2 = 16, n = 256$) and $m = 92$ black points. In [12], Taillard and Gambardella have considered 126 instances similar to Tai256c for $3 \leq m \leq 128$.

In this work, we have constructed a new, extra-large size benchmark GPP instance to enable the creation of colour patterns of the superior quality. The dimensions of the square grid are increased to 32 and 32 (i.e., $n_1 = n_2 = 32$). Thus, the size of the



problem (n) and the number of available halftones for a single RGB colour increase to 1024 (2^{10}). In contrast to the instances Tai64c, Tai256c, where the value of m is linked to the particular problem instance, our instance consists of the distance matrix only — so that the parameter m can be integrated within the optimization software and the software user can easily switch between various m 's. (For the convenience, the values b_{kl} of the distance matrix from formula (2) are scaled by a scaling factor 100000, which does not affect the formulation of the problem in any way.) To the best of our knowledge, this instance appears to be the largest benchmark instance of the GPP/QAP currently available, and it constitutes new interesting practical application, which has not been studied yet. (An electronic copy of the instance can be freely requested by contacting one of the authors.)

3 Algorithmic Solution of the GPP: A Hybrid Genetic-Evolutionary Algorithm

For the solution of the grey pattern problem, various optimization techniques — both exact and heuristic — can be employed. The exact algorithms are suitable only for small-sized problems [3]. For larger problem instances, heuristic methods are used [3, 6, 7, 9, 12]. In this work, a population-based hybrid genetic-evolutionary algorithm (HGEA), which has already been proven to be extremely efficient [7], is applied. A high-level pseudocode of HGEA is presented in Fig. 3. The main features of the algorithm are as follows¹.

- The algorithm HGEA operates directly with permutations which are associated with individuals' chromosomes. Our algorithm is a hybrid algorithm and it is mainly based on two intertwined processes, "interevolution" (global genetic-evolutionary process) and "intraevolution" (local process of improvement of the separate solutions (individuals)). As a heuristic improvement (intraevolutionary) algorithm, we employ an iterated tabu search (ITS) [8] based procedure (a series of runs of the self-contained tabu search procedure coupled with embedded random mutations) (see Fig. 4). The ITS procedure is applied in three situations: initial population construction, post-recombination (every offspring produced by a recombination procedure is subject to improvement), and post-restart improvement. The important aspect of the ITS optimizer is that it can find different resulting solutions even though the starting solution is the same.
- The size of the population of solutions is equal to PS , where the precise value of PS can be chosen by the algorithm user, but it is suggested that this value be not large. The initialization of the population is carried out in two phases. In the first phase, the population individuals (i.e., the corresponding chromosomes) are created in a random way. At the second phase, all the population members undergo extensive improvement process thanks to the ITS procedure.

¹ We are omitting both the explanation of the principles of functioning of genetic algorithms (GAs) and the details of the specialized, elaborated hybrid genetic-evolutionary algorithm. For thorough studies of the essence and principles of GAs, the interested readers are referred to [10]; whereas the detailed description of HGEA can be found in [7].

- We make usage of the specialized recombination procedure, which efficiently respects the specific structure of the GPP. This recombination principle was originally used by Drezner in [3], whereas we employ our own implementation [7]. The number of recombinations per one generation (i.e., iteration of interevolution) is directed by the parameter N_{offspr} , whose value can vary between 1 and PS . The parents for recombination are selected randomly.
- The mutation operation is implemented in a rather unconventional way. Mutations are based on controlled random interchanges of single elements of a permutation and are tweaked in the ITS procedure, instead of being a direct part of the genetic algorithm. The mutation process is regulated by the mutation rate (strength) parameter, rather than a mutation probability. Usually, there are N_{offspr} calls to the ITS procedure per single generation, which results in $N_{\text{offspr}} \times Q$ mutations per generation.
- For the population replacement, we apply a generational update strategy. Remind that, at ever generation, N_{offspr} offspring are produced. The population replacement is then carried out by removing N_{offspr} worst individuals and restoring the original population size, PS . In this way, the "elitism" is maintained. Additionally, a special kind of "hot restart" takes place to reorganize the similar chromosomes if the genetic variability is lost. The restarts are triggered in the cases when the population diversity (entropy) falls down below the specified threshold (for details, see [7]).

The genetic-evolutionary algorithm is continued until a pre-defined number of generations (iterations of interevolutions), N_{gen} , have been performed.

procedure HybridGeneticEvolutionaryAlgorithm;

//input: n – the problem size (total number of the grid points), m – the density parameter (number of black points),
 // B – distance matrix, PS – population size, N_{gen} – # of generations, N_{offspr} – # of offspring per generation
 //output: p^* – the best solution found

create the random initial population P of size PS ($P \subset \Pi_n, |P| = PS$);

apply intraevolution procedure to all population members;

for *interevolution_iteration* := 1 **to** N_{gen} **do begin** //interevolutionary process

for *recombination_iteration* := 1 **to** N_{offspr} **do begin** // N_{offspr} offspring are created at each iteration

$p^* := \underset{p \in P}{\text{argmin}} z(p)$; //the best so far solution is archived (saved)

 randomly select parents $p', p'' \in P$ for recombination;

 apply recombination procedure to p', p'' , get the offspring p''' ;

 apply intraevolution procedure to p''' ; $P := P \cup \{p'''\}$

endfor;

get new population P from the existing population ($|P| = PS$);

if population diversity is less than a predefined threshold **then begin**

 apply random mutations to all members of the current population, except the best one;

 apply intraevolution procedure to every mutated member;

if $z(\underset{p \in P}{\text{argmin}} z(p)) < z(p^*)$ **then** $p^* := \underset{p \in P}{\text{argmin}} z(p)$

endif

endfor.

Fig. 3. High-level pseudocode of the hybrid genetic-evolutionary algorithm

```

procedure Intraevolution (Iterated tabu search);
//input:  $p$  – current solution,  $\mu_{min}$ ,  $\mu_{max}$  – minimum and maximum mutation rates,
//       $Q$  – # of iterations of intraevolution,  $\tau$  – # of tabu search iterations,
//       $h_{low}$ ,  $h_{high}$  – lower and higher tabu tenures
//output:  $p^\diamond$  – (improved) solution


---


  apply tabu search procedure to  $p$ , get (improved) solution  $p^\nabla$ ;
  //the tabu search procedure is controlled by the parameters  $\tau$ ,  $h_{low}$ ,  $h_{high}$ 
   $p^\diamond := p^\nabla$ ;  $\mu := \mu_{min} - 1$ ; //  $\mu$  is the current mutation rate
  for intraevolution_iteration := 1 to  $Q$  do begin
    if  $\mu < \mu_{max}$  then  $\mu := \mu + 1$  else  $\mu := \mu_{min}$ ; //updating the mutation rate
    apply random mutation procedure to  $p^\nabla$ , get mutated solution  $p^-$ ;
    apply tabu search procedure to  $p^-$ , get (improved) solution  $p^\nabla$ ;
    if  $z(p^\nabla) < z(p^\diamond)$  then begin
       $p^\diamond := p^\nabla$ ; //the best improved solution is archived
       $\mu := \mu_{min} - 1$  //the mutation rate is reset to its minimum value
    endif
  endfor.
  
```

Fig. 4. High-level pseudocode of the intraevolution (iterated tabu search) procedure

4 Computational Experiments

4.1 Experimental Setup

The extensive computational experiments have been carried out using x86 series computer with an Intel 3.1 GHz four kernel processor, with 4GB RAM and 64-bit MS Windows operating system.

We have experimented with the above-mentioned extra-large size benchmark GPP instance, where the instance size (n) is equal to 1024. Remind that the grid size is 32×32 , and the value of the density parameter m varies between 1 and 512. (It is well enough to obtain analytical solutions only for $m = 1, 2, \dots, n/2$, as pointed out in Sect. 2.3.)

Table 1. Main controlling parameters of hybrid genetic-evolutionary algorithm

Parameter	Value	Remarks
Population size, PS	10	
Number of generations (iterations of interevolution), N_{gen}	500	
Number of offspring per generation, N_{offspr}	10	
Minimum and maximum mutation rates, μ_{min} , μ_{max}	$[0.05n,$ $0.07n]$	n is the size of the problem
Number of iterations of intraevolution, Q	30	
Number of tabu search iterations, τ	$0.5n$	
Lower and higher tabu tenures, h_{low} , h_{high}	$[0.03n,$ $0.05n]$	

Regarding the hybrid genetic-evolutionary algorithm, it is programmed in Pascal programming language (using Free Pascal Compiler (FPC)). In addition, the IDE package Qt Creator is used for converting the analytical solutions to the graphical frames.

The algorithm's main controlling parameters are shown in Table 1. (The particular values of the parameters were established during preliminary experimentation.)

4.2 Results of the Computational Experiments

The experiments were organized in such a way that, for each value of m , 10 independent runs (single executions) of the algorithm HGEA with different random number generator's seeds are performed. The best found solution p^* (with respect to the objective function z (see formula (3))) among 10 runs is considered as a final (resulting) solution. Since no other solutions of this problem are not available, we regard our obtained solutions (permutations) p^* and the corresponding objective functions values $z(p^*)$ as the best known solutions (BKS) and best known values (BKV) of the objective function, respectively. For the sake of brevity, we report only the best known values of the objective function (see Tables 2, 3). The BKVs for all 512 m 's are given.

Although many solutions in Tables 2, 3 are conjectured to be possibly optimal (or at least pseudo-optimal/near optimal), still it is, of course, too early to say for sure that all the produced solutions are optimal ones. Anyway, these solutions are currently the only source of reference ("reference solutions") for other researchers to assess their algorithms.

The fortunate aspect in regards to solving the GPP is that the analytical solutions can be quite easily transformed (mapped) to their "visual counter-parts". Thus, the analytical solutions can correspondingly be represented as graphic colour patterns (images) and the quality of the produced patterns can alternatively be evaluated from the "user's aesthetics point of view" (see Fig. 5).

The following colours (according to the RGB colour model) were used for the frames (images) depicted in Fig. 5:

- ✓ grey frames in Fig 5 (a, b, c, d, e, f): foreground: black (R (red) = 0, G (green) = 0, B (blue) = 0); background: white (R = 255, G = 255, B = 255);
- ✓ colour frames in Fig 5 (g, h, i, j, k, l): foreground: black (R = 0, G = 0, B = 0); background: blue (a tone of the blue colour) (R = 175, G = 203, B = 255).

In the graphical illustrations, every square of the frame is physically consisting of $100 = 10 \times 10$ pixels for the visibility convenience. In turn, $1024 = 32 \times 32$ squares constitute a single grid and, in addition, each 1024-square-grid is usually replicated 8 times horizontally and 8 times vertically — so that a fine-looking image emerges.

This is a very small fraction of the colour patterns potentially available. More colour patterns variants will be available at the web page: <http://www.personalas.ktu.lt/~alfmise/>.

Table 2. Best known solutions for the grey pattern problem ($n = 1024, m = 1, 2, \dots, 258$)

m^\dagger	BKV*	m^\dagger	BKV*	m^\dagger	BKV*	m^\dagger	BKV*	m^\dagger	BKV*	m^\dagger	BKV*
1	0	44	2042792	87	9563920	130	23460170	173	44367334	216	72260190
2	390	45	2147200	88	9818424	131	23897592	174	44935036	217	72990332
3	1954	46	2260650	89	10074140	132	24335656	175	45517412	218	73726832
4	3908	47	2373506	90	10331422	133	24773832	176	46093494	219	74467984
5	9488	48	2482832	91	10600710	134	25213730	177	46684562	220	75201458
6	15882	49	2607474	92	10871062	135	25653062	178	47284674	221	75959354
7	24290	50	2730510	93	11138470	136	26091040	179	47871440	222	76713866
8	32808	51	2857088	94	11411510	137	26536474	180	48489950	223	77485740
9	45844	52	2988998	95	11679880	138	26980086	181	49061450	224	78228880
10	60310	53	3120248	96	11944352	139	27426740	182	49657684	225	78977922
11	75878	54	3257234	97	12237102	140	27873238	183	50264826	226	79757678
12	91852	55	3398018	98	12523996	141	28319430	184	50881508	227	80520900
13	114040	56	3535048	99	12814056	142	28761578	185	51505020	228	81287994
14	136706	57	3684478	100	13105672	143	29211334	186	52115178	229	82061894
15	160770	58	3829950	101	13398398	144	29649520	187	52740836	230	82837128
16	185552	59	3984538	102	13691306	145	30118164	188	53351568	231	83613898
17	218392	60	4136400	103	13988062	146	30588480	189	53962538	232	84406568
18	251618	61	4291962	104	14288780	147	31065948	190	54575284	233	85226822
19	288006	62	4447434	105	14593444	148	31546098	191	55185346	234	86032736
20	324794	63	4604860	106	14899130	149	32025690	192	55788864	235	86829778
21	365546	64	4762688	107	15216394	150	32508848	193	56452088	236	87618540
22	407406	65	4949042	108	15540204	151	32992712	194	57086766	237	88445972
23	451448	66	5132250	109	15859540	152	33479148	195	57739124	238	89239062
24	496888	67	5312762	110	16177106	153	33968988	196	58392270	239	90079542
25	549180	68	5493398	111	16505302	154	34461110	197	59055984	240	90875504
26	603368	69	5675784	112	16837956	155	34957410	198	59711606	241	91698908
27	659044	70	5868614	113	17174378	156	35451236	199	60357328	242	92523578
28	716280	71	6061636	114	17508602	157	35944108	200	61005880	243	93371894
29	777436	72	6253544	115	17849756	158	36437606	201	61656140	244	94187252
30	837798	73	6451748	116	18191920	159	36933614	202	62313154	245	95044544
31	907090	74	6658646	117	18535442	160	37426912	203	62979360	246	95865322
32	975008	75	6866464	118	18902942	161	37947342	204	63648372	247	96720682
33	1050792	76	7077272	119	19272770	162	38464394	205	64329116	248	97531736
34	1125558	77	7287952	120	19631156	163	38982914	206	65021762	249	98361638
35	1203646	78	7497962	121	20001764	164	39500236	207	65721964	250	99225594
36	1281132	79	7708934	122	20370638	165	40025416	208	66422364	251	100062350
37	1368444	80	7919112	123	20746696	166	40550536	209	67136312	252	100898116
38	1456842	81	8147012	124	21117234	167	41080848	210	67852496	253	101733670
39	1547598	82	8363950	125	21484868	168	41606752	211	68585494	254	102566006
40	1638808	83	8600584	126	21852518	169	42140990	212	69320690	255	103399158
41	1736236	84	8839620	127	22218924	170	42674036	213	70064396	256	104232704
42	1834074	85	9079818	128	22581376	171	43220914	214	70794386	257	105247082
43	1935946	86	9322672	129	23021790	172	43788356	215	71528284	258	106260632

[†] m — grey density (number of "black points");
^{*} BKV — best known value of the objective function.

Table 3. Best known solutions for the grey pattern problem ($n = 1024, m = 259, 260, \dots, 512$)

m^\dagger	BKV ‡	m^\dagger	BKV ‡	m^\dagger	BKV ‡	m^\dagger	BKV ‡	m^\dagger	BKV ‡	m^\dagger	BKV ‡
259	107273354	302	153146596	345	205167090	387	264325810	429	329073668	471	399837320
260	108286116	303	154286278	346	206544390	388	265800982	430	330767540	472	401592882
261	109299348	304	155427952	347	207925194	389	267231448	431	332330274	473	403351666
262	110334994	305	156547878	348	209234904	390	268732412	432	333972866	474	405337064
263	111323160	306	157695530	349	210612178	391	270226114	433	335584244	475	406875344
264	112363736	307	158882820	350	211922934	392	271722660	434	337231610	476	408927990
265	113383802	308	159999648	351	213304876	393	273215384	435	338872018	477	410409038
266	114381534	309	161157378	352	214686716	394	274667266	436	340435998	478	412182568
267	115407098	310	162344014	353	216044260	395	276205888	437	342143456	479	414134596
268	116418498	311	163515706	354	217445916	396	277692530	438	343794314	480	415733856
269	117469062	312	164641724	355	218756522	397	279195142	439	345446648	481	417519180
270	118477824	313	165838280	356	220146686	398	280691684	440	347099218	482	419302686
271	119497906	314	167026084	357	221526988	399	282217722	441	348768590	483	421092758
272	120561446	315	168188912	358	222888300	400	283719254	442	350325606	484	422883164
273	121617698	316	169391790	359	224268836	401	285244434	443	352087302	485	424678088
274	122661492	317	170558574	360	225646838	402	286756590	444	353718272	486	426473544
275	123707558	318	171760456	361	227020504	403	288278564	445	355428484	487	428272184
276	124737088	319	172950064	362	228390592	404	289827134	446	356954184	488	430071632
277	125778830	320	174127764	363	229827232	405	291340218	447	358612252	489	431876322
278	126851898	321	175343494	364	231201722	406	292871228	448	360270272	490	433683572
279	127903394	322	176518174	365	232597170	407	294394624	449	361968220	491	435492454
280	128957972	323	177756358	366	233971762	408	295973512	450	363665156	492	437303524
281	130021358	324	178959204	367	235390342	409	297489616	451	365361310	493	439118116
282	131092290	325	180158842	368	236843924	410	299050002	452	367244164	494	440933678
283	132151576	326	181377560	369	238269552	411	300582666	453	369017804	495	442752278
284	133228938	327	182598340	370	239660062	412	302164308	454	370457992	496	444570032
285	134303894	328	183826382	371	241079294	413	303727338	455	372157104	497	446397066
286	135383862	329	185039942	372	242516090	414	305271642	456	373863658	498	448224550
287	136469240	330	186245648	373	243950266	415	306884748	457	375575430	499	450053654
288	137551508	331	187480294	374	245386948	416	308431548	458	377289052	500	451883116
289	138648032	332	188702624	375	246816000	417	310003030	459	379005274	501	453718668
290	139728662	333	189934056	376	248257914	418	311591866	460	380724964	502	455554546
291	140837408	334	191155546	377	249712758	419	313155496	461	382449898	503	457391626
292	141877516	335	192376124	378	251167654	420	314755774	462	384176484	504	459230104
293	143027452	336	193623722	379	252561134	421	316363990	463	385902750	505	461073188
294	144139856	337	194844852	380	254068150	422	317917076	464	387628048	506	462916382
295	145231736	338	196104848	381	255542570	423	319480056	465	389368514	507	464761614
296	146333878	339	197399218	382	256989148	424	321094312	466	391105794	508	466607612
297	147484950	340	198600932	383	258436576	425	322673988	467	392843892	509	468457260
298	148589638	341	199891460	384	259897258	426	324312270	468	394587900	510	470307298
299	149717898	342	201181102	385	261381896	427	325910730	469	396332818	511	472158510
300	150871056	343	202519622	386	262850744	428	327490484	470	398083462	512	474010112
301	152006034	344	203834574								

$^\dagger m$ — grey density (number of "black points");
 ‡ BKV — best known value of the objective function.

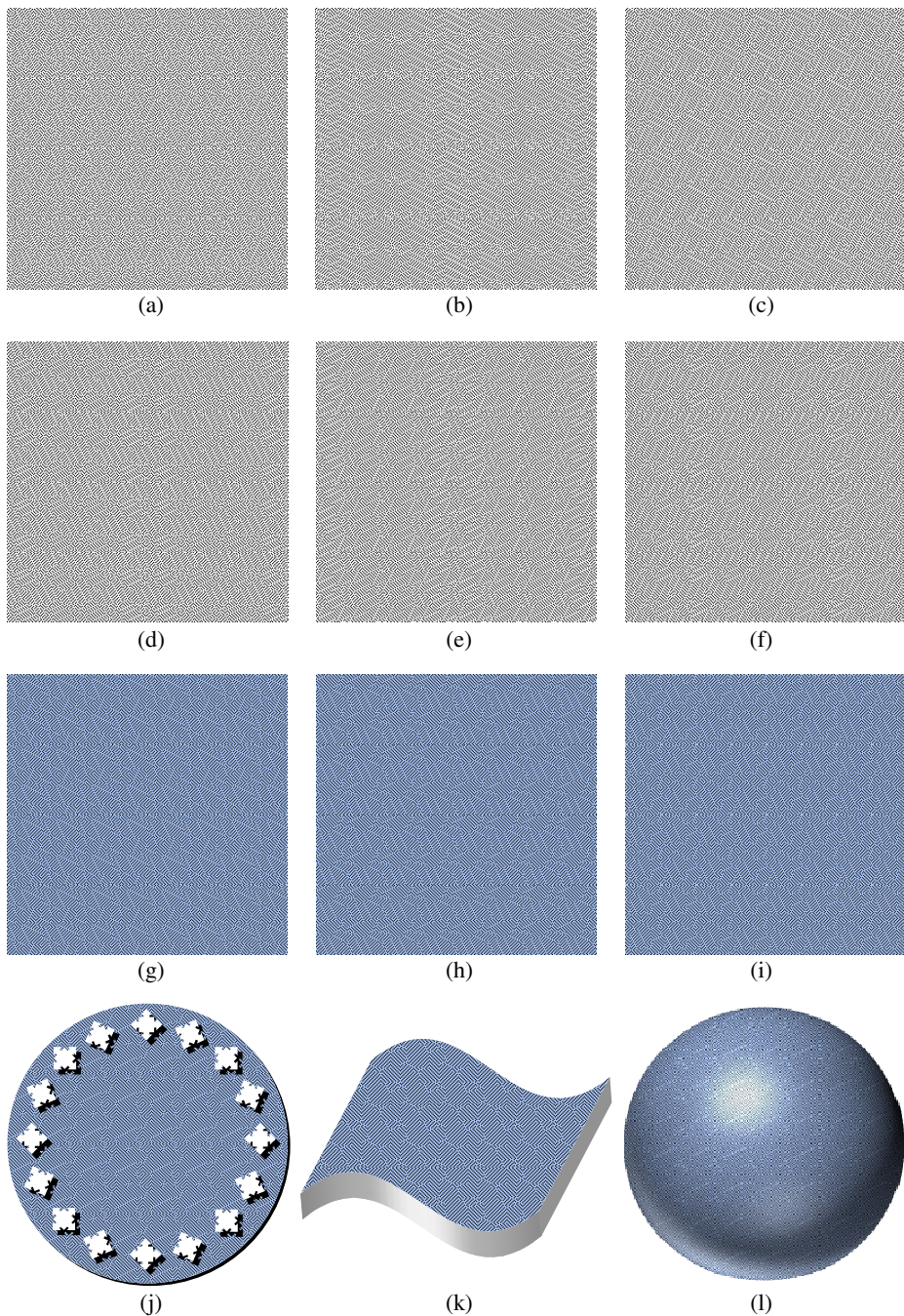


Fig. 5. Examples of grey and colour frames on 2D and pseudo-3D surfaces ($n = 1024$): a) $m = 316$, b) $m = 317$, c) $m = 318$, d) $m = 319$, e) $m = 320$, f) $m = 321$, g) $m = 326$, h) $m = 327$, i) $m = 328$, j) $m = 333$, k) $m = 334$, l) $m = 335$

5 Concluding Remarks

In this paper, we were interested in the automatic computational generation of the high-quality colour patterns (digital halftone textures and images). As a formal model of algorithmic design of the digital halftones, the grey pattern problem (GPP) has been used.

New extra-large scale instance (data set) of the GPP is constructed so that the creation of more perfect, superior-quality colour patterns is enabled. The heuristic genetic-evolutionary algorithm is applied for effectively solving the instance constructed. The best known analytical solutions achieved during the extensive experiments with this algorithm are provided. As a confirmation of the quality of the analytical solutions produced, we also give the visual representations of fine-looking halftone patterns and the reader can judge about the perfection of the images obtained.

We hope that the presented approach would be helpful for the engineers of modern creative multimedia. It may also be suitable as a prototype of the computer-aided design software for the architects and/or designers of abstract/computer graphic arts.

References

1. Burkard, R.E., Karisch, S., Rendl, F.: QAPLIB – a quadratic assignment problem library. *J. Glob. Optim.* 10, 391–403 (1997), <http://www.seas.upenn.edu/qaplib> (cited June 29, 2013)
2. Čela, E.: *The Quadratic Assignment Problem: Theory and Algorithms*. Kluwer, Dordrecht (1998)
3. Drezner, Z.: Finding a cluster of points and the grey pattern quadratic assignment problem. *OR Spectrum* 28, 417–436 (2006)
4. Kang, H.R.: *Digital Color Halftoning*. In: Dougherty, E.R. (ed.). *SPIE/IEEE Series on Imaging Science & Engineering*. The SPIE Optical Engineering Press/IEEE Press, Bellingham/Piscataway (1999)
5. Lau, D.L., Arce, G.R.: *Modern Digital Halftoning*, Sec. Ed. In: Liu, K.J.R. (ed.). *Signal Processing and Communications Series*. Marcel Dekker, New York-Basel (2008)
6. Misevičius, A.: Experiments with hybrid genetic algorithm for the grey pattern problem. *Informatica* 17, 237–258 (2006)
7. Misevičius, A.: Generation of grey patterns using an improved genetic evolutionary algorithm: Some new results. *Inform. Technol. Contr.* 40, 330–343 (2011)
8. Misevičius, A.: An implementation of the iterated tabu search algorithm for the quadratic assignment problem. *OR Spectrum* 34, 665–690 (2012), doi:10.1007/s00291-011-0274-z
9. Misevičius, A., Rubliauskas, D.: Performance of hybrid genetic algorithm for the grey pattern problem. *Inform. Technol. Contr.* 34, 15–24 (2005)
10. Sivanandam, S.N., Deepa, S.N.: *Introduction to Genetic Algorithms*. Springer, Heidelberg (2008)
11. Taillard, E.D.: Comparison of iterative searches for the quadratic assignment problem. *Locat. Sci.* 3, 87–105 (1995)
12. Taillard, E.D., Gambardella, L.M.: *Adaptive memories for the quadratic assignment problem*. Techn. Report. IDSIA-87-97, Lugano, Switzerland (1997)
13. Ulichney, R.A.: *Digital Halftoning*. MIT Press, London (1987)

Design of Visual Language Syntax for Robot Programming Domain

Ignas Plauska and Robertas Damaševičius

¹ Kaunas University of Technology, Centre of Real Time Computer Systems,
Studentų 50, LT-51368, Kaunas, Lithuania

² Kaunas University of Technology, Software Engineering Department,
Studentų 50-415, LT-51368, Kaunas, Lithuania
ignas.plauska@ktu.lt, robertas.damasevicius@ktu.lt

Abstract. The paper discusses the development of the visual language syntax based on the application of sound methodological principles, a visual communication model, a visual syntax model, a formal description of syntax based on visual grammar metalanguage (an extension of BNF) and ontology of visual signs (graphemes). The syntax of an illustrative visual language VisuRobo for the mobile robot programming domain is presented.

Keywords: visual programming, visual language, visual communication model, visual syntax model, visual metalanguage.

1 Introduction

Software design and development involves high-level *cognitive processes* such as assimilating, constructing and sharing domain knowledge and making decisions. Cognition is based on the developer's *mental models* [1], which provide a structure for organization of domain knowledge within the developer's mind. The task of the programmer is to map a mental model of a solution of a domain-specific problem to a system of computer-readable signs, i.e. a program written in a programming language. Many types of programming languages exist. General purpose programming languages are usually domain-independent and are good for general problem solving. On the other hand, domain-specific programming languages (DSLs) are tailored towards a specific application domain, and are based only on the relevant concepts and features of that domain. A DSL allows domain experts to express high-level concepts succinctly using a notation tailored to a set of specific domain problems. Therefore, DSLs can be considered as a medium of communication that allows to bridge the gap between the mental model and the problem domain systems and, consequently, to cut the distance from ideas to products in software engineering.

The complexity of modern software engineering and its application domains such as robotics stimulated the move from one-dimensional string grammar based textual languages to visual languages which use non-linear graphical notations. The problem transcends just using graphical symbols for programming and includes *visual*

knowledge engineering [2] and *visual software engineering* [3]. Currently, visual programming languages (VPLs) are used in many different application fields such as the development of graphical user interfaces (GUIs), teaching computer science, and model-driven development [4]. In particular, Unified Modeling Language (UML) [5] is a widely known example of a visual software engineering language that is used for modelling and specifying software-intensive systems.

The main motivations for visual languages are as follows: 1) higher level of abstraction, which is closer to the user's mental model and involves manipulation of visual elements rather than machine instructions [6], 2) higher expressive power characterized by two (or more) dimensional relations between visual elements [7], and 3) higher attractiveness to non-professional or novice programmers motivated by simpler description of complex things. Other advantages of VPLs include economy of concepts required to program (i.e., smaller program size), concreteness of programming process, explicit depiction of relationships between program entities, and immediate visual feedback [8].

Visual programming allows using a conceptual model that is tailored to the mental process of the user rather than constraints of the programming language syntax or the target platform. Indeed, visual notations allow for the description and understanding of complex systems, such as concurrent and/or real-time systems, for which traditional textual descriptions are inadequate [9]. Robots are good examples of such complex systems which require having knowledge of multiple domains such as mechatronics, analogue and digital electronics, embedded software, real-time systems, kinematics, communication protocols and control algorithms, etc. The robotics domain is characterized by heterogeneity and diversity of robots and their different capabilities. Furthermore, a large variety of existing robotic platforms requires having high-level methods for general modelling and reasoning about what different robots are capable of doing. The examples of VPLs designed for the robot programming domain, are Microsoft Visual Programming Language (MVPL) and Lego NXT-G language. These languages have been criticized for lack of flexibility and usability [10]. Other challenges to these languages come from targeting users that are usually not formally trained in programming or software engineering (e.g., children, or robot hobbyists), and include unusual features that do not seem to be shared by a majority of users.

The contribution of this paper is the design of the visual language syntax for the robot programming domain based on the application of sound methodological principles including a model of visual communication, a formal description of syntax based on using visual metalanguage and an ontology of signs.

The structure of the remaining parts of the paper is as follows. Section 2 presents a Visual Communication Model as a foundation of the proposed visual programming language syntax. Section 3 describes the modelling of visual language grammar using visual meta-syntax. Section 4 provides a case study in developing syntax for the *VisuRobo* language. Finally, Section 5 presents evaluation of results and conclusions.

2 Elements and Models of Visual Communication

Using a VPL to develop a visual program (diagram) is a form of visual communication (conversation or dialogue) between a designer and a user. In general, any language is a formal system of signs described by a set of grammatical rules to communicate some meaning. In particular, a VPL is a system of *visual* signs (i.e., primitive graphical elements, graphemes or pictograms), which represent *domain-specific* concepts, processes, or physical entities, and uses *more than one dimension* of space to convey semantics [8]. Spatial arrangement of graphical elements together with a semantic interpretation provides a space of communication for users of a VPL [11]. Below we present an analysis of known models of visual communication.

2.1 Shannon-Weaver's and Berlo's Models of Communication

A model of communication proposed by Shannon and Weaver [12] consists of a sender (a source), a message, a code (a language or other set of symbols or signs that can be used to transmit a message), a channel (the path on which the message travels), noise (or interference), and a receiver. The sender selects a message and encodes it into a signal that is sent over the communication channel to the receiver. The receiver decodes the transmitted signal and interprets the message. In the process, the message may be corrupted by noise that can be psychological (arising from the receiver's attitudes, biases or beliefs), physical (coming from the environment of communication) or semantic (caused by the receiver's misinterpretation). The model has been criticized for disregarding the semantic content of message and the simplistic interpretation of concept of information. Furthermore, it implies that human communication is similar to machine communication such as sending a signal in computer systems [13].

Berlo's Model of Communication [14] extended the Shannon and Weaver's model with perceptory (hearing, seeing, touching, etc. channels), structural (content, elements, structure of message) and social (skills, culture, attitudes) elements of communication, but provided no technical details how the model could be implemented.

2.2 Semiotic Engineering

Semiotic engineering [15] interprets the communication channel as a human-computer interface (HCI). Designers (senders) send their message to users (receivers) through the interface using signs, which associate domain entities with their meaning and representation. The interface acts as a meta-message from a designer to users. This meta-message conveys the designers' interpretation of the domain problem and provides users with artefacts to support users in solving the problem. The meta-message is defined using a metalanguage, and a metalanguage consists of signs that signify relationships of interface elements to each other and to domain entities they represent. The difficulties with semiotic engineering arise from the lack of practical approaches how to deal with the interpretability problem: the designer aims to formulate a concise singular message (that can be interpreted only one way) while the users may derive multiple and often conflicting interpretations of the message.

2.3 Visual Language Model

A visual language model proposed by Hari Narayan *et al.* [16] focused on the human use of visual languages and described three objects of interest to investigation of such languages: a computational system, a cognitive system, and the visual display as a communication medium (or channel). Visual representations that encode and convey information appear on the visual display and require visual perception, comprehension and reasoning on the cognitive side, as well as visual parsing, interpretation, and program execution on the computational side. The authors themselves recognize that the model is not full as a more complete taxonomy complemented with a formal system to define semantics is required.

2.4 Theory of Visual Display and Visual Variables

Theory of Visual Display [17] elaborated on the structure of visual display as the communication channel. The main elements are: 1) objects – basic units of meaning, 2) regions – provide context for objects, and 3) relations – connect objects or regions.

Bertin [18] analysed the main elements of visual objects. These could be encoded graphically using 8 visual variables which provide a visual alphabet for constructing visual notations as follows: shape, size, orientation, pattern, colour, hue and two spatial dimensions (vertical and horizontal). Notation designers can create graphical symbols by combining the variables together in different ways.

2.5 Physics of Notations

Moody [19] proposed a framework of methodology for developing cognitively-effective visual languages. The framework defines a visual notation that consists of a set of graphical symbols (visual vocabulary), a set of compositional rules (visual grammar), and definitions of the meaning of each symbol (visual semantics). Visual vocabulary includes 1D graphic elements (lines), 2D graphic elements (areas), 3D graphic elements (volumes), textual elements (labels) and spatial relationships. A valid expression (diagram) in a visual notation consists of visual symbols (tokens) arranged according to the rules of the visual grammar. Visual vocabulary and visual grammar together form the visual syntax of the notation. The language metamodel defines mapping of visual symbols to the constructs they represent. The approach has been defined as a scientific theory that allows both understanding how and why visual notations communicate as well as improve their ability to communicate [19].

2.6 Ontological Engineering

Ontological engineering provides methodologies for building ontologies: formal representations of concepts within a domain and the relationships between concepts. Ontology defines the vocabulary of a problem domain and a set of constraints on how terms can be combined to model the domain in a declarative way [20]. The language

ontology is the description of what the primitives of a language are able to represent in terms of domain phenomena, i.e., it is the representation of a conceptualization of the domain in terms of the language’s vocabulary [21].

2.7 Proposed Model of Visual Communication

Based on the discussed models of visual communication and visual modelling, we propose our Model of Visual Communication (Fig. 1). The model has sound formal and engineering foundations. It defines the composition of language’s syntax and ensures “low noise” communication due to using shared domain and sign ontologies.

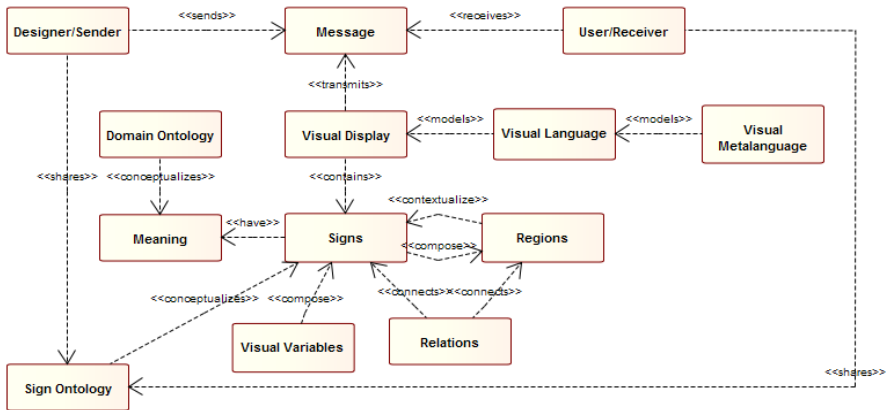


Fig. 1. Model of Visual Communication

A Visual diagram, which describes a specific domain process or a solution of domain problem, is a Message. The Message is sent by a Designer, transmitted via a Visual Display as a communication channel and received by the User(s). Visual Display is a metaphor that facilitates the rapid transfer of an effective mental model into the user’s head [22]. The elements of Visual Display are Signs (basic units of domain knowledge), Regions that provide context to Signs and Relations that connect Signs to Regions. The arrangement of Signs is defined by Visual Language that models Visual Display, whereas Visual Language is modelled by Visual Metalanguage. Signs are chosen to signify a shared meaning of domain that is conceptualized by Domain Ontology, whereas their graphical representation is conceptualized by Sign Ontology shared both by Designer and User(s). The use of ontologies as shared conceptualizations of knowledge minimize the “noise” (i.e., misunderstanding, etc.) introduced during the communication process and allows logical reasoning and inference over signs as representation of meaning. Therefore, ontologies can be employed as a tool for visual or diagrammatic reasoning [23].



3 Modelling the Syntax of Visual Language

A VPL is an artificial system of communication that uses visual elements. Any visual language can be characterized by three main elements: lexical definition (symbol vocabulary), syntactical definition (grammar), and semantic definition. VPLs differ from textual programming languages by the type of used symbols and the type of their relation to each other.

3.1 Lexical Definition

First, a set of symbols is extended from a set of characters (e.g., ASCII, Unicode) to a set of any images. Formally, a visual symbol S_v is defined as a quadruple $S_v = (I, C, M, A)$, where I is the image that is shown to the user of the language; C is the position of the symbol in the visual sentence that defines its context, i.e., the relation of the visual symbols to other symbols; M is the semantic meaning of the visual symbol; and A is the set of actions that are performed when the symbol is activated.

We propose using ontological engineering methods for modelling and specifying a vocabulary of the language. The symbol vocabulary formally can be defined as ontology of symbols (signs) $O = (D, R)$, where D is some domain, and $R \subseteq D^n$ is a set of relations defined in D . Visual language L_v can be defined as $L_v \subseteq S_v^*$, $S_v \in O$. The relations between symbols can be: taxonomic (the symbols belong to the same group or category of symbols), mereologic (one symbol is a part of other symbol), positional (the symbols are always used together though do not form a separate symbol).

Second, the sequencing of symbols is extended from one-dimensional to multi-dimensional. If we have a visual language L_v then a visual sentence of a language L_v is a spatial arrangement of visual symbols specified according to the syntax rules (grammar) of L_v . The definition of language grammar implies the need for a meta-grammar that defines the structural composition of the syntax grammar rules.

3.2 Syntactical Definition

In the classic definition of generative string grammars, a grammar $G(L)$ of language L is defined as $G(L) = (N, T, P, S)$, where N is a finite set of nonterminal symbols (grammar variables), none of which appear in strings formed from G , T is a finite set of terminal symbols (grammar constants), $T \cap N = \emptyset$; P is a finite set of production rules that map from one string of symbols to another in the form of $(N \cup T)^* N (N \cup T)^* \rightarrow (N \cup T)^*$, and S is the start symbol, $S \in N$. In a visual language, the ordering of visual symbols is non-linear, therefore, the visual production rules include visual relations R as follows: $((N \cup T)R)^* N (R(N \cup T))^* \rightarrow (N \cup T)^*$.

In practice, the syntax of textual languages is usually defined by meta-syntax notation such as the Extended Backus-Naur-Form (EBNF) (ISO/IEC 14977). To define a visual grammar of a visual language, the definition of string grammar must be

extended to include visual symbols and visual relations, which indicate the spatial arrangement of the elements of productions such as connection relation to connect visual symbols with adjacent regions through links or arrows, and geometric relations (containment, horizontal/vertical and left/right concatenations, etc.). In one of such extensions, Picture Description Language (PDL) [24], each terminal symbol is labelled at its head and its tail, and the coincidence operators between primitives are defined. Ledley [25] also proposed a set of topological operators as relations between terminal symbols. The summary of visual relation operators is given in Table 1.

Table 1. Visual relation operators

Operator	Interpretation	Operator	Interpretation
$S_1 + S_2$		$S_1 \rightarrow S_2$	
$S_1 \times S_2$		$S_1 \uparrow S_2$	
$S_1 - S_2$		$S_1 \odot S_2$	
$S_1 * S_2$			

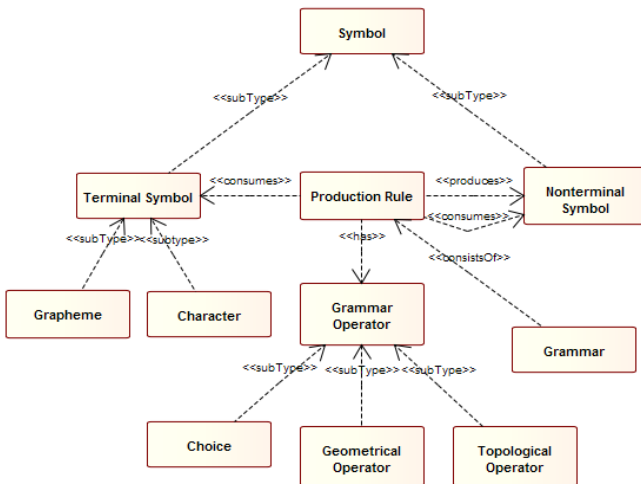


Fig. 2. Visual Syntax model



3.3 Visual Syntax Model and Meta-syntax of Visual Language

The proposed Visual Syntax Model is summarized in Fig. 2. The model defines two types of symbols: terminals (characters or graphemes) and non-terminals. Production rules describe how a sequence of terminal and nonterminal symbols can be consumed by the parser of the language to produce terminal symbols as governed by the grammar operators. Grammar operators are: *choice* operator (selection of production rules out of a set of alternatives), *geometrical* operators (2D production), *topological* operators (symbol containment/concatenation) and *abstraction* operators (symbol instantiation between different levels of abstraction beyond the 2D visual space).

Considering the above, we propose the following meta-syntax based on EBNF:

```

<syntax> ::= <rule> | <rule> <syntax>
<rule> ::= "<" <rule-name> ">" "==" <expression> <EOL>
<expression> ::= <term-list> | "(" <term-list> ")"
                | <term-list> <operator> <expression>
<operator> ::= "|" | "+" | "-" | "x" | "*" | "→" | "↑" | "⊙"
<term-list> ::= <term> | <term> <term-list>
<term> ::= <grapheme> | "<" <rule-name> ">"
<grapheme> ::= <icon> | <icon> "(" <attribute-list> ")"
<attribute-list> ::= <literal> | <literal> <attribute-list>
<literal> ::= ' ' ' <text> ' '

```

4 Case Study: Modelling Syntax of Visual Language

As a case study of the proposed approach, we analyse and present results of modelling a visual programming language *VisuRobo* for the mobile robotics domain. The designer faces two main challenges when developing a visual language: 1) selection (construction) and design of a visual vocabulary for the developed language, and 2) selection of the meta-modelling language for describing the syntax of the language.

4.1 Visual Vocabulary of the Language

When designing a vocabulary the main concerns are as follows:

1) *Understandability*. The designers want to communicate ideas using visual language with as less noise as possible. This requires that the mental models of designers must be “shareable”, i.e., the symbols used to communicate meaning must be easily recognizable and interpretable. The meaning of symbols must be known intuitively.

2) *Usability*. The symbols must be designed using a principled methodology (see, e.g., Moody’s “Physics of notations” [19]), and their usability must be thoroughly evaluated using quantitative metrics [10] and/or qualitative surveys.

For our design task we have decided to select a subset of symbols from a set of road traffic sign system rather than to design our own set of symbols. The advantages of such choice are as follows: 1) Traffic signs are a part of our everyday life and are simple, easily recognizable, legible and understandable [26]; 2) Traffic signs were designed to be usable in different contexts of use (day/night, static/dynamic

environment) and contain a minimum amount of information (or “conceptual baggage” [27]) required to communicate a message; 3) Traffic signs are an international standard (*Vienna Convention on Road Signs and Signals*); 4) Ontology of road signs is available [28]; 5) Meaning of traffic signs are easily transferrable to a mobile robotics domain using the analogy principle [29], which allows the user to metaphorically reinterpret familiar signs in another context of use based on the similarity of vehicle driving and mobile robotics domains.

The selection of signs for *VisuRobo* is based on the Robot Programming Ontology [30], which defines the main actions of a mobile robot. Ontological description allows to define their graphical representation formally and ontology tools (such as Protégé) provide capabilities for checking consistency of formal description and reasoning over the meaning of signs. An example of the description of the PARKING sign in the Road Sign Ontology [28] is given in Fig. 3. The ontology provides the classification of road signs and defines their graphical composition in terms of colour, shape and represented symbols. The description of the Parking sign states that the sign shall have blue background, white border, a rectangular shape and the “P” symbol on it.

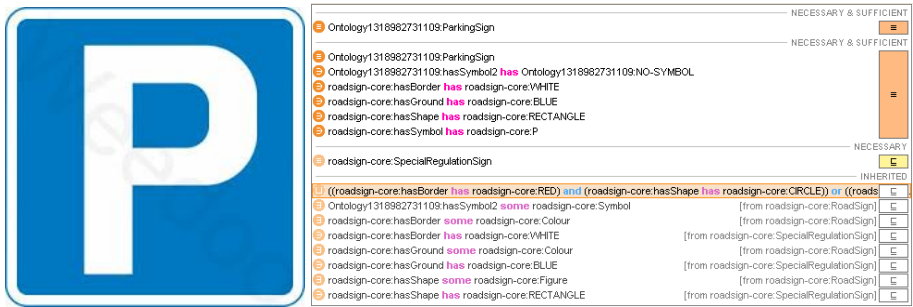













Fig. 3. Parking sign and its assertions derived from the Road Sign Ontology as seen in Protege

In Table 2, we summarize the visual graphemes of the *VisuRobo* language. We consider each sign as a metaphor; therefore the description is given following the Barr’s [31] relations (parts) of metaphor as follows: 1) *Definition*: the meaning of the metaphor itself. 2) *Designer’s Interpretation*: the meaning of the grapheme according to the designer’s mental model. 3) *Match*: how metaphor in the mental model matches to its realization, i.e., the evaluation of the designer’s interpretation. 4) *User interpretation*: the meaning of the grapheme according to the user’s mental model. 5) *Success*: how the user’s interpretation matches the definition of grapheme, i.e., the evaluation of user’s interpretation.

Visual variables [18] that are used to communicate meaning are *shape*, *colour* and *orientation*. The signs for the begin/end of a program have an octagonal shape, the signs for moving (driving) have a round shape, and the signs for temporary stopping or complex nested operation have a rectangular shape. The colours are used sparingly: *green* for starting command, *black* for road and sensors, *red* for stopping command, *yellow* for conditional execution, and *blue* for other commands. The orientation of arrows on driving signs shows the direction of movement. Textual notation is used for communicating simple meaning (“Start”, “Stop”, “Park”). Good visibility is ensured by high contrast between background, foreground and text colour.

Table 2. Graphemes (visual tokens) of language and their interpretation

Grapheme	Definition	Designer's interpretation	User's interpretation	Match/Success
	Road	Defines a path of execution for the robot	Drive, proceed to next command	The metaphor of execution is captured and transferred to user well
	(non standard sign)	Defines an entry point to the program	Start driving robot	Analogy to "STOP" sign is used (shape), but colour is changed to "prohibitive" red to "permissive" green
	Stop	Defines an end point to the program	Robot stops	Analogy between the concepts of stopping a vehicle and finishing a program is used
	Ahead only	Instructs a robot to drive forward for a set amount of time	Robot drives forward	Perfect match
	(non standard sign)	Instructs a robot to drive backward for a set amount of time	Robot drives backward	Analogy with driving forward is used to create a sign for "driving backward"
	Turn left	Instructs a robot to turn left using given the turn angle	Robot turns left	Perfect match
	Turn right	Instructs a robot to turn right using given the turn angle	Robot turns right	Perfect match
	Roundabout	Repetitive statement	Loop	Metaphor of driving roundabout matches well with repetition and looping concepts
	Recommended speed	Set a speed of driving	Robot's driving speed, in percents	Good match (though some unclarity in measurement units is noted)
	Parking	Instructs a robot to stop driving and wait for the given amount of time	Pause, parking	The meaning of sign is understood intuitively well
	Repair	Instructs a robot to execute a mission until the sensor command returns true	Robot performs some operations	An analogy between repair workshop and complexity of robot mission is observed
	Speed camera	Reads ultrasound sensor value and returns true if sensor value is higher than given threshold value	Any sensor	Metaphor of "Camera" as on observing device is extended to all kinds of sensors
	Y-intersection	Defines two paths of execution the selection thereof is defined by value returned by the sensor command	Conditional execution	Decision on taking different roads matches well with metaphor of conditional execution
	(non standard sign)	Defines an intersection of two paths of execution	Merging of execution paths	Analogy with the Y-intersection sign is used to create a sign for a merger of execution paths

4.2 Specification of Grammar Rules

The language is based on a metaphor of road traffic and follows a set of assumptions:

1. The behaviour of a robot is represented by a road metaphor.
2. Each road may consist of an unlimited number of sub-roads a robot must cover.
3. Road defines an independent path of execution.
4. Each road has its speed limitations valid until next intersection with another road.
5. Intersection means the end of incoming roads and beginning of outgoing roads.
6. There are two types of intersections. Join intersection has 2 incoming roads and one outgoing road. Split intersection has one incoming road and 2 outgoing roads.
7. Each road has only one beginning and only one end.
8. While on road the robot can execute a number of missions.
9. Each mission consists of implementing a pre-defined algorithm until specific condition is met.
10. Conditions are defined by the value returned from a sensor.

The syntax of the *VisuRobo* language is defined using EBNF-like syntax extended with visual tokens, geometrical relation primitives and topological operators (Fig. 4).

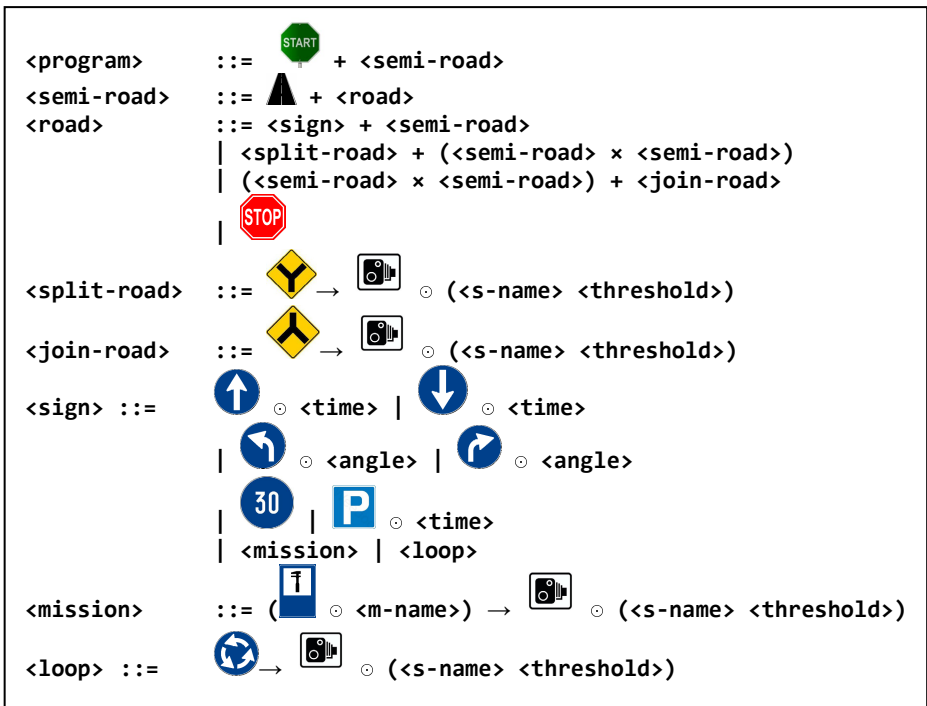


Fig. 4. Syntax description using visual extension of BNF



5 Conclusions

1. Visual Communication Model, which is based on the combination and amalgamation of ideas from [12, 14-19] as well as elements of ontological engineering. The model defines the construction of visual language's syntax and ensures "low noise" communication due to shared domain and sign ontologies.
2. Visual Syntax Model for describing grammar of VPLs and meta-syntax of visual metalanguage. The metalanguage is an extension of EBNF with visual symbols and nonlinear composition of rules that allows to define the syntax of VPL.
3. *VisuRobo*, an illustrative visual robot programming language, serves as a proof-of-concept for the proposed Visual Communication Model, Visual Syntax Model and visual metalanguage. The language uses a metaphor of road driving for a mobile robot and uses a subset of visual signs adopted from the Road Sign Ontology to facilitate transfer of meaning between language designers and users, and to deal with the sign interpretability problem.
4. The evaluation of the language syntax according to Barr [31] and Bertin [18] is given. The familiarity of visual signs and the context of their use allow to transfer the designer message to the users without significant semantic misinterpretations while the consistent use of visual variables (shape, colour and orientation) allows to create a simple, usable and attractive vocabulary of the language.

Acknowledgement. The work described in this paper has been carried out within the framework of the Operational Programme for the Development of Human Resources 2007-2013 of Lithuania „Strengthening of capacities of researchers and scientists" project VP1-3.1-ŠMM-08-K-01-018 „Research and development of Internet technologies and their infrastructure for smart environments of things and services" (2012-2015), funded by the European Social Fund (ESF).

References

1. Gentner, D., Stevens, A.L. (eds.): *Mental Models*. Lawrence Erlbaum Associates (1983)
2. Eisenstadt, M., Domingue, J., Rajan, T., Motta, E.: *Visual Knowledge Engineering*. IEEE Transactions on Software Engineering 16(10), 1164–1177 (1990)
3. Zhang, K., Kong, J., Cao, J.: *Visual Software Engineering*. Wiley Encyclopaedia of Computer Science and Engineering (2008)
4. Zhang, K.: *Visual Languages and Applications*. Springer (2007)
5. Booch, G., Rumbaugh, J., Jacobson, I.: *The Unified Modeling Language User Guide*. Addison Wesley Longman Publishing Co., Inc., Redwood City (1999)
6. Bentrad, S., Meslati, D.: *Visual Programming and Program Visualization – Towards an Ideal Visual Software Engineering System*. IIIT- ACEEE Int. Journal on Information Technology 1(3), 56–62 (2011)
7. Myers, B.A.: *Taxonomies of Visual Programming and Program Visualization*. Visual Languages and Computing 1(1) (1990)
8. Burnett, M.: *Visual Programming*. In: J. Webster (ed.), *Encyclopedia of Electrical and Electronics Engineering*. John Wiley & Sons (1999)

9. Deufemia, V.: A Grammar-based Approach to Specify and Implement Visual Languages. PhD Dissertation, University of Salerno (2002)
10. Plauska, I., Damaševičius, R.: Usability Analysis of Visual Programming Languages Using Computational Metrics. In: Proceedings of the IADIS International Conference on Interfaces and Human-Computer Interaction 2013, Prague, Czech Republic, pp. 63–70 (July 2013)
11. Lakin, F.: Visual grammars for visual languages. In: Proc. of the Sixth National Conference on Artificial Intelligence AAAI 1987, vol. 2, pp. 683–688. AAAI Press (1987)
12. Shannon, C.E., Weaver, W.: The mathematical theory of communication. University of Illinois Press, Urbana (1949)
13. Mortensen, C.D.: Communication: The Study of Human Communication. In: Communication Models, ch. 2, McGraw-Hill Book Co. (1972)
14. Berlo, D.K.: The Process of Communication. Holt, Rinehart, and Winston (1960)
15. Souza, C.S.: The Semiotic Engineering of Human-Computer Interaction. MIT Press (2005)
16. Hari Narayanan, N., Hubscher, R.: Visual language theory: Towards a human computer interaction perspective. In: Marriott, K., Meyer, B. (eds.) Visual Language Theory, pp. 87–128 (1998)
17. Tartre, M.: Theory of Visual Display (2013), <http://www.metaperture.com>
18. Bertin, J.: Semiology of Graphics: Diagrams, Networks, Maps. ESRI Press (2010)
19. Moody, D.L.: The “Physics” of Notations: Toward a Scientific Basis for Constructing Visual Notations in Software Engineering. IEEE Trans. Soft. Eng. 35(6), 756–779 (2009)
20. Devedzic, V.: Understanding Ontological Engineering. Communications of the ACM 45(4), 136–144 (2002)
21. Guizzardi, G.: On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models. In: Proc. of conference on Databases and Information Systems IV: Selected Papers from the 7th International Baltic Conference DB&IS 2006, pp. 18–39. IOS Press (2007)
22. Blackwell, A.F.: The reification of metaphor as a design tool. ACM Transactions on Computer-Human Interaction (TOCHI) 13(4), 490–530 (2006)
23. Glasgow, J., Hari Narayanan, N., Chandrasekaran, B.: Diagrammatic Reasoning: Cognitive and Computational Perspectives. MIT Press, Cambridge (1995)
24. Shaw, A.: A Formal Picture Description Scheme as a Basis for Picture Processing Systems. Inf. Control (14), 9–52 (1969)
25. Ledley, R.: Programming and Utilising Digital Computers. McGraw-Hill (1962)
26. Ng, A.W.Y., Chan, A.H.S.: Cognitive Design Features on Traffic Signs. Engineering Letters 14(1), 13–18 (2007)
27. Anderson, B., Smyth, M., Knott, R.P., Bergan, M.S., Bergan, J., Alty, J.L.: Minimising conceptual baggage: making choices about metaphor. In: Proc. of Conference on People and Computers IX (HCI 1994), pp. 179–194. Cambridge University Press (1994)
28. Pousa, M., Motto, O., Carasusán, E.: Road Sign Ontology (2011), <https://raw.githubusercontent.com/ecarasusan/roadsign/master/roadsign.owl>
29. Breitman, K.K., Barbosa, S.D.J., Casanova, M.A., Furtado, A.L.: Conceptual modeling by analogy and metaphor. In: Proc. of the 16th ACM Conference on Information and Knowledge Management, Lisbon, Portugal, pp. 865–868 (2007)
30. Plauska, I.: Ontology for Robot Programming Domain. In: IVUS, pp. 51–56 (2013)
31. Barr, P., Noble, J., Biddle, R.: A semiotic model of user-interface metaphor. In: Liu, K. (ed.) Virtual Distributed and Flexible Organisations, pp. 189–216. Kluwer Academic (2003)

Testing Stochastic Systems Using MoVoS Tool: Case Studies

Kenza Bouaroudj, Djamel-Eddine Saidouni, and Ilham Kitouni

MISC Laboratory, University Constantine 2, 25000, Algeria
{bouaroudj, saidouni, kitouni}@misc-umc.org

Abstract. MoVoS tool is an implementation of testing theory based on stochastic refusals graph. It allows the automatic extraction of test cases from specification of stochastic systems. Those systems are modeled by Maximality based Labeled Stochastic Transition System “MLSTS”.

In This paper, we present the application of MoVoS tool on two cases studies to valid it. Those case studies permit us to illustrate the functionality of tool and to show that this tool can deal with large system.

Keywords: formal testing models, refusals graphs, maximality semantics, canonical tester.

1 Introduction

Nowadays, software and hardware are getting more and more complex. They usually consist of a variety of components which can be produced by different manufacturers. This leads to compatibility problems between different products.

In order to unify development processes, formal testing techniques [6], [7] provide systematic procedures to ensure implementations conformity between product and its specification.

Formal testing theory has been studied for a long time. Based on finite state machines at the beginning, it has been extended to conformance testing of transition systems. Test generation methods have been developed for both of them. Now, various framework and Tools for test generation have been developed for various languages and models, both untimed (e.g., [9], [13]) and timed (e.g., see [10],[11],[14],[15]).

In this paper, the system is represented by the “Maximality-based Labeled Stochastic Transition System (MLSTS)” model where actions elapse in time and their durations depend on the probabilistic distributed function. This model is based on maximality semantics [16]. MLSTS is a true concurrency semantic model based on the representation of maximal events. Thus, it is well suitable for modeling real time, concurrent, stochastic and distributed systems.

The main of this paper is to present a tool for testing stochastic systems. The steps used in this tool to generate test cases are as follows: First, the Stochastic Refusals Graph (SRG) is generated by several transformations on MLSTS. Then, the SRG

allows us the construction of canonical tester and test cases by several transformations on it. Finally, the sender part of ABP protocol and cell production will be tested using this tool, in order to illustrate the functionality of tool and to show that this tool can be applied to test a large system.

This paper is organized as follows: section 2 introduces MoVoS tool functionality and recall the theory used in MoVoS tool. In section 3, two case studies are presented which are sender parts of ABP protocol and cell production. In section 4 the results of testing those examples are presented, Section 5 concludes the paper and gives some perspectives.

2 MoVoS-Tool

2.1 MoVoS Tool Functionality

MoVoS is a tool for generating test cases from specification. This tool is implemented with MATLAB. Its functional view is sketched in Fig. 1. The input is a specification of the intended behavior of the system under test. The output is a set of test cases.

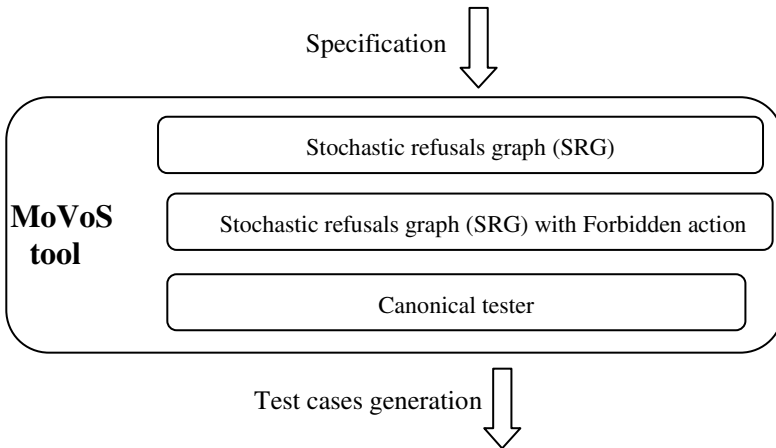


Fig. 1. Functionality of MoVoS tool

The behavior of the specification and implementation are modeled by maximality based labeled stochastic transition system (MLSTS). In this tool [3], the specification can be modeled by stochastic process algebra S-LOTOS.

2.2 Testing Theory in MoVoS Tool

Maximality Based Labeled Stochastic Transition System (MLSTS)

The MLSTS model [1], [2] characterizes the stochastic temporal properties of concurrent systems, using arbitrarily distributed duration of actions [1].

MLSTS transitions are labeled by action and a probabilistic distribution function f is attributed to this action.

The basic idea is to use the clock variables (random variables) for materializing maximal events and to keep track of durations. They permit the control and the observation of time elapsing. For this purpose, clock is associated to each action which represents action's start and appears in the resulting state.

When a clock is reset, it takes a random value which depends on the probability distributed function of the action duration.

Clock countdown synchronously time, when a clock expires (i.e. reaches the value 0), causally dependent actions are enabled (Arous, Saidouni & Ilić, 2011).

As illustration, let $E=(a;b;stop)[](a;stop \parallel b;stop)$ be a behavior expression and $F=\{(a,f),(b,f_1),(c,g)\}$

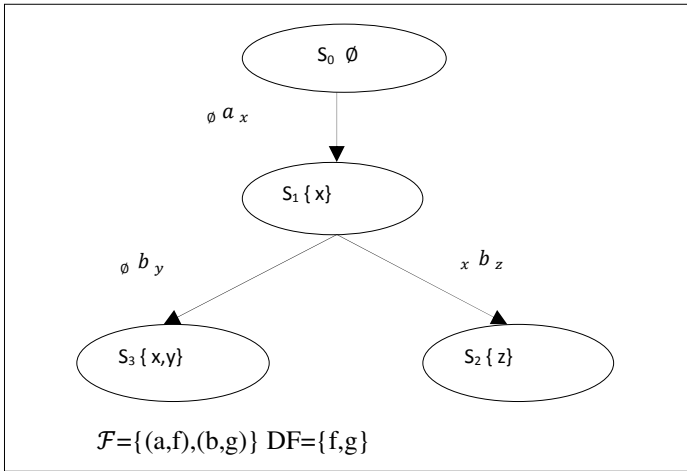


Fig. 2. MLSTS associated to E

The maximality-based labeled stochastic transition system associated to E is given by Fig. 2. In the initial state, no action has begun its execution; the initial state is then labeled by the empty set. Starting from this state, the actions a and b can start. Their corresponding transitions are identified respectively by clocks x and y . x and y take random variable which depends respectively on the probability distribution function f and g . The state S1, labeled by the set $\{x\}$, means that the action a is potentially running at this state. The transition identified by clock y corresponds to the start of the left b. The state S3, labeled by the set $\{x,y\}$, shows that the two actions a and b may progress simultaneously. From the state S1, two scenarios are possible: either the left action b begins its execution, leading to the state S3; or the right action b leads to the state S2. The right action b can start only if action a ends its execution. This causality between the executions of a and b is captured by the set $\{x\}$ associated with the transition leading the system from the state S1 to the state S2. In the resulting state, only the action b may be under execution. That's why; this state is labeled by $\{z\}$.



Definition 1 (MLSTS). Let DF a subset of $(\mathfrak{R} \rightarrow [0,1])$ be finite set of probabilistic distributed functions. H is a finite set of clocks identifiers.

A maximality based Labeled Stochastic Transition System is a structure $(\Omega, \lambda_S, \mu_S, \Psi_S, \xi_S)$ over Act , where: $\Omega = \langle S, s_0, T \rangle$ is a transition system with; S a countable set of states, s_0 is the initial state and T a countable set of transitions specifying the states change.

α and β are two applications of T in S such that for all transition t : $\alpha(t)$ is the origin of the transition and $\beta(t)$ its goal.

- $\lambda_S: T \rightarrow Act$ is a labeling function.
- $\mathcal{F}: Act \rightarrow DF$ This function associates to each action a probabilistic function distribution that governs its durations.
- $\Psi_S: S \rightarrow P(H)$; This function associates to each state a finite set of clocks related to its maximum events.
- $\mu_S: T \rightarrow P(H)$; This function associates to each transition a finite set of clocks. This set corresponds to its direct causes.
- $\xi_S: T \rightarrow H$; This function associates to each transition, a clock identifying its occurrence and counting its duration.

Such that $\psi(s_0) = \emptyset$ and for any transition t , $\mu(t) \subseteq \psi(\alpha(t)), \xi(t) \notin \psi(\alpha(t)) - \mu(t)$ and $\psi(\beta(t)) = (\psi(\alpha(t)) - \mu(t)) \cup \{\xi(t)\}$. For more details, the reader is referred to [1]

Each transition $t \in T$ has the following form $t = (S, Mi, (a, f), x, S')$, where: $\Psi_S(t) = x$, $\mu_S(t) = Mi$, $\lambda_S(t) = (a, f)$, $\alpha(t) = S$, $\beta(t) = S'$ such that $Mi \in \mathcal{M}$ and \mathcal{M} is set of clocks names.

Stochastic Refusals Graph (SRG)

In this section, we present Stochastic Refusals Graphs (SRGs) [5].

Definition 2: A Stochastic Refusals Graph $srg = (\Omega', \lambda_S, \mu_S, \Psi_S, \xi_S)$ is a deterministic bi-labeled graph structure of MLSTS with:

$\Omega' = \langle S, s_0, \Delta, Ref_{SRG} \rangle$ a transition system over a set of actions $Q \in Act$. S is a finite set of states.

$s_0 \in S$ is the initial state.

$\Delta \in (S \times Q \times S)$ is a transition relation.

$Ref_{SRG}: G \rightarrow P(P(\bar{Q}U\bar{Q}))$ is an application that associates for any $s \in S$ a set of refusals, where :

$\bar{Q} = \{\bar{a}(X) : a \in Q, X \subseteq H\}$ and

$\bar{Q} = \{\bar{a}(X) : a \in Q, X \subseteq H\}$

The semantics of the set $P(P(\bar{Q}U\bar{Q}))$ is as follows:

$\bar{a}(X) \in Ref_{SRG}(g)$: is a permanent refusal. It means that an action may be refused permanently in the state s . This refusal is caused by the indeterminism present in the system. Permanent refusal is possible but not certain. The certitude will take place after the expiration of clocks (X).

$\bar{\bar{a}}(X) \in Ref_{SRG}(g)$: Means that action a is refused until the expiration of clocks (X). This type of refusals occurs since actions have duration.

Stochastic Refusals Graph Generation

Let $mlsts = (\Omega, \lambda, \mu, \xi, \psi)$ be an MLSTS where, $\Omega = \langle S, S_0, T \rangle$. The construction of the SRG is done during the determinization of $mlsts$ as follow:

Algorithm :SRG of MLSTS

Input : $mlsts = (S, s_0, T, \lambda, \mu, \xi, \psi)$

Data A : set of states

B: set of transitions

Begin

A := $\{s_0\}$

Repeat

B := \emptyset

For each $s \in A$ do

Mark (s)

For all deterministic transition t and state s is treated do */* \(\alpha(t)=s\) Ref is not calculated yet */*

Ref = Ref $\cup \{\bar{a}(M) \mu(t) = M \neq \emptyset\}$

end

B := B $\cup \{t \in \text{out}(s) \mid t \text{ non-determinist}\}$

A := (A $\cup \text{non-mark}(\text{succ}(A))$) - (mark

(A))

For each $B_a(s) \in B$ with $|B_a(s)| \geq 2$ do

$s' = \text{create-new-state}(S)$ */* s' is not in S */*

S := S $\cup \{s'\}$

$\psi := (\psi \mid s' \rightarrow \text{Norm}(\bigcup_{1 \leq i \leq n} \psi(s_i)))$ with $s_i \in \text{succ}(s)$

/ ψ is extended to the new state s' */*

$t' := \text{create-new-transition}(T)$ */* t' is not in*

T **/*

add-transition($T, t', s, s', B_a(s)$)

A := A $\cup \{s'\}$

T := extend-transition-with-duplicate

($\text{succ}(s), s', t'$)

End for

End for

Until A = \emptyset

Output : SRG($mlsts$) = (S, s_0 , T, λ, μ, ξ, ψ) */* note that, the SRG($mlsts$) resulting is deterministic */*

Algorithm 2 extend-transition-duplicate ($\text{succ}(s), s', t'$)
function

Output T: set of all transitions

Begin

For each $s_i \in \text{succ}(s)$ do

For each $s_k \in \text{succ}(s)$ do */* s_k different from s_i */*

T' = $\{t \mid \alpha(t) = s_i\}$

T'' = $\{t_k \mid \alpha(t_k) = s_k\}$

For each $t = (s_i, M, a, x_i, s'_i) \in T'$ do

For each $t_k = (s_k, M_k, a, x_i, s'_k) \in T''$

```

do
  T := T ∪ {t}
  α(t) := (α | t → s')
  μ := (μ | t → M [xi/x]) // * x is equal to ξ(t')* //
  Ref = Ref ∪ { {ā(M) μ(t) = M ≠ ∅ } ∪
  {ā(M) μ(tk) = M ≠ ∅ } }
  End for
End for
End for
End for
End

```

Canonical Tester

A canonical tester [5] for a specification with respect to conformance relation is able to detect every implementation that disagrees with the specification according to the $conf_{srg}$ relation.

The conformance relation $conf_{srg}$ is based on permanent and temporary refusals in addition to forbidden actions. Forbidden actions at a given state s , noted $Forb(s)$, is the set of actions that are not allowed at this state. Formally $Forb(s) = Act - out(s)$.

$Out(s) = \{ s \in S : \exists s' \in S \text{ after } \sigma \text{ and an atom } M(a, f)_x, s \xrightarrow{M(a, f)_x} s' \}$.

This relation is defined as follow:

Definition 3 (implementation relation) Let I and S be MSLTSs

So,

$$I \text{ conf}_{srg} S \stackrel{\text{def}}{=} \forall \sigma \in Traces(S) \left\{ \begin{array}{l} (Forb(I, \sigma) \subseteq Forb(S, \sigma)) \text{ and} \\ (Ref_{srg}(I, \sigma) \subseteq Ref_{srg}(S, \sigma) \text{ such that } x_I \leq x_S) \end{array} \right.$$

On canonical tester, three verdicts are used which are pass, inconc and fail. At every step of the test computation, if a state is reachable so, it is decorated by pass verdict. The inconclusive verdict incon is produced by the non-determinism present in the system, and captured by permanent refusals set. Fail is a new state introduced to canalize transitions which are not allowed. This case corresponds to two kinds of actions:

- Forbidden actions
- Temporary refusals corresponding to actions for which direct causes doesn't respect the timed constraints.

The framework to generate canonical tester takes as input SRG with forbidden action then, it generate the tester.

SRG with forbidden action defined as follow

Definition 4

Let $srg(s) = (\Omega', \lambda_s, \mu_s, \psi_s, \xi_s)$.

The extended transition system of Ω' is $\Omega' e=S,0,\Delta,Ref_e$ Such that:

$Ref_e(g) = (Ref_{sr_g}(s) \cup Forb(s))$ So, the $sr_g_e(S) = (\Omega'_e, \lambda_S, \mu_S, \psi_S, \xi_S)$ is named extended stochastic refusals graph (ESRG).

A test cases is a path of canonical tester, it is defined as follow:

Definition 5

A test cases is an MLSTS

$TC = (S, s_0, \Delta, Verdict, \lambda, \mu, \xi, \psi)$ such that:

- $s_0 \in S$ the initial locality
- $Verdict = \{\text{pass, inconc, fail}\}$

The verdict fail is found in the new state created in order to canalize transitions which are not allowed.

3 Cases Studies

3.1 The Sender Part of ABP Protocol

ABP Protocol

ABP (Alternating Bit Protocol) is a connection-less protocol for transferring messages in one direction between a pair of protocol entities. It is a simple form of the Sliding Window Protocol with a window size of 1 [18], [11]. The name of this protocol stems from the fact that each message is augmented with an additional bit. This protocol uses one-bit sequence number (which alternates between 0 and 1) in each message and an acknowledgment to determine whether the message must be retransmitted. g

The Alternating bit Protocol (ABP) has attracted the interest of the research community due the fact that it has interesting aspects to verify, to prove and to test. For instance the authors [19] have proved its correctness, [18] has verified ABP using formal verification technique, [15] has tested ABP by generating successes graph. Also, its performance has been studied several times.

The aim of this section is to illustrate the use of MoVoS tool; we will focus on sender part.

The Model

The alternating bit protocol consists of a sender S, a receiver R, a channel K from S to R and a channel L from R to S; its architecture is as follow:

The sender S and receiver R communicate via the same lossy communications mediums (L and K). So messages may be lost. The basic principle is to stamp messages with a one bit sequence number. When a protocol entity sends message (either data or acknowledgment) with sequence number b . the next message it receives should be $\neg b$. If the sequence number is not as expected, the protocol entity concludes that the message has been lost and retransmits.

In this paper, we will focus on the sender part of ABP protocol. To explain how the tool work.

The Sender Part of ABP Protocol

The Sender part of ABP is responsible for accepting messages from the application and sending them via the Medium to the Receiver. The MLSTS of sender is presented in Fig .3.

3.2 Cell Production

The production cell case study is an attempt to define a realistic industrial application. It was developed by FZI in Karlsruhe as part of the Korso Project [17]

The aim of that project was to show the benefits of formal methods, with the production cell being used as vehicle for comparing different approaches. To date, it has been described in over thirty formalisms, It was modeled by timed automata in [8] and with stochastic process algebra PEPA in [12]

The production cell is composed of six elements: feed belt, deposit belt, press, robot with two arms A and B, elevating rotary table and sensor. Fig. 4 shows the components of the production cell.

The purpose of the cell is to take metal plates (banks) from the feed belt. Then, they will be pressed and moved to deposit belt. Plates are moved by robot. Arm A of the robot takes a plate from the table (which itself must twist and rise when it gets a single plate) and places the plate on the press. When the plate has been pressed, Arm B of the robot carries the plate to the deposit belt. The arms of the robot are fixed (with respect to each other) so the robot controller must coordinate its operations on the two arms.

In this system, the ultimate destination of the bank is the feed belt, so this cycle is repeated indefinitely. In a more realistic setting the bank would be moved to the next production phase. The production cell is modeled by MLSTS but, it is too large to show here, since the complete model is obtained by concurrent execution of the all components (feedbelt, table, press, robot, deposit belt).

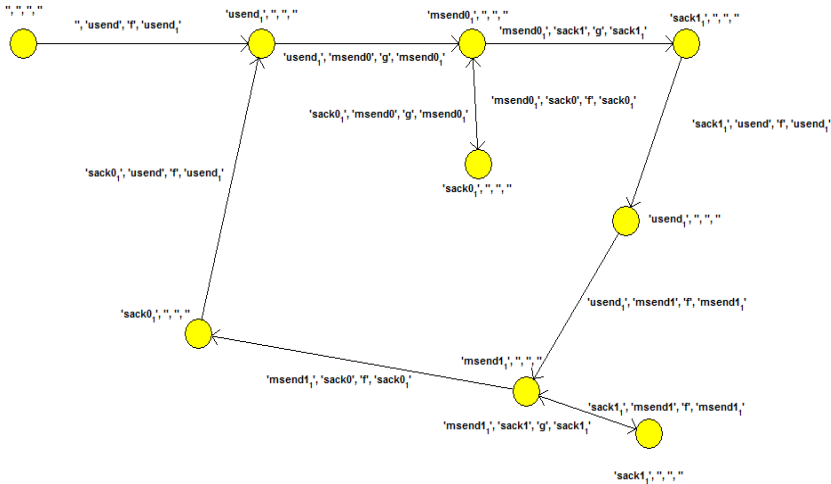


Fig. 3. MLSTS of sender part of ABP protocol



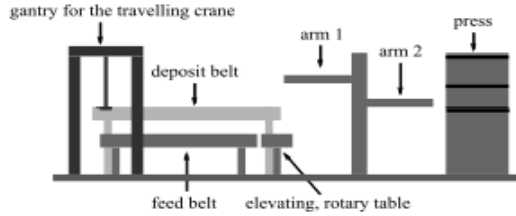


Fig. 4. The components of the production cell

3.3 Testing Step

Testing the Sender Part of ABP Protocol

The steps to generate test cases with MoVoS tool are done as follows:

- SRG Generation

We generate stochastic refusal graph using the tool. SRG of ABP sender is represented by Fig. 5 we notice that every state is decorated by refusals. We take the example of the S2.

S2 is decorated as follow:

```
{'msend0', '(sack1,g),(sack0,f)', '' }
```

That's means, the sender can't receive correct acknowledge or not, until data are send to receiver (i.e. msend0 reach 0).

- Canonical Tester Generation

Fig. 6 represents canonical tester. It is generated with MoVeS Tool.

We notice that every state is decorated with verdicts (pass, incon) and new state is created which is decorated with fail verdicts. The new state Fail is introduced to canalize transitions labeled by actions which are not permitted, such as those on the Forb set, or transition labeled by actions which are offered without respecting their clocks.

For example in Figure.16, after msend0 we can't found the action msend1 (msend1 is in forbset); so the transition labeled by msend1 go to the locality fail. The framework used to generate this tester is presented in section 5.2.

- Test Cases Generation

Fig. 7 is test cases. Test case is path of canonical tester. This one is generated randomly. If we take the close look at the test case generated by the MoVeS tool and depicted in Fig. 7, we can see that it examines one behavior of specification. More precisely, it starts by giving the message that will be sending, then the message is transmitting to receiver, finally the sender receives a correct acknowledgment. We notice that every state is decorated with pass verdict, that's means system successes after doing each correct action.

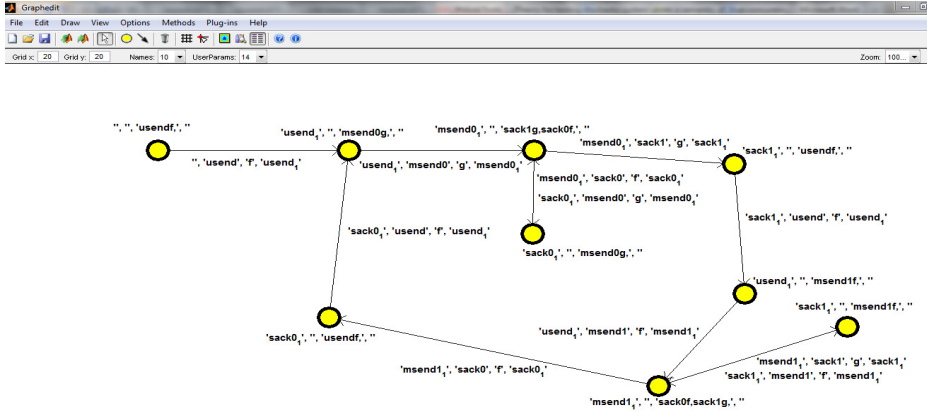


Fig. 5. SRG of the sender part of ABP protocol

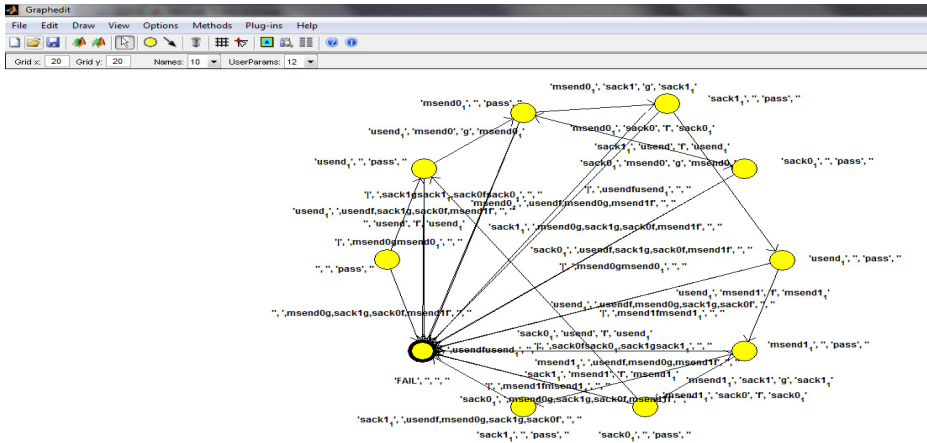


Fig. 6. Canonical tester of the sender part of ABP protocol

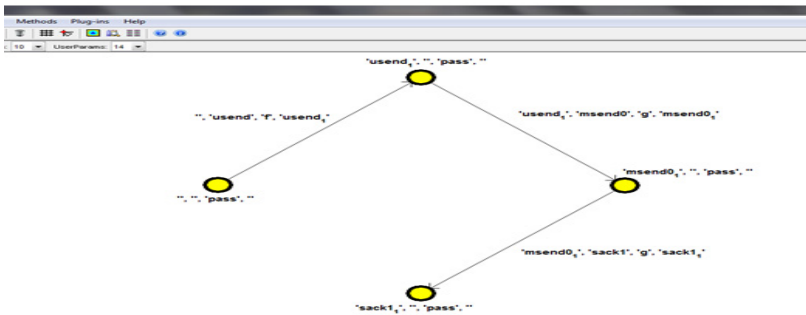


Fig. 7. Test cases of the sender part of ABP protocol

Testing Cell Production

The hall behavior of the product cell is tested using the MoVoS tool. The Canonical tester obtained is too large to show here. The results are summarized in table 1. The purpose of this case study is to show that this tool can deal with large systems

Table 1. Test cases result

	Feed belt	Deposit belt	Table	Robot	Press	Complete System
State	6	4	6	23	7	1459
Transition	6	4	6	29	7	4280
Case Study	17	11	17	68	20	3203

4 Conclusion and Future Work

In this paper, we have presented the principles of MoVoS tool, its underlying theory and algorithms. At the end we have presents two cases study, the first one is the sender part of ABP protocol which permit us to illustrate the way the tool work. The second one allows us to show that the tool deals with large systems.

Much other works remain to be done; we plan to complete this architecture by strategy for choosing which of test cases are sufficient for insuring some completeness guarantees. A related problem is how to select test suites with some good coverage measure to reduce them.

References

1. Arous, M., Saidouni, D.E., Ilić, J.M.: Maximality Semantics based Stochastic Process Algebra for Performance Evaluation. In: 1st IEEE International Conference on Communications, Computing and Control Applications (CCCA 2011), Hammamet, Tunisia, March 3-5 (2011), IEEE Catalog Number: CFP1154M-ART, ISBN: 978-1-4244-9796-6
2. Arous, M., Saidouni, D.E., Ilić, J.M.: Addressing State Space Explosion Problem in Performance Evaluation Using Maximality-based Labeled Stochastic Transition Systems. In: to appear in 2nd International Conference on Computer and Software Modeling - ICCSM 2012, Cochin, India, October 20-21. IPCSIT (2012)
3. Arous, M., Bouaroudj, K., Saïdouni, D.E.: An environment for modeling and verifying / testing stochastic systems. Issues of JATIT 50 (appear in April 2013)
4. Belinfante, A., Feenstra, J., de Vries, R.G., Tretmans, J., Goga, N., Feijs, L., Mauw, S., Heerink, L.: Formal test automation: A simple experiment. In: 12th Int. Workshop on Testing of Communicating Systems. Kluwer (1999)
5. Bouaroudj, K., Kitouni, I., Hachichi, H., Saidouni, D.E.: Extending Refusal Testing by Stochastic Refusals for Testing Non-deterministic Systems. IJCSI 9(5) (September 2012)
6. Brinksma, E.: A theory for the derivation of tests. In: Aggarwal, S., Sabnani, K. (eds.) Proceedings of the 8th IFIP Symposium on Protocol Specification, Testing and Verification (PSTV 1988). North-Holland (1989)

7. Tschaen, V.: 6 test generation algorithms based on preorder relations. In: Broy, M., Jonsson, B., Katoen, J.-P., Leucker, M., Pretschner, A. (eds.) *Model-Based Testing of Reactive Systems. LNCS*, vol. 3472, pp. 151–171. Springer, Heidelberg (2005)
8. Burns, A.: How to verify a safe real time: the application of model checking and timed automata to the production cell case study. *Real Time Systems Journal* 24(2), 135–152 (2003)
9. Clarke, D., Jéron, T., Rusu, V., Zinovieva, E.: STG: A symbolic test generation tool. In: Katoen, J.-P., Stevens, P. (eds.) *TACAS 2002. LNCS*, vol. 2280, pp. 470–475. Springer, Heidelberg (2002)
10. Jalali, V., Borujerdi, M.R.M.: A hybrid information retrieval system for medical field using meSH ontology. In: Prasad, S.K., Routray, S., Khurana, R., Sahni, S. (eds.) *ICISTM 2009. CCIS*, vol. 31, pp. 31–40. Springer, Heidelberg (2009)
11. Hessel, A., Larsen, K.G., Nielsen, B., Pettersson, P., Skou, A.: Time-optimal real-time test case generation using UPPAAL. In: Petrenko, A., Ulrich, A. (eds.) *FATES 2003. LNCS*, vol. 2931, pp. 114–130. Springer, Heidelberg (2004)
12. Holton, D.R.W.: A PEPA Specification of an Industrial Production Cell. *The Computer Journal* 38(7), 542–551 (1995)
13. Jard, C., Jéron, T.: Tgv: theory, principles and algorithms, a tool for the automatic synthesis of conformance test cases for non-deterministic reactive systems. *Software Tools for Technology Transfer (STTT)* 10 (2004)
14. Nielsen, B., Skou, A.: Automated test generation from timed automata. In: Margaria, T., Yi, W. (eds.) *TACAS 2001. LNCS*, vol. 2031, pp. 343–357. Springer, Heidelberg (2001)
15. Nielsen, B.: Specification and test of real-time systems: PhD thesis, Aalborg University (April 2000)
16. Saïdouni, D.E., Courtiat, J.P.: Prise en compte des durées d'action dans les algèbres de processus par l'utilisation de la sémantique de maximalité. In: *Proceedings of CFIP 2003*, Hermes, France (2003)
17. Tyszberowicz, S.S.: How to implement a safe real-time system: The OBSERV implementation of the production cell case study. *Real Time Systems* 15(1), 61–90 (1998)
18. Kamrul, H.T.: Formal verification of the Alternating Bit Protocol. In: *6th International Conference on Computer & Information Technology, ICCIT (2003)*
19. Groote, J.F., Springintveld, J.: Focus points and convergent process operators. A proof strategy for protocol verification. *Journal of Logic and Algebraic Programming* 49(1/2), 31–60 (2001)

Two Scale Modeling of Heterogeneous Solid Body by Use of Thick Shell Finite Elements

Dalia Čalnerytė and Rimantas Barauskas

Kaunas University of Technology, Faculty of Informatics, Kaunas, Lithuania
dalia.calneryte@stud.ktu.lt, rimantas.barauskas@ktu.lt

Abstract. An elasticity parameters evaluation for homogeneous material is considered in this paper if parameters of consisting materials are known in micro scale. The thick shell formulation for homogeneous orthotropic material is discussed and total Lagrangian formulation for the 4-node thick shell element in implicit and explicit analysis is described. The results of the thick shell model are compared with the results of 3D model and LS-Dyna shell model with the same loading.

Keywords: Multi-scale modeling, Total Lagrangian formulation, 4-node thick shell element.

1 Introduction

Multi-scale finite element analysis is widely used for modeling and simulation of the physical behavior of materials, the internal structure of which is non-homogeneous and/or architecturally complex. The main idea of multi-scale modeling is to analyze the same physical phenomena or behavior in different length scales. The models of different scales are used to represent the behavior of the same object, however, with different level of minuteness. Appropriate transfer of behavioral features among the models must be ensured. Different assumptions are used for creation of models in each length scale. All the materials traditionally considered as homogeneous are in fact heterogeneous at micro-scale. Traditionally used isotropic, orthotropic, anisotropic behavioral models of materials are based on experimentally known data. As a rule, we consider that they do not require any further analysis at micro-scale. In the real world majority of materials are composites, where the parameters of materials are known only in micro scale. That is the reason why multi-scale modeling is used for evaluation of equivalent material parameters in upper scale. Equivalent parameters are the elasticity parameters of homogeneous body which has the same behavior as heterogeneous body. Sometimes in engineering computations traditional orthotropic models with properly adjusted parameters are employed, however, they may serve only as very rough estimations of the reality.

The equivalent parameters of a material are evaluated according to the rules in [2] and [3]. Obviously this is not the only way to evaluate material parameters – the method using asymptotic homogenization is presented in [9]. Applying this method only

one periodic cell is analyzed with specific periodicity boundary conditions and asymptotic expansion of displacement fields. Moreover the material parameters can be evaluated by mechanical approach with respect to material share in the model. All these methods work fine when linear elasticity is analyzed. The problem arises when there is material or geometrical non-linearity.

In this work, we concentrate on elaboration of thick shell finite elements suitable for multi-scale computations. Shell elements are convenient in computations of upper scales if the dimensions of a body are significantly small in one direction compared with others. Three main formulations in analysis of geometrical non-linearity as total Lagrange, updated Lagrange and co-rotational formulations may be employed. In the total Lagrange formulation the reference configuration is the initial state of the element and in the updated Lagrange formulation the reference configuration is the last known state [5, 10]. In the co-rotational formulation the reference configuration translates and rotates with the element [10]. The total Lagrange formulation is used in this paper for the formulation of the thick shell finite element. The stiffness tensor of the material is obtained by means sequential multi-scale coupling. The homogenized mechanical stiffness constants of the structure are obtained by performing the FE analysis of the mechanical behaviour of the micro-cube. Pure stress components of the micro-cube are created by prescribing the necessary displacements of the sides of the micro-cube. The micro-level finite element model is employed for computing stresses within the micro-cube. MATLAB mathematical software environment and finite element software LS-DYNA were employed.

2 Evaluation of Equivalent Parameters for 3D Solid Model

Homogenization of the composite material to linear elastic material is considered, where the elasticity tensor is used in order to relate stresses and strains in accordance with the generalized Hooke's law [1]:

$$\boldsymbol{\sigma} = \mathbf{D}\boldsymbol{\varepsilon}, \quad (1)$$

where \mathbf{D} is 6×6 elasticity matrix, $\boldsymbol{\sigma} = \{\sigma_x \ \sigma_y \ \sigma_z \ \tau_{xy} \ \tau_{yz} \ \tau_{zx}\}^T$, $\boldsymbol{\varepsilon} = \{\varepsilon_x \ \varepsilon_y \ \varepsilon_z \ \gamma_{xy} \ \gamma_{yz} \ \gamma_{zx}\}^T$ are the stress and strain tensors in Voigt's notation respectively. In order to evaluate equivalent parameters of the 3D thick shell model it is assumed that there is zero stress in normal direction ($\sigma_z = 0$). Hence the inverse Hooke's law for 3D model may be separated into two systems [2, 3]:

$$\begin{Bmatrix} \varepsilon_x \\ \varepsilon_y \\ \gamma_{xy} \end{Bmatrix} = \begin{bmatrix} S_{11} & S_{12} & 0 \\ S_{21} & S_{22} & 0 \\ 0 & 0 & S_{44} \end{bmatrix} \begin{Bmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{Bmatrix}; \quad (2)$$

$$\begin{Bmatrix} \gamma_{yz} \\ \gamma_{zx} \end{Bmatrix} = \begin{bmatrix} S_{55} & 0 \\ 0 & S_{66} \end{bmatrix} \begin{Bmatrix} \tau_{yz} \\ \tau_{zx} \end{Bmatrix}, \quad (3)$$

where S_{ij} is a component in the i th row and j th column of the compliance matrix $\mathbf{S} = \mathbf{D}^{-1}$.

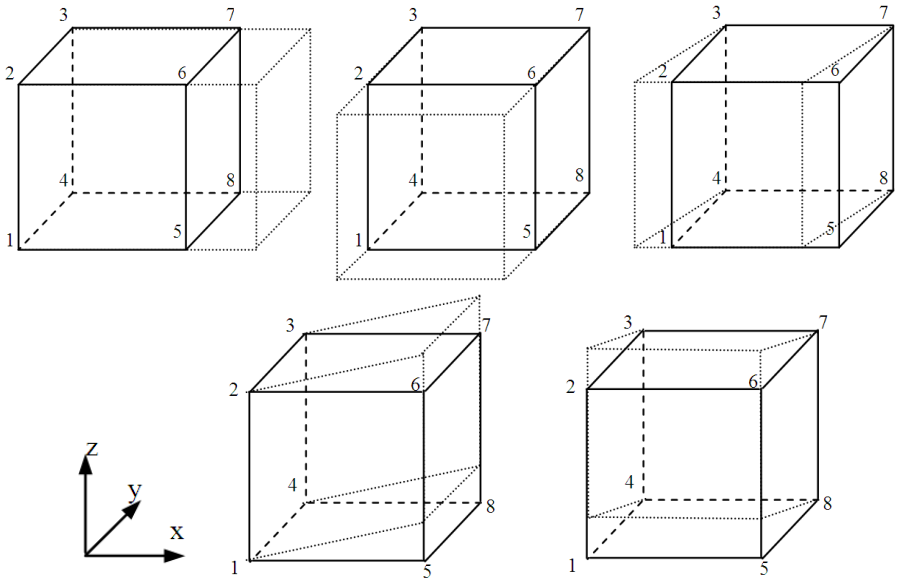


Fig. 1. Schemes for modeling pure stress in LS-DYNA

In order to evaluate the homogenized material parameters of the heterogeneous material pure stresses are simulated according to the schemes in Fig. 1 for the 3D solid finite model of the micro-cube which represents detailed heterogeneous micro-structure.

Table 1. Pure stress assumptions and formulas for parameters evaluation

Assumptions	Formula
$\sigma_x \neq 0, \sigma_y = 0, \tau_{xy} = 0, \tau_{yz} = 0, \tau_{zx} = 0$	$E_x = \frac{\sigma_x}{\epsilon_x}, \nu_{xy} = -\frac{\epsilon_y}{\epsilon_x}$
$\sigma_x = 0, \sigma_y \neq 0, \tau_{xy} = 0, \tau_{yz} = 0, \tau_{zx} = 0$	$E_y = \frac{\sigma_y}{\epsilon_y}, \nu_{yx} = -\frac{\epsilon_x}{\epsilon_y}$
$\sigma_x = 0, \sigma_y = 0, \tau_{xy} \neq 0, \tau_{yz} = 0, \tau_{zx} = 0$	$G_{xy} = \frac{\tau_{xy}}{\gamma_{xy}}$
$\sigma_x = 0, \sigma_y = 0, \tau_{xy} = 0, \tau_{yz} \neq 0, \tau_{zx} = 0$	$G_{yz} = \frac{\tau_{yz}}{\gamma_{yz}}$
$\sigma_x = 0, \sigma_y = 0, \tau_{xy} = 0, \tau_{yz} = 0, \tau_{zx} \neq 0$	$G_{zx} = \frac{\tau_{zx}}{\gamma_{zx}}$

If material is isotropic it is enough to know Young's modulus and Poisson's ratio. Orthotropic material for the thick shell element is defined by 6 parameters: E_x , E_y is Young's modulus in index direction, ν_{xy} is Poisson's ratio in index plane, G_{xy} , G_{yz} , G_{zx} is shear modulus in index plane (Poisson's ratio ν_{yx} is redundant because of the symmetry of elasticity matrix ($\nu_{xy}E_y = \nu_{yx}E_x$)):

$$\mathbf{D} = \begin{bmatrix} \frac{E_x}{1-\nu_{xy}\nu_{yx}} & \frac{E_x\nu_{yx}}{1-\nu_{xy}\nu_{yx}} & 0 & 0 & 0 \\ \frac{E_y\nu_{xy}}{1-\nu_{xy}\nu_{yx}} & \frac{E_y}{1-\nu_{xy}\nu_{yx}} & 0 & 0 & 0 \\ 0 & 0 & G_{xy} & 0 & 0 \\ 0 & 0 & 0 & \kappa G_{yz} & 0 \\ 0 & 0 & 0 & 0 & \kappa G_{zx} \end{bmatrix} \quad (4)$$

Where $\kappa=5/6$ is a shear correction factor and its purpose is to improve shear displacement approximation.

3 4-node Thick Shell Element

Any shell element can be defined by material properties, nodal point coordinates, shell mid-surface normals and shell thickness at each mid-surface node. The thick shell element is a degenerated three dimensional solid element with integration over its mid-surface. Any point of the thick shell may be related to the top and bottom surfaces of the element:

$$\begin{Bmatrix} x \\ y \\ z \end{Bmatrix} = \sum N_k(\xi, \eta) \cdot \left(\frac{1+\zeta}{2} \begin{Bmatrix} \tilde{x}_k \\ \tilde{y}_k \\ \tilde{z}_k \end{Bmatrix}_{top} + \frac{1-\zeta}{2} \begin{Bmatrix} \tilde{x}_k \\ \tilde{y}_k \\ \tilde{z}_k \end{Bmatrix}_{bottom} \right) \quad (5)$$

Where $N_k(\xi, \eta)$ is a shape function of the k th node and ζ is a linear coordinate in the thickness direction.

For convenience the previous equation can be rewritten in respect to the mid-surface coordinates and a vector connecting upper and lower points. This vector is a product of shell thickness h_k at the k th node and a unit vector \mathbf{v}_3 in the direction normal to the mid-surface [1]:

$$\begin{Bmatrix} x \\ y \\ z \end{Bmatrix} = \sum N_k(\xi, \eta) \cdot \left(\begin{Bmatrix} \tilde{x}_k \\ \tilde{y}_k \\ \tilde{z}_k \end{Bmatrix} + \frac{1}{2} \zeta h_k \mathbf{v}_3 \right) \quad (6)$$

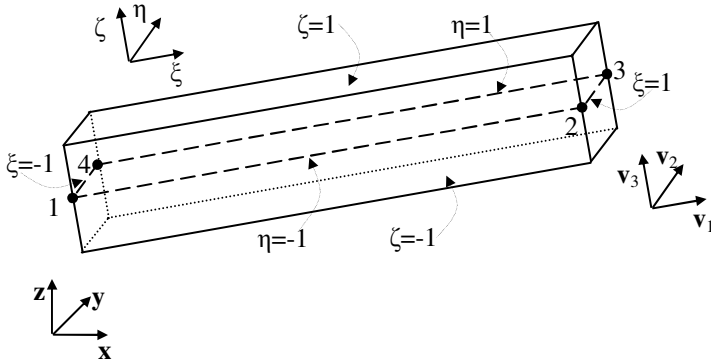


Fig. 2. Thick shell element

The analyzed element is isoparametric – shape functions map every quadrilateral element to square and are used to interpolate the element coordinates and displacements. Shape function for the k th node of 4-node thick shell element is bi-linear Lagrange polynomial:

$$N_k(\xi, \eta) = \frac{1}{4}(1 + \xi_k \xi)(1 + \eta_k \eta), \quad k = 1, 2, 3, 4 \tag{7}$$

Due to the fact that the strain in the thickness direction is assumed to be 0, the displacements at each node of the thick shell is uniquely defined by three Cartesian components of the mid-surface node displacement and two rotations about orthogonal directions defined by vectors $\mathbf{v1}$ and $\mathbf{v2}$ normal to $\mathbf{v3}$:

$$\tilde{\mathbf{u}}_k = \{\tilde{u}_k \quad \tilde{v}_k \quad \tilde{w}_k \quad \tilde{\alpha}_k \quad \tilde{\beta}_k\}^T \tag{8}$$

It is evident that a coordinate vector \mathbf{x} in a Cartesian system may be defined by

$$\mathbf{x} = x\mathbf{e}_x + y\mathbf{e}_y + z\mathbf{e}_z \tag{9}$$

Where $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ are base vectors. Vectors $\mathbf{v1}$ and $\mathbf{v2}$ in Fig. 2 may be constructed with the following formulas [1]:

$$\mathbf{v1} = \frac{\mathbf{e}_x \times \mathbf{v3}}{|\mathbf{e}_x \times \mathbf{v3}|}, \quad \mathbf{v2} = \frac{\mathbf{v3} \times \mathbf{v1}}{|\mathbf{v3} \times \mathbf{v1}|} \tag{10}$$

The displacements of any point of the thick shell may be written in respect of the mid-surface displacements [1]:

$$\begin{Bmatrix} u \\ v \\ w \end{Bmatrix} = \sum N_k(\xi, \eta) \cdot \left(\begin{Bmatrix} \tilde{u}_k \\ \tilde{v}_k \\ \tilde{w}_k \end{Bmatrix} + \frac{1}{2} \zeta h_k [\tilde{\mathbf{v}}1_k \quad -\tilde{\mathbf{v}}2_k] \begin{Bmatrix} \tilde{\alpha}_k \\ \tilde{\beta}_k \end{Bmatrix} \right) \tag{11}$$



A 2x2 Gauss integration rule is used for numerical integration of the 4 node element in plane and a 2 point Gauss integration rule is used for numerical integration through thickness.

4 Total Lagrangian Formulation in FEM

Total Lagrangian formulation relates 2nd Piola–Kirchhoff stress to Green–Lagrange strain and all variables of the body are referred to the initial configuration [5].

4.1 Implicit Analysis

Finite element discretization of total Lagrangian formulation for a single element (\mathbf{R} – vector of nodal forces and moments) for i th iteration of implicit analysis [5]:

$$\mathbf{K}\Delta\mathbf{u}^{(i)} = \mathbf{R} - \mathbf{F} \tag{12}$$

Where $\mathbf{F} = \int_{V_0} \mathbf{B}_L^T \hat{\mathbf{S}} dV$, $\hat{\mathbf{S}}^T = \{\sigma_x \ \sigma_y \ \sigma_z \ \tau_{xy} \ \tau_{yz} \ \tau_{xz}\}$, \mathbf{F} is a vector of nodal

internal forces and moments, $\Delta\mathbf{u}^{(i)}$ is an increment of nodal displacements in i th iteration and stiffness matrix \mathbf{K} is a sum of linear \mathbf{K}_L and non-linear \mathbf{K}_{NL} parts [4]:

$$\mathbf{K}_L = \int_{V_0} \mathbf{B}_L^T \mathbf{D} \mathbf{B}_L dV, \mathbf{K}_{NL} = \int_{V_0} \mathbf{B}_{NL}^T \mathbf{S} \mathbf{B}_{NL} dV \tag{13}$$

Where $\mathbf{S}^T = \begin{bmatrix} \sigma_{xx} \mathbf{I}_3 & \sigma_{xy} \mathbf{I}_3 & \sigma_{xz} \mathbf{I}_3 \\ \sigma_{xy} \mathbf{I}_3 & \sigma_{yy} \mathbf{I}_3 & \sigma_{yz} \mathbf{I}_3 \\ \sigma_{xz} \mathbf{I}_3 & \sigma_{yz} \mathbf{I}_3 & \sigma_{zz} \mathbf{I}_3 \end{bmatrix}$ and \mathbf{I}_3 – 3x3 identity matrix, \mathbf{D} – elasticity tensor,

\mathbf{B}_L is a matrix such that $\boldsymbol{\varepsilon} = \mathbf{B}_L \mathbf{u}$ and $\boldsymbol{\varepsilon}$ is Green – Lagrange strain, \mathbf{u} – vector of nodal displacements. Usually \mathbf{B}_L is a sum of two matrices. For non-linear part \mathbf{B}_{NL} can be written:

$$\mathbf{B}_{NL} = \begin{bmatrix} | & N_{kx}' \cdot \mathbf{I}_3 & \rho_{k,x} \cdot \mathbf{v1} & -\rho_{k,x} \cdot \mathbf{v2} & | \\ \dots & N_{ky}' \cdot \mathbf{I}_3 & \rho_{k,y} \cdot \mathbf{v1} & -\rho_{k,y} \cdot \mathbf{v2} & | \dots \\ | & N_{kz}' \cdot \mathbf{I}_3 & \rho_{k,z} \cdot \mathbf{v1} & -\rho_{k,z} \cdot \mathbf{v2} & | \end{bmatrix} \tag{14}$$

Where $\rho_{k,x} = \frac{h}{2} (N_{kx}' \cdot \zeta + N_k \cdot \zeta_x')$, $\rho_{k,y} = \frac{h}{2} (N_{ky}' \cdot \zeta + N_k \cdot \zeta_y')$, $\rho_{k,z} = N_k$, $k = 1,2,3,4$.

Displacements after i th iteration is a sum of displacements after $(i-1)$ th iteration and increment of displacements in i th iteration:

$$\mathbf{u}^{(i)} = \mathbf{u}^{(i-1)} + \Delta\mathbf{u}^{(i)} \tag{15}$$



4.2 Explicit Analysis

The global system of discretized equations of motion at the n th time step is [7]:

$$\mathbf{M}\ddot{\mathbf{u}}_n + \mathbf{K} \cdot \mathbf{u}_n = \mathbf{R}_n \quad (16)$$

Where \mathbf{u}_n is a vector of nodal displacements at the n th time step, \mathbf{M} – a mass matrix, \mathbf{K} – stiffness matrix non-linearly dependent on strains, \mathbf{R}_n – vector of nodal (active) forces and moments at the n th time step.

The diagonal mass matrix is required in calculations of explicit analysis. Firstly the total element mass is evenly distributed among the four element nodes. Then the rotational nodal masses m_θ are calculated by scaling the translational mass m_t at the node by factor α [8]:

$$m_t = \rho \frac{A}{4} h, \quad m_\theta = \alpha \cdot m_t, \quad \alpha = \frac{h^2}{12} \quad (17)$$

Where ρ is the density of material, A is the area of element, h is the thickness of shell, m_t is used for calculating translational accelerations, m_θ is used for calculating rotational accelerations.

Instead of the product of stiffness matrix and displacements the vector of internal forces may be evaluated [7]:

$$\mathbf{K} \cdot \mathbf{u}_n = \mathbf{F}_n = \int_{V_0} (\mathbf{B}_L^T)_n \hat{\mathbf{S}}_n dV \quad (18)$$

Displacements at the $(n+1)$ th time step Δt are explicitly computed using central difference formula [7]:

$$\mathbf{u}_{n+1} = \Delta t^2 \mathbf{M}^{-1} (\mathbf{R}_n - \mathbf{F}_n) + 2\mathbf{u}_n - \mathbf{u}_{n-1} \quad (19)$$

5 Numerical Experiments

5.1 Evaluation of Material Parameters

The heterogeneous material with periodic microstructure is considered in micro scale. This material consists of two isotropic materials composed as shown in Fig. 3 and called fiber and matrix materials. Each material is defined by Young's modulus and Poisson's ratio and density additionally for the explicit analysis in Table 2. The fibers lie along the x axis in the model.

Table 2. Material parameters of 3D model

	Fiber material	Matrix material
Young's modulus, E	$73.1 \cdot 10^9 \text{ N/m}^2$	$3.45 \cdot 10^9 \text{ N/m}^2$
Poisson's ratio, ν	0.22	0.35
Density, ρ	1830 kg/m^3	900 kg/m^3

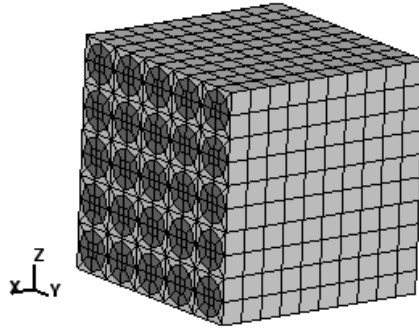


Fig. 3. 3D model used for parameters evaluation

The equivalent parameters for homogeneous material are evaluated by the scheme described in earlier sections. Density of homogeneous material is calculated as weighted mean considering the material share in the model.

Table 3. Material parameters of thick shell model

Young's modulus, E_x	$44.3 \cdot 10^9 \text{ N/m}^2$
Young's modulus, E_y	$14.4 \cdot 10^9 \text{ N/m}^2$
Poisson's ratio, ν_{xy}	0.32
Shear modulus, G_{xy}	$4.43 \cdot 10^9 \text{ N/m}^2$
Shear modulus, G_{yz}	$4.05 \cdot 10^9 \text{ N/m}^2$
Shear modulus, G_{zx}	$4.94 \cdot 10^9 \text{ N/m}^2$
Density, ρ	1432.7 kg/m^3

5.2 Bending Test

The initial geometry of structure is plane and the one end of the structure is constrained in all directions and rotations. The structure in Fig. 4 is $1m$ length, $0.5m$ width and its thickness (h) is $0.1m$. The out of plane loading is applied in the free end of the structure. It is one of the tests for finite element proposed in [6].

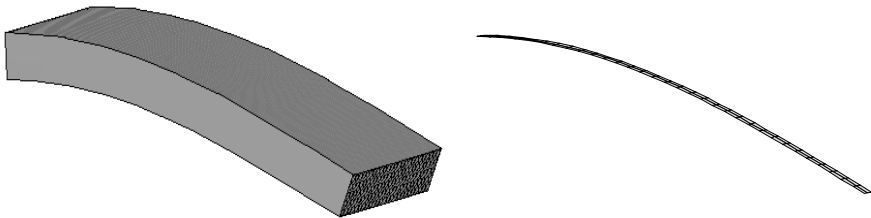


Fig. 4. Deformed configurations of 3D model and thick shell

Implicit Analysis

As it is shown in Table 4 the displacements linearly depend on loaded force when the loading value is small. In this case the relative difference between displacements of 3D model and thick shell elements is constant and displacements of thick shells are greater in absolute value. It should be noticed that displacements were estimated in only one corner mid-surface node of 3D element and the heterogeneity could cause disagreement.

Table 4. Displacements of the free end in z direction (m)

Loading (N)	3D model (m)	Shell (m)	Shell (LS-DYNA) (m)
1e1	1.34e-6	1.74e-6	5.19e-6
1e4	1.34e-3	1.74e-3	5.19e-3
1e5	1.34e-2	1.74e-2	5.19e-2
2e5	2.68e-2	3.47e-2	1.02e-1
4e5	5.35e-2	6.90e-2	1.97e-1
1e6	1.32e-1	1.65e-1	4.10e-1
2e6	2.50e-1	3.00e-1	5.90e-1

Explicit Analysis

In explicit analysis the out-of-plane force evolves linearly according to time.

Table 5. Displacements of the free end in z direction

t (s)	Loading (N)	3D model (m)	Shell (m)	Shell (LS- DYNA) (m)
0.0001	1e4	3.82e-6	6.24e-6	3.87e-6
0.0003	3e4	6.53e-5	8.12e-5	8.07e-5
0.0005	5e4	2.39e-4	2.78e-4	3.20e-4
0.0007	7e4	5.60e-4	6.32e-4	7.71e-4
0.0009	9e4	1.05e-3	1.17e-3	1.46e-3
0.001	1e5	1.37e-3	1.52e-3	1.94e-3

It is obvious that in explicit analysis the displacements do not change linearly though loading force evolves linearly. In addition, displacements of the models differ more than 10% at each moment except for the 3D and LS-DYNA models at the first step. It is important to notice that explicit analysis was performed with a time step $\Delta t = 10^{-5} s$ and the geometrical non-linearity of models was small.

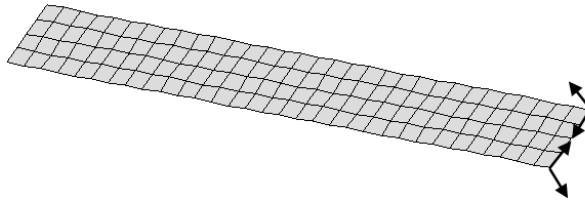
The displacements of models differ more than 15% at the first two moments. Distinct from the **Table 5**, in **Table 6** geometrical non-linearity is large and the displacements of 3D model and both shell models also do not differ more than 15% at last four moments.

Table 6. Displacements of the free end in z direction

t (s)	Loading (N)	3D model (m)	Shell (m)	Shell (LS- DYNA) (m)
0.0001	1e7	4.92e-3	6.23e-3	3.87e-3
0.0003	3e7	6.62e-2	7.73e-2	7.79e-2
0.0005	5e7	2.28e-1	2.22e-1	2.48e-1
0.0007	7e7	4.36e-1	3.94e-1	4.30e-1
0.0009	9e7	6.42e-1	5.87e-1	6.19e-1
0.001	1e8	7.65e-1	7.03e-1	7.34e-1

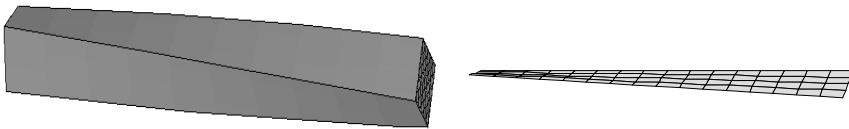
5.3 Twisting Test

The equal forces in y and z directions are loaded in the free end of the beam as shown in Fig. 5. The length of an arrow is F. It is one of the tests for finite element proposed in [6].

**Fig. 5.** Loading for twisting test

Implicit Analysis

The initial geometry of beam in Fig. 6 is plane and the one end of the beam is constrained in all directions and rotations. The beam is $1m$ in length, $0.1m$ in width and its thickness (h) is $0.1m$.

**Fig. 6.** Deformed configurations of 3D model and thick shell**Table 7.** Displacements of the corner of thick shell and midsurface of 3D model

F (N)	3D model (m)		Shell model (m)		Shell (LS-DYNA)(m)	
	y	z	y	z	y	z
1e3	8.54e-6	8.25e-5	1.62e-6	7.19e-5	6.12e-7	7.93e-5
1e4	8.99e-5	8.36e-4	2.08e-5	7.28e-4	1.19e-5	8.06e-4
1e5	1.50e-3	9.36e-3	7.70e-4	7.96e-3	8.89e-4	9.16e-3
2e5	4.95e-3	2.06e-2	3.19e-3	1.70e-2	4.69e-3	2.06e-2

The displacements of twisted models differ significantly in y direction as shown in Table 7 though displacements in z direction differ less than 10% for 3D and LS-DYNA shell models. Like for the bending test here displacements were estimated in only one corner mid-surface node of 3D element. In addition, the loading is sensitive for heterogeneity.

The initial geometry of the structure in Fig. 7 is plane and the one end of the structure is constrained in all directions and rotations. The structure is $1m$ in length, $0.5m$ in width and its thickness (h) is $0.1m$.

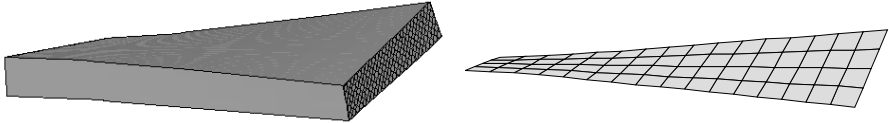


Fig. 7. Deformed configurations of 3D model and thick shell

Table 8. Displacements of the corner of thick shell and midsurface of 3D model

F (N)	3D model (m)		Shell model (m)		Shell (LS-DYNA) (m)	
	y	z	y	z	y	z
1e4	1.08e-4	1.52e-3	3.70e-5	1.54e-3	3.64e-5	1.68e-3
5e4	6.10e-4	7.74e-3	2.77e-4	7.84e-3	2.99e-4	8.65e-3
1e5	1.41e-3	1.59e-2	8.07e-4	1.60e-2	9.28e-4	1.78e-2

The displacements of twisted models differ significantly in y direction as shown in Table 8 though displacements in z direction differ less than 15%.

6 Final Remarks

The homogenized elasticity parameters for heterogeneous material were evaluated in this paper by modeling pure stresses. The 4-node thick shell element with equivalent parameters was implemented. The results of 3D model of heterogeneous structure, thick shell model with 4-node elements and shell model in LS-DYNA were compared.

For bending test the results of shell model in LS-DYNA differed the most compared with the displacements for 3D heterogeneous structure in implicit analysis. Though the differences do not exceed 15% when non-linearity is large in explicit analysis.

Two structures were tested for twisting. Displacements in y direction differed significantly for both structures with all analyzed loadings and difference of displacements in z direction exceeded 20% only with large loading.

In summary evaluation of equivalent elasticity parameters for heterogeneous material described in this paper can be used to analyze behavior of body with composite material only approximately. Moreover thick shell element formulation used for body modeling in upper scale is rather primitive and does not avoid problems such as shear locking. However such element is valuable because of its simple implementation and low computational cost.

References

1. Zienkiewicz, O.C., Taylor, R.L.: The Finite Element Method for Solid and Structural Mechanics. Elsevier, Oxford (2005)
2. Barbero, E.J.: Finite element analysis of composite materials. CRC Press (2008)
3. Kaw, K.A.: Mechanics of Composite Material. CRC Press (2006)
4. Dvorkin, E.N., Bathe, K.J.: A continuum mechanics based four-node shell element for general nonlinear analysis. Eng. Comput. 1, 77–88 (1984)
5. Bathe, K.J.: Finite Element Procedures. Prentice Hall (1996)
6. Macneal, R.H., Harder, R.L.: A Proposed Standard Set of Problems to Test Finite Element Accuracy. Finite Element in Analysis and Design 1, 3–20 (1985)
7. Miller, K., Joldes, G., Lance, D., Wittek, A.: Total Lagrangian explicit dynamics finite element algorithm for computing soft tissue deformation. Communications in Numerical Methods in Engineering 23, 121–134 (2007)
8. Tabiei, A., Tanov, R.: Sandwich shell finite element for dynamic explicit analysis. International Journal for Numerical Methods in Engineering 54, 763–787 (2002)
9. Pinho-da-Cruz, J., Oliveira, J.A., Teixeira-Dias, F.: Asymptotic homogenization in linear elasticity. Part I: Mathematical formulation and finite element modeling. Computational Materials Science 45, 1073–1080 (2009)
10. The, L.H., Clarke, M.J.: Co-rotational and Lagrangian formulations for elastic three-dimensional beam finite elements. Journal of Constructional Steel Research 48, 123–144 (1998)

Development in Authentication of AODV Protocols to Resist the Attacks

Ahmad Alomari

Faculty of Mathematics and Computer Science, University of Bucharest, Bucharest, Romania
alomari.jordan@gmail.com

Abstract. Mobile Ad hoc networks (MANETs) are new wireless networks with a self-configuring and self-maintaining network characterized as dynamic topology. This work will discuss the existing approaches that have intended to create a defense against various attacks at different layers. Routing is a very important function in MANETs, as described earlier. Making routing protocols efficient often increases the security risk of the protocol and allows a single node to significantly impact the operation of the protocol because of the lack of protocol redundancy. We focus on our scheme on authentication between the nodes and we choose the on-demand routing protocol like Ad Hoc On-Demand Distance Vector (AODV) protocol to apply this scheme to, which it depends on hash function, hash lock and random number generation and also on secret value and time stamp. This scheme is used to produce security and authentication environment between the nodes in Mobile Ad Hoc Network.

Keywords: routing protocol, hash function, AODV, black hole attacks.

1 Introduction

A mobile ad hoc network is a collection of wireless mobile nodes forming a temporary network without the aid of any established infrastructure or centralized administration. This new type of self-organizing network combines wireless communication with a high-degree node mobility. Unlike conventional wired networks, they have no fixed infrastructure (base stations, centralized management points and the like) [1]. The union of nodes forms an arbitrary topology. Because the nodes are mobile, the network topology may change rapidly and unpredictably over time. The network is decentralized; all network activity, including discovering the topology and delivering messages, must be executed by the nodes themselves; that is, routing functionality will be incorporated into mobile nodes. In [2], and [3], threshold cryptography has been proposed to provide a reliable, distributive key management for MANET by exploiting some nodes as a trust anchor for the rest of the network.

Security in a MANET is an essential component for basic network functions like packet forwarding and routing, network operation can be easily attack if countermeasures are not embedded into basic network functions at the early stages of their design. Unlike networks using dedicated nodes to support basic functions like

packet forwarding, routing and network management, in ad hoc networks those functions are carried out by all available nodes. This very difference is at the core of the security problems that are specific to ad hoc networks. As opposed to dedicated nodes of a classical network, the nodes of an ad hoc network cannot be trusted for the correct execution of critical network functions. Several routing protocols for ad hoc networks have been developed to produce a secure environment between the nodes. In our schemes we can apply this in most of the kinds of the routing protocol, and we choose the AD hoc on-demand Distance Vector (AODV) since it is the most popular of the routing protocols and used widely [4]; we focus in this paper on the authentication between the nodes, because the networks accessible by authorized nodes in Mobile ad hoc networks, for short MANET, have become a very important research area over the last past years. The structure of a MANET consists from mobile nodes which can act as a sender, and a forwarder which is used for messages. Our accent will be put on the unique feature of these protocols, feature which is represented by the ability to trace routes in spite of dynamic topology. The attacks which can exist on ad-hoc network can be passive attacks and active attacks.

In this paper, security in mobile ad-hoc networks represents a fold problem. The first problem is represented by the security of the routing protocols which enable the nodes to communicate with each other and the second problem refers to protection of the data which is traveling through the network on routes established by the routing protocols.

The paper has been organized in sections. Section I is the introduction, section II we make a security analysis of the most attacks of routing protocols; section III contains a review of AODV; section IV speaks about our scheme: the main idea of this scheme is to use the hash function, hash lock, secret value and time stamp to increase the authentication between the nodes when they start communicating in the ad hoc network.

2 Security Analyses

The structure of ad hoc networks make them very vulnerable to many types of attacks such as passive eavesdropping, active interfering, impersonate, black hole, data tampering and one of the most important attack on which is very difficult to create a security solution, denial of service. The detection of compromised mobiles in a large scale ad hoc network is threatened by [5]:

- The mobiles are constantly interpreted as nodes;
- The protocols implemented are cooperative in nature;
- There is lack of fixed infrastructure and a central concentration point in the place where intrusion detection system is able to collect audit data;
- There is no distinction made between a normal node and an anomaly which can be found in wireless networks.
- Because MANET is an open environment, all nodes are able to access data from the communication range.

The idea of making AODV secure represents a real challenge, because first of all we need to understand security attributes and mechanisms. Security is viewed as a structure composed from mixture of processes, procedures and systems. All this components ensure confidentiality, authentication, availability, integrity, access control and non-repudiation [6]. The triad CIA (*confidentiality, integrity and authentication*) which can be applied in our solution means:

- *Confidentiality* is obtained (should be obtained) by preventing the unauthorized nodes to access data.
- *Authentication* is used to ensure the identity of source as well as neighbor nodes to prevent a node from having an unauthorized access to resources and confidential information, as well as to stop it from having interfering operations with the rest of the nodes.
- *Integrity* is very important because it helps to prevent malicious nodes from altering data and resending it.
- Regarding the *repudiation* [6] [7], the node sends a message and it cannot deny that the message was sent by it.

In the following, we will go through different types of attacks, illustrating how they act. We mention that, some of the attacks have been presented in real life, and we were able to see the experiments and how the components react at those attacks.

1. Impersonation.

In these kinds of attacks the attacker is able to join the network by spoofing as an innocent node, that's why these attacks are also called spoofing attacks. The network is overtaken by several such nodes which will conduct malicious behavior such as obstructing proper routing by injecting false routing packets into the network or by modifying routing information.

2. Eavesdropping.

It is a passive attack that takes place by snooping on transmitted data on a legitimate network with the goal of collecting information, such as topology of the network, geographical location or optimal routes in the network. The attack is very hard to detect because it takes place while the data is being transmitted from one node to another.

3. Wormhole attack.

It is one of the most sophisticated and severe attacks in MANETs. The attacker connects two parts (which can be found at a specified distance) of the network and after this he tunnels the messages received in one part of the network to the other.

4. Black hole attack.

The striker lures the traffic of the network in such a way that it compromises the node and forms a black hole, putting the opponent at the centre [8]. In this attack, malicious nodes trick all their neighboring nodes to attract all routing packets to them.

In MANET there is no fixed network topology [9] [10]. The nodes in this kind of network need to discover themselves. In MANET routing protocols, nodes need to announce their presence to other fellow nodes as well as to know the presence of their next nodes in the network.

3 Review on Ad Hoc on Demand Distance Vector (AODV)

AODV is a reactive unicast routing protocol for mobile ad hoc network. AODV only needs to maintain the routing information about the active path [6] [4].

This protocol can be called a pure on-demand route acquisition system; nodes that do not lie on active paths, neither maintain any routing information nor participate in any periodic routing table exchanges. Further, a node does not have to discover and maintain a route to another node until the two must communicate, unless the former node is offering its services as an intermediate forwarding station to maintain connectivity between two other nodes.

The Ad Hoc O-Demand distance Vector (AODV) has two phases. Now we give a brief explanation of these phases:

- Route discovery

In this phase, when the node needs to determine the route to the destination node, it floods the network with a RREQ message. If the destination is not in range of the source node, it broadcasts a RREQ message to the destination by its neighboring nodes.

The intermediate nodes store reverse routes back to the original node. Each destination point maintains a single sequence number increased monotonically, which serves as a logical time at that node. The sequence numbers are used to ensure that routes are only updated with newer ones. When a destination node receiving the request, the node generates a RREP and sends it through the intermediate nodes.

- Route Maintenance

This phase is carried out whenever there is a broken link between nodes. A node can detect the broken link by monitoring the link actively or the promiscuous node passively [11]. When a break occurs along an active path, the node upstream of the break (i.e., closer to the source node) invalidates the routes to each of the destinations in its routing table and then generates a route error (RERR) message.

4 Proposed Methods for Using RFID Protocol in Mobile

The proposed method uses hash functions but digital signatures can also be used.

To have a strong security, the scheme of SAODV (Security AODV) represents an extension of AODV in which the digital signatures hash chain mechanisms are used to accomplish the security on different levels, by incorporating a digital signature in each node for authentication and integrity in routing messages like RREQ, RREP and RRER. The signature is checked by the neighbor nodes which receive the message.

4.1 Enhancement Authentication of AODV protocol (EA-AODV) to Resist the Attack

Access control mechanisms are frequently based on public key cryptographic primitives or symmetric key primitives requiring secure key distribution. In our work

the hash locks are simple access control mechanism based on one-way hash function. Every node in the hash lock scheme will be equipped with a hash function. This scheme is appropriate for secret keys or both public/private keys.

Now every node has the key and can compute the hash value of the key, the hash output is desired as the metaID of the node, for example: node A computes the hash function for its key then the metaID for node A is $H(K_a)$. The node will store the metaID, and every node stores its metaID and also the all metaIDs for all nodes in the MANET (mobile ad hoc network).

If node A wants to send packet to node G, first we use this scheme to be sure that node G is the authenticated node (destination node). We can summarize this scheme by these steps (see figure 1):

- Node A sends query to the destination node (node G) this query: who are you?;
- Node G sends his metaID to node A when he receives the query;
- When node A receives the metaID of node G it starts search in his metaIDs storage which one has the same value;
- Node A sends to node G his key. With this step the source node (A) assures that the node G belongs to the network and node G also authenticates of node A.

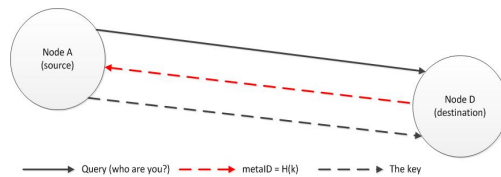


Fig. 1. Hash lock scheme

In this scheme the attacker can track the node belong to any network and can send many inquiries to the node and the node will respond to these queries with the same value (as in Figure 2) and because the attacker has the same value in each response he can track the key or know the hash function which will make it weak and vulnerable to many attacks such as impersonation attack, eavesdropping attack, black-hole and wormhole attacks.

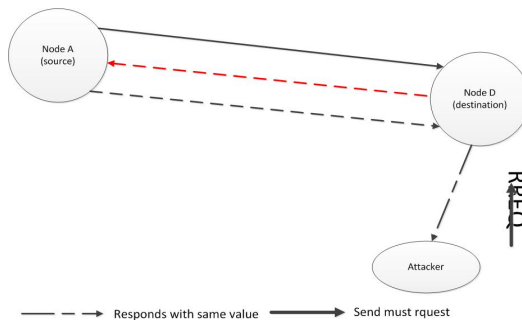


Fig. 2. Attack on the hash lock authentication scheme

So to improve previous scheme we use random number generation in this scheme to prevent the attackers tracking the routing protocols.

4.2 Applying Random Number and Hash Function in Our Scheme for Authentication in AODV

Every node here stores all the IDs of the nodes in the mobile ad hoc network.

We present a practical heuristic based one-way hash function. We also offer a theoretically stronger base on pseudo-random function (PRF). The scheme proposes an improvement protocol to fix the flaws from the last scheme. The improvement protocol is secure with merits of privacy protection, resisting counterfeit attack and obtaining mutual authentication.

As in hash lock scheme the node are equipped with one-way hash function, but now also have the random number generator. If two nodes want to communicate in the mobile ad hoc network we can apply this scheme to authenticate between them and we explain this scheme by the following steps as shows in figure 3.

Step 1: The source node here generates a random number R_s and sends it with the query to the destination node with its identification ID_s by the routing discovery phase.

Step 2: When the destination node receives the message from the source node, it computes $h(ID_s \oplus R_s) \oplus h(ID_d)$ and sends it to the source node, directly if it is in the range of the source node or by neighbor nodes if it is not in the range of source node.

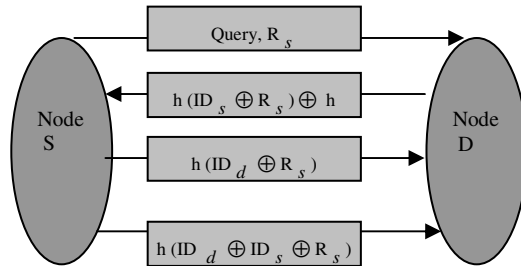


Fig. 3. Random number with hash function scheme

Step 3: When the source node receives $h(ID_s \oplus R_s) \oplus h(ID_d)$ value, the source node computes hash function of the $ID_d \rightarrow h(ID_d)$ and compares this value with the result of $h(ID_d) = \{ h(ID_s \oplus R_s) \oplus h(ID_d) \} \oplus h(ID_s \oplus R_s)$, if the two values are equal then the verification succeed. After that the source node computes $h(ID_d \oplus R_s)$ and sends it to the destination node.



Step 4: After the destination node receive $h(ID_d \oplus R_s)$ and if the verification holds then the authentication of the source node is done. In final, the destination node computes $h(ID_d \oplus ID_s \oplus R_s)$ and when the source receives this message and the verification process holds then the source node authenticates the destination node.

In this scheme we use the exclusive XoR operation and hash function and this makes the computation cost quite low. The scheme is secure and has capability to protect the privacy. Here the destination and source nodes change the values $h(ID_s \oplus R_s) \oplus h(ID_d)$, $h(ID_d \oplus R_s)$ and $h(ID_d \oplus ID_s \oplus R_s)$ on each authentication process, which makes it hard for the adversary to find this values and then trace the source and destination nodes by eavesdropping messages.

After that, if the nodes in the ad hoc networks want to send or receive packets they will use hash function and digital signature in every routing request (RREQ) and routing replay (RREP).

4.3 Using Timestamp with Secret Value to Protect the Privacy

This scheme proposes the use of random numbers with timestamp in the nodes to protect the privacy of the nodes. Random numbers are protected with secret values, and the use of hash functions to prevent forgery and copying. Also we use the time frame to provide a solution to prevent the replay attacks, and the protection of variable values makes it possible to solve synchronization problems.

In this scheme we still use the same parameters like in the last two schemes but also we add new parameters: Ts: time stamp; SVn: n-th secret value, where $n=1, 2, 3, \dots$

The proposed scheme consists of four stages and provides security through timestamp and random numbers. Also providing anonymity can protect the privacy of nodes. We use also the secret value (SV), which is the exchange of messages in a secure way between any source and destination node in a network. This exchange can be made in two ways:

1. Secret Value (SV) Distribution with Confidentiality and Authentication:

We can use the public key to exchange the secret value to provide protection against both active and passive attacks. We begin at a point when it is assumed that node S and node D have exchanged public keys.

2. Diffie-Hellman Key Exchange:

The Diffie-Hellman algorithm efficiency depends on the hardness of computing discrete logarithms.

The proposed scheme gives an improvement for the authentication between the nodes on the insecure channel in the network.

This scheme is based on the challenge response method using a static identifier and one way randomized hash function. Furthermore our scheme uses increasing timestamp to make the response more identifiable and anonymous. We can explain this stage as can be seen in figure 4.

Step 1: When any node in the Ad Hoc Network wants to communicate with other node in the same network or in another one as shown in figure 4 which explains the

authentication process between the source (S) and the destination (D) node before they start send and receive the important data. The source node generates a time stamp T_s and sends the timestamp with query to the destination node by the routing discovery phase. The timestamp of the source must be more than the value of the timestamp the destination node ($T_d < T_s$) has.

Step 2: When the destination receives the last message from the source node and makes sure the value $T_d < T_s$, then the destination node generates random number R_d after that the destination computes $M_d = h(ID_d) \oplus SV \oplus h(SV \parallel R_d \parallel Ts)$ and sends it to with R_d to the source node.

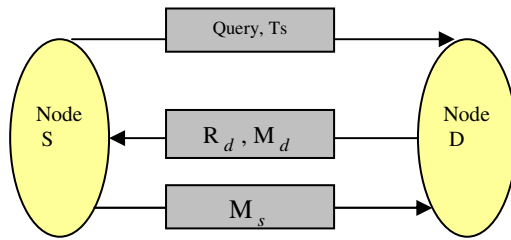


Fig. 4. Timestamp with secret value authentication

Step 3: When the source node receives values by the routing request from the destination node by using the same routing path if it still available or another routing path if not available. The source node starts to make the verification process by computing first this value $h(SV \parallel R_d \parallel Ts)$ and after that computing $h(ID_d) = M_d \oplus SV \oplus h(SV \parallel R_d \parallel Ts)$, after that the source node checks if $h(ID_d)$ belongs to the destination or not. In final the source node computes $M_s = h(ID_d \parallel SV \parallel R_d \parallel Ts)$ value and sends it to the destination node.

Step 4: When the destination received the last value from the source node, it starts the verification process by computing the same value $h(ID_d \parallel SV \parallel R_d \parallel Ts)$ if it matches the authentication process succeeded and the timestamp $T_d \leftarrow T_s$ is updated.

Hash chain is used to verify the integrity of the hop count field of RREQ and RREP messages by allowing each node that receives the message to verify that the hop count has not been modified by malicious nodes. Hash chain consists of applying repeatedly a one-way hash function for a number of seeds.

Here we still use Destination Sequence Number (DSN) as in AODV to ensure that all routes are loop free and routing information is proper and valid. During the process of forwarding, the RREQ packets, the intermediate nodes record the address of the neighbor from whom the first copy of broadcast message is received in their routing tables. This helps to establish a reverse path [7] [12].

AODV uses the hop count parameter to determine the shortest path between two nodes. A malicious node can set false hop counts and wrong values of the sequence number. This leads to redirection of network traffic or to a DoS attack.

When a path is not available for the destination, a route request packet is flooded along the network. The RREQ contains the following fields: source address, request ID, source sequence number, destination address, destination sequence number, hop-count. The request ID is increased every time the source node sends new RREQ, so the pair (ID request, source address) defines a unique RREQ. On receiving RREQ message each node checks the request ID and the source address. If the node has already received a RREQ with the same pair of parameters the new RREQ packet will be ignored. Otherwise the RREQ will be forwarded (broadcast) or replied (unicast) with a RREP message:

- If the node doesn't have a route to the destination or it has one that is not updated, the RREQ will be re-broadcasted with increased hop-count;
- If the node has a path with a sequence number greater than or equal to that of RREQ, RREP message will be generated and sent back to the source.

The number of RREQ messages that a node can send per second is limited.

AODV has an optimization using an expanding ring (ESR) technique for the flooding RREQ messages. Every RREQ has a time to live (TTL) value that specifies the number of times it should return this broadcasted message. This value is set to a predetermined value in the first transition and increased at retransmissions, which appear when there is no reply. Such flooding used TTL large enough - larger than the diameter of the network - to reach all nodes in the network, so as to ensure the successful discovery of the path in only one round of flooding. However, this time delay approach causes high flow of unnecessary broadcast messages. This can be solved with an optimally chosen set of TTLs.

4.4 Operation on AODV

Every time a node initiates a RREQ or a RREP message, it performs the following operations [5]:

- Creates random number called seed;
- Sets the Max Hop Count field to the Time To Live value (from the IP header);
- Where Max Hop Count = Time to live;
- Seed value is used to sets the Hash field;

Hash = seed

- Sets the Hash Function field to determine the hash function that it intends to use.

In our scheme we still use these parameters, which come from "Secure Ad hoc On-Demand Distance Vector (SAODV)" proposed by Manel Guerrero Zapata [5]. We still calculate Top Hash by hashing seed Max Hop Count times:

Top Hash = $h(\text{Max Hop Count})(\text{seed})$

Where:

- H is a hash function.
- $H_i(x)$ is the result of application the hash function, h, Xi times.

Each time a node receives RREQ or RREP, it performs the following operations to verify the hop count:

Applies the hash function h *Max Hop Count* minus *Hop Count* times to the value in the *Hash* field, if the result is equal to the value contained in the Top Hash field, the verification has been done:

$$\text{Top Hash} = h^{\text{Max Hop Count} - \text{Hop Count}} (\text{Hash})$$

Here we can apply this operation with our schemes by use concatenated between them:

$h(\text{ID}_s \oplus \text{R}_s) \oplus h(\text{ID}_d) \parallel \text{Top Hash}; [\text{Top Hash} = h(\text{Max Hop Count})]$ (from destination to source)

Or when we use random number with secret value and timestamp scheme it will be:

$\text{R}_d, \text{M}_d \parallel \text{Top Hash}; [\text{Top Hash} = h(\text{Max Hop Count})]$ (from destination to source)

And

$h(\text{ID}_d \oplus \text{R}_s) \parallel \text{Top Hash}; [\text{Top Hash} = h(\text{Max Hop Count})]$ (from Source to destination).

Or when we use random number with secret value and timestamp scheme it will be:

$\text{M}_s \parallel \text{Top Hash}; [\text{Top Hash} = h(\text{Max Hop Count})]$ (from Source to destination)

A node applies the hash function to the Hash value in the Signature Extension to account for the new hop $\text{Hash} = h(\text{Hash})$, before re-broadcasting a RREQ or forwarding a RREP.

The value of the hash function refers to the hash function used to calculate the hash. Hash Function, Max Hop Count, Top Hash, and Hash fields are transmitted in the Signature Extension. After hash schemes we use digital signatures [5] [12] to protect the integrity of the non-mutable data in the RREQ and RREP.

When the node is receiving a RREQ message, it first verifies the signature before creating or updating a reverse route to the source of the RREQ. If the RREQ was received with a Double Signature Extension, then the node will also store the signature for the RREP and the lifetime (which is the 'reverse route lifetime' value) in the route entry. An intermediate node will reply to a RREQ with a RREP only if it fulfills the AODV's requirements and the node has the corresponding signature and old lifetime to put into the Signature and Old Lifetime fields of the RREP Double Signature Extension. Otherwise, it will rebroadcast the RREQ as it has no cached route. When the destination receives a RREQ, it will reply with a RREP with a Single Signature Extension. When a node receives a RREP, it first verifies the signature before creating or updating a route to that host. If the signature verification is successful, it will store the route with the signature of the RREP and the lifetime. Otherwise the RREP is discarded [13] [14].

5 Conclusion

In this paper, we present the most important attacks on the routing protocols and how to resist to these attacks in a secure way. Also we propose a set of schemes to increase the security between the nodes by enhancing and improving the authentication and confidentiality between the nodes. The proposed idea uses hash functions but also digital signatures can be used. The first method proposed by us uses crypto-based identifiers, for short CBID; after that we used the hash function and random hash function. Our solution expands the security scope and provides more authentication service between the nodes in MANET.

References

1. Subharthi, P., Pan, J., Jain, R.: Architectures for the future networks and the next generation Internet. *Computer Communications Journal* (2010)
2. Stoleru, R., Wu, H., Chenji, H.: Secure Neighbor Discovery in Mobile Ad Hoc Networks. In: Eighth IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (2011)
3. Haboub, R., Ouzzif, M.: Secure and Reliable Routing in Mobile Ad Hoc Networks. *International Journal of Computer Science & Engineering Survey (IJCSES)* 3 (2011)
4. Narra, H., Cheng, Y., Çetinkaya, E.K., Rohrer, J.P., Sterbenz, J.P.: Destination-sequenced distance vector (DSDV) routing protocol implementation in ns-3. In: 4th International ICST Conference on Simulation Tools and Techniques (2011) ISBN: 978-1-936968-00-8
5. Secure Ad hoc On-Demand Distan Mobile Networks Laboratory Nokia Research Center FIN-00045 NOKIA GROUP, Finland ce Vector Routing Manel Guerrero Zapata manel.guerrero-zapata@nokia.com
6. Mala, C.R., Shetty, S., Padmashree, S., Elevarasi, E.: Wireless Ad hoc Mobile Networks. In: National Conference on Computing Communication and Technology, pp. 168–174 (2010)
7. Das, A.R., Perkins, C.E., Royer, E.M.: Performance Comparison of Two On-demand Routing Protocols for Ad Hoc Networks (2012)
8. Mistry, N., Jinwala, D.C.: Improving AODV Protocol against Blackhole Attacks. In: IMECS 2010, Hong Kong (2010)
9. Deswal, S., Singh, S.: Implementation of Routing Security Aspects in AODV. *International Journal of Computer Theory and Engineering* 2 (2010)
10. Montero-Castillo, J., Palomar, E.: Cooperation in Ad Hoc Network Security Services. In: The Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (2011)
11. Zahedi, K., Samad Ismail, A.: Route Maintenance Approach for Link Breakage Prediction in Mobile Ad Hoc Networks. *International Journal of Advanced Computer Science and Applications* 2 (2011)
12. Kumar, D., Ojha, D.B., Kumar, A.: Securing MANETs by Q Routing Protocol. *International Journal of Engineering Research and Applications* 2(6) (2012)
13. Karlsson, Dooley, Pulkkis: Routing Security in Mobile Ad-hoc Networks. *Issues in Informing Science and Information Technology* 9 (2012)
14. Panaousis, E.A., Ramrekha, T.A., Millar, G.P., Politis, C.: Adaptive and Secure Routing Protocol for Emergency Mobile Ad Hoc Network. *International Journal of Wireless and Mobile Network* 2 (2010)

Evaluation of Open Source Server-Side XSS Protection Solutions

Jonas Ceponis¹, Lina Ceponiene², Algimantas Venckauskas¹, and Dainius Mockus¹

¹ Kaunas University of Technology, Computer Department, Studentu str. 50,
LT-51368, Kaunas, Lithuania

² Kaunas University of Technology, Information System Department, Studentu str. 50,
LT-51368, Kaunas, Lithuania

{jonas.ceponis,lina.ceponiene,algimantas.venckauskas}@ktu.lt,
dainius.mockus1987@gmail.com

Abstract. Web protection against XSS attacks is an indispensable tool for implementing reliable online systems. XSS attacks can be used for various malicious actions and stealing important information. Protection may be implemented both on user computer and on server side. In this work we have analyzed the server side protection solutions. These solutions must ensure appropriate level of security and at the same time should not considerably increase page response time. The aim of this paper is to determine the most effective and safe free tools for protection against XSS attacks for web pages created using PHP, ASP.NET and Java technologies.

Keywords: XSS attacks, server-side protection, response time.

1 Introduction

Nowadays a lot of personal information is stored online – in various internet portals or using internet storage services. More and more financial and electronic commerce operations are performed using internet. These circumstances have caused an increased amount of crime attacks in internet. One of the most popular type of threats in web applications is cross site scripting – XSS [7, 9, 24]. Developers of internet servers, web pages and browsers integrate various defense techniques for protection against cross site scripting attacks. In most cases these defense techniques check input data and usually take some time to perform this operation [11]. Consequently protecting against cross site scripting increases internet page response time [17]. In our work we have analyzed available free solutions created using PHP, Java and ASP.NET technologies for protecting websites against XSS attacks and impact of analyzed solutions on page response time.

The rest of the paper is organized as follows. Section 2 presents types of XSS attacks. Section 3 talks about possible and existing solutions for protection from XSS attacks. In section 4 testing of free PHP, ASP.NET, and Java tools for websites are presented. Section 5 ends the paper with conclusions.

2 XSS Attacks

There exists a multitude of various threats for information systems: buffer errors, cross-site scripting (XSS), various authentication issues, inappropriate management of permissions, privileges and access control, cross-site request forgery (CSRF), cryptographic issues, code and SQL injection [10], format string and numeric errors, race conditions, etc. [19]. In addition to well-known vulnerabilities listed above, new types, such as HTTP parameter pollution are discovered [2].

XSS attacks are usually implemented using JavaScript programming language. HTML, Flash or any other type of programming language that can be executed in the user browser can also be used for implementing XSS attack [1]. XSS is a special attack type as it is pointed against users, not servers. The attackers usually choose a reliable website; therefore such websites should be more carefully protected from XSS-type attacks. For the first time this type of attack has been discovered in 2000 and since then their number has experienced a significant growth [23]. According to [23] XSS vulnerabilities represent more than 10 percent of all detected vulnerabilities starting from 2008.

XSS attack using malicious code insertion into web application becomes possible when:

- data is entered into the system from unreliable sources (unauthorized users);
- data is entered by one user and then the page content is displayed to the other users, without the proper scanning of input data. Such situation can occur in comments, forums, etc.

XSS attack bypasses the security mechanisms that are integrated in modern browsers; can extract important data and use them in a variety of actions on behalf of another person; can monitor and send user actions (pressed keys, etc.) to the attacker. XSS attacks can be divided into 3 groups: non-persistent, persistent (caused by server-side code) and DOM-based (caused by client-side code).

Non-persistent XSS attacks are usually used in dynamic web pages [28]. Attacks of this type are performed using user data input forms, pop-ups or fake pages where users are requested or persuaded to enter various personal data (often passwords or other sensitive information). With the help of social engineering and proper use of human factors users can be forced to click on the malicious link and enter required information.

Persistent code insertion is a type of attack which uses the malicious code entered into web page input fields and stored in database [36]. When another user views a web page, stored malicious code is sent to the user browser and executed. This type of attack becomes possible if website does not properly check input data. Persistent or stored XSS vulnerabilities can be used for a very effective multiple attack type, since the number of attacks depends on the number of page hits.

DOM-based XSS vulnerability is based on the Document Object Model [15]. This type of attack typically changes web page view in the browser. This vulnerability occurs in the user computer, not on the server side. Since web browser has a number

of permissions in user computer operating system, it is possible to run other processes with the same permissions as browser process during this kind of attack; therefore these attacks can do a lot of damage. The consequences can be even worse, if the user is logged on as the administrator.

Increasing amount of information and operation types in internet raised the number of XSS attacks. XSS attack is performed when browser opens the site on the user side – at this moment a malicious code insertion into the site is attempted. XSS attacks are usually performed in websites which use applications with forms for data input from the user side. The use of applications with data input features in websites makes them more dynamic and gives users more freedom to a wide variety of actions. Therefore it is unreasonable to protect from XSS attacks just by avoiding application usage. Instead various protection tools should be used.

3 Protection from XSS Attacks

Currently almost every website has dynamic content part where user can write comments, interact with other users or just perform a search. All these features enable malicious users to execute XSS attacks. Website programmers should take precautionary measures to protect web users from cross site scripting attacks. There are three ways of protecting against XSS attacks categorized by the location of implementation: server side protection, user computer (browser) protection, and hybrid protection.

Software firewall is usually installed for implementing server-side protection from XSS attacks [8]. Firewall checks all communication between the user computer and the server. Program-level firewall can inspect queries from user computer to server, check the server responses to user requests, or can do both of these steps at a time. The most effective protection is guaranteed by testing incoming and outgoing information flow, but it can considerably slow down the server and increase page response time.

Protection against code insertion attacks can be implemented in the user computer [6, 31]. This protection is implemented by installing some software tools. Such tools are JavaScript code verification component in internet browser or intrusion detection software – IDS [3, 4, 21]. As well as server-based protection, user side protection can be implemented using program-level firewall. But in such case computer user needs to install and configure the protection system.

Hybrid XSS protection is implemented both on server and on user computer side [22]. Server side software is responsible for the initial content inspection and safety rule creation. Protection part in user side computer responds to the set of safety rules and makes final inspection of the content.

Regardless of protection type, the site must be protected according to several key criteria. First, in order to protect the site, complete input data verification is always required. Before using user-entered data anywhere and anytime, protection mechanism must verify whether they meet the data type and structure requirements. This operation must be carried out in the user browser, before transmitting the data to server; and on the server side, before storing data in database or otherwise using it. The

entered value must be checked to make sure it fits in the range of possible values. The user entered data must also be cleared from potential code insertion options; compatibility with the set of security policy rules must be ensured. User uploaded files must be checked for the size and format limitations. In order to ensure that user cannot insert and execute code in the website, it is necessary to filter out invalid characters that can change HTML code of the page [12, 25].

XSS vulnerability occurs due to lack of filtering of user-entered information. When malicious user finds a way to insert code, he can access the site content, session cookies, or other information. Almost all XSS attack opportunities originate from the fail of checking HTML input and output. Attention should be especially paid to HTML tags (< and >) and other special characters. No matter how simple it sounds, almost every detected XSS vulnerable site is characterized by inability to remove the brackets from the input, or the inability to encode these brackets in the output. The purpose of XSS attack is usually a user data theft, in order to gain access to confidential information; access to paid information for free; tracking users visited pages; changing the users browser settings, etc. It is therefore very important to ensure effective website security.

Currently a variety of tools and methods for web content protection from XSS attacks is used [5, 16, 20, 30, 32, 33, 37]. To fully protect against cross site scripting attacks, user can simply turn off the program code execution on the user side. But in this case functionality of website can be limited and this is inconvenient for the user. Developers of web browsers increasingly strengthen protection integrated into application and restrict suspicious page code execution. Computer users also are free to install a wide range of plug-ins that can enhance the security. Browsers are used according to user individual preferences and needs, therefore, convincing the person to choose more secure browser is almost impossible. Moreover, all modern browsers already have at least a minimum of protection against code insertion attacks.

But efforts to prevent code insertion attacks only from user side are not enough. The main and the biggest security problems arise on the server side; therefore the main focus should be on site developers. Each dynamic content site must use comprehensive filtering of user entered data, as well as verification of information sent to the user. Sometimes it is even necessary to use the content encryption. Web developers must choose the appropriate security levels and methods to assess the full protection from the risks and dangers on the server side. This is not an easy task, and it can be implemented using various ways: from basic programming language commands (such as commands in PHP language: `htmlspecialchars()`, `strip_tags()` or `utf8_decode()`) to server side API for web content encryption or encoding libraries. Currently many such libraries are developed and used. They are implemented using different programming languages and can be either proprietary or open source.

The main objective of our work is to determine which of free server-side content filtering libraries are enough for protection and at the same time do not significantly affect website response time. For that purpose we analyzed several free protection solutions for each of three different platforms – ASP.NET, Java, and PHP.

4 Case Study Experiments and Discussion

In order to perform testing of XSS protection tools we created three test websites. Websites were developed using different platforms and run on localhost:

1. ASP.NET (.NET Framework 4) server and C # programming language. For data storage site uses MS SQL Server 2008 Express database.
2. Java programming language (Spring Framework 3.0.2). For data storage site uses MySQL 5.5.24 database.
3. Apache 2.2.22 webserver and PHP 5.3.13 programming language. For data storage site uses a MySQL 5.5.24 database.

Our test websites have the form for entering comments and the list of already entered comments. Comment forms are found in nearly every dynamic webpage and are available to everyone, so often become the target of malicious attacks. Comments are stored in database. Site response times were recorded using developer tools integrated in Google Chrome browser (ver. 24.0.1312.57). Computer used for testing had Intel Core i7-2600 3.4 GHz CPU, 8GB RAM and 1TB SATA3 HDD. Computer had Windows 7 Professional 64-bit operating system installed with additional software for website support mentioned above.

Initial site testing was performed without any protection. Afterwards the sites were tested with integration of different security tools. Response time was recorded during each of these tests. Two types of response time were separated: viewing comments response time and saving comments response time. Viewing comments response time was measured during retrieval of comments from the database and processing them before showing on the user screen. Saving comments response time was measured during saving a newly entered comment with filtering of the data using selected security measures.

The measurements were performed sequentially, one after another. During testing the computer did not perform any other tasks and measurement results were not influenced by any external factors. Before each iteration the browser cache was cleared. Measurements were carried out under the same conditions; the same number of entries were displayed and stored during all experiments.

Response time cannot be the only one measure of protection – a simple library that performs only the minimum security checks might work quickly, but would not ensure the required security level. Just a few checks in such minimalistic library cannot ensure that the site is protected from all types of attacks. One can easily predict that website with no additional security library or with only simple programming language commands for protection will work fastest. But it is absolutely necessary to evaluate reliability of each library in terms of security. In our work the straightforward way to check the reliability of the protection library was used – we reviewed the functionality and decided which type of the attacks the tool is able to protect from. For this purpose we evaluated selected protection tools by such criteria:

- Character encoding filter. All web application data entered by any user must be examined for the encoding character set. For example, a hexadecimal encoding can

be used to encode malicious URL. Inconsistent data must be removed or properly encoded.

- HTML code formatting. HTML code of web application must be cleaned and properly formatted, e.g. missing or mismatched end tags must be detected and corrected.
- Attribute filtering. Many filtering tools adequately sanitize tags in entered data, but that is not enough. Attributes must be also checked for the presence of potentially dangerous commands. For example, malicious expressions could be inserted into a style attribute in an anchor tag; when such page is rendered, dangerous code is executed.
- Code insertion scanning. Code insertion into web application data is the perfect means for XSS attack. It is therefore necessary to check whether the user entered data does not contain any code.

In our work we consider that the protection tool has an appropriate level of security only if it meets all the criteria mentioned above.

4.1 ASP.NET Website Testing

Websites used for testing were implemented using different technologies; therefore the tools for protecting each website are also different. For protecting ASP.NET (C#) website we have analyzed the following open-source tools:

- ASP.NET command (`HttpUtility.HtmlEncode()`).
- Microsoft Anti-Cross Site Scripting Library (AntiXSS) – Microsoft library for HTML code filtering and invalid characters detection and removal [20].
- OWASP AntiSamy - OWASP library for code filtering [26].
- TidyManaged - HTML/XHTML/XML markup parser & cleaner based on Tidy library [34].

Analysis of these tools was based on four main criteria which we consider ensuring the proper level of protection from XSS attacks (character encoding filter, HTML code formatting, attribute filtering, code insertion scanning). Analysis results are presented in Table 1. The tool which meets all four functionality criteria is considered secure from XSS attacks.

Table 1. Security libraries for ASP.NET functionality rating

Functionality / Library	Character encoding filter	HTML code formatting	Attribute filtering	Code insertion scanning
ASP.NET command	Partially	No	No	Partially
AntiXSS	Yes	Yes	Yes	Yes
OWASP AntiSamy	Yes	Yes	Yes	Yes
TidyManaged	Yes	No	Yes	Yes

Afterwards the protection tools were tested on our experimental website. The first test was performed without any protection. Then each tool was tested and response times were recorded. Results of the response time testing for ASP.NET website are presented in Fig. 1.

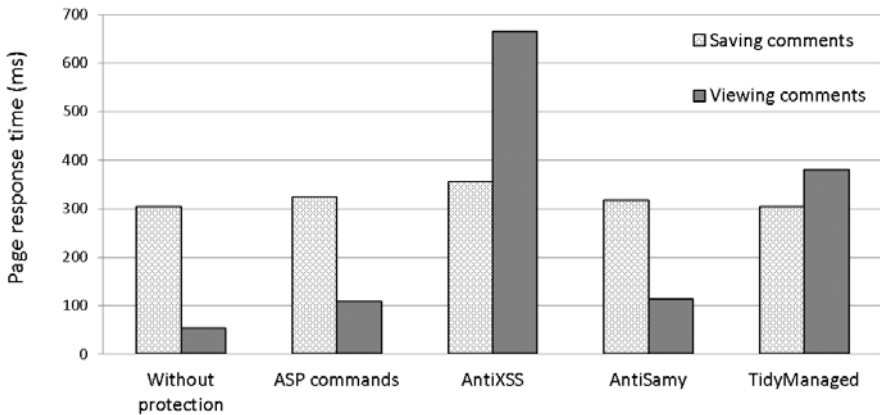


Fig. 1. Site response time measurement results for ASP.NET (C#) website

The longest response time was acquired using AntiXSS tool, and the shortest, as expected, was acquired using ASP.NET command. But ASP.NET command (`HttpUtility.HtmlEncode()`) cannot ensure the required security level. In the analysis of protection functionality (table 1) we found that the AntiXSS and OWASP AntiSamy source filter library can perform the same functions, but OWASP AntiSamy response time (Fig. 1) is much better than AntiXSS.

Consequently the best security library according to functionality and response time criteria is OWASP AntiSamy. OWASP AntiSamy library provides a reasonable level of security and only slightly increases page response time.

4.2 Java Website Testing

The tools for securing Java based website from XSS attacks were also analyzed in our work. We have included the following free tools in our analysis:

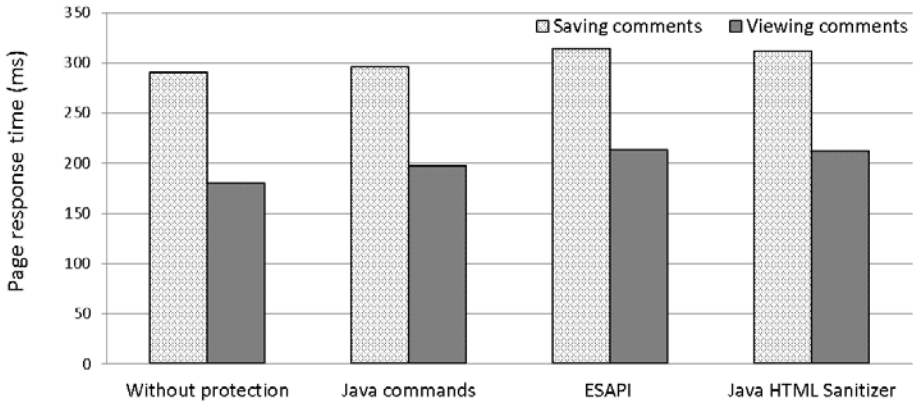
- Java commands (`escapeXml()`, `replaceAll()`).
- ESAPI - web application security control library to write lower-risk applications [35].
- OWASP Java HTML Sanitizer - HTML Sanitizer written in Java for protecting web application against XSS [27].

Although different technologies are used for implementing protection tools, the main criteria for ensuring security and protection against XSS attacks remain the same for all the tools. In our work the same criteria are used for comparing the security functionality of ASP.NET, Java and PHP tools. Analysis results for Java based security tools are presented in Table 2.

Table 2. Security libraries for Java functionality rating

Functionality / Library	Character encoding filter	HTML code formatting	Attribute filtering	Code insertion scanning
Java command	Partially	No	No	Partially
ESAPI	Yes	No	Yes	Yes
Java HTML Sanitizer	Yes	Yes	Yes	Yes

Response time testing for all analyzed security tools was also performed. Results for response time testing in Java website are presented in Fig. 2.

**Fig. 2.** Site response time measurement results for Java website

Test results show that both Java commands and the two analyzed security libraries increased response times only slightly in comparison with response times of website without any protection.

From the Table 3 we can see that OWASP Java HTML Sanitizer ensures the best level of security. This library only slightly increased the page response time. According to the obtained results it can be stated that OWASP Java HTML Sanitizer is the excellent free tool for protection from XSS attacks for Java based websites.

4.3 PHP Website Testing

The tools for protecting PHP based website against XSS attacks were also analyzed in our work. These tools are:

- PHP commands (`htmlspecialchars()`, `strip_tags()`, `utf8_decode()`).
- HTMLPurifier – library for filtering HTML code and removing invalid characters. This tool is able to correct the HTML code, elementary errors [13].
- SafeHTMLChecker – HTML code filtering library [29].
- htmLawed – HTML code filtering library. It has plugins for popular content management systems and is widely used there [14].

- Kses – PHP HTML/XHTML code filter. It removes unwanted elements and attributes from the test block [18].

All these protection tools were analyzed based on our selected four main security criteria. The security criteria are the same as used in Java and ASP.NET tools analysis. Analysis results are presented in Table 3.

Table 3. Functionality of security libraries for PHP

Functionality / Library	Character encoding filter	HTML code formatting	Attribute filtering	Code insertion scanning
PHP commands	Partially	No	No	Partially
HTMLPurifier	Yes	Yes	Yes	Yes
SafeHTMLChecker	No	No	Yes	Yes
htmlLowed	Partially	No	Yes	Yes
Kses	Partially	No	Partially	Yes

Response time testing experiment was performed using PHP website without any protection and using website protected with analysed security tools. Results of response time testing in PHP website are presented in Fig. 3.

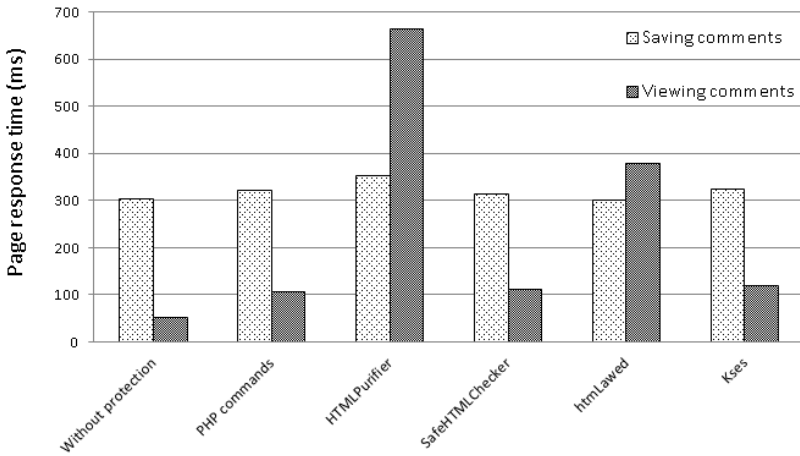


Fig. 3. Site response time measurement results for PHP website

The results show clearly that all security methods and libraries increased web page response times – the shortest page response time is in absence of any protection. Further analysis shows that HTMLPurifier library was the slowest tool and increased response time most of all. PHP commands is the least time consuming solution. SafeHTMLChecker library showed the smallest response time increase among tested libraries.



Analysis of protection capabilities of security libraries shows that the SafeHTMLChecker library (which had best response time in our test) cannot guarantee safety of HTML code – it does not perform verification of HTML code and does not check HTML code formatting.

Since the goal of our work is to offer the protection that ensures an adequate level of security for website, the protection library must meet all functionality requirements listed in table 3. As we can see in table 3, there is only one security library which meets all listed requirements – HTMLPurifier. But this library has an important drawback – in site response time testing it showed the longest site response time. In our future work we are planning to optimize HTMLPurifier without compromising its functionality. Our goal is to decrease site response time with HTMLPurifier as much as possible leaving the same protection functions as it has now.

5 Conclusion

New XSS attacks are constantly created and targeted to different sites vulnerabilities. Protection against these attacks is an extremely difficult task. Protection can and should be conducted not only in the user computer but also on server side. A variety of commercial and free tools and methods for protection against XSS attacks are available.

In our work we evaluated performance and security level of available free tools for server-side XSS protection implemented using different technologies. We created experimental websites using PHP, ASP.NET and Java. Websites were tested by measuring site response time. The protection tools were evaluated not only by response times, but also according to several functionality criteria. Simple programming language commands used for protection was the least time consuming solution, but they are unable to provide required security level.

Finally, according to the result of analysis and experimental testing, we have selected the best protection tool for each technology. In ASP.NET based website we recommend using OWASP AntiSamy. In Java website the best results were achieved using OWASP Java HTML Sanitizer. Analysis of security tools for PHP based websites showed controversial results – the only tool which was considered secure enough for protection from XSS attacks also had the longest response time in experimental testing. This tool is HTMLPurifier library. Therefore we have decided to aim our future work towards optimization of HTMLPurifier without compromising its functionality.

References

1. Acker, S., Nikiforakis, N., Desmet, L., Joosen, W., Piessens, F.: FlashOver: automated discovery of cross-site scripting vulnerabilities in rich internet applications. In: Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. ACM (2012)

2. Balduzzi, M., Gimenez, C., Balzarotti, D., Kirda, E.: Automated discovery of parameter pollution vulnerabilities in web applications. In: Proceedings of the 18th Network and Distributed System Security Symposium (2011)
3. Bates, D., Barth, A., Jackson, C.: Regular expressions considered harmful in client-side XSS filters. In: Proceedings of the 19th International Conference on World Wide Web, pp. 91–100. ACM (2010)
4. Brooks, M.: Bypassing Internet Explorer's XSS Filter. Traps of Gold-Defcon (2011)
5. Bugeja, J., Price, G.: A Pragmatic, Policy-Driven Framework for Protection Against Cross-Site Scripting. Royal Holloway Series (2012)
6. Curtsinger, C., Livshits, B., Zorn, B.G., Seifert, C.: ZOZZLE: Fast and Precise In-Browser JavaScript Malware Detection. In: USENIX Security Symposium (2011)
7. FireHost Inc.: Cross-Site Scripting Attacks Up 160% in Final Quarter of 2012 (2013), <http://www.firehost.com/company/newsroom/web-application-attack-report-fourth-quarter-2012>
8. Galan, E., Alcaide, A., Orfila, A., Blasco, J.: A multi-agent scanner to detect stored-XSS vulnerabilities. In: Proceedings of the International Conference for Internet Technology and Secured Transactions, pp. 1–6 (2010)
9. Grossman, J., Hansen, R., Petkov, P.D., Rager, A., Fogie, S.: XSS Attacks: Cross-Site Scripting Exploits and Defense. Syngress (2007)
10. Hidhaya, S.F., Geetha, A.: Intrusion Protection against SQL Injection and Cross Site Scripting Attacks Using a Reverse Proxy. In: Thampi, S.M., Zomaya, A.Y., Strufe, T., Alcaraz Calero, J.M., Thomas, T. (eds.) SNDS 2012. CCIS, vol. 335, pp. 252–263. Springer, Heidelberg (2012)
11. Hooimeijer, P., Livshits, B., Molnar, D., Saxena, P., Veanes, M.: Fast and precise sanitizer analysis with BEK. In: Proceedings of the 20th USENIX Conference on Security (2011)
12. Hope, P., Walthor, B.: Web Security Testing Cookbook: Systematic Techniques to Find Problems Fast. O'Reilly Media, Inc. (2008)
13. HTML Purifier, <http://htmlpurifier.org>
14. htmLawed, http://www.bioinformatics.org/phplabware/internal_utilities/htmLawed/index.php
15. Klein, A.: DOM-based Cross-Site Scripting of the Third Kind, <http://www.webappsec.org/projects/articles/071105.html>
16. Korscheck, C.: Automatic Detection of Second-Order Cross-Site Scripting Vulnerabilities. Diploma Thesis, Wilhelm-Schickard-Institut für Informatik University at Tübingen (2010)
17. Kotha, R., Prasad, K., Naik, D.: Analysis of XSS attack mitigation techniques based on platforms and browsers. In: SEA, CLOUD, DKMP, CS & IT, vol. 5, pp. 395–405 (2012)
18. kses, <http://sourceforge.net/projects/kses/>
19. Lundeen, R., Ou, J., Rhodes, T.: New Ways I'm Going to Hack Your Web App. Blackhat AD (2011)
20. Microsoft Anti-Cross Site library V4.2, <http://www.microsoft.com/en-us/download/details.aspx?id=28589>
21. Hamada, M.H.A.: Client Side Action Against Cross Site Scripting Attacks. Degree of Master in Information Technology, Islamic University Faculty of Information Technology (2012)
22. Nadji, Y., Saxena, P., Song, D.: Document Structure Integrity: A Robust Basis for Cross-site Scripting Defense. In: Network and Distributed System Security Symposium (2009)
23. National Institute of Standards and Technology: CVE and CCE Statistics Query Page, <http://web.nvd.nist.gov/view/vuln/statistics>

24. Nunan, A.E., Souto, E., dos Santos, E.M., Feitosa, E.: Automatic Classification of Cross-Site Scripting in Web Pages Using Document-based and URL-based Features. In: Proceedings of ISCC, pp. 702–707 (2012)
25. Open Web Application Security Project: XSS (Cross Site Scripting) Prevention Cheat Sheet, [https://www.owasp.org/index.php/XSS_\(Cross_Site_Scripting\)_Prevention_Cheat_Sheet](https://www.owasp.org/index.php/XSS_(Cross_Site_Scripting)_Prevention_Cheat_Sheet)
26. OWASP AntiSamy Project, https://www.owasp.org/index.php/Category:OWASP_AntiSamy_Project
27. OWASP Java HTML Sanitizer Project, https://www.owasp.org/index.php/OWASP_Java_HTML_Sanitizer_Project
28. Pelizzi, R., Sekar, R.: Protection, usability and improvements in reflected XSS filters. In: Proceedings of the 7th ACM Symposium on Information (2012)
29. SafeHTMLChecker, http://doc.b2evo.net/v-1-9/evocore/_blogs-inc_misc-_htmlchecker.class.php.html
30. Saxena, P., Molnar, D., Livshits, B.: Scriptgard: Preventing script injection attacks in legacy web applications with automatic sanitization. Tech. rep., Microsoft Research (2010)
31. Selvamani, K., Duraisamy, A., Kannan, A.: Protection of Web Applications from Cross-Site Scripting Attacks in Browser Side. International Journal of Computer Science and Information Security 7, 229–236 (2010)
32. Shar, L.K., Tan, H.: Automated removal of cross site scripting vulnerabilities in web applications. Information and Software Technology 54, 467–478 (2012)
33. Tibom, P.: Incapsula vs. CloudFlare. Security Review & Comparison (2012)
34. TidyManaged, <https://github.com/markbeaton/TidyManaged>
35. The OWASP Enterprise Security API, <https://www.owasp.org/index.php/ESAPI>
36. Wang, Y., Li, Z., Guo, T.: Program Slicing Stored XSS Bugs in Web Application. In: Proceeding of the 5th IEEE International Symposium on Theoretical Aspects of Software Engineering, pp. 191–194 (2011)
37. Weinberger, J., Saxena, P., Akhawe, D., Finifter, M., Shin, R., Song, D.: A systematic analysis of XSS sanitization in web application frameworks. In: Atluri, V., Diaz, C. (eds.) ESORICS 2011. LNCS, vol. 6879, pp. 150–171. Springer, Heidelberg (2011)

Minimization of Numerical Dispersion Errors in Finite Element Models of Non-homogeneous Waveguides

Andrius Krisciunas and Rimantas Barauskas

Department of System Analysis, Kaunas University of Technology,
Studentu Str. 50–407, LT–51368 Kaunas, Lithuania
{andrius.krisciunas, rimantas.barauskas}@ktu.lt

Abstract. The paper presents the approach for the reduction of numerical errors, which are inherent for simulations based on wave propagation models in discrete meshes. The discrete computational models always tend to generate errors of harmonic wave propagation velocities in higher frequency ranges, which can be treated as numerically-induced errors of dispersion curves. The result of the errors is the deterioration of the shapes of simulated waves as the time of simulation increases. The presented approach is based on the improvement of the matrices of elements of the finite element model by means of correction of the modal frequencies and modal shapes of an individual element. The approach developed by the authors earlier and proved to work in the case of a uniform waveguide now has been demonstrated to be valid for simulations of waves in networks of waveguides. The non-reflecting boundary conditions can be implemented by combining synthesized and lumped mass elements in the same model. The propagating wave pulses can be satisfactorily simulated in comparatively rough meshes, where only 6-7 finite elements per wavelength are used.

Keywords: finite elements, wave propagation, modal synthesis, modal errors.

1 Introduction

The short-waves and pulses propagation simulations are encountered in various engineering applications, such ultrasonic measurement techniques oriented for defects and impurities detection inhomogeneous structures, hydraulic pressure pulses propagation in large pipeline networks, etc. The concept of the short-wave considered in this paper concept relies on the comparison of the length scales of the shapes of propagating waves and of the model of the propagation environment. We assume that the wavelength is hundreds or thousands times shorter than the characteristic length of the propagation environment. One of the most important problems, which occur in designing and implementing finite element (FE) models of the wave propagation, is huge dimensionalities of the models and there for every high demands for computing resources. This is exceptionally important in simulations of short waves propagation. In order to achieve areas on able accuracy of the computation extremely dense finite element meshes are necessary. A highly refined FE mesh in its turn requires very small time integration steps, which increase the computation time even more.

Generally, the dimensionality of the computational models is reduced as rougher meshes are applied. The measure for roughness of the mesh is the number of elements per characteristic wavelength. Unfortunately, rough meshes tend to increase the simulation errors, which exhibit themselves as severe deterioration of the shapes of propagating wave pulses as the time of simulation increases. It is well known that this happens due to the errors of representation of wave propagation velocities of different harmonic components of the propagating pulse. In any discrete model of wave propagation waves of different frequencies propagate with slightly different velocities than they should in reality, and the magnitude of an error depends on the frequency of the wave. The relationship of the wave velocity against the wave frequency or against the wavelength is called the dispersion curve, therefore the errors under consideration are often referred to as numerical dispersion errors or phase velocity errors.

Already in early 1980 it was noticed that solutions provided by the models using lumped and consistent mass matrices tend to generate essentially different patterns of the deterioration of propagating wave pulses because of errors of modal frequencies of the individual finite elements of the structure [1]. The weighted average of the consistent and lumped mass matrices is referred to as the generalized mass matrix, which enables to obtain same accuracy of the overall model with less number of elements per wavelength. In 2004 the element matrix synthesis technique was proposed for modification of modal shapes and frequencies of an individual element such that after assembling the element matrices into structural matrices the overall model would generate minimal possible phase velocity errors [2]. The performance of the method was examined in the case of 1D homogenous structures consisting of coextensive elements. About 80% of the natural frequencies of the overall model generated errors less than 2% compared with exact solution, and only 6-7 elements per wavelength were enough to obtain same accuracy as in models based on generalized mass matrices with 17-18 nodes per wavelength. In [3] it was demonstrated that for 2D homogenous structures assembled of identical elements based on synthesized mass matrices the method worked properly and maintained similar accuracy. In order to analyze parts of large models as particular sub-models the non-reflecting boundary conditions at the cut boundaries were applied [4].

In this paper the performance of the models based on the synthesized matrices was examined in the case of wave propagation in non-homogeneous branched 1D structures. As a sample structure the FE model for transient pressure wave simulation in fluid pipe network was constructed. Several finite element models suited for fluid and gas flow transients have been reported. In [5] the finite element model of the flow in the pipe with laminar frequency-dependent friction was developed. In [6] the formulation of the fluid-structure interaction included axial vibration model of the pipe aiming at better estimating the relative velocity of the fluid against the pipe wall. Most reported finite element models of transient pipe flow were based on simplified systems of governing equations, in which convection and (or) non-linear terms can be neglected. A characterization of different options of existing transient models and approaches to their solution has been provided in [7]. The approach developed in this work treats the finite element model of the sample pipe work structure as a standard structural dynamic equation in terms of pressure variables and their first and second-order time derivatives, where the flow velocities compared with pressure wave speed are very small and can be neglected. The results obtained in his work justify the

validity of the synthesized matrices approach in the case of non-homogeneous structures, as well as reveals certain difficulties when examining the non-reflecting boundary conditions.

2 Modal Synthesis Technique

A linear dynamic finite element model used for simulation of vibrations and wave propagation can be always presented as

$$[M]\{\ddot{U}\} + [C]\{\dot{U}\} + [K]\{U\} = \{R(t)\} \quad (1)$$

where $[M]$, $[C]$, $[K]$ are structural mass, damping and stiffness matrices, $\{U(t)\}$ is the nodal displacement vector, $\{R(t)\}$ is the lumped forces vector.

By assuming that the damping forces are very small the proportional form of the damping matrix as $[C] = \alpha[M]$ is employed. The form of equation (1) is the same for an individual element, as well as, for the finite element structure, the only difference being the dimensionality of vectors and matrices.

Mass and stiffness matrix could be expressed using modal synthesis technique

$$[M] = ([Y]^T)^{-1}[Y]^{-1} \quad (2.1)$$

$$[K] = ([Y]^T)^{-1}[\text{diag}(\omega_1, \omega_2, \dots, \omega_n)][Y]^{-1} \quad (2.2)$$

where $[Y]$ and $\text{diag}(\omega_1, \omega_2, \dots, \omega_n)$ are shape functions and modal frequencies of non damped structures. Synthesized matrices of elements $[\tilde{M}_{el}]$ and $[\tilde{K}_{el}]$ obtained by properly modifying their modal frequencies and modal shapes. The goal of the modification is that the computational domains assembled of the synthesized element matrices would generate minimal possible phase velocity errors. It accomplished by taking first N exact modal shapes and modal frequencies of a domain and modifying them by means of properly selected coefficient vectors $\{a^\omega\}$ and $\{a^y\}$ as

$$[\text{diag}(\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_N)] = [\text{diag}(\omega^2)]\{a^\omega\} \quad (3.1)$$

$$[\{\tilde{y}_1\}, \{\tilde{y}_2\}, \dots, \{\tilde{y}_N\}] = [Y]\{a^y\} \quad (3.2)$$

where $\{\tilde{y}_1\}, \{\tilde{y}_2\}, \dots, \{\tilde{y}_N\}$ and $\text{diag}(\omega_1, \omega_2, \dots, \omega_n)$ are synthesized element shape functions and modal frequencies. The first N nearly exact modal shapes and frequencies can be obtained by presenting the volume occupied by an individual element by means of a highly refined model of dimensionality n . N . Coefficients $\{a^\omega\}$ and $\{a^y\}$ are computed by minimizing modal frequency error as target function

$$\min_{\{a^\omega\}, \{a^y\}} \Psi = \sum_{i=1}^N \left(\frac{\hat{\omega}_i - \hat{\omega}_{i0}}{\hat{\omega}_{i0}} \right)^2 \quad (4)$$

where $\hat{\omega}_i$ – modal frequency of i -th mode of domain assembled of synthesized elements, $\hat{\omega}_{i0}$ – exact value of the modal frequency of i -th mode. The modal frequency error of the joined domain is minimized by employing the gradient descend method, where sensitivity functions $\frac{\partial \Psi}{\partial \{a^y\}}$ and $\frac{\partial \Psi}{\partial \{a^\omega\}}$ are employed.

3 A Pressure Impulse Propagation FE Model as an Example of a Branched 3D Structure of Uni-dimensional Waveguides

The basic set of uni-dimensional flow equations, which contains the continuity equation and the linear momentum conservation equation reads as

$$\left\{ \begin{array}{l} \frac{\delta m}{\delta t} + \frac{\delta(mv)}{\delta x} = 0; \end{array} \right. \quad (5.1)$$

$$\left\{ \begin{array}{l} \frac{\delta(mv)}{\delta t} + \frac{\delta(mv^2)}{\delta x} + A \frac{\partial p}{\partial x} + \frac{f}{D} \frac{mv|v|}{2} + mg \sin a = 0. \end{array} \right. \quad (5.2)$$

where δx – differential element of the fluid in the pipe of uniform cross-section, p – is the fluid pressure, $m = \rho A$ is the mass of the fluid of mass density ρ per unit length of the pipe of cross-sectional area A , v is the velocity of the fluid flow, a is the angle of the pipe to against the horizontal, g is the free-fall acceleration, $f = \frac{0.3614}{RE}$ is the friction coefficient for turbulent flow, where $RE = \frac{\rho D |v|}{\mu}$ is the Reynolds number (μ – the dynamic viscosity of the fluid). The standard FE model for pipeline flow cannot be directly gained from (5) equation. At low values of flow velocity compared to pressure wave propagation speed and nearly incompressible fluid, equation (5) can be expressed as a standard structural dynamic equation without non-linear and convection terms as

$$\left\{ \begin{array}{l} \frac{\delta^2 p}{\delta t^2} + \frac{f}{D} |v| \frac{\delta p}{\delta t} - \frac{\tilde{K}}{\rho_0} \frac{\delta^2 p}{\delta x^2} = 0; \\ \frac{\delta v}{\delta t} = -\frac{1}{\rho_0} \frac{\delta p}{\delta x} - \frac{f}{D} \frac{v|v|}{2} - g \sin a. \end{array} \right. \quad (6)$$

where $\tilde{K} = K(1 + \frac{KD}{hE})^{-1}$ is the equivalent bulk modulus of the pipe, which combines the bulk stiffness of the fluid and the radial expansion stiffness of the pipe, K is the bulk modulus of the fluid, E is the stiffness modulus of the pipe material, D and h are the diameter and the wall thickness of the pipe.

In this paper we examine pressure wave propagation in non-damped systems where and dynamic equation of (6) can be transform into equation (1), where the element matrices read as

$$[M^e] = A \int_0^L [N]^T [N] dx \approx \frac{AL}{3} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \approx \frac{AL}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (7.1)$$

$$[K^e] = \frac{A}{L} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \quad (7.2)$$

$$\{R^e\} = -A\tilde{K} \left(\frac{P^*}{L} - g \sin a \right) \begin{Bmatrix} 1 \\ -1 \end{Bmatrix} \quad (7.3)$$

$$[C^e] = 0 \quad (7.4)$$

where in (6.3) p^* – is the pressure created by a pump, if the pump is assumed to work within the region presented by this particular finite element.

The modal synthesis technique described in section 2 of this work is applied to the subdomains formed of elements with matrices (7.1) and (7.2) in order to obtain the synthesized matrices exhibiting the improved dynamic performance.

4 Numerical Investigation

The verification of the model assembled of synthesized elements is carried out by analyzing its behavior in sample situations and by comparing the obtained results against the solutions obtained in highly refined (>35 nodes per wavelength) meshes of elements employing the generalized mass matrices. In this work models of synthesized elements are assembled of 10-node elements, where a^ω and a^y parameters are computed by solving problem (4) for first 89% modal frequency errors of the refined model. For the analysis of wave propagation in non-homogenous waveguides different types of pipes were used in the sample structures. The integration in time was performed by means of the central difference method. The physical parameters of the model were selected corresponding to the waveguide as the water-filled pipeline with bulk modulus $K = 2.2^9 (N/m^2)$, mass density $\rho = 995 (kg/m^3)$ and dynamic viscosity $\mu = 5.47^{-4} (N * s/m^2)$. Evaluations were accomplished by comparing modal frequency errors and the propagating pressure impulse shapes obtained by using the models of synthesized elements and the models of elements based on the generalized mass matrices containing the same number of elements per wave length. Element number N per wavelength was selected experimentally by modeling the pressure pulse in a uniform 2700 (m) length pipe of 91 node with Young's modulus $E = 2.1^{11} (N/m^2)$, diameter $D = 0.1 (m)$, wall thickness $h_1 = 0.0035 (m)$ which pressure wave speed $C = 1112.7 (m/s)$. The excitation pressure pulse was generated at the left hand end of the pipe as the pressure overshoot $10^5 (Pa)$ as a half-sine pulse of duration $dt = 0.147 (s)$.

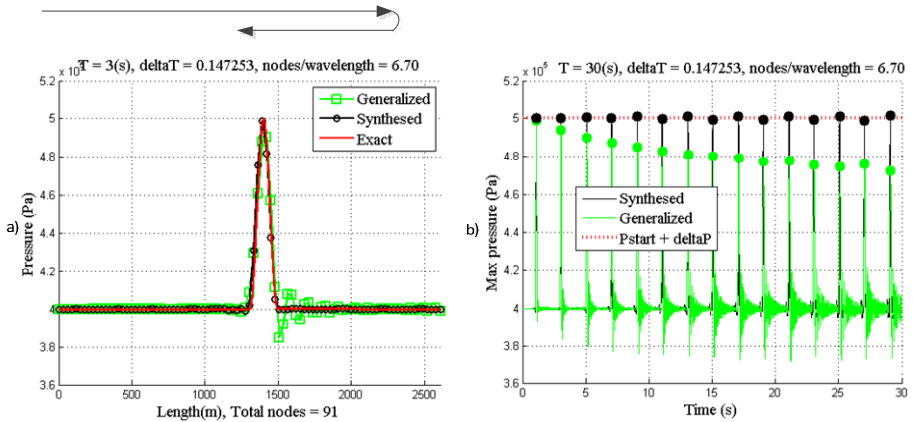


Fig. 1. a) Pressure pulse after 3s. simulation. b) Pressure in 47 node for 30s simulation.

Figure 1a represents the pressure pulse after 3 (s) in the case of 6.7 elements per wavelength. The synthesized model (black line) performs in close coincidence with the exact solution (red line), while the model based on generalized mass matrix elements (green line) generates significant errors in front of the main pulse. As the simulation time is increased, the accumulation of errors is evident. Figure 1b represents the vibration of a selected node (in this example node 47) during 30 (s) simulation, where small circles mark the time instances as the front of the impulse arrives at the node.

The model of sequentially connected pipes of two different types (diameters $D_1 = 0.1(m)$, $D_2 = 0.05(m)$, wall thickness $h_1 = 0.0035 (m)$, $h_2 = 0.0025 (m)$ and Young's modulus $E = 2.1^{11} (N/m^2)$) has been employed as an example of the non-homogenous model.

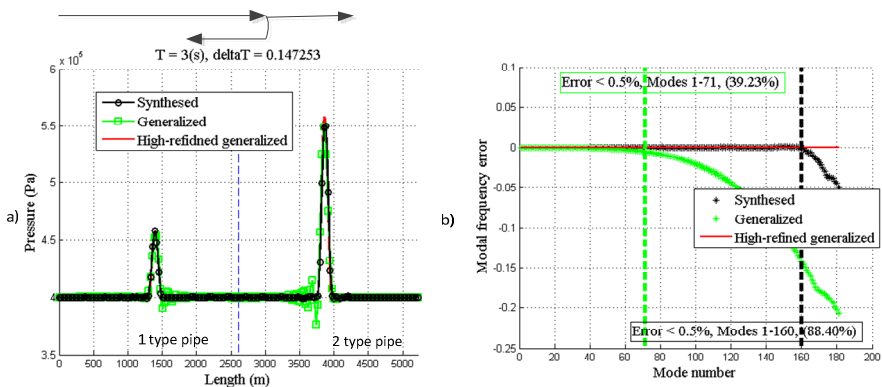


Fig. 2. a) Pressure impulse after 3(s). b) Model frequency error.

Figure 2a represents the pressure pulse shape after 3(s), while the pulse was partially reflected at the intersection of the pipes. The model based on synthesized elements, 6.7 elements per wavelength, provided satisfactory results compared with those obtained in a highly refined mesh (~42 nodes per wavelength) model based on generalized mass matrices. The model based on the generalized mass matrices with 6.7 nodes in wavelength generated significant errors due to the numerical dispersion. In Figure 2b modal frequency errors produced by different models are compared, where dashed line marks boundary of the modes region with the modal frequency error less than 0.5%. It can be seen that in the synthesized model the error less than 0.5% is ensured for about 89% of the total number of modes.

The branched non-homogenous model composed of uni-dimensional waveguide segments with non-reflecting boundary conditions has been analyzed (Figure 3a).

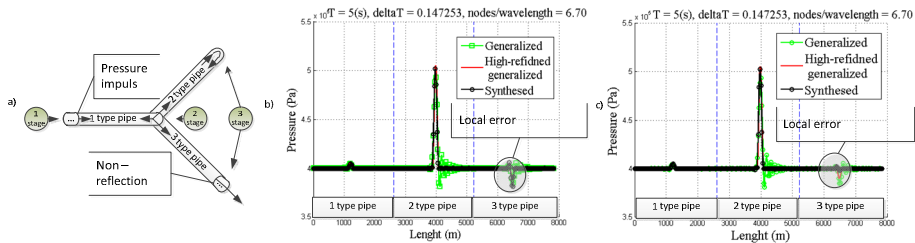


Fig. 3. a) Branched inhomogeneous model b) Local error caused boundary non reflection in model of synthesized elements c) Local error in model with one small lumped element before boundary non reflection condition implementation

There are 3 types pipes (diameters $D_1 = 0.1(m)$, $D_2 = 0.08(m)$, $D_3 = 0.05(m)$, wall thickness $h_1 = 0.0035(m)$, $h_2 = 0.003(m)$, $h_3 = 0.0025(m)$ and Young modulus $E = 2.1^{11}(N/m^2)$) used in model. At stage 1 the pressure wave actuated by $10^5(Pa)$ half-sine pulse in the left-hand end of the pipe type 1. At stage 2 (after 2(s)) the impulse is partially reflected, and distributed to pipes of types 2 and 3. At stage 3 (after 4(s)) at the end of pipe type 2 the impulse reflects and comes back while at the end of pipe type 3 the non- reflecting boundary condition is implemented. Figure 3b represents the situation after 5(s), where pressure impulse has traveled $\sim 6.8km$ (wave speeds are different in the pipes of different types). As can be seen, the junctions between different segments of the model based on synthesized matrices do not introduce any marked disturbances in the simulation results. This means that the synthesized matrices approach is working properly for non-homogeneous waveguide network models. The non-reflecting boundary condition applied directly to the synthesized matrices model does not work properly and generate significant local error up to 13% of the impulse height. However, the error could be easily reduced up to $\sim 1.5\%$ of impulse amplitude by adding a short-length lumped mass matrix finite element just before the point where the non-reflecting boundary condition was implemented. Thus the overall performance of the model was even better than obtained by the highly refined model based on the generalized mass matrices (Figure 3c). Local errors could not be completely eliminated. Probably, they were caused by the combination of elements of two different types (synthesized and lumped) within the same model.

5 Conclusion

The approach for the reduction of the simulation errors of propagating pulses in uni-dimensional waveguides has been investigated in the case of branched non-homogeneous structures with implemented non-reflecting boundary conditions. The overall approach based on synthesized mass matrices has been earlier verified for the case of uniform waveguides. The results of this work demonstrated that the approach is valid for wave propagation simulations in branched structures joined of 1D segments possessing different characteristic wave propagation speeds. Non-reflecting boundary conditions applied directly for synthesized models generate significant errors up to 13% of the pulse height. The addition of a small lumped element just before the non-reflection boundary point enabled to reduce the error significantly, where only a small local error remained. One of possible application areas of the developed model is the simulation of the pipeline leakage monitoring system, the purpose of which is to determine the location of the leakage (pressure drop pulse) based on pressure variations registered by meters arranged at various places of the pipeline.

References

1. Mullen, R., Belytschko, T.: Dispersion analysis properties of finite element semi-discretizations of the two-dimensional wave equations. *International Journal for Numerical Methods in Engineering* 18, 1–29 (1982)
2. Barauskas, R., Barauskiene, R.: Highly convergent dynamic models obtained by modal synthesis with application to short wave pulse propagation. *International Journal for Numerical Methods in Engineering* 61(14), 2536–2554 (2004)
3. Barauskas, R.: On highly convergent 2D acoustic and elastic wave propagation models. *Communications in Numerical Methods in Engineering* 22(3), 225–233 (2006)
4. Wolf, J.P., Song, C.: *Finite-element modelling of unbounded media*. Wiley, Chichester (1996)
5. Shu, J.-J.: A finite element model and electronic analogue of pipeline pressure transients with frequency-dependent friction. *Journal of Fluids Engineering* 125(1), 194–199 (2003)
6. Kochupillai, J., Ganesan, N., Padmanabhan, C.: A new finite element formulation based on the velocity of flow for water hammer problems. *International Journal of Pressure Vessels and Piping* 82(1), 1–14 (2005)
7. Osiadacz, A.J.: Different Transient Flow Models-Limitations, Advantages, And Disadvantages. In: *PSIG Annual Meeting* (1996)

Novel Method to Generate Tests for VHDL

Vacius Jusas and Tomas Neverdauskas

Software Engineering Department, Kaunas University of Technology,
Studentu St. 50, LT-51368, Kaunas, Lithuania

{vacius.jusas,tomas.neverdauskas}@ktu.lt

Abstract. Verification is the most crucial part of the chip design process. Test benches, which are used to test VHDL code, need perform efficiently and effectively. We present an algorithm that achieves high code coverage by analyzing the finite state machine (FSM), and control flow graph (CFG) that are constructed from the source code. The symbolic execution of VHDL (Very-high-speed integrated Hardware Description Language) code is used as well. These three elements are combined into framework (TestBenchGen) written in Python programming language and evaluated against ITC'99 benchmark suite.

Keywords: Finite state machines, control flow graphs, hardware verification, test generation.

1 Introduction

Growing advances in very large scale integration (VLSI) technology have led to an increased level of complexity in modern hardware systems. Such complex systems have many technological challenges. At very top of the list is verification, which takes 40 to 70 [1] per cent of the total development effort for the design.

Hardware verification is the process of evaluating to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase. It is recognized as the largest task in silicon development, and as such has the biggest impact on the time to market. Late detection of design errors typically results in higher costs due to the associated time delay as well as loss of production.

With the emergence of complex high-performance microprocessors, functional test generation has become an essential verification step. With the ever-growing demand for greater performance and faster time to market, coupled with the exponential growth in hardware size, verification has become increasingly difficult.

VHDL commonly used with field-programmable gate arrays (FPGA) to generate and configure logic gates. FPGA [2] emerges broad variety of applications and industries including requiring very high reliability such as medicine, defense, military and space.

In this paper, we present a novel method to achieve high coverage results in hardware verification process and use this method to generate test for ITC'99 benchmark suite at high level of abstraction.

2 Background

The simulation is still the most widely used form of design verification, because the formal verification is possible either for smaller projects or small parts of larger projects only [2]. But the simulation has to use test suites to validate the design functionalities. Many different approaches are used in order to generate test cases for design verification.

Hardware designers then perform extensive simulations for what they call “behavioral verification”, an activity a software engineer might term “validation”, or “software testing”. Because VHDL is similar to a high-level programming language, we can apply software assurance techniques to a hardware design in order to identify and remove faults. These faults need to be detected through the use of test benches. Test bench automation through the generation of test patterns and test cases increases the efficiency and effectiveness of behavioral verification.

Coverage achieved during verification is the single most important parameter in determining the quality of verification results. In conventional simulation based verification, coverage of a set of tests or a test suite is measured using various metrics such as code coverage, toggle coverage, FSM coverage [3]. Code coverage measures the fraction of statements in the RTL source code executed or covered while simulating the test suite. Since code coverage can be easily related to the RTL code and reporting it adds little overhead to simulation, it is the most popular coverage metric.

2.1 VHDL Structure

The VHDL description of the device consists of two parts: entity and architecture. The entity represents the interface of the device, and the architecture is used to code the functional implementation of the device [4]. Different levels of functional implementation can be used. The most frequently used description levels are the following: behavioral, register transfer level (RTL), and structural. The behavioral architecture body of entity describes its function in an abstract way and the concurrent statements in it are limited to process statements, subprogram calls and signal assignments. The process statements are further made up of sequential statements that are much like the kinds of statements we see in a conventional programming language such as statements evaluating expressions, statements assigning values to variables (variable-assignment statements), conditional execution statements (if-then-else, case, etc.), repeated execution statements (loops) and subprogram calls. In addition, there is the signal assignment statement, which is unique to hardware modeling languages. This statement is similar to variable assignment statement, except that it causes the value on a signal to be updated at some future time.

2.2 Finite State Machine

Finite state machine (FSM) is based on quintuple [4]: $N=(S, \Sigma, q_0, F, \delta)$, where S is a finite, non-empty set of states, Σ – finite, non-empty set of input symbols, q_0 – initial

state $q_0 \in S$, F – a (possible empty) set of final states and δ – transition function $\delta : S \times \Sigma \rightarrow \mathcal{P}(S)$. Each transition is labeled with a condition that needs to be satisfied for reaching next state.

In VHDL FSM is a sequential logic circuit which visits states of some finite set, the process of visiting depends upon the values of the inputs and the previous state. The state transitions are synchronized by a clock. Unlike the regular sequential circuit, the state transition of the FSM is more complicated and the sequence exhibits no simple, regular pattern, as in a counter or shift register.

In a synchronous FSM, the transition is controlled by a clock signal (mostly rising) and can occur only at edge of the clock. The main application of an FSM is to implement operations that are performed in a sequence of steps. A large hardware system usually involves complex tasks or algorithms, which can be expressed as a sequence of actions based on system status and external commands. An FSM can function as the control circuit (known as the control path) that coordinates and monitors the operations of other units (known as the data path) of the system.

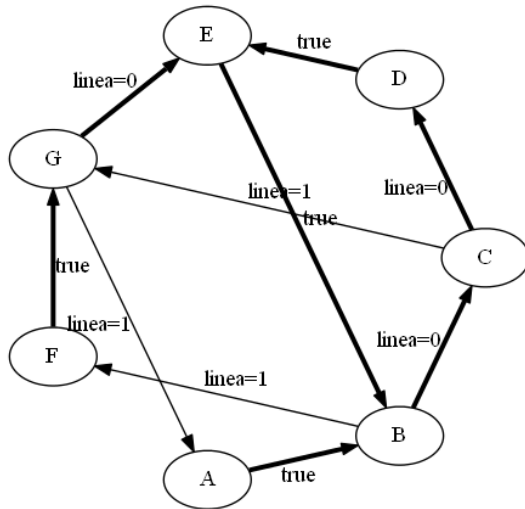


Fig. 1. Graphical representation of FSM in B02

FSM’s can also be used in many simple tasks, such as detecting a unique pattern from an input data stream [5] or generating a specific sequence of output values. So, it is very important part of VHDL semantics and can be used as main part (but not the only) for functional test generating.

In VHDL FSM are mainly represented as *if-else* or *case-when* code structures. The difference is in the derivation of the next-state logic, which should be implemented according to a state diagram. The FSM of B02 circuit from ITC99 benchmark suite is presented in Fig. 1.



2.3 Control Flow

Control flow is the order function calls, instructions and statements are executed or evaluated when a program is ran. Many programming languages have what are called control flow statements, which are used to determine what section of code is run in a program at a given time.

In behavioral descriptions of VHDL, the main statement is the process statement. The process statement can appear in the body of an architecture declaration. The body of the process statement includes sequential statements like those found in software programming languages and it can be implemented as control flow [6].

2.4 Symbolic Execution

Symbolic execution is one of the many techniques that is be used to automate software testing by generating test cases that achieve high coverage of program executions. A significant scalability challenge for symbolic execution is how to handle the exponential number of paths in the code.

SE is an extension of normal execution, providing the normal computations as special case. Computational definitions for the basic operators of the language are extended to accept symbolic inputs and produce symbolic formulas as output [7]. The *state* of a symbolically executed program includes the symbolic values of program variables, a *path condition* (PC) and a program counter, representing next statement to be executed. The path condition is a (quantifier-free) *BOOLEAN* formula over the symbolic inputs. It accumulates constraints which the inputs must satisfy in order for an execution to follow the particular associated path [8]. A *symbolic execution tree* characterizes the execution paths followed during the symbolic execution of a program. The nodes represent program states and the edges represent transitions between states. The main difference between CFG and symbolic execution is that SE produces all possible execution paths of program. Result of symbolic execution is Boolean formula that is solved with SMT solver to provide concrete values.

There are cases of using SE in software testing [9] but in hardware verification known research still is in progress. Our approach differs from other known approaches since we used FSM in test generation process.

It's important not to confuse symbolic execution with symbolic simulation [10].

3 Framework

Test generation framework “TestBenchGen” combines methods described in previous sections into novel methodology. Basic framework structure is presented in Fig. 2.

The flow of the algorithm to generate tests is presented in Fig. 3. First, VHDL source file is read by grammar parser and formal structure of code is formed. Second, FSM is generated. In the next step, initial state q_0 and next state q_1 of FSM are used as two starting nodes. In between them, all nodes of CFG are loaded by creating virtual function.

All the parts of framework are created in Python programming language.

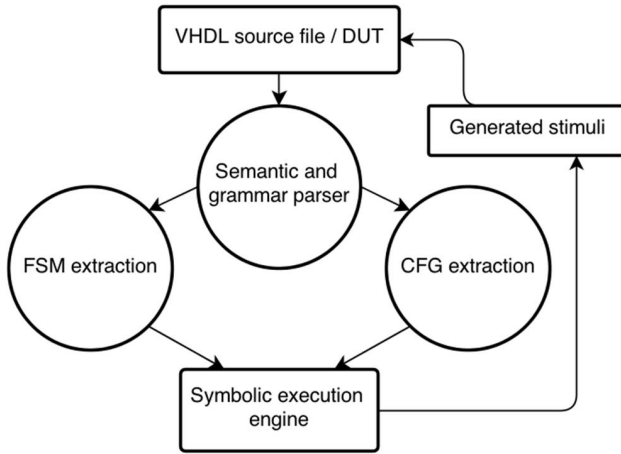


Fig. 2. “TestBenchGen” structure

Virtual function is a transformation of VHDL code (manually) to Python in such way that binds all local variables, which exist between q_0 and q_1 , to this function parameters list. Virtual function body consists of all the programming code that is provided in CFG between q_0 and q_1 states. Virtual function is executed symbolically and concrete values are computed by SMT library Z3 [11]. That result is used to generate benchmarks (why not stimuli, benchmark is new concept) if next state of FSM does not exist. Otherwise, next states in FSM q_1 and q_2 are used next.

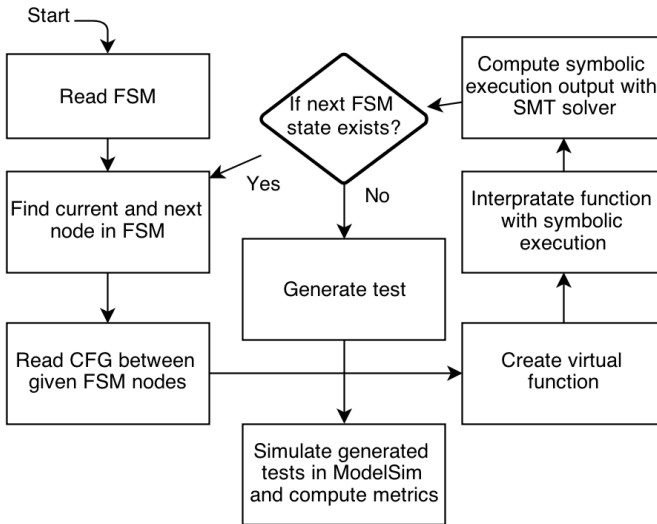


Fig. 3. “TestBenchGen” test generation algorithm

4 Evaluation

As mentioned in previous chapter step after code extraction is FSM generation. The FSM is characterized by the number of states and the number of transitions. Detailed information on FSM, which was obtained for the benchmarks of ITC'99 suite, is presented in Table 1. The last column in Table 1 named "paths" represents total count of elementary circles [12] called paths.

Table 1. FSM Metrics

Circuit	FSM's	FSM states	FSM transitions	Paths
b01	1	8	24	24
b02	1	7	10	5
b03	1	3	2	1
b04	1	2	3	1
b05	1	5	8	3
b06	1	7	13	7
b07	1	7	13	4
b08	1	4	9	3
b09	1	4	8	4
b10	1	11	24	10
b11	1	9	38	7
b12	-	-	-	-
b13	4	8,4,4,10	10,7, 6, 90	3,4,3,11
b14	-	-	-	-
b15	2	8, 10	27, 35	17,8

Next step is control flow graph extraction. In our framework, each process is treated as separate control flow graph $G = (V, E)$. Each statement in a process is a node $v \in V$ in the control flow graph and the edges $e \in E$ represent the control flow among statements. We add an edge (e_{a1}, e_{a2}) if the statement $a1$ is executed immediately after the statement $a2$.

Our framework supports branch statements (*Case*, *If / Else*) in VHDL. For each branch a node is introduced with edge connection to parent element. An start and an end node will be added as unique entry and exit points of the process. In a control flow graph (Fig. 2), each node represented as a rectangular block matches a straight-line code without any branching. Directed edges are used to represent jumps in the

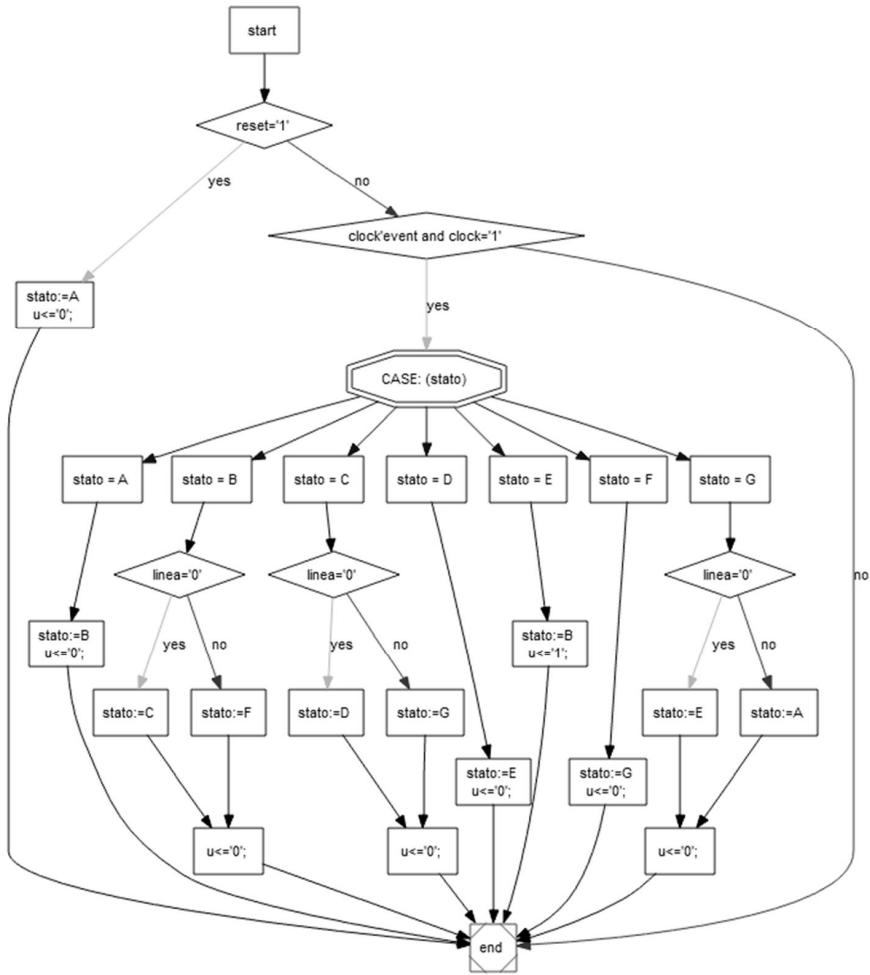


Fig. 4. Graphical CFG representation of B02

control flow. Branch operations are shown as diamond. Excerpt of source code, which matches the graph shown in Fig. 4, is presented in Fig. 5. This source code is from ITC'99 benchmark B02 scheme.

```

if reset='1' then
...
elsif clock'event and clock='1' then
  case stato is
    when A =>
      ...
    when B =>
      if linea='0' then
        ...
      else
        ...
      end if;
    ...
    when C =>
      if linea='0' then
        ...
      else
        ...
      end if;
    ...
    when D =>
      ...
    when E =>
      ...
    when F =>
      ...
    when G =>
      if linea='0' then
        ...
      else
        ...
      end if;
    ...
  end case;
end if;

```

Fig. 5. Control flow of B02 in VHDL

Detailed information about each CFG of ITC'99 benchmark suite circuits is provided in Table 2. Cyclomatic Complexity [13] directly measures the number of linearly independent paths. Number of concurrent statements shows count of different processes. Last column in Table 2 presents count of different possible execution paths between start and end nodes.

Table 2. CFG Metrics

Circuit	Cyclomatic Complexity	Number of concurrent statements	Unique flows in CFG between start and end
b01	18	1	18
b02	12	1	12
b03	18	1	18
b04	13	1	40
b05	17	3	6; 8002; 14
b06	16	1	28
b07	14	1	14

Table 2. (continued)

b08	10	1	10
b09	10	1	10
b10	28	1	33
b11	23	1	23
b12	20	4	16; 3; 1; 104
b13	11	5	14; 10; 9; 20; 14
b14	163	1	20043
b15	38	3	48; 128; 18

In order to evaluate generated benchmarks against ITC'99 ModelSim Student Edition is used. *TestBenchGen* achieves very high result of coverage. Detailed results of B03 evaluation for each earlier presented metric are provided in Table 3. The evaluation of generated test stimuli uses 5 different code coverage measurements.

Statement coverage measures the number of executable statements within the model that have been executed during the simulation run. In most verification cases statement coverage is used as minimum goal [14].

Branch coverage [15], sometimes also referred to as decision coverage, measures the coverage of *if* and *case* statements that affect the control flow of the HDL execution.

Focused expression coverage (FEC) — a row based coverage metric which emphasizes the contribution of each expression input to the expression's output value. FEC measures coverage for each input of an expression. If all inputs are fully covered, the expression reaches 100% FEC coverage. In FEC, an input is considered covered only when other inputs are in a state that allows it to control the output of the expression. Further, the output must be seen in both 0 and 1 states while the target input is controlling it. If these conditions occur, the input is said to be fully covered. The final FEC coverage number is the number of fully covered inputs divided by the total number of inputs.

FSM coverage shows the ability to reach all the states and traverse all possible paths through a given state machine. There are two types of coverage for FSM:

- State coverage – all states of an FSM are hit during simulation.
- State transition coverage – FSM transition among all states that are achievable in simulation.

Table 3. B03 detailed results

Coverage Type	Bins	Hits	Misses	Coverage (%)
Statement	57	57	0	100,00%
Branch	26	24	2	92,30%
FEC Condition	4	4	0	100,00%
FSM State	3	3	0	100,00%
FSM Transition	6	6	0	100,00%
Average				98,07%

TestBenchGen was used to generate tests from B01 to B09 excluding B05 because of parallel processes included in this benchmark. *TestBenchGen* currently does not support test generation of simultaneous executed VHDL processes. All coverage metrics are combined as average and presented in Table 4.

Presented results shows how effective based on given metrics TestBenchGen can create tests. Most of the cases when coverage is below 100% respond to branch coverage measurements which are impossible to reach in control flow. Reaching absolute 100% for automated method can be impossible task because of variety of small “special cases” in code, control flow or symbolic execution tree.

Table 4. B01-B09 (except B05) results

Benchmark	Combined Coverage (%)
B01	96,23%
B02	94,30%
B03	98,07%
B04	97,12%
B06	97,43%
B07	95,04%
B08	96,11%
B09	97,63%
Total average	96,49%

In comparison with other quite a new work [16] our method outperforms it for the same ITC'99 benchmark circuits since our method uses FSM information. If we compare with quite different technique such as high-level decision diagrams [17] our method outperforms it too, therefore authors [17] measure only statement and branch coverage.

5 Conclusions and Future Work

We presented our framework that uses the most common structures of VHDL and combines different methods in order to generate test stimuli for verification of VHDL code. The proposed approach allows obtaining high coverage of test stimuli using various metrics and confirms the importance of FSM usage in testing process.

Another important research task is to evaluate parallel processes that are present in several benchmarks (for example, in B05).

References

1. Bareiša, E., Jusas, V., Motiejūnas, K., Šeinauskas, R.: The use of a software prototype for verification test generation. *Information Technology and Control* 37, 265–274 (2008)
2. Uros Legat, A.B., Novak, F.: On line self recovery of embedded multi-processor SOC on FPGA using dynamic partial reconfiguration. *Information Technology and Control* 41 (2012)
3. Jou, J.-Y., Liu, C.: Coverage analysis techniques for hdl design validation. In: *Proc. Asia Pacific CHip Design Languages*, pp. 48–55 (1999)
4. Lee, D., Yannakakis, M.: Principles and methods of testing finite state machines-a survey. *Proceedings of the IEEE* 84, 1090–1123 (1996)
5. Van Lunteren, J.: High-performance pattern-matching for intrusion detection. In: *IEEE INFOCOM* (2006)

6. Rahmouni, M., Jerraya, A.A.: Formulation and evaluation of scheduling techniques for control flow graphs. In: European Design Automation Conference, with EURO-VHDL, Proceedings EURO-DAC 1995, pp. 386–391 (1995)
7. King, J.C.: Symbolic execution and program testing. *Commun. ACM* 19, 385–394 (1976)
8. Khurshid, S., Păsăreanu, C.S., Visser, W.: Generalized Symbolic Execution for Model Checking and Testing. In: Garavel, H., Hatcliff, J. (eds.) TACAS 2003. LNCS, vol. 2619, pp. 553–568. Springer, Heidelberg (2003)
9. Cadar, C., Godefroid, P., Khurshid, S., Păsăreanu, C.S., Sen, K., Tillmann, N., et al.: Symbolic execution for software testing in practice: preliminary assessment. In: Proceedings of the 33rd International Conference on Software Engineering, pp. 1066–1071 (2011)
10. Kolbi, A., Kukula, J., Damiano, R.: Symbolic RTL simulation. In: Proceedings of the Design Automation Conference, pp. 47–52 (2001)
11. de Moura, L., Bjørner, N.S.: Z3: An efficient SMT solver. In: Ramakrishnan, C.R., Rehof, J. (eds.) TACAS 2008. LNCS, vol. 4963, pp. 337–340. Springer, Heidelberg (2008)
12. Jusas, V., Neverdauskas, T.: FSM Based Functional Test Generation Framework for VHDL. In: Skersys, T., Butleris, R., Butkiene, R. (eds.) ICIST 2012. CCIS, vol. 319, pp. 138–148. Springer, Heidelberg (2012)
13. Gill, G.K., Kemerer, C.F.: Cyclomatic complexity density and software maintenance productivity. *IEEE Transactions on Software Engineering* 17, 1284–1288 (1991)
14. Harris, I.G.: Fault models and test generation for hardware-software covalidation. *IEEE Design & Test of Computers* 20, 40–47 (2003)
15. Tasiran, S., Keutzer, K.: Coverage metrics for functional validation of hardware designs. *IEEE Design & Test of Computers* 18, 36–45 (2001)
16. Liu, L., Vasudevan, S.: Efficient validation input generation in rtl by hybridized source code analysis. In: Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 1–6 (2011)
17. Minakova, K., Reinsalu, U., Chepurov, A., Raik, J., Jenihhin, M., Ubar, R., et al.: High-level decision diagram manipulations for code coverage analysis. In: 11th International Biennial Baltic Electronics Conference, BEC 2008, pp. 207–210 (2008)

Empirical Analysis of the Test Maturity Model Integration (TMMi)

Kerli Rungi¹ and Raimundas Matulevičius²

¹ Playtech Estonia OÜ,
Vanemuise 7, 51001, Tartu, Estonia
kerli@playtech.com

² Institute of Computer Science, University of Tartu,
Liivi 2, 50409, Tartu, Estonia
rma@ut.ee

Abstract. Testing is an essential part of the software development lifecycle for gaining trust in the quality of the delivered product. Concerns have been raised over the maturity of existing test processes and techniques, and the desire to improve has been acknowledged. Even though there are test process improvement models available on the market, the guidelines on how to use them are unsatisfactory. This paper describes the empirical analysis of Test Maturity Model integration (TMMi) through a single-object case study of test process assessment and improvement planning conducted in Playtech Estonia's Casino unit. An evaluation on the performance of TMMi is presented, raising also some concerns over its applicability in agile environments. Improvement possibilities for TMMi are described, which contribute to the potential enhancement of the framework.

Keywords: Test Process Improvement, TMMi, Quality Assurance, Testing.

1 Introduction

The role of software is continuously growing – according to International Software Testing Qualification Board (ISTQB) [9], the amount of software in consumer products doubles every 24 months together with the growth of application complexity. In order to ensure that the growth of complexity and size of software systems does not bring along a regression in quality, we need to assess our current test processes and focus on improving them to comply with the rising expectations.

There exist a number of test maturity models, for example, Test Maturity Model integration [15], Business Driven Test Process Improvement Model [12], Critical Testing Processes [1], and others. However, choosing the most suitable model for a specific organisational context still remains a challenging activity. In this paper, we investigate the problems that an organisation is experiencing when it uses a test maturity model and the difficulties related to the targeted context. We analyse how the Test Maturity Model integration (a.k.a., TMMi) [15] is applied in a single-object case study of test process assessment and improvement planning conducted in Playtech Estonia's Casino unit, an organisation focusing on the software development of online gaming solutions. More specifically, we consider the following questions:

1. How suitable is TMMi as a test process assessment and improvement model?
2. How the TMMi could potentially be improved?

To answer the above questions, we research the existing test improvement models on the market and analyse the application of the TMMi model. We provide an overview of the test process maturity concept of the framework and suggest how the described best practices could be considered in the organisation's improvement process. A test process assessment case study, which consists of a survey and staff interviews, is conducted with the goal to identify the informal maturity level for the Casino unit according to the TMMi model. This serves as an input for the proceeding test process improvement planning case study where additional staff interviews are conducted. As a result of the assessment and improvement case study, we show the importance of understanding the reasons and objectives for test process improvement in consideration with the needs of the organisation. Identifying the best model is primarily dependent on defining organisation-side requirements for an improvement framework. Finally, improvements to the studied model are suggested based on the gained experience, which contribute to the potential enhancement of the framework.

As part of the research process we identified several studies (see details in [10]) that have been conducted with the goal to choose the best test process improvement model based on comparisons of a list of models available on the market. However, these studies have been conducted among a small quantity of improvement models, whereas we focus on a more comprehensive list of models. Furthermore, we conduct an empirical study on a software development organisation to validate the chosen model. As a result, we present improvement suggestions for enhancing the model. This aspect has not been addressed in any of the aforementioned papers.

The paper is structured into six sections. Section 2 presents the literature study among different test process improvement models and presents TMMi as the model of our choice. Section 3 describes a survey conducted in Playtech Estonia's Casino unit. It is followed by the test process improvement planning detailed in Section 4. Section 5 presents the evaluation of the usability of TMMi and suggests improvement for the model. Finally, in Section 6 we conclude our paper.

2 Literature Study of Test Process Improvement Models

In this section, firstly, we identify the desired characteristics of the improvement framework from Playtech Estonia's Casino unit's perspective and use them as the basis to survey the test maturity models. Secondly, we give a detailed overview of the TMMi model chosen for the further analysis.

2.1 Surveyed Test Process Improvement Models

The following parameters were identified as desirable characteristics for a test process improvement framework from the Casino unit's perspective and were used as the basis for the literature study: (i) maturity measurement; (ii) terminology; (iii) compliance to standards; (iv) assessment method; and (v) availability.

The listed models were studied and compared: (i) Test Maturity Model integration [15]; (ii) Business Driven Test Process Improvement Model [12]; (iii) Critical Testing Processes [1]; (iv) Systematic Test and Evaluation Process [5]; (v) Test Improvement Model [4]; (vi) Testability Support Model [13]; (vii) Testability Maturity Model [2]; (viii) Testing Assessment Programme [13]; (ix) Software Quality Rank [3]; and (x) Test Organisation Maturity [6]. Table 1 presents the comparison of the surveyed models based on the five decision parameters. We have selected Test Maturity Model integration (TMMi) for further analysis because:

- Maturity measurement – TMMi uses an evolutionary staged model of maturity levels, which allows evident assessment of the current state of maturity and provides a clear improvement path for gaining a higher maturity level;
- Standards – being a complementary model to CMMI [11], the assessment procedure experience of TMMi and related assessment results will help to conduct any future evaluations according to CMMI. Plans for CMMI driven evaluation for Playtech Estonia are currently under discussion;

Table 1. Comparison of surveyed models (“—” indicates that no apparent reference could be obtained about the corresponding characteristic)

Criteria Model	Maturity measurement	Terminology	Standards	Assessment method	Availability
STEP	None	IEEE Glossary of Software Engineering Terminology Standard	IEEE Test Documentation Standard, IEEE Unit Testing Standard	Qualitative interviews and quantitative metrics	Publicly available
TAP	5 levels	—	CMM, Crosby's quality maturity scheme	Questionnaire	No public info
TIM	5 levels	—	CMM, TMM*	Questionnaire	Questionnaire not publicly available
TMM*	3 levels	—	CMM	Scorecard questionnaire with 20 testability factors	No public info
TSM	3 levels	—	—	—	No public info
CTP	None	—	—	Qualitative evaluation to a set of metrics and factors per each process area	Publicly available
SQR	5 ranks	—	CMM	—	No public info
TMMi	5 levels	ISTQB Standard Glossary of Terms Used in Software Testing	CMMI, ISO/IEC 15504-2	ISO/IEC 15504-2 compliant TAMAR	Publicly available
TPI Next	4 levels	Tmap Next	—	Test Maturity Matrix	Publicly available
TOM	Value from 20 to 100	—	None	Test Organisation Maturity Questionnaire	Publicly available

- Terminology – majority of the QA team in the Casino unit is certified on ISTQB Foundation level and therefore the ISTQB Standard Glossary of Terms Used in Software Testing [8] is the basis of the used testing terminology;
- Assessment method – TMMi provides a comprehensive guideline on how to conduct both a formal and informal assessment, including the instructions on how to combine an assessment team;
- Availability – both the TMMi framework and its assessment model requirements are publicly available and free of charge.

2.2 Test Maturity Model Integration

TMMi, published in 2007, is the successor of Test Maturity Model (TMM) that was produced by the Illinois Institute of Technology in 1996 [16]. It is a complementary model to CMMI process improvement method. There are five levels in the TMMi framework, where each level follows the former in an evolutionary manner and every maturity level except the default level 1 consists of three to five process areas:

1. Initial
2. Managed: (i) Test Policy and Strategy, (ii) Test Planning, (iii) Test Monitoring and Control, (iv) Test Design and Execution, (v) Test Environment;
3. Defined: (i) Test Organisation, (ii) Test Training Program, (iii) Test Lifecycle and Integration, (iv) Non-functional Testing, (v) Peer Reviews;
4. Measured: (i) Test Measurement, (ii) Product Quality Evaluation, (iii) Advanced Reviews;
5. Optimisation: (i) Defect Prevention, (ii) Quality Control, (iii) Test Process Optimisation.

Each process area consists of a set of generic and specific goals. In order to achieve a goal, a set of practices have been provided, describing the activities that are important for achieving the associated goal. Goal satisfaction is used in assessments as the basis for deciding if a process area has been achieved. In order to provide flexibility to the organisation, it allows having alternative practices or alternative implementation of the practices listed in the framework. As TMMi is a staged model, the level of capability on a maturity level is its lowers achievement rating. All of the goals for the relevant maturity level and preceding levels need to be satisfied in order to achieve the maturity level. If not, the achievement ranking only provides an indication of capability on a particular maturity level.

Together with the TMMi Reference Model, the TMMi Foundation has defined the requirements for conducting assessments within the TMMi Assessment Method Application Requirements (TAMAR) document [14]. TAMAR conforms to the international standard ISO/IEC 15504-2 [7].

3 Test Process Assessment Using TMMi

The goal of conducting this assessment is to evaluate the usability of TMMi for test process maturity assessment in Playtech Estonia's Casino unit. An informal

assessment is conducted against the full scope of TMMi framework in correspondence with TAMAR. As a result, non-official maturity rating against TMMi is produced and the assessment is utilized internally for gaining a rough understanding of the organisational testing maturity.

3.1 Assessment Team

Playtech Corporation (Playtech) is a supplier of online gaming software. The environment of the current case study is Playtech Estonia's Casino unit located in Tartu, Estonia, and its Quality Assurance (QA) team in particular. The Casino QA team is independent from the development department and is divided into three sub-teams, each consisting of a team leader and five to six engineers with a total of 21 QA professionals. There are four different roles – QA Manager, QA Team Leader, Senior QA Engineer and QA Engineer. ISTQB Standard Glossary of Terms Used in Software Testing, which is also the source for base terminology in the TMMi model, is the basis of the used testing terminology in the Casino unit. Using the same terminology should aid the team in understanding the framework's contents and therefore no TMMi specific training has been conducted. The team's responsibility includes both manual and automated testing activities. The testing is conducted on multiple levels – component, unit, system and acceptance testing level [8].

According to TAMAR, an informal assessment should be led by an experienced assessor. However, the assessor does not need to be formally accredited to perform the role. In addition, according to TAMAR, the assessment team for informal assessment can consist of a single person, including the lead assessor. One of the authors of this paper (also acting as a Casino QA Manager), was chosen as the assessment team leader (a.k.a., lead assessor) and conducted this current assessment.

3.2 Assessment Survey

The goal of the assessment survey was to collect data about the goal fulfilment of TMMi. The questionnaire was produced based on the full contents of the TMMi framework and included statements describing important testing practices. Assessment statements were divided into 16 categories according to the process areas of TMMi. Every category included 1 to 5 statements, each of which represents a specific mandatory goal of a process area, resulting in a total of 49 statements.

The wording of statements was adjusted to comply with the vocabulary used in the organisation. Examples and clarifications were provided to aid the respondents interpret those statements where considered relevant. Test terms and definitions were provided from ISTQB Standard Glossary of Terms Used in Software Testing.

Each of the statements was answered by choosing the most accurate answer in the context of Playtech Estonia's Casino unit from the following list:

- Yes – the statement is clearly understandable for the respondent and is at least partially fulfilled, hence the statement is true;
- No – the statement is clearly understandable for the respondent and the described activity does not take place, hence the statement is false;

- Don't know – the respondent does not understand the topic described in the statement and has no knowledge if the practice taking place or not.

The assessment survey was distributed to the Casino QA team electronically via the company intranet for a period of one week. Each respondent was requested to report back the time used to fill in the questionnaire. The results of the assessment survey served also as an input for succeeding staff interviews.

3.3 Staff Interviews

Staff interviews were conducted to obtain clarifications on assessment survey statements in order to avoid any potential misunderstandings and collect evidence on the successful satisfaction of specific goals of TMMi. In order to retain an achievable amount of goals and survey statements we limited the focus of interviews to process areas of maturity level 2 and 3. Both levels contain five process areas with a total of 35 statements. We excluded the statements for which the response was “Yes” by all the interviewees and additional supportive evidence was also provided by the assessor herself. As a result, 17 statements were left out from the interview scope, leaving 18 statements for further investigation as part of the staff interviews.

Three representatives were chosen from among the survey respondents, each of whom represented a different role – QA engineer, Senior QA Engineer and QA Team Leader. They were chosen to represent both the slowest and fastest response time of filling the survey, but their detailed responses to the survey statements were not analysed before the selection process to avoid bias based on their answers.

Face-to-face interviews were performed between the assessor and each of the selected participants. In addition to collecting evidence, we requested feedback on the clarity of the assessment survey's purpose, overall clarity of its contents and suggestions on improvement activities.

3.4 Data Analysis Method

Assessment survey data was submitted electronically and was exported to tabular format for further analysis. Descriptive statistics [17] were used to characterise the data regarding assessment survey response times and answer distribution to visualise central tendency and dispersion. The overall percentage of “Yes” responses across all respondents was calculated per each survey statement to determine the achievement level for each of the corresponding goals from TMMi. To ensure the consistency of assessment results the measurement scale defined in TAMAR was:

- Not Achieved – little of no evidence found of compliance. Process achievement score in the range from 0% to 15%;
- Partially Achieved – some evidence found of compliance, but includes weaknesses and is incomplete. Achievement score in the range from 16% to 50%;
- Largely Achieved – significant evidence found of compliance, but still includes some minor weaknesses. Achievement score in the range from 51% to 85%;
- Fully Achieved – consistent convincing evidence found of compliance. Process achievement score in the range from 86% to 100%.

After the maturity rating for each goal was obtained, the rating of each process area was determined. The rating of each process area is equivalent to the lowest rating of the goals that support the process area. The rating of the maturity level is equivalent to the lowest rating of the process areas that support the maturity level.

As a result of staff interviews, the achievement level of each goal was increased, decreased or retained, depending on the collected evidence. Based on interview results, the maturity levels of each process area were revisited and adjusted. Finally, an overall understanding of the organisational testing maturity was obtained.

3.5 Results

Assessment Survey Results. The survey received responses from 19 out of 21 Casino QA members, which is ~90% of the team. The mean time spent on filling the survey was approximately 35 minutes. Out of 49 survey statements, the minimum amount of “Yes” replies was 23 and maximum 45. There were no respondents with 0 “Yes” responses. A total of 4 respondents gave 0 of “No” responses. The widest response ranges and deviation values observed in the case of QA Engineers probably indicate that due to their shorter work experience, they are probably the least knowledgeable about the overall spectrum of test practices in the organisation. Although it would have seemed probable to see a correlation between the respondents’ role and their “Yes” percentage (e.g., more experienced team member giving higher amount of “Yes” answers), we did not identify any such correlation.

We determined the achievement level of each maturity level of TMMi based on the “Yes” responses of the statements in the conducted assessment survey. Table 2 represents the results according to which the achievement rating of the maturity level 2 is “not achieved” and that of the other maturity levels is “partially achieved”. Due to TMMi framework’s staged structure and the “not achievement” rating for maturity level 2, we can conclude that based on the assessment survey results, Playtech Estonia’s Casino unit is on maturity level 1.

Table 2. Achievement level of maturity based on the assessment survey

Maturity Level	Key Process Area	Achievement level	Achievement level of Maturity
2 - Managed	Test Policy and Strategy	not achieved	not achieved
	Test Planning	largely achieved	
	Test Monitoring and Control	partially achieved	
	Test Design and Execution	partially achieved	
	Test Environment	largely achieved	
3 - Defined	Test Organisation	largely achieved	partially achieved
	Test Training Program	fully achieved	
	Test Lifecycle and Integration	partially achieved	
	Non-functional Testing	largely achieved	
	Peer Reviews	largely achieved	
4 - Measured	Test Measurement	partially achieved	partially achieved
	Product Quality Evaluation	partially achieved	
	Advanced Reviews	partially achieved	
5 - Optimisation	Defect Prevention	largely achieved	partially achieved
	Quality Control	partially achieved	
	Test Process Optimisation	largely achieved	

Staff Interview Results. The evidence collected during the staff interviews was compared with the results from the assessment survey. All of the 17 statements that were excluded from the interview scope received 79% or higher score – 76.5% of them have a rating higher than 85%, which corresponds to the “fully achieved” achievement level, and 23.5% of them have a rating between 79% and 85%, which indicates the achievement level of “largely achieved”. Based on the gathered evidence, we can claim that the overall “Yes” response ratings and the achievement level derived from them for the corresponding statements are justified.

As a result of staff interviews, the achievement level of 7 out of 18 statements was changed – for 2 statements lowered and for 5 increased by one level. Process area achievement levels were recalculated and as a result two (Non-functional Testing, Peer Reviews) out of ten process areas received a lower achievement. This, however, did not change the overall maturity measurement of level 1 as obtained based on the results of the conducted assessment survey. Test policy and strategy remained our weakest process area and is the only one with a “not achieved” rating.

In terms of general questions regarding the clarity of the purpose and contents of the survey, the staff interview participants highlighted the following aspects:

- The fact that a “Yes” response should be selected when the process described by the survey statement takes place at least partially was missed by few. As a result, those respondents were overtly strict in their responses;
- The wording of several statements was believed to be ambiguous, grammatically difficult and, therefore, resulted in misunderstandings which were clarified during the staff interviews;
- A general concern was raised over the strict approach and requirements of TMMi framework, which assumes the existence of formal documentation and processes. That might have a counterproductive effect on the unit’s testing processes.

3.6 Threats to Validity

When planning and conducting the assessment, several potential threats to its validity were identified. The following threats were acknowledged:

- Assessment result objectiveness is under risk due to the assessor being also a member of the Casino QA team and due to the limitations of the assessment survey’s answering scheme. To mitigate the imposed risk, the assessment participants are chosen from all available roles within the Casino QA team. It was also explicitly emphasised to the respondents that there are no right or wrong answers and that answers should be based on the actual observations;
- Study limited only to one organisation might be insufficient to make wide-scale conclusions. The goal of this assessment is to evaluate the performance of the TMMi model from the perspective of Playtech Estonia’s Casino unit. We acknowledge this inherent risk of a limited empirical study;
- Selection of the assessment team might be potentially unilateral and therefore poses a threat of one-sided view on the process and activities. The current study is focused on test process assessment and therefore it seemed logical that the employees closest to the testing activities are part of the assessment team.

Involving people from outside the QA team might have broadened the spectrum of perception of the testing process and would have given more insight on the overall unit-wide knowledge of existing testing practices. However, due to the current development approach, we believe that employees outside the QA team are not aware of all the test process details that the TMMi model focuses on. So we decided to not include people from outside the QA team;

- Interpretation on achievement levels and the fulfilment of different goals might not be fully aligned with the expectations of the TMMi Foundation due to insufficient guidelines. On the other hand, it has been emphasised by the TMMi authors that it is only a collection of best practices as opposed to being a strict standard. Thus applying organisation can customise them according to its needs.

4 Test Process Improvement Planning Using TMMi

The goal of improvement planning is to identify improvement activities on the areas identified as weaker during the maturity assessment. Additionally, we evaluate the fitness of those improvements to the dynamic nature of business in Playtech. The long-term objective is to establish more efficient and productive test processes [10].

4.1 Improvement Scope and Improvement Proposals

When defining the current improvement scope, we were guided by the following aims: (i) keep the scope achievable in terms of its size, with small and clear step by step improvement action items; (ii) assess the suitability of the expected practices of TMMi in the context of the organisation; and (iii) determine the most beneficial ones. Thus, we focused on the less mature areas of TMMi level 2 and identified five goals across three process areas with the score of “partially achieved” or lower (Table 3).

Table 3. Least Mature TMMi Level 2 Goals and related Process Areas

Process Area	Survey statement/goal	Achievement level
Test Policy and Strategy	1.1. A test policy is defined and agreed upon with stakeholders within the Casino unit.	partially achieved
	1.2. A unit-wide test strategy is introduced and implemented. The test strategy identifies and defines the test levels to be performed.	partially achieved
	1.3. A set of goal-oriented test process performance indicators is defined and deployed. Those indicators, such as number of defects found or test coverage, measure the quality of test process.	not achieved
Test Monitoring and Control	3.2. Actual product quality is monitored against the quality measurements defined in the plan and the quality expectations of the customer/user or other stakeholders.	partially achieved
Test Design and Execution	4.2. During test design, the test procedures and test cases are developed and prioritised, including the intake test. Test data is created and the test execution schedule is defined.	partially achieved

For each of the five identified goals, we thoroughly analysed the expected practices and sub-practices listed in the TMMi model documentation to determine those that would suit our organisation's business objectives, culture and needs. We also consider the opinions of the QA team members who participated in assessment staff interviews. Additionally, we attempted to avoid introducing too much formality and bureaucracy into the currently existing work practices (e.g., instead of detailed test policy and strategy documents we proposed a lighter web-based approach).

4.2 Evaluation of Improvement Proposals through Staff Interviews

According to Wohlin *et al* [17] the process improvement proposals are very hard to evaluate without direct human involvement. For evaluating the proposed improvement activities, staff interviews were conducted. The same three Casino QA team members were involved as during the staff interviews during the TMMi based test process maturity assessment. The following questions were asked from the interviewees about the improvement recommendations for each of the TMMi goals:

1. Do you agree with the suggested adaption for Playtech Estonia's Casino unit?
2. Do you see any benefit in implementing this goal/requirement in the suggested way? Why or why not?
3. What other approach would you suggest in order to comply with this goal apart from or in addition to the suggested adaption?

As a result of the feedback obtained during staff interviews, we refined the improvement proposals for all TMMi goals, although for some the changes were rather small-scale. Test policy documentation was replaced with a lighter suggestion and a QA mission statement definition was proposed instead. Formal definitions of entry and exit criteria for test levels and suspension and resumption criteria for testing activities were removed. Decisions on starting and stopping testing activities were based on the currently actual situation, considering priorities, commitments, timelines, resource availability and so forth, so such document-based approach did not receive support as it was not considered to add any significant practical value. Test process performance indicators were defined (e.g. defect trends, progress measurements) and a regular analysis approach of the collected data was proposed with the aim of process improvement in order to: (i) reduce the amount of development and production defects, (ii) improve the on-time delivery capability; (iii) increase defect closure efficiency; (iv) increase testing efficiency; and (v) increase delivery quality. Throughout all the improvement proposals, emphasis was also placed on incorporating them into the new employee training program to contribute to the institutionalisation of the process, since this is the spirit of generic goals in TMMi.

At the end of each interview, one additional question was asked: "Which out of the five discussed TMMi goals would you consider the most beneficial in terms of efficiency and productivity improvement?" To our surprise, the goal represented by the statement 4.2 was chosen by all interviewees. Its practical value in terms of an efficient resource utilisation and detecting business-critical defects as early as possible in the project's lifecycle were highlighted as main reasons. We will consider this feedback when planning the implementation of improvement activities.

4.3 Threats to Validity

The following threats to validity should be considered when reviewing these results:

- Lack of cooperation from higher management poses a risk of improper focus and prioritisation of improvement activities. The process areas selected for improvement planning as part of the case study might not be the most important ones for the organisation. Therefore, during the staff interviews we focused on finding out which goals are believed to be the most beneficial and profitable in order to prioritise between the selected ones;
- Selection of the assessment team might be potentially unilateral and therefore poses a threat of one-sided view on the process and activities. Focus during the evaluation of the improvement proposals was only on the opinions of QA team employees. Stakeholders outside of the QA team could have been involved, especially representatives from Product Management side. Unfortunately, due to limited interest from higher management side, we decided to start using the bottom-up approach and involve the business side representatives once we have defined the baseline for the improvement process;
- Our interpretation of acceptable alternative practices to the ones described in TMMi might be inadequate. Although the TMMi framework emphasises that organisations should choose improvement areas based on their best judgement and needs, we could have consulted TMMi Foundation about the potential alternative practices per goal and whether our current selection of activities is acceptable from their side and compliant to TMMi framework requirements. Not doing so questions the reliability of the proposed improvement activities and might eventually cause us to not reach the desired maturity level. This will be addressed as part of the assessment summary that will be sent to TMMi Foundation as a result of the informal assessment and improvement areas identification.

5 Performance Evaluation of TMMi

The following section summarises the experiences gathered as a result of the empirical study conducted on Playtech Estonia's Casino unit using TMMi model. The view on whether and how the TMMi model could be improved is presented.

5.1 Evaluation of Performance and Suitability of TMMi

The comprehensive and detailed documentation of TMMi framework in terms of the reference model, assessment method and data submission requirements initially convinced us that it provides sufficient information and support in assessing and improving an organisation. However, when we reached the practical activities, some shortcomings and challenges were revealed.

Achievement level rating instructions are incomplete in TAMAR. TAMAR was lacking an essential part of information – even though an achievement rating scale was described in the document, there was no explanation on the minimum rating that must be obtained in order to achieve a goal or its related process area. TMMi

Foundation was contacted with the corresponding inquiry and as an outcome we learned that only “largely achieved” and “fully achieved” assessments result in a satisfied rating. This raised concerns over the completeness of TMMi documentation and whether there are any other important framework requirements which have not been explicitly mentioned and could mislead the users of the framework.

Documentation defects were noticed in TAMAR, which raised concerns over its reliability and correctness. TMMi Foundation was contacted and a confirmation was received that the issues raised were errors in the document and allegedly the corresponding info was also forwarded to the authors of TAMAR for correction. This, however, raised concerns regarding whether the quality improvement of our test process can be based on a framework that includes defects in its documentation.

The assessment method requirements are not straightforward in terms of achievement level calculation. One of the most difficult parts of the empirical study was to clearly define the TMMi based assessment method. We realised that TAMAR leaves room for interpretation for grading and even though it defines the requirements considered essential to TMMi based assessment methods, it does not provide any examples of accredited approaches. Considering no public availability of accredited methods, we can assume that TMMi follows strict accreditation rules for assessment methods to ensure that the calculation mechanisms are equivalent in all of them and there is no disparity in the assessments of various assessors.

TMMi reference model is lacking information on acceptable practices. TMMi model mostly focuses on the best practices that should be implemented without going into much detail on how they could be implemented. Hardly any references are provided concerning acceptable alternatives. This left a lot of room for interpretation regarding the alternative practices and suggested improvement activities, raising a risk of non-compliance to the actual expectations of TMMi.

The study of the TMMi model and the feedback received as part of staff interviews continuously brought up the seemingly bureaucratic nature of the framework – a lot of emphasis is put on formal documenting, reporting, logging, and recording activities as part of its expected practices. Concerns were raised by employees that the focus on such an amount of formalities might have a counterproductive effect and instead of increasing the efficiency will increase the efforts on maintaining the documentation. Formally documented strict processes might limit our delivery capability if following them is against internal practices and does not fit the actual situation.

The staged maturity model does not provide sufficient flexibility for the organisation. As a result of the assessment process, we noticed that we are more mature in many process areas of higher maturity levels than lower levels. It became clear that the maturity of process areas depends highly on the nature, business, and needs of the organisation. If a continuous model of TMMi had been available, we could have been able to define step-by-step approach to process improvement, taking into account the specific needs of the organisation on the relevant process areas.

In the light of the potential adoption of an agile development approach SCRUM [8] within the next six months, we decided to briefly also examine TMMi framework in the context of agile test process assessment and improvement:

- There is no direct reference to agile practices within the TMMi reference documentation, contrary to CMMI, which TMMi is largely based upon. In the

current form it is more suitable for organisations working according to more traditional models, such as the V-model [8];

- The balance between the formal test documentation expected by the TMMi model and the lack of focus on comprehensive documentation in agile practices is doubtful. The TMMi framework seems to put a lot of effort and emphasis in formal documentation as opposed to the Agile Manifesto, where working software is appreciated over comprehensive documentation.
- Test Organisation process area under TMMi level 3 emphasises the requirement of an independent test team and claims that testers should not be considered as developers and they should report to management independent of the development management. This, in the authors' opinion, conflicts with the concept of agile teams, where all the team members are considered as developers and should work together to create synergy between their roles and act as a self-organising team.

5.2 Improvement Possibilities for TMMi

The experience gained from the empirical study of TMMi brought attention to several aspects of the framework that could be improved:

- Unlike CMMI, TMMi is currently only available in a staged representation form. The availability of a continuous model would allow the organisation to pick the process areas that are believed to bring the greatest benefit and to arrange improvement activities based on capability levels rather than maturity levels;
- To facilitate the usage of TMMi framework for informal assessments by potentially inexperienced assessors, the existing contents should be improved: (i) the mistakes and shortcomings of assessment requirements document TAMAR should be fixed to avoid misinterpretation and to ensure reliability and correctness of the provided guidelines; (ii) more examples of acceptable alternatives to the expected practices represented in the TMMi reference documentation should be added; and (iii) emphasis of the formality of practices should be redirected to the practical activities themselves and their consistency;
- More focus on people and communication aspects should be incorporated within the model. We believe that this would help to improve the way the organisation communicate with its stakeholders, being able to talk in the language of quality and risk for making informed decisions and helping them to see value in practices such as test policy and strategy definition;
- To ensure that TMMi would also be the preferred test process improvement model in agile organisations, additional information should be added to the model on how to interpret the expected practices in agile context. Publishing agile-focused assessment method requirements is also a potential alternative.

6 Conclusions and Future Work

In the current paper we analysed how the TMMi helps to assess the test process maturity and what potentially could be improved during this process. More specifically we have conducted the empirical study [10] within the Playtech Estonia's Casino unit. This led to the following observations:

- *It is important to define detailed requirements from the organisation side to the tool and the expected benefits gained from its usage.* Our study showed that we were not sufficiently thorough when defining the required characteristics for an improvement model and that the lack of cooperation with the management hindered us from considering their expectations;
- *Only a few test process improvement models have comprehensive information available publicly and free of charge.* Even for informal assessments it might be necessary to invest in additional handbooks or services by an experienced assessor, which of course also carries significant financial cost;
- *Incomplete assessment method requirements and defects in the TMMi document examples were revealed, which is a concern for correctness and reliability.* As we observed, TAMAR does not provide an actual assessment method to use for achievement level calculation. It is possible that when an organisation decides to conduct an informal assessment in-house before an external formal assessment is performed, the probability of misinterpretation of expected practices might lead to a failed attempt to reach a certain formal maturity accreditation;
- *Understanding the goal for improvement is important.* The goals in our study were driven by the internal desires of the QA team and were not initiated by the management, posing a risk that the focus is not on the same areas as expected from the business stakeholders. During the course of the study, we discovered that a continuous test improvement model would have probably given us more flexibility when considering the opinion of employees and needs of the organisation in terms of increasing efficiency and productivity;
- *The availability of a continuous model would potentially allow for more flexibility to the organisation when choosing the improvement path.* As TMMi is only available in a staged representation, all of its process areas are strictly divided between maturity levels. A continuous model would allow the organisation to pick the process areas that are believed to bring the greatest benefit and arrange improvement activities based on capability levels rather than maturity levels;
- *Interpretation guidelines for agile environments should be incorporated to TMMi similar to CMMI.* Having in mind TMMi's promise to be usable regardless of the software development methodology in the organisation, more information on acceptable alternative practices should be provided as part of the reference model. The current focus of formality and bureaucracy can be misleading and overwhelming for an inexperienced model user, especially considering the fast nature of agile development approach. In the current format, TMMi emphasises several practices which seemingly conflict with the values of Agile Manifesto.

Finally, we do not believe that we made the wrong choice when choosing TMMi as the test process improvement model for a software development organisation. Although we identified various shortcomings of the performance of TMMi, considering the wide spectrum of organisations, their work methodologies and needs, the choice of the most appropriate model highly depends on their improvement objectives and financial capabilities. TMMi can definitely be considered as a valuable

source of best practices when planning improvements, however, as we saw, it includes weaknesses and there are also other models on the market to consider.

As for the future work, we plan to carry through the implementation of improvement suggestions defined as part of the process described in Section 4. Other more theoretical topics related to the area could be: (i) publish a continuous representation of TMMi; (ii) improve the model's applicability in agile environments; and (iii) definition of a test process improvement model selection framework.

Acknowledgments. We would like to thank Taavi Ilp for proof-reading the draft of this paper and for providing insights on the content improvement.

References

1. Black, R.: Critical Testing Processes: An Open Source, Business Driven Framework for Improving the Testing Process. Rex Black Consulting Services, <http://www.rbc-us.com/images/documents/critical%20testing%20processes.pdf> (last visited June 19, 2013)
2. Daughtrey, T.: Fundamental Concepts for the Software Quality Engineer. ASQ Quality Press (2001)
3. Eldh, S., Punnekkat, S., Hansson, H.: Experiments with Component Tests to Improve Software Quality. ISSRE, Industrial Track (2007)
4. Ericson, T., Subotic, A., Ursing, S.: TIM – A Test Improvement Model (1997), <http://www.lucas.lth.se/events/doc2003/0113A.pdf> (last visited June 19, 2013)
5. Gelperin, D., Hetzel, B.: The Growth of Software Testing. Communications of the ACM 31(6) (1988)
6. Gerrard Consulting: Test Organisation Maturity Questionnaire v2 (2013), <http://gerrardconsulting.com/tom/tom200.pdf> (last visited June 19, 2013)
7. ISO/IEC 15504-2:2003 Information Technology – Process Assessment, Part 2 – Performing an Assessment (2003)
8. International Software Testing Qualification Board: Standard Glossary of Terms Used in Software Testing Version 2.2, <http://www.istqb.org/downloads/finish/20/101.html> (last visited June 19, 2013)
9. International Software Testing Qualification Board: Certified Tester Expert Level Syllabus – Improving the Testing Process (2011), <http://www.istqb.org/downloads/finish/18/12.html> (last visited June 19, 2013)
10. Rungi, K.: Empirical Analysis of Test Maturity Model Integration (TMMi). Master thesis, University of Tartu (2013)
11. Software Engineering Institute: CMMI – Capability Maturity Model Integrated (2012), <http://cmmi.institute.com> (last visited June 19, 2013)
12. Sogeti Netherland, B.V.: TPI Next – Business Driven Test Process Improvement. UTN Publishers (2009)
13. Swinkels, R.: A Comparison of TMM and other Test Process Improvement Models. Technical Report, Frits Philips Institute (2000)

14. TMMi Foundation: TMMi Assessment Method Application Requirements (TAMAR) Version 2.0 (2009), <http://www.tmmi.org/pdf/TMMi.TAMAR.pdf> (last visited June 19, 2013)
15. TMMi Foundation: Test Maturity Model Integration (TMMi) Release 1.0 (2012), <http://www.tmmi.org/pdf/TMMi.Framework.pdf> (last visited June 19, 2013)
16. van der Ven, R.: Models to Improve your Test Process. Capgemini (2012), <http://www.nl.capgemini.com/expertise/publicaties/models-to-improve-your-testprocess/?d=7B2DF59E-1CBF-D9C9-C4D2-3DFAD200B65B> (last visited June 19, 2013)
17. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering. Springer, Heidelberg (2012)

Behavior Analysis of Real-Time Systems Using PLA Method

Dalius Makackas, Regina Miseviciene, and Henrikas Pranevicius

Kaunas University of Technology, Faculty of Informatics, Studentu 56-443, Kaunas, Lithuania
{dalius.makackas, regina.miseviciene,
henrikas.pranevicius}@ktu.lt

Abstract. The paper deals with a behavior analysis task of real-time system specified by the PLA method. An algorithm for creating a reachable state graph is used while solving for the task. The algorithm evaluates intervals of time when the defined system events occur. An approach based on the algorithm for the reachable state graph generation is presented within this paper. The suggested approach is illustrated by an example.

Keywords: real-time system, analysis, trajectory modeling, reachable state graph.

1 Introduction

A real-time system's accuracy depends not only on the logical result of computations, but also on the time at which the results are produced [2]. For a design of this type of system, high security, reliability and performance requirements are raised. Various formal methods for describing such systems functioning are applied. The most commonly used are following formal notations: Time Petri Nets [6-7, 19], Discrete Event System Specification (DEVS) [2, 5, 20], Timed Automata [1, 4, 8], Piece-linear Aggregate (PLA) [16, 18], Finite State Machine [21] and others. Such formal specifications of real-time systems can be analyzed by functionality or behavior. A functional analysis is performed by creating a system simulation model. A behavioral analysis examines all the possible trajectories of the system, while checking whether the specification is made correctly. As the real-time systems interact with their environment in real-time, time properties are very important. Thus, more recently, considerable research efforts have been devoted to verification of the time properties. There are various verification techniques [9, 10, 12, 15]. However, conventional verification methods do not perform a full inspection of real-time systems. Their main drawback is that the traditional verification methods underestimate the system performance over time, or analyzes only system whose operation time is deterministic. Functioning of such systems is described only by one trajectory. However, many operations of real-time systems depend on a certain interval. Thus, describing such systems by a single trajectory is not possible, infinitely many endings of the operation in time.

The operations may result in any precisely specified time interval. Thus, real time systems can have a number of operating trajectories. Verification of these trajectories is problematic, because of the need to generate and verify all the possible modes.

A goal of this article is to present a novel approach for creation of reachable state graph of operating trajectories for behavior analysis. The approach is based on the algorithm for the reachable state graph generation. The algorithm permits precise evaluation of specified time intervals for operations. The algorithm is designed for real-time systems specified by Piece-linear aggregate method [16].

The remainder of this paper is organized with the following approach. The next section provides a formal definition of Piece-linear aggregate; Section 2 describes real-time system functioning trajectories; Section 3 provides a reachable state graph creation algorithm; and an illustrative example is proposed in Section 4. Conclusions are presented in the last section.

2 Piece-Linear Aggregate Specification Method

The paper analyzes real-time systems, specified for by Piece-linear aggregate method [16].

A system specified by the Piece-linear aggregate method is understood as a set of interacting piece-linear aggregates. Each aggregate is defined by a set of states $Z = \{z_1, z_2, \dots\}$, a set of input signals $X = \{x_1, x_2, \dots\}$, a set of output signals $Y = \{y_1, y_2, \dots\}$, a set of internal E'' and external E' events, a set of transition $H : E \times X \rightarrow Z$ and output $G : E \times Z \rightarrow Y$ operators.

The aggregate method generates time-point sequences $T = \{t_0, t_1, \dots\}$ and state $\{z(t_0), z(t_1), \dots\}$ transitions in these time points. The state $z(t) = (v(t), z_v(t))$ consists of two components: discrete $v(t)$ and continuous $z_v(t)$. Each element $w_i(t)$ of a continuous component $z_v(t) = (w_1(t), w_2(t), \dots)$ indicates a time when an event e_i occurs. The event changes j elements of discrete and continuous component of state according to the law: $h_j^v(t) = h_j^v(t, z(t))$, $h_j^w(t) = h_j^w(t, z(t))$.

In the aggregate model it is also defined the concept of the operation. This function takes the following values:

$$O_e = O_e(t) = O(e, t) = \begin{cases} 1, & \text{it is active at time } t; \\ 0, & \text{it ended at time } t; \\ -1, & \text{it is pasive at time } t. \end{cases}$$

Each operation is linked with continuous component. If the operation O_e is active then value of continuous component is $w_e(t) > t$; if the operation is passive, it is not known when the next event will occur and continuous component is $w_e(t) < t$; if the operation is ended at time t then $w_e(t) = t$.

The Piece-linear aggregate specification method can be used for real-time system specification. This method is described in detail by Russian scientists N. Buslenko

and I. Kovalenko [3]. Professor H. Pranevicius proposed a modification by adding to the method control sequences, which built in comfortable assumptions of these models in computer systems realization. The Piece-linear aggregate specification is used for two purposes: to create simulation models and to validate and to verify the system. Validation and verification is based on creation of a reachable state graph. The essence of the reachable state method consists in the fact that, with the aggregate system specifications, the system generates a set of all possible trajectories. Then, the trajectories are analyzed in respect of properties under investigation.

3 Real-Time System Functioning Trajectories

Real-time systems are defined as follows: “It is an environment that responds to random external events. The respond to a particular event is a set of actions; each of them must be carried out in certain time constraints” [11, 13, 14].

Based on the definition, real-time system has the strict, fixed temporary conditions. The actions must be carried out under the defined conditions. Real-time systems are divided into two categories: real-time systems with strict requirements and real-time systems with probabilistic requirements. This article explores the systems with strict requirements. They must ensure that the appropriate actions will be carried out strictly within the prescribed time interval.

The system is investigating by analyzing functioning trajectories $S_0, e_1(I_1), S_1, e_2(I_2), S_2, \dots$, where I_i is time interval.

For example (Fig. 1), if the system contains two active operations O_1 and O_2 they can be ended by the relevant events e_1 and e_2 . The event e_1 can occur in the interval $I_1 = (t_i + \alpha_1; t_i + \beta_1)$ and the event e_2 - in the interval $I_2 = (t_i + \alpha_2; t_i + \beta_2)$.

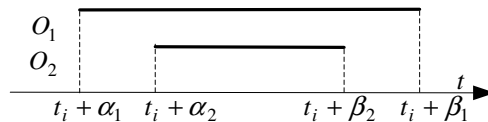


Fig. 1. Time intervals of active operations

In this case there are three time intervals:

1. If an event e_1 will occur at time $t_m \in (t_i + \alpha_1; t_i + \alpha_2)$, then the second event e_2 will occur at time $t_{m+1} \in (t_i + \alpha_2; t_i + \beta_2)$.
2. If an event e_1 will occur at time $t_m \in (t_i + \alpha_2; t_i + \beta_2)$, then the second event e_2 cannot occur before the first. The second event will occur at time $t_{m+1} \in (t_m; t_i + \beta_2)$.
3. If an event e_2 will occur at time $t_m \in (t_i + \alpha_2; t_i + \beta_2)$, then the first event e_1 will occur at the time $t_{m+1} \in (t_m; t_i + \beta_1)$.

There are three possible trajectories of system functioning:

- $S_0, e_1(t_i + \alpha_1; t_i + \alpha_2), S_1, e_2(t_i + \alpha_2; t_i + \beta_2), S_2;$
 - $S_0, e_1(t_i + \alpha_2; t_i + \beta_2), S_1, e_2(t_m; t_i + \beta_2), S_2,$ where $t_i + \alpha_2 < t_m < t_i + \beta_2;$
 - $S_0, e_2(t_i + \alpha_2; t_i + \beta_2), S_1^*, e_1(t_m; t_i + \beta_1), S_2^*,$ where $t_i + \alpha_2 < t_m < t_i + \beta_1;$
- Graphically this is illustrated in a tree-like structure (Fig. 2).

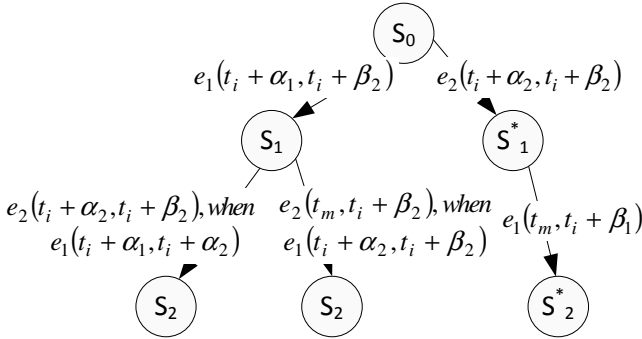


Fig. 2. Tree-like structure of example

4 Reachable State Graph Creation Algorithm

All functioning trajectories must satisfy the following statements. All the statements are proven in [17].

Statement 1. If $w_e(t)$ can take any value in the interval (α, β) , then an event e can occur at any time $t_m \in (\alpha, \beta)$.

Statement 2. If the system is at the state s , then the next event e_i will occur at time $t \in (\min_i \alpha_i, \min_i \beta_i)$. According to this definition (Fig. 3) $\alpha = \min_{1 \leq i \leq n} \alpha_i$ and $\beta = \min_{1 \leq i \leq n} \beta_i$.

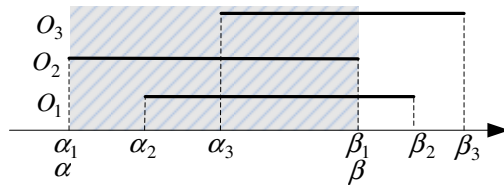


Fig. 3. Graphic depiction of operation ending intervals

Fig. 3 shows that in the interval (α_1, α_2) may finish only operation O_2 ; in the interval (α_2, α_3) - operations O_1 and O_2 ; in the interval (α_3, β_1) - operations O_1, O_2 and O_3 . In this case $\alpha = \alpha_1$ and $\beta = \beta_1$.

Statement 3. Suppose that in a state s at time t' the operation O_j was active. If at the end of the operation O_i ($i \neq j$) at time t_m operator $H(e_i)$ did not change the continuous component $w_j(t)$, then the system will move to a state where the continuous component $w_j(t)$ satisfies the condition: $\max\{t_m, \alpha_j\} < w_j(t_m) < \beta_j$.

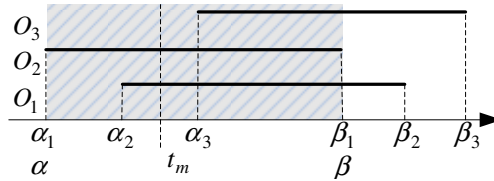


Fig. 4. Active operations range adjustment

Statement 4. The newly generated operation may fall either outside or inside of the relevant range (α, β) (Fig. 5). The earlier mentioned definitions should be evaluated in the both intervals.

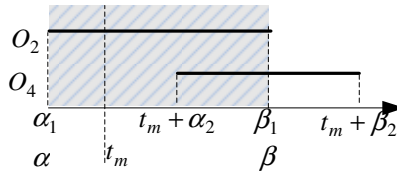


Fig. 5. The newly generated operation falls inside of the relevant interval

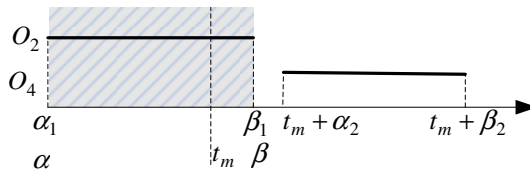


Fig. 6. The newly generated operation falls outside of the relevant interval

According to these definitions, a state graph is formed according to algorithm presented in Fig. 7.



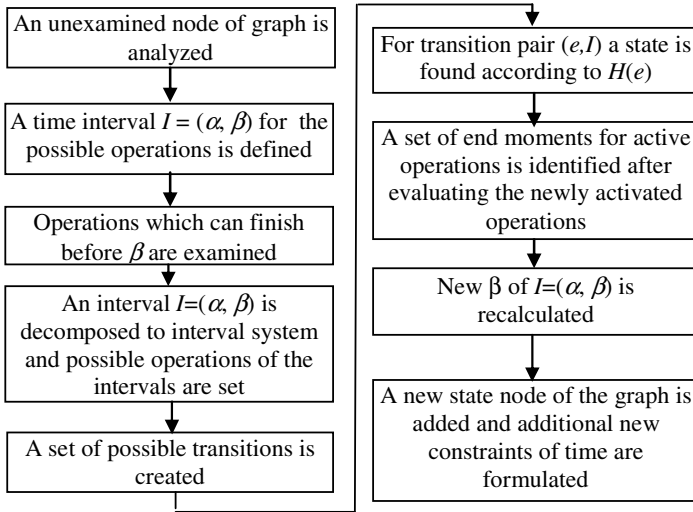


Fig. 7. A flowchart of the state graph analysis algorithm

5 Reachable State Graph Creation Example

A service system consists of one input and two service devices (Fig. 8). Service application messages, arriving to the system, are placed in a queue. When one of the devices becomes available for the service the message is passed to him. If both devices are available, the message is transmitted to the first device.

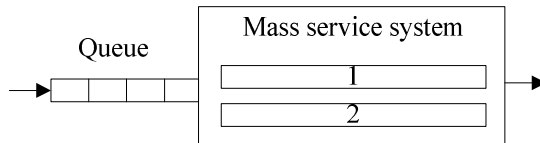


Fig. 8. Two-channel mass service system

The system specification consists of the components:

- a set of inputs $X = \emptyset$ and a set of outputs $Y = \emptyset$;
- a set of events $E = E' \cup E''$, where $E' = \emptyset$; $E'' = \{e_1, e_2, e_3\}$, e_1 - a new message arrived, e_2 - a first channel service is completed, e_3 - a second channel service is finished;
- controlling sequences $e_1 \mapsto \alpha_0, \alpha_1, \alpha_2 \dots$, $e_2 \mapsto \beta_0, \beta_1, \beta_2 \dots$, $e_3 \mapsto \gamma_0, \gamma_1, \gamma_2 \dots$;
- a discrete component $v(t) = (n(t))$, where $(n(t))$ is a number of messages in a queue.
- a continuous component $z_v(t) = (w(e_1, t), w(e_2, t), w(e_3, t))$;
- a parameter s - is a maximum length of the queue.



- time limitations on the duration of operations are these: $4 < \alpha_i < 6$, $3 < \beta_i < 5$, $2 < \varphi_i < 4$, $\forall i = 1, 2, \dots$.

Transition operators are as follows:

$H(e_1)$:

$$n(t_m) = \begin{cases} n(t_m - 0) + 1, & t_m < w(e_2, t_m - 0) \wedge t_m < w(e_3, t_m - 0) \wedge n(t_m - 0) < s; \\ 0, & \text{otherwise;} \end{cases}$$

$$w(e_1, t_m) = t_m + \alpha_m;$$

$$w(e_2, t_m) = \begin{cases} t_m + \beta_m, & t_m > w(e_2, t_m - 0); \\ w(e_2, t_m - 0), & \text{otherwise;} \end{cases}$$

$$w(e_3, t_m) = \begin{cases} t_m + \gamma_m, & t_m < w(e_2, t_m - 0) \wedge t_m > w(e_3, t_m - 0); \\ w(e_3, t_m - 0), & \text{otherwise;} \end{cases}$$

$H(e_2)$:

$$n(t_m) = \begin{cases} n(t_m - 0) - 1, & n(t_m - 0) > 0; \\ 0, & \text{otherwise;} \end{cases}$$

$$w(e_2, t_m) = \begin{cases} t_m + \beta_m, & n(t_m - 0) > 0; \\ w(e_2, t_m - 0), & \text{otherwise;} \end{cases}$$

$H(e_3)$:

$$n(t_m) = \begin{cases} n(t_m - 0) - 1, & n(t_m - 0) > 0; \\ 0, & \text{otherwise;} \end{cases}$$

$$w(e_3, t_m) = \begin{cases} t_m + \gamma_m, & n(t_m - 0) > 0; \\ w(e_3, t_m - 0), & \text{otherwise;} \end{cases}$$

A generation of a reachable state graph is carried out in accordance to the algorithm presented in Fig. 7.

Step 1. The generation of the reachable state graph starts from the initial state. The state S consists of three components: a discrete component $\nu(t)$, a continuous component $z_\nu(t)$ and a set of time constraints R :

$$:1: (0; (t_0 + 4, t_0 + 6), \emptyset, \emptyset; R_0), \text{ where } R_0 = \emptyset.$$

The first interval $I = (\alpha, \beta)$ is defined according to formulas $\alpha = \min\{t_0 + 4\} = t_0 + 4$ and $\beta = \min\{t_0 + 6\} = t_0 + 6$ (Fig. 9). Operations, which may finish in the interval first of all, are found. According to PLA specification only one operation O_1 is active. Since there is only one operation, using a transition operator $H(e_1)$ we find the next state: $(0; (t_1 + 4, t_1 + 6), (t_1 + 3, t_1 + 5), \emptyset)$.

Check if the new activated operation will not end earlier than $\beta = t_0 + 6$. Since the condition is satisfied $\min\{t_0 + 6, t_1 + 3\} = t_0 + 6 = \beta$, a new activated operation can not finish before the examined interval. The next state is as follows:

$$:2: (0; (t_1 + 4, t_1 + 6), (t_1 + 3, t_1 + 5), \emptyset; R_{11}), \text{ where } R_{11} = R_0 \cup \{t_0 + 4 < t_1 < t_0 + 6\}.$$

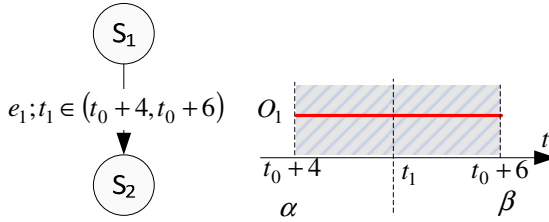


Fig. 9. Reachable state graph fragment $S_1 \rightarrow S_2$ and a transition e_1, t_1

Step 2. The next interval $I = (\alpha, \beta)$ is defined by formulas $\alpha = \min\{t_1 + 3, t_1 + 4\} = t_1 + 3$ and $\beta = \min\{t_1 + 5, t_1 + 6\} = t_1 + 5$.

Operations which may finish in the interval first of all, are found. They are two operations (O_1 and O_2). The interval $I = (\alpha, \beta)$ is separated (Fig. 10) into two intervals $\{t_1 + 3, t_1 + 4\}$ and $\{t_1 + 4, t_1 + 5\}$.

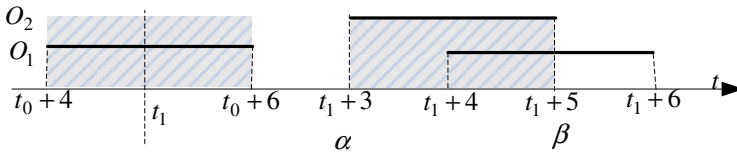


Fig. 10. Separated two intervals $\{t_1 + 3, t_1 + 4\}$ and $\{t_1 + 4, t_1 + 5\}$

The possible transitions there are three:

$$(e_2, t_2 \in (t_1 + 3; t_1 + 4)), (e_2, t_2 \in (t_1 + 4; t_1 + 5)), (e_1, t_2 \in (t_1 + 4; t_1 + 5))$$

Step 2.1. A transition ($e_2, t_2 \in (t_1 + 3; t_1 + 4)$) is analyzed first of all (Fig. 11).

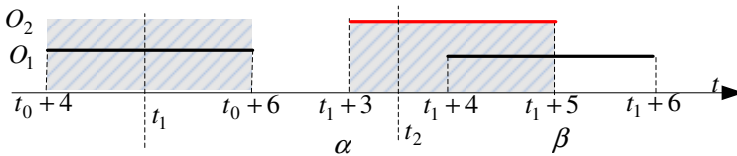


Fig. 11. An event $(e_2, t_2 \in (t_1 + 3; t_1 + 4))$

Using a transition operator $H(e_2)$ the next state is defined: $(0; (t_1 + 4, t_1 + 6), \emptyset, \emptyset)$. Since the operation O_1 after the event remained active, we have to recalculate the end of the interval in such a way: $(\max\{t_2, t_1 + 4\}; t_1 + 6) = (t_1 + 4, t_1 + 6)$.

The third state is as follows (Fig. 12):

$$3: (0; (t_1 + 4, t_1 + 6), \emptyset, \emptyset; R_{21}), \text{ where } R_{21} = R_{11} \cup \{t_1 + 3 < t_2 < t_1 + 4\}$$

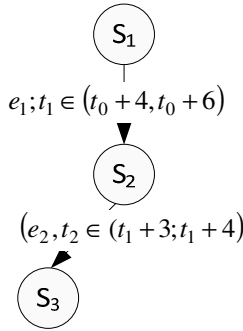


Fig. 12. A fragment of the reachable state graph (a transition $S_2 \rightarrow S_3$)

Step 2.2. The second transition $(e_2, t_2 \in (t_1 + 4; t_1 + 5))$ is analyzed next (Fig. 13).

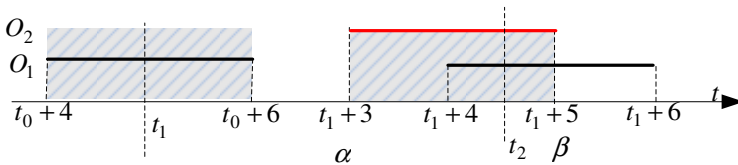


Fig. 13. An event $(e_2, t_2 \in (t_1 + 4; t_1 + 5))$

Using transition a transition operator $H(e_2)$ the next state is defined: $(0; (t_1 + 4, t_1 + 6), \emptyset, \emptyset)$. Since an operation O_1 after the event remained active, we have to recalculate the end of the interval in such a way: $(\max\{t_2, t_1 + 4\}, t_1 + 6) = (t_2, t_1 + 6)$.

The forth state is as follows (Fig. 14):

$$4: (0; (t_2, t_1 + 6), \emptyset, \emptyset; R_{22}), \text{ where } R_{22} = R_{11} \cup \{t_1 + 4 < t_2 < t_1 + 5\}$$

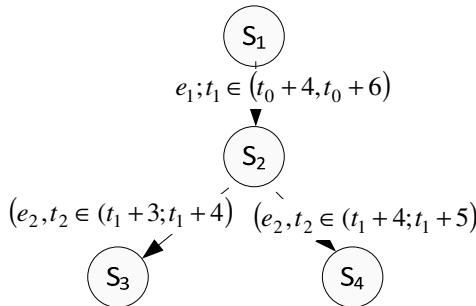


Fig. 14. A fragment of the reachable state graph (a transition $S_2 \rightarrow S_4$)

Step 2.3. The third transition $(e_1, t_2 \in (t_1 + 4; t_1 + 5))$ is analyzed next (Fig. 15).

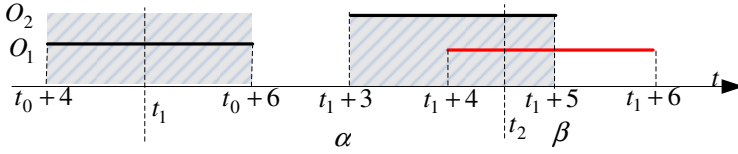


Fig. 15. An event $(e_1, t_2 \in (t_1 + 4; t_1 + 5))$

Using the transition operator $H(e_1)$, the next state is defined: $(0; (t_2 + 4, t_2 + 6), (t_1 + 3, t_1 + 5), (t_2 + 2, t_2 + 4))$.

Since an operation O_2 after the event remained active, we have to recalculate the end of the interval in such a way: $(\max\{t_2, t_1 + 3\}, t_1 + 5) = (t_2, t_1 + 5)$. The fourth state is as follows (Fig. 16):

$$S: (0; (t_2 + 4, t_2 + 6), (t_2, t_1 + 5), (t_2 + 2, t_2 + 4); R_{23}), \quad \text{where}$$

$$R_{23} = R_{11} \cup \{t_1 + 4 < t_2 < t_1 + 5\}.$$

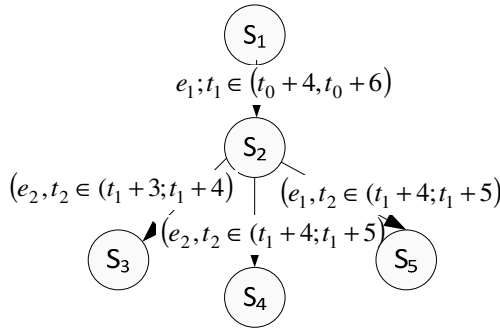


Fig. 16. A fragment of the reachable state graph (a transition $S_2 \rightarrow S_5$)

6 Conclusions

Conventional verification methods do not perform a full analysis of real-time systems as the traditional verification methods underestimate the system performance over time, or analyze only system whose operation time is deterministic. However, many operations of real-time systems depend to a certain interval and may result in any precisely specified time interval.

This paper presents a novel approach for creation of a reachable state graph. While creating the reachable state graph an algorithm is used. The algorithm permits to evaluate intervals of time when the defined system events occur. When the reachable state graph is made then various properties can be verified: dead ends, inefficient cycles, reachability and so on.

Acknowledgements. The work described in this paper has been carried out within the framework the Operational Programme for the Development of Human Resources

2007-2013 of Lithuania „Strengthening of capacities of researchers and scientists“ project VP1-3.1-ŠMM-08-K-01-018 „Research and development of Internet technologies and their infrastructure for smart environments of things and services“ (2012- 2015), funded by the European Social Fund (ESF).

References

1. Aceto, L., Bouyer, P., Burgueno, A., Larsen, K.G.: The power of reachability testing for timed automata. *Theoretical Computer Science* 300, 411–475 (2003)
2. Bonhomme, P.: Scheduling and control of real-time systems based on a token player approach. *Discrete Event Dynamic Systems* 23, 197–209 (2013)
3. Buslenko, N.P., Kalashnikov, V.V., Kovalenko, I.N.: *Lectures on the Theory of Complex Systems*. Sov. Radio, Moscow (1973) (in Russian)
4. Dang, Z., Ibarra, O.H., Kemmerer, R.A.: Generalized discrete timed automata: decidable approximations for safety verification. *Theoretical Computer Science* 296, 59–74 (2003)
5. David, R., Alla, H.: On hybrid Petri nets. *Discrete Event Dynamic Systems* 11, 9–40 (2001)
6. Ding, Z., Jiang, C., Zhou, M.: Design, Analysis and Verification of Real-Time Systems Based on Time Petri Net Refinement. *ACM Transactions on Embedded Computing Systems (TECS)* 12 (2013)
7. Ghomri, L., Alla, H.: Modeling and analysis using hybrid Petri nets. *Nonlinear Analysis: Hybrid Systems* 1, 141–153 (2007)
8. Gómez, R.: Model-checking timed automata with deadlines with Uppaal. *Formal Aspects of Computing* 25, 289–318 (2013)
9. Halbwachs, N., Poy, Y.E., Roumanoff, P.: Verification of real-time systems using linear relation analysis. *Formal Methods in System Design* 11, 157–185 (1997)
10. Knorrack, D., Apvrille, L., Pacalet, R.: Formal system-level design space exploration. *Concurrency and Computation: Practice and Experience* 25, 250–264 (2013)
11. Kopetz, H.: *Real-time systems: design principles for distributed embedded applications*. Springer Science+ Business Media (2011)
12. Krena, B., Vojnar, T.: Automated formal analysis and verification: an overview. *International Journal of General Systems* 42, 335–365 (2013)
13. Laplante, P.A.: *Real-Time Systems Design and Analysis*. Wiley-IEEE Press (2004)
14. Lorin, H., Deitel, H.M.: *Operating Systems*. Longman Higher Education (2009)
15. Mekki, A., Ghazel, M., Toguyeni, A.: Validation of a New Functional Design of Automatic Protection Systems at Level Crossings with Model-Checking Techniques. *IEEE Transactions Intelligent Transportation Systems* 13, 714–723 (2012)
16. Pranevicius, H.: Complex systems formalization and analysis (in Lithuanian). *Technologija, Kaunas* (2008)
17. Pranevicius, H., Raudys, S., Rudzionis, A., Ratkevicius, K., Sakalauskaite, J., Makackas, D.: Agent system models. *Mokslo aidai, Vilnius* (2008)
18. Pranevicius, H., Miseviciene, R.: Verification of piece-linear aggregate specifications. *Kaunas, Technologija* (2006)
19. Renganathan, K., Bhaskar, V.: Performance evaluation and model checking in systems modeled as Hybrid Petri nets. *Applied Mathematical Modelling* 36, 3941–3947 (2012)
20. Saadawi, H., Wainer, G.: Principles of Discrete Event System Specification model verification. *Simulation* 89, 41–67 (2013)
21. Yin, Y., Liu, B., Ni, H.: Real-time embedded software testing method based on extended finite state machine. *Systems Engineering and Electronics* 23, 276–285 (2012)

Measuring the Performance of Process Synchronization with the Model-Driven Approach

Vladislav Nazaruk and Pavel Rusakov

Riga Technical University, Latvia

{vladislaavs.nazaruks,paveis.rusakovs}@rtu.lv

Abstract. Concurrent and parallel computing can be used to effectively speed up computations. However, the overall performance gain depends on the way how the concurrency has been introduced to the program. Considering process synchronization as one of the most important aspects in concurrent programming, this paper describes the possibility of application of the model-driven approach in the concept of a software framework for measuring the performance of process synchronization algorithms. Such framework could help determine and analyze the features of specific synchronization techniques in different cases, depending on different input parameters.

Keywords: concurrent computing, process synchronization, performance of algorithms, model-driven approach.

1 Introduction

Concurrent computing is defined as a form of computing which supposes that a computational unit (i. e., a computer program) is divided into several computational units (processes, threads) which can be executed not only sequentially, but also simultaneously (or concurrently). The main purpose of such dividing is to be able for the system to execute other units of a program while one (or some) of units of this program is idle (i. e., waiting for some external event). This means that the system is able to dynamically reallocate processor time between threads when some thread becomes idle. Due to given advantages concurrent computing nowadays has become sufficiently widespread. [7]

Such processor time reallocation allows to decrease the time needed for the program to execute, compared to equivalent non-concurrent program. In addition, on systems with more than one processing unit in them (for example, in a multi-processor system or in systems with a multi-core processor), the presence of several threads allows to execute more than one thread at each time moment — this is also the opportunity to decrease execution time of a program. Therefore the main goal of using concurrent computing is to reduce the execution time of a program or, in other words, to speed up the computations. [1]

One of the main aspects of concurrent computing is that in order to accomplish a corporate goal for a program, its threads should communicate with each other in a

certain way. All communications in a concurrent program are done by means of inter-process (or inter-thread) communication mechanisms. Process synchronization is one of most important types of inter-process communication; its aim is to assure a specific coherence of execution of actions between several threads (or processes).

As process synchronization mechanisms control the execution of threads, these mechanisms, operated by their own logic, pause and resume different threads. Such intermediation of synchronization mechanisms in the execution of other parts of a program obviously impacts the execution speed. As there exist a number of different process synchronization algorithms, and the same concurrent computational task can be implemented in different ways by using different synchronization algorithms, the analysis of the impact of process synchronization mechanisms on the execution speed of concurrent programs, and the development of the corresponding model-driven framework are fairly topical issues; they are in the focus of this paper.

The goal of this paper is to show the possibility of application of the model-driven approach in the measurement of the performance of process synchronization, by defining a basic concept of a software framework for measuring such performance. This framework will be abbreviated as PSPMF throughout the paper. The performance is opposite to the rate, how much the synchronization itself influences the execution time of the program. Such framework can help determine and analyze the features of specific synchronization techniques in different cases, depending on different input parameters: a number of threads, a pattern of thread interaction (including both possible thread interaction mechanisms and specific cases of thread interaction), a pattern of waiting for external events etc.

In the paper, there are suggested and analyzed methods how an impact of process synchronization mechanisms (or, in other words, the performance of synchronization mechanisms) on the execution speed for different situations could be measured and interpreted. By applying these methods in practice, for example, it would be possible to predict some issues concerning the overall performance of a given concurrent program when using different synchronization algorithms; this, in its turn, can help develop guidelines for selecting more appropriate process synchronization algorithms in specific or more general situations.

In Section 1, there is given an introduction to the research. Section 2 shows the difference of the current work comparing it to related works. In Section 3, the context of the performance measuring framework is described. In Section 4, there is described a model of a concurrent system, as well as the Model Driven Architecture which can be used for implementing the framework. Section 5 shows how data obtained by the framework can be used further. In Section 6, the detailed specification of the part of the framework is given. Section 7 gives conclusions and defines some aspects of possible further work.

2 Related Works

The measurement of the performance of process synchronization is not a new area of study. For instance, there exist many scientific works, including [10], [11], which

describe the issue of measuring the performance of inter-process communication algorithms.

There are a number of works as well which propose the use of frameworks for measuring the performance of specific systems. For example, the works [8], [9] propose framework for building a specific benchmark suite and a framework for multi-processor performance characterization.

However, this paper shows an idea of applying a model-driven approach for building a framework to measure the performance of process synchronization algorithms. The authors believe that such a method allows building a more flexible platform for analysing different process synchronization algorithms.

3 Context of the Framework

Before specifying the design of the framework, possible use cases and other requirements for it are needed to be defined. In this section, such requirements are identified and described.

Let CS represent any concurrent system. If we consider CS as a black box, then this system can be defined by a specific transformation (T) which transforms given input data (I) to specific output data (O):

$$CS(I; O): I \xrightarrow{T} O. \quad (1)$$

As CS is defined as a concurrent system, it is assumed that this system is multi-process (or multi-threaded), and it uses process synchronization in order to coordinate execution of the processes. Let the transformation T itself be implemented in CS by means of several processes P_i . Then the task of the PSPMF can be formally defined as following: to measure the expenses of process synchronization in a specific execution case of CS , providing as a result a specific numerical value.

By saying “in a specific execution case”, there is meant that the result of the measurement can be different in different execution cases. The possible difference can be induced by the following factors:

- there are changes in input data I ,
- the system depends not only on its input data I , but also on its persistent internal state, which has been changed;
- the system is executed in a different environment;
- the system or its execution environment is non-deterministic.

Typical ways of using the measurement results need the results obtained from several concurrent systems with identical resulting behaviour of transformations: i. e. if $T(CS_1) = \dots = T(CS_k)$ for a set of concurrent systems CS_i . In other words, this means that it has sense only to compare the performance of synchronization algorithms when they solve one and the same task, but using different strategies and their implementations for the synchronization.

The main use case of the results obtained by the use of the PSPMF is to see, which concurrent system (combining with its environment) can afford better performance. Moreover, the results should not only be used for arranging concurrent systems in a specific order, but also should give quantitative information about the degree of difference between the measurements.

Therefore, the framework should be reliable enough in order to make acceptable conclusions despite possible influencing factors mentioned above. In order to reach that, these factors are to be described now.

The first factor (changes in input data) is a natural factor for possible change in the result of the performance measurement. Moreover, this factor can affect measurement results significantly. In order to minimize the possible effect of this factor, the comparison of measurement results should only be done for systems with the same input data. Therefore, the framework should be used to analyze concurrent systems of the same behaviour, and only on the same input data set. Therefore (considering the framework is denoted by F), the framework should take an input data I as a parameter, as well as the concurrent system CS itself:

$$F(CS; I): \langle CS; I \rangle \rightarrow M, \quad (2)$$

where M is a result of a measurement.

The second possible influencing factor is a presence of an internal state of a system, which is stored in the system between its executions, and which can also affect the output. Although we can simplify the situation and introduce the restriction for the PSPMF that the system should not have its persistent states, we will simply consider the value of this state at the moment when the system is started as a part of input data. Then, when using the framework, it is needed to compare concurrent systems not only with equal inputs, but also with equal initial states.

The next influencing factor is the possibility of executing the system in a different execution environment. This factor can be considered from different perspectives, depending on the abstraction level of the model of the concurrent system (starting from abstract model and ending with its specific implementation in a specific execution environment – on a specific software and hardware platform). Therefore, the full PSPM framework should be able to handle concurrent systems specified on different abstraction levels – for each abstraction level of the model of the concurrent system being analyzed. This implies that the framework should be organized in several layers. The organizational structure of the framework, which depends on such abstraction levels of the system model, as well as these levels themselves, in more detail are discussed in Section 4.

The fourth influencing factor, that the system or its execution environment can be stochastic, is one of the main issues needed to be resolved. One of the possible ways how to obtain more or less constant and objective measurement results is to use Monte Carlo method [2]. However, there is needed a deeper analysis to identify possible influencing parameters of stochastic events and their characteristics. For the rest contents of the paper, all systems being mentioned, as well as their execution environments, are to be considered non-stochastic, i. e. deterministic.

4 Model of a Concurrent System

The most abstract level of a system model is a mathematical model which formally and fully describes the system, but at the same time not considering any implementation details specific to any execution environment (or software and hardware platforms). The least abstract level of a system model is an implementation specific model, which fully describes not only the logic of the system, but also the execution environment of the system.

Depending on the necessity, any number of intermediate models of a system can be introduced. Such flexible granularity of model abstraction levels is reasonably important especially in the context of the performance analysis of process synchronization algorithms.

This can be explained in the following way. From one point of view, all algorithms (including all process synchronization algorithms) are primarily defined in an abstract way and therefore irrespectively to their specific implementation. However, even such their definition lets them to be analyzed thoroughly, usually by finding a minimum, a maximum, an average values, and a standard deviation of values of a given criterion (for example, a memory use, a time needed for an algorithm to complete) [3]. Therefore, some kind of analysis of process synchronization algorithms could be based only on the most abstract model of the corresponding concurrent system.

From another point of view, the main role in the performance of a system is played by the execution environment. This is because exactly the execution environment determines precise expenses for performing each operation of a system algorithm. In other words, the execution environment defines a mapping m , which maps each type of an atomic operation o to its scale $m(o)$ (for example, the execution time of the operation).

Let us assume that there is only one control flow in the system S (i. e., the system consists of only one thread). Let us have a specific estimation (it is not important here, is it a minimum, a maximum or an average estimation) of operations to be performed during the execution of S : an operation $o_i \in S$ will be performed $w(o_i)$ times. Then we can make an estimation M of a specific performance criteria (defined by m) of the whole system S using the following formula:

$$M(S) = \sum_{o_i \in S} m(o_i) \cdot w(o_i). \quad (3)$$

Equation (3) gives a *static* evaluation of performance, i. e. to obtain the evaluation, there is no need to have knowledge of an execution environment when analyzing the system on a more abstract layer of it. In other words, a static evaluation of performance here is dependent on the evaluation obtained at the previous level of abstraction (this corresponds to evaluations of $w(o_i)$, which for a single-threaded system can be made at the higher abstraction level).

However, if we have a concurrent system, then (3) or a similar equation cannot be used anymore in order to *statically* estimate the performance: now $w(o_i)$ becomes

dependent on $m(o_j)$ ¹, and, therefore, it is impossible to determine $w(o_i)$ at a higher abstraction level, without the knowledge of an execution environment. In other words, now we are used to have a *dynamic* estimation, i. e. we need *to execute the system* in its environment (it does not matter, whether it is a real or a simulated environment).

This conclusion ends the explanation of necessity of models of a concurrent system at different levels of abstraction.

Therefore, the PSPM framework should be able to:

- specify a concurrent system in an abstract way – i. e. without any reference to a possible execution environment;
- specify all possible execution environments of a specific concurrent system;
- to the extent possible automatically detail a model of a higher abstraction level to a lower abstraction level (i. e. automate making a model more specific).

These three above mentioned desires regarding to the framework perfectly fit to a philosophy of the Model-Driven Architecture (MDA). MDA is defined as “an approach to IT system specification that separates the specification of functionality from the specification of the implementation of that functionality on a specific technology platform” [4].

MDA specifies three default models of a system corresponding to the three viewpoints on a system: computation independent, platform independent and a platform specific [12]. An implementation specific viewpoint could be thought as a low-level program source code for a specific system.

MDA is based on model transformations—the process of converting one model to another within the same system—starting from PIM to ISM [12] (see Figure 1).



Fig. 1. Model transformations in MDA

Table 1 shows the possible connection between MDA models and specific abstraction levels of models of data stored in the PSPM framework. This correspondence is to be developed and implemented in the further research.

¹ To be more precise, we can imagine that for each thread in a concurrent system we have an additional operation o_{wait} : to wait (for a specific event). As each thread cannot predict, how many times and when exactly it can be put into the waiting state (and possible answers to this questions are absolutely not predictable and mostly depend on the external environment), an adequate estimation of $w(o_{wait})$ cannot be obtained without knowledge about the execution environment.

Table 1. Connections between MDA models and data stored in PSPMF

Model	Contains information
Computation Independent Model (CIM)	—
Platform Independent Model (PIM)	Logic of the system described in an abstract language
Platform Specific Model (PSM)	PIM + all details how the system will use specific features of execution environment
Implementation Specific Model (ISM; source code)	Source code that can be straight executed on a machine
Platform Model	Detailed information about execution environments: software, hardware and other technological platforms

5 Further Application of Data Collected by Framework

When the framework is used to collect a known amount of data concerning the performance of process synchronization algorithms for a number of different situations, these data are needed to be analysed. The aim of such analysis is to discover patterns that are specific to these collected data. Thus, *data mining* techniques should be applied.

Data mining can help discover weak and strong sides of a specific process synchronization algorithm. The advantage of applying data mining techniques to data obtained by the multi-level framework is supposed to be the interconnected performance tendencies for the same algorithm at different abstraction levels.

6 Detailed Specification of the Framework

In this section, the part of PSPM framework that is not connected to model-driven approach is being specified in detail. There are described several performance metrics, factors the metrics can depend on, and some general considerations about the ways of measurement. This section is based on the authors' work [7].

6.1 Performance Metrics

In order to measure the performance of process synchronization algorithms, it is necessary to introduce some requisite metrics, or measures of some properties, of a concurrent program.

For ease of describing, let all metrics be divided into two classes: observational metrics and analytical metrics. Observational metrics are metrics which are obtained by direct measurement of properties; and analytical metrics are metrics obtained by calculations over observational and/or other analytical metrics.

Let performance metrics also be divided into two classes, depending on their scope: thread-level metrics and system-level metrics. Thread-level metrics are specific to each thread of a concurrent program, and are obtained considering the

corresponding thread outside the context of other threads. On the contrary, system-level metrics are specific to a program in general.

Before discussing specific metrics, some general assumptions which will allow formalizing the results, should be formulated. They are the following [7]:

- given a computer program, all metrics are affected only by some input parameters (defined later in this paper) — that is, they are not affected by some random factors;
- the problem is finite (and, therefore, completion time of the program is also finite);
- all threads start their execution simultaneously.

Now let us consider thread-level metrics. In Figure 2, there is schematically shown lifetime of a thread. One can define two principal possible states of a thread: performing effective actions (s_{eff} — all actions a thread performs in isolation) and synchronizing with other threads (s_{syn}). These states usually alternate with each other; and the alternation could occur an unlimited number of times. The s_{syn} state has its two sub-states: executing instructions needed to provide the synchronization ($s_{\text{syn:ex}}$) and waiting for synchronization to complete ($s_{\text{syn:wait}}$).

Let us assign to each state s_a the total time t_a the thread is in this state. Analyzing the thread-level metrics, it is clear that for each thread effective time t_{eff} is constant; however, $t_{\text{syn:ex}}$ and $t_{\text{syn:wait}}$ are both variable, and for better performance they should be minimized. It is also important to say that the time $t_{\text{syn:ex}}$ is dependent only on the synchronization mechanisms (including their implementation) used in the thread; however $t_{\text{syn:wait}}$ is dependent both on the synchronization mechanisms (excluding their implementation) used in the thread, and as well as the behaviour of other threads — which is far less predictable. [7]

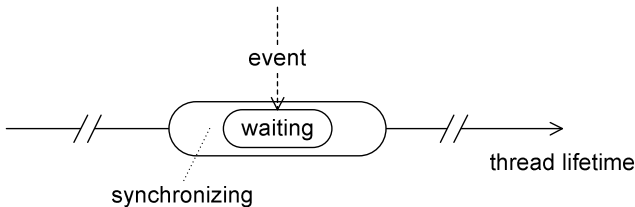


Fig. 2. Thread lifetime

Considering system-level observational metrics, there could be defined overall execution time (T_{ex}) — a time interval between the beginning and the end of execution of the entire concurrent program. Other system-level metrics could be defined as follows [7]:

- total effective time (T_{eff}) — the sum of effective time (t_{eff}) for all threads;
- total waiting time ($T_{\text{syn:wait}}$) — the sum of waiting time ($t_{\text{syn:wait}}$) for all threads; etc.

When measuring overall performance of synchronization mechanisms in a concurrent program, the following analytical system-level metrics are fairly significant [7]:

— effectiveness rate = $\frac{T_{\text{eff}}}{T_{\text{ex}}}$; (4)

shows how effectively was executed actual work; if a number of processing units (or cores in a processing unit) is n , the best values of this measure converge to $\frac{1}{n}$, however, this is possible only when threads almost do not depend on each other;

— synchronization execution rate = $\frac{T_{\text{syn:ex}}}{T_{\text{eff}}+T_{\text{syn}}}$; (5)

shows the rate of operating costs for executing instructions in synchronization algorithms; takes values in interval $[0; 1)$ (0 is the best, 1 is the worst); this metric depends only on synchronization algorithms used in separate threads and their implementation;

— synchronization waiting rate = $\frac{T_{\text{syn:wait}}}{T_{\text{eff}}+T_{\text{syn}}}$; (6)

numerical interpretation is similar to the previous metric; however, unlike the previous, this metric shows a holistic result which depends not only on summation of independent results of separate threads, but also on a way that different threads interact with each other.

6.2 Factors that Metrics Can Depend On

When working with metrics, it is important to know nearly all factors (inputs) that metrics can depend on. In this section, the authors of this paper tried to identify such factors.

Firstly, factors on which can depend performance metrics should be split into two parts: program-dependent and environment-dependent. The latter metrics usually imply some significant properties of program execution environment (usually hardware), including:

- a number of processing units (or number of cores in a processing unit),
- type and architecture of a processing unit,
- performance of a processing unit; etc.

Program-dependent factors mostly depend on a structure and synchronization logic of threads and include the following:

- number of threads,
- structure of each thread: how much effective work a thread should do before each act of synchronization,
- pattern of thread interaction,
- pattern of waiting for external events,
- synchronization mechanisms used in each synchronization time; etc.

6.3 General Considerations About Approaches to Measurement

When there are defined basic performance measures of process synchronization algorithms and main factors which influence the results of these measures, it is needed to define a way how the measurements could be done. Potentially there are two different approaches to such measurements: observation of real systems and simulation (see [5]).

The first approach consists of taking a real computer program (or writing a “synthetic” program with needed properties), adding to its code some time measurement routines which will measure the time intervals t_{eff} , $t_{\text{syn:ex}}$, $t_{\text{syn:wait}}$ and T_{ex} , and running the program (possibly, multiple times to obtain higher measurement precision). This approach has the following characteristics:

- measurements can be done relatively easy (by inserting in a program a relatively simple measurement code);
- the precision of measurements will suffer due to the measurement code will interfere with the parts of base program;
- time needed to obtain results could be reasonably large due to real-time execution; moreover, many execution times would be necessary if there is a need to obtain the measurements for different input parameters;
- there will be rather hard to generalize the obtained results due to a large number of influencing factors.

The second approach consists of modelling a concurrent system taking as inputs some most important general properties (factors) that could describe the system (however, some amount of work should be done to select from a list of properties those which are most appropriate), and simulating its execution, where measurement results are fixed by the simulating environment. This approach could be characterized by the following facts:

- to apply this approach, there is needed for a strong mathematical model of a structure and operational logic of thread synchronization (preferably, such model could be based on Petri nets; see [6]);
- the results will be obtained much faster due to non-real-time execution of the model;
- it would be easier to see the correlation between input information and measures, and for the analysis of the results, mathematical methods could be used (results would be more statistic-oriented) — this is mostly due to the minimization of influencing factors.

7 Conclusions

The gain of applying concurrent and/or parallel computing depends not only on the environment of the program, but also on the program itself. During the research there was defined a basic concept of a framework for measuring the performance of process

synchronization algorithms which uses a model-driven approach. In this paper, the framework was primarily described and analyzed from several points of view, including the structure and the functionality of the framework. The authors decided to take as a base for working with models of concurrent systems the Model-Driven Architecture; the framework can be implemented within the principles of MDA. After applying the framework, data mining techniques can be used to perform deeper performance analysis.

The main direction for further work is to the extent possible fully integrate the concept of the framework being developed with an MDA approach; letting use all the advantages of the last. Other possible directions for further work include the following:

- to refine and formalize requirements to the framework;
- to maximally simplify the input data set (i. e., factors that metrics can depend on) in order to easier understand the dependencies between these input data and the measures;
- to formalize the description of such input data set and to define requirements for software which can simulate a specific interaction of threads and measure needed metrics.
- to implement test software that can model the execution of concurrent programs given a specific input data set;
- to analyze the performance of widely-used process synchronization algorithms in different use-cases.

References

1. Downey, A.B.: The Little Book of Semaphores, 2nd edn. (2008), <http://greenteapress.com/semaphores/downey08semaphores.pdf>
2. Metropolis, N., Ulam, S.: The Monte Carlo Method. *Journal of the American Statistical Association* 44(247), 335–341 (1949)
3. Knuth, D.E.: The Art of Computer Programming, 3rd edn., vol. 1, p. 670. Addison-Wesley, Reading (1997)
4. MDA Guide Version 1.0.1 (June 2003), <http://www.omg.org/cgi-bin/doc?omg/03-06-01.pdf>
5. Hartmann, S.: The World as a Process: Simulations in the Natural and Social Sciences (2005), <http://philsci-archive.pitt.edu/2412/1/Simulations.pdf>
6. Winskel, G., Nielsen, M.: Models for Concurrency (1993), <http://www.daimi.au.dk/PB/463/PB-463.pdf>
7. Nazaruk, V., Rusakov, P.: Methods for Analyzing Performance of Process Synchronization Algorithms. In: Proceedings of the 53rd International Scientific Conference of Daugavpils University (2012)
8. Nanda, A.K.: A Framework for Multiprocessor Performance Characterization and Calibration (1992), <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.54.6805&rep=rep1&type=pdf>

9. Balakrishnan, V.: A Framework for Performance Evaluation of Parallel Discrete Event Simulators (1993), <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.22.7949&rep=rep1&type=pdf>
10. Bagrodia, R.: Process synchronization: design and performance evaluation of distributed algorithms. IEEE Transactions on Software Engineering 15(9), 1053–1065 (1989)
11. Wright, K., Gopalan, K., Kang, H.: Performance Analysis of Various Mechanisms for Inter-process Communication (2007), <http://osnet.cs.binghamton.edu/publications/TR-20070820.pdf>
12. Truyen, F.: The Fast Guide to Model Driven Architecture. The Basics of Model Driven Architecture (2006), http://www.omg.org/mda/mda_files/Cephas_MDA_Fast_Guide.pdf

Author Index

- Abd Ghani, Abdul Azim 146
Afonin, Andrej 57
Alomari, Ahmad 334
- Barauskas, Rimantas 322, 357
Bareiša, Eduardas 272
Barisas, Dominykas 272
Bernotaityte, Gintare 134
Bouaroudj, Kenza 310
Budnikas, Germanas 82
Butkiene, Rita 134
Butleris, Rimantas 114
- Čalnerytė, Dalia 322
Čeidaitė, Gintarė 261
Ceponiene, Lina 345
Ceponis, Jonas 345
- Damaševičius, Robertas 297
- Fedorov, Roman 93
- Greibus, Mindaugas 186
Gudoniene, Daina 102
Gulić, Marko 22
Guogis, Evaldas 285
Gurbuz, Tarkan 102
- Ivanovienė, Irma 1
- Javdani Gandomani, Taghi 146
Jusas, Vacius 365
- Kitouni, Ilham 310
Krisciunas, Andrius 357
Kulvietis, Genadijus 57
- Laukaitis, Algirdas 70
Lenčiauskas, Vaidotas 34
- Magdalenic, Ivan 22
Makackas, Dalius 82, 392
Malčius, Edmundas 34
Markievicz, Irena 173
Maskeliunas, Rytis 249
- Matulevičius, Raimundas 376
Md. Sultan, Abu Bakar 146
Mickevičiūtė, Eglė 114
Miseviciene, Regina 82, 392
Misevičius, Alfonsas 285
Mockus, Dainius 345
- Nazaruk, Vladislav 403
Nemuraite, Lina 122, 134
Neverdauskas, Tomas 365
Niemi, Tapio 159
Niinimäki, Marko 159
Nummenmaa, Jyrki 159
- Packevičius, Šarūnas 272
Paradauskas, Bronius 134
Paramonov, Viacheslav 93
Pavalkis, Saulius 122
Plauska, Ignas 297
Pranevicius, Henrikas 392
- Raškinis, Gailius 249
Ratkevičius, Kastytis 249
Raudys, Aistis 34
Rimas, Jonas 1
Roszkowska, Agata 11
Roszkowski, Jerzy 11
Rudžionis, Algimantas 249
Rudžionis, Vytautas 249
Rungi, Kerli 376
Rusakov, Pavel 403
Rutkauskiene, Danguole 57, 102
Ruzhnikov, Gennagy 93
- Saariluoma, Pertti 159
Saidouni, Djamel-Eddine 310
Sakalauskas, Leonidas 222
Savulioniene, Loretta 222
Sharif, Khaironi Yatim 146
Shumilov, Alexandr 93
Stanevičienė, Evelina 285
- Tamosiunaite, Minija 173
Telksnys, Laimutis 186, 261
Thamir, Alaskar 198

Thanisch, Peter 159

Theodoulidis, Babis 198

Venckauskas, Algimantas 345

Vitkute-Adzgauskiene, Daiva 173

Vrdoljak, Boris 22

Wachnik, Bartosz 46

Zhang, Zheyang 159

Žilinskas, Antanas 236

Zulzalil, Hazura 146