

Galina Setlak, Krassimir Markov  
(editors)

**Business and Engineering  
Applications of Intelligent and  
Information Systems**

I T H E A<sup>®</sup>

Rzeszow - Sofia

2011

**Galina Setlak, Krassimir Markov (ed.)**

**Business and Engineering Applications of Intelligent and Information Systems**

ITHEA®

2011, Rzeszow, Poland; Sofia, Bulgaria,

ISBN: 978-954-16-0053-5 (printed)

ISBN: 978-954-16-0054-2 (online)

ITHEA IBS ISC No.: 23

First edition

Printed in Poland

Recommended for publication by The Scientific Council of the Institute of Information Theories and Applications FOI ITHEA

This issue contains a monograph that concern actual problems of research and application of information technologies, especially the new approaches, models, algorithms and methods for information modeling to be used in business and engineering applications of intelligent and information systems.

It is represented that book articles will be interesting for experts in the field of information technologies as well as for practical users.

**© All rights reserved.**

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. The use of general descriptive names, registered names, trademarks, 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.

**Copyright © 2011**

© 2011 ITHEA® – Publisher; Sofia, 1000, P.O.B. 775, Bulgaria. [www.ithea.org](http://www.ithea.org) ; e-mail: [info@foibg.com](mailto:info@foibg.com)

© 2011 Galina Setlak, Krassimir Markov – Editors

© 2011 For all authors in the book.

® ITHEA is a registered trade mark.

**ISBN: 978-954-16-0053-5 (printed)**

**ISBN: 978-954-16-0054-2 (online)**

C\o Jusautor, Sofia, 2011

## PREFACE

The concept "Business informatics" became popular about twenty years ago when the usual concepts had lost their generality, abstractness and actuality. Concepts such as "automation of the company's management", "information service of the decision making", "information modeling of the micro-economical processes", and so on, had caused theoretical rationalizing of the whole information basics and searching of a new unified concept.

The traditional understanding of the concept "Business Informatics" is as an interdisciplinary discipline, which is aimed to study information structures, operations and processes that are inherent to the business and support their automation. Building of adequate information models of business activities, based on common theoretical foundation, leads to more clear and deep understanding of these activities and thence to their optimizing and automating. The concept of "Business Informatics" we may define as a science about the unity of Business Information, Business Information Subjects, and Business Information Interaction.

**Business Information:** Important task of the Business Informatics is to propose relatively complete investigation and classification of the business information. The development of the different companies as well as of the changing of the business environment on principle does not allow complete description. As a rule the scientists are contented by the systematizing of the formally defined information which is used by the community structures - taxation authorities, statistics, etc. Usually out of consideration are types of information created by the company in the environment and vice versa – by the environment in the company. Very important scientific area of business informatics is discovering regularities in the large volumes of stored data. This is well known area of Data Mining and Knowledge Acquisition as well as of the corresponded knowledge discovery and knowledge-based intelligent systems;

**Business Information Subjects:** One and the same business information can be perceived in a different manner from different business subjects. In general the business information subjects may be divided in two main groups – which belong to the company and the rest, which belong to the environment: partners, clients, rivals and neutrals, which in given moment may change their type. The concepts which cover this variety of type are Intelligent Agents and Multi-agent Systems, Decision Making Support, etc.

**Business Information Interaction:** The investigation of the business information subjects is determined by the necessity of provide qualitative automated service of their interaction. The important information need to be collected, stored, processed and distributed to corresponded decision makers in appropriate mode suitable for quick and non-vague perceiving, in one hand, and ensuring correct information interaction in the frame of the company as well as in the environment, from other hand. Considerable role in this area has the Natural Language Processing.

It is clear, in the practice we could not separate the parts of business informatics. Every intelligent application has to contain elements that cover and to serve all of them. In addition, the developing intelligent and information systems is closely connected with many theoretical and engineering disciplines. It is impossible here to point all of them but several are outlined.

We express our thanks to all authors of this monograph as well as to all who support its publishing.

---



---

## TABLE OF CONTENTS

Preface .....	3
Table of Contents .....	4
Index of Authors .....	6
<b>KNOWLEDGE DISCOVERY &amp; KNOWLEDGE-BASED SYSTEMS</b>	
<b>Discovering Knowledge in Parliamentary Elections</b>	
Jerzy Hołubiec, Grażyna Szkatuła, Dariusz Wagner .....	7
<b>Tacit Knowledge as a Resource for Organizations and its Intensity in Various Value Creation Models</b>	
Sumeer Chakuu .....	18
<b>Knowledge Management as Active Labour Market Policy Development Factor</b>	
Tatjana Bilevičienė, Eglė Bilevičiūtė .....	27
<b>DATA MINING, KNOWLEDGE ACQUISITION</b>	
<b>Database Server Usage in the Social Networks Analysis</b>	
Katarzyna Hareźlak .....	38
<b>Analysing and Visualizing Polish Scientific Community</b>	
Piotr Gawrysiak, Dominik Ryzko .....	47
<b>Arsima Model</b>	
Vitalii Shchelkalin .....	57
<b>Fuzzy Sets: Math, Applied Math, Heuristics? Problems and Interpretations</b>	
Volodymyr Donchenko .....	69
<b>Rough Set Methods in Analysis of Chronologically Arranged Data</b>	
Piotr Romanowski .....	81
<b>About Multi-variant Clustering and Analysis High-dimensional Data</b>	
Krassimira Ivanova, Vitalii Velychko, Krassimir Markov, Iliya Mitov .....	91
<b>NATURAL LANGUAGE PROCESSING</b>	
<b>Social Context as Machine-Processable Knowledge</b>	
Alexander Trousov, John Judge, Mikhail Alexandrov, Eugene Levner .....	104
<b>Folksonomy - Supplementing RICHE Expert Based Taxonomy by Terms from Online Documents (Pilot Study)</b>	
Aleš Bourek, Mikhail Alexandrov, Roque Lopez .....	115
<b>Classification of Free Text Clinical Narratives (Short Review)</b>	
Olga Kaurova, Mikhail Alexandrov, Xavier Blanco .....	124
<b>AUTOMATED TRANSFORMATION OF ALGORITHMS</b>	
<b>Models of the Process of an Analysis of XML-Formatted Formulae of Algorithms</b>	
Volodymyr Ovsyak, Krzysztof Latawiec, Aleksandr Ovsyak .....	136
<b>Modelling and Control of Computational Processes Using Max-Plus Algebra</b>	
Jerzy Raszka, Lech Jamroz .....	146
<b>INTELLIGENT AGENTS AND MULTI-AGENT SYSTEMS</b>	
<b>Tactical Management of Supply Chain with Agent Based Modeling and Simulation</b>	
Jacek Jakiela, Paweł Litwin, Marcin Olech .....	156

<b>Semantically Rich Educational Word Games Enhanced by Software Agents</b> Boyan Bontchev, Sergey Varbanov, Dessislava Vassileva .....	167
--	-----

## **ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE**

<b>Macromodeling for VLSI Physical Design Automation Problems</b> Roman Bazylevych, Marek Pałasiński, Lubov Bazylevych .....	178
<b>Artificial Intelligence in Monitoring System</b> Lucjan Pelc, Artur Smaroń, Justyna Stasieńko .....	189
<b>Database and Knowledge Base as Integral Part of the Intelligent Decision Support System, Created for Manufacturing Companies</b> Monika Piróg-Mazur, Galina Setlak .....	202
<b>Study of Integration Algorithm and Time Step on Molecular Dynamic Simulation</b> Janusz Bytnar, Anna Kucaba-Piętal .....	211
<b>Information Systems for Metrology</b> Roman A. Tabisz, Łukasz Walus .....	221

## **BUSINESS INTELLIGENCE SYSTEMS**

<b>Business discovery – a New Dimension of Business Intelligence</b> Justyna Stasieńko.....	234
--	-----

## **INTELLIGENT APPLICATIONS: MEDICAL AND DIAGNOSTIC SYSTEM**

<b>Performance of Computer-Aided Diagnosis Techniques in Interpretation of Breast Lesion Data</b> Anatoli Nachev, Mairead Hogan, Borislav Stoyanov.....	245
--	-----

## **MECHANICAL ENGINEERING**

<b>Description of Surfaces Having Stratified Functional Properties</b> Wiesław Graboń .....	255
--	-----

---

---

**INDEX OF AUTHORS**

<b>Alexandrov Mikhail</b>	104, 115, 124	<b>Markov Krassimir</b>	91
<b>Bazylevych Lubov</b>	178	<b>Mitov Iliya</b>	91
<b>Bazylevych Roman</b>	178	<b>Nachev Anatoli</b>	245
<b>Bilevičienė Tatjana</b>	27	<b>Olech Marcin</b>	156
<b>Bilevičiūtė Eglė</b>	27	<b>Ovsyak Aleksandr</b>	136
<b>Blanco Xavier</b>	124	<b>Ovsyak Volodymyr</b>	136
<b>Bontchev Boyan</b>	167	<b>Pałasiński Marek</b>	178
<b>Bourek Aleš</b>	115	<b>Pelc Lucjan</b>	189
<b>Bytnar Janusz</b>	211	<b>Piróg-Mazur Monika</b>	202
<b>Chakuu Sumeer</b>	18	<b>Raszka Jerzy</b>	146
<b>Donchenko Volodymyr</b>	69	<b>Romanowski Piotr</b>	81
<b>Gawrysiak Piotr</b>	47	<b>Ryżko Dominik</b>	47
<b>Graboń Wiesław</b>	255	<b>Setlak Galina</b>	202
<b>Harężlak Katarzyna</b>	38	<b>Shchelkalin Vitalii</b>	57
<b>Hogan Mairead</b>	245	<b>Smaroń Artur</b>	189
<b>Hołubiec Jerzy</b>	7	<b>Stasieńko Justyna</b>	189, 234
<b>Ivanova Krassimira</b>	91	<b>Stoyanov Borislav</b>	245
<b>Jakiela Jacek</b>	156	<b>Szkatuła Grażyna</b>	7
<b>Jamroź Lech</b>	146	<b>Tabisz Roman</b>	221
<b>Judge John</b>	104	<b>Troussov Alexander</b>	104
<b>Kaurova Olga</b>	124	<b>Varbanov Sergey</b>	167
<b>Kucaba-Piętal Anna</b>	211	<b>Vassileva Dessislava</b>	167
<b>Latawiec Krzysztof</b>	136	<b>Velychko Vitalii</b>	91
<b>Levner Eugene</b>	104	<b>Wagner Dariusz</b>	7
<b>Litwin Paweł</b>	156	<b>Walus Łukasz</b>	221
<b>Lopez Roque</b>	115		

---

## Knowledge Discovery & Knowledge-Based Systems

---

### DISCOVERING KNOWLEDGE IN PARLIAMENTARY ELECTIONS

Jerzy Hołubiec, Grażyna Szkatuła, Dariusz Wagner

**Abstract:** *The aim of the paper is to present the new methodology of building the knowledge base of parliamentary elections. The knowledge base can be used for analysis the rules describing electorate preferences during voting process. Two case studies-Polish parliamentary elections of 2001 and 2007 - illustrate the considerations.*

**Keywords:** *parliamentary election, knowledge base.*

---

#### Introduction

---

Political parties, taking part in parliamentary elections, present their socio-political programmes to the society. The programmes, presented before elections, can be considered as the set of promises. After the election representatives of some parties are elected to the parliament, others are not.

The following question can be formulated. Can we get some information on electorate preferences during electorate campaign? Can we find what they are?

Analysis of socio-political programmes of each party, participating in elections, makes it possibly to identify sets of attributes to be used to describe elements of their programmes. Having the set of attributes and their values for all the parties participating in parliamentary elections, knowledge base can be construct. Knowledge base comprise all the information needed to characterize the election campaign.

---

#### Analysis of voting behaviour

---

Problems concerned with the analysis of electorate preferences as well as forecasting outcomes of parliamentary elections have been investigated by sociologists and political scientists for a long time. There is an extensive literature on statistical methods making it possible to anticipate attitudes of voters as well as results of parliamentary elections on the basis of opinion polls. However, the number of papers – at least those known to the authors – on applying methods of artificial intelligence for such a purpose is rather small. It should be emphasized that the authors were inspired to investigate the possibility of using these methods in such a case by discussions during preparation of the paper.

Conclusion resulted in the necessity to define examples being considered as well as attributes to describe them. Moreover, the values taken by attributes introduced were to be established. It was evident that political parties and groupings taking part in electoral campaign had to be treated as examples.

The problem of choosing attributes has been much more complicated. Initially, it was assumed that attributes were to be connected with election programmes of political parties and groupings taking part in election. Special

attention was paid to those elements of the programmes under consideration that were related to basic areas of economy and social life, such as tax system, health care system, economic growth.

The set of attributes describing electoral promises of political parties and groupings taking part in the electoral campaign of 1977 was assumed as the reference point. From sociological analyses it results that electorate preferences are not concerned with elements of socio-economic programmes of political parties and grouping only, but – to a considerable degree – their medial image is of significance. In keeping with results of these analyses the initial set of attributes was enlarged for the case of 2001 election. Moreover, political scientists suggested that for election of 2005 one has to take into account attributes corresponding to watchdogs and declare values. Table 1 presents the complete set of attributes applied for the description of political parties and groupings in election of 1997, 2001 and 2005. In each column the attributes used in the analysis of a particular election are given. It follows from this table that the significance of attributes undergoes changes. Some were used only once, others in all three cases considered.

Table 1. The set of attributes

	<i>Attribute</i>	1997	2001	2006
1.	<i>Economic policy</i>		X	X
2.	<i>Corporate tax</i>	X		
3.	<i>Personal income tax</i>	X	X	X
4.	<i>Agriculture</i>		X	X
5.	<i>Regional policy</i>		X	X
6.	<i>Unemployment</i>		X	X
7.	<i>Social security</i>	X		X
8.	<i>Health care system</i>		X	X
9.	<i>Education and research</i>	X	X	X
10.	<i>Internal safety</i>	X	X	X
11.	<i>Foreign policy</i>			X
12.	<i>Attitude towards the European Union</i>		X	X
13.	<i>Local democracy</i>	X		X
14.	<i>Character of the State</i>			X
15.	<i>Secret services</i>	X		X
16.	<i>Attitude towards the abortion law</i>	X		X
17.	<i>Attitude towards women</i>			X
18.	<i>Declared watchwords and values</i>			X
19.	<i>Declared political orientation</i>	X		X
20.	<i>Organisation of the electoral campaign</i>	X	X	X
21.	<i>Forms of the electoral campaign</i>			X
22.	<i>Reception of the electoral campaign by voters</i>			X



23	<i>Orientation of the electoral campaign</i>			X
24.	<i>Experience in governmental and parliamentary activities</i>		X	X
25.	<i>Results of the previous election</i>			X
26.	<i>Visible leader</i>		X	X
27.	<i>Election to Parliament</i>	X	X	X

In the analyses carried out by the authors the following problems, related to the character of decision attribute, were investigated.

For the decision attribute 1 the partition of examples into following classes is accomplished: Y1.1 – political parties and groupings having entered the Parliament; Y1.2 – political parties and groupings not having entered the Parliament.

Taking into account some specific situations that can occur in real life, sometimes it is reasonable to divide the first class Y1.1 into two subclasses: Y1.1.1 – political parties and groupings having entered the Parliament in a strong position, i.e. having a large number of seats; Y1.1.2 – political parties and groupings having entered the Parliament in a weak position, i.e. having a small number of seats.

In the case of decision attribute 2 the partition into three classes is carried on: Y2.1 – political parties and groupings that locate their position on the left side of the political scene; Y2.2 – political parties and groupings that locate their position in the middle of the political scene; Y2.3 – political parties and groupings that locate their position on the right side of the political scene.

Due to some difficulties with determining and adequate position of a particular party of grouping active on the Polish political scene, the partition into four classes can be useful: Y2.1 – parties or groupings that define their political views as leftist; Y2.2 – parties or groupings that define their political views as liberal; Y2.3 – parties or groupings that define their political views as rightist; Y2.4 – parties or grouping whose political views cannot be explicitly defined as belonging to one of three classes distinguished.

In the case of third decision attribute the number of classes is equal to the number of political parties and groupings taken into consideration.

For each of the decision attributes mentioned, it is possible to determine classification rules to a particular class.

If the decision attribute 1 is used, then obtained classification rules can be applied to determine: electorate preferences and programme or medial image differences among political parties or groupings having entered the Parliament (taking into account the number of seats received) and those that failed.

In the case of decision attribute 2, the classification rules generated can be applied to determine in detail programme or medial image differences among political parties and groupings belonging to given classes.

If the third decision attribute is used, then classification rules obtained can be applied to determine similarities or differences in programme or medial image among political parties and groupings considered.

The algorithm described above is relatively simple and efficient.

In second and third case, can be applied the approach based on the rough set theory, forwarded by Pawlak (1982, 1991).

The rules formed can be used in classification of new examples (i.e. ones that have not appeared in the learning process) for which class membership is not known. Such a classification is carried out through verification of fulfilment of conditions in the conditional parts of the rules.

The rules can also be used to identify the dependencies existing in the information set of the examples that have not been previously known explicitly. They can help in understanding and explaining the existing relations between attributes or class definitions.

---

### **The Case study 1. The Polish Parliamentary election 2001**

---

The considerations concern the example of the elections to the Polish Parliament, which took place on September 23<sup>rd</sup>, 2001.

In the situation considered it is the set of eight political parties: SLD, UP, PO, SO, PiS, PSL, LPR, AWSP and UW. Only first six of these political organizations ultimately entered the Parliament. The attributes used to describe examples are presented below. The value that a given attribute can assume is given in brackets.

*a*<sub>1</sub>: *unemployment* {1 – to make the Labour Code more flexible; 2 – to take actions making working personnel more mobile and to decrease cost of establishing new work places; 3 – to start public and interventional works; 4 – other proposals}.

*a*<sub>2</sub>: *education and research* {1 – to increase governmental spending on education and research; 2 - free education at all the levels; 3 – to stop elimination of rural schools and to establish vocational colleges in small cities; 4 – to provide common access to Internet and learning of foreign languages}.

*a*<sub>3</sub>: *personal income tax* {1 – to simplify the personal income tax system and to introduce linear personal income tax in the future; 2 – to introduce progressive personal income tax, tax on stock exchange transactions and tax on capital; 3 – to decrease the lowest rate of personal income tax and to establish pro-family policy; 4 – other proposals}.

*a*<sub>4</sub>: *economic policy* {1 – government control of all strategic and monopolistic enterprises; 2 – government control of chosen strategic and monopolistic enterprises; 3 – restructurization of the public finance sector; 4 – government support of inexpensive housing projects}.

*a*<sub>5</sub>: *health care system* {1 – to improve the existing health-care system and to increase governmental spending on this system; 2 – to eliminate the existing health-care system and to establish a new one; 3 – other proposals}.

*a*<sub>6</sub>: *agriculture and regional policy* {1 – to protect the Polish agriculture against foreign competition; 2 – to make the industrial-agricultural complex a fly wheel of economy; 3 – to develop non-agricultural activities in villages and to promote infrastructural investments}.

*a*<sub>7</sub>: *internal safety* {1 – to increase the efficiency of judiciary system; 2 – to make the Penal Code more repressive; 3 – to roll into one municipal guard and police; 4 – to develop citizen self-defence; 5 – other proposals}.

*a*<sub>8</sub>: *attitude towards the European Union* {1 – to enter the European Union under advantageous conditions; 2 – pronounced backing the accession to the European Union; 3 – stout resistance to the accession to the European Union}.

*a*<sub>9</sub>: *election to the Parliament* {1 – yes; 2 – no}.

*a*<sub>10</sub>: *experience of government matters and services* {1 – yes; 2 – no}.

*a*<sub>11</sub>: *visible leader* {1 – yes; 2 – no}.

*a*<sub>12</sub>: *organization of the electoral campaign* {1 – professional, 2 – not professional}.

Table 2 presents the values of attributes taken into account for particular political parties and groupings. It can be considered as the knowledge base.

*Table 2 Knowledge base describing the Polish parliamentary election of 2001*

Attributes Parties	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	$a_9$	$a_{10}$	$a_{11}$	$a_{12}$
SLD-UP	1	2	2	2	2	3	1	2	1	1	1	1
PO	1	4	1	3	2	2	3	2	1	2	1	1
SO	3	2	4	1	2	1	2	3	1	2	1	1
PiS	2	3	3	4	1	3	2	1	1	1	1	2
PSL	2	3	2	2	2	1	5	1	1	1	2	2
LPR	4	1	3	1	3	2	1	3	1	2	2	1
AWSP	2	4	3	4	1	3	2	1	2	1	2	2
UW	1	1	1	3	3	3	4	2	2	1	2	2

The decision attribute  $a_9$ : “election to the Parliament” divides the set of examples into two classes in the following manner:

- class  $Y_{yes}$ : contains the political parties having entered the Parliament,
- class  $Y_{no}$ : contains the political parties not having entered the Parliament.

### **The Case study 2 – the Polish Parliamentary election 2007**

The set of six political parties is taken into account: PO, PiS, LiD, PSL, Samoobrona and LPR. Only first four of these political parties ultimately entered the Parliament.

As a result of the analysis of the electoral campaign to the Polish Parliament 2007 thirty four attributes were identified. Twenty two attributes ( $a_1, \dots, a_{22}$ ) characterize programmes of political parties taken into account. Seven attributes ( $a_{23}, \dots, a_{29}$ ) are used to describe characteristics of political parties as well as watchwords and values presented by them. Four attributes ( $a_{30}, \dots, a_{33}$ ) characterize the electoral campaign. The last attribute  $a_{34}$  describes results of the election. It is considered as the decision one. The detailed description of attributes taken into account and their values is given in Appendix.

Table 3 presents the values of the attributes for particular political parties.

*Table 3. Knowledge base describing the Polish parliamentary election of 2007*

Attributes Parties	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	$a_9$	$a_{10}$	$a_{11}$	$a_{12}$	$a_{13}$
PO	2	2	1	2	1	1	1	1	1	1	1	1	2
PiS	1	1	1	1	1	2	2	3	3	2	1	2	1
LiD	3	2	2	2	2	3	2	2	2	2	2	3	1
PSL	2	2	1	2	2	1	2	3	3	3	3	1	2
Samoobrona	3	3	2	3	1	2	3	3	3	3	2	1	2
LPR	1	1	1	1	3	2	3	3	3	1	1	2	3

Attributes Parties	a <sub>14</sub>	a <sub>15</sub>	a <sub>16</sub>	a <sub>17</sub>	a <sub>18</sub>	a <sub>19</sub>	a <sub>20</sub>	a <sub>21</sub>	a <sub>22</sub>	a <sub>23</sub>	a <sub>24</sub>	a <sub>25</sub>	a <sub>26</sub>	a <sub>27</sub>
PO	2	1	1	1	1	1	1	1	1	1	2	1	1	2
PiS	1	1	1	1	2	1	2	2	2	1	2	1	2	1
LiD	2	2	2	3	1	2	3	1	3	2	2	2	1	3
PSL	2	3	1	1	3	1	3	3	2	1	1	2	2	3
Samoobrona	3	3	1	2	1	3	3	3	2	2	2	2	3	1
LPR	3	3	3	1	1	3	3	2	2	2	2	2	3	1

Attributes Parties	a <sub>28</sub>	a <sub>29</sub>	a <sub>30</sub>	a <sub>31</sub>	a <sub>32</sub>	a <sub>33</sub>	a <sub>34</sub>
PO	1	1	1	1	1	2	1
PiS	2	2	1	1	1	1	1
LiD	2	3	2	3	3	3	2
PSL	2	2	2	2	2	2	2
Samoobrona	2	4	3	2	2	1	3
LPR	3	5	3	2	1	1	3

---

## Conclusion

---

The paper presents a new approach to the analysis of electoral campaign.

The application of voting procedures makes it possible to determine which political parties taking part in the election campaign are the winners and which are the losers. The main goal of every political party is to enter the parliament, in other words to be supported by large number of voters. To accomplish such a strategy, political parties have to work out and present their socio-political programmes.

The analysis of socio-political programmes presented by all the parties taking part in the election campaign, makes it possible to identify attributes that can be used to characterize these programmes. Having determined the set of attributes and their values it is possible to form knowledge base comprising all the information needed to characterize the election campaign under consideration.

In the paper results of the Polish parliamentary election of 2001 and 2007 are used, as the illustration of proposed methodology how to build a knowledge base.

Making use of this knowledge base one can make an attempt to answer the following question: why some parties received the large number of votes but other ones – not?.

In the situation considered in the paper, the decision attributes  $a_{34}$ : “election to the Parliament” divides the set of examples (i.e. the political parties) into three disjoint classes in the following manner:

- class  $Y_1$ : contains the political parties having entered the Parliament with large support,
- class  $Y_2$ : contains the political parties having entered the Parliament with small support,
- class  $Y_3$ : contains the political parties not having entered the Parliament.

The process of formation of a class description on the basis of the set of examples having certain common properties, which distinguish a given class from the other ones. These descriptions can be represented in the form of rules of "IF *certain conditions are fulfilled* THEN *membership in a definite class takes place*" type. It means, that the algorithm of machine learning can be applied. Using machine learning analysis one can receive the answer for above mentioned question.

---

## **Bibliography**

---

- [Brams, 2008] Brams S. The presidential election game. New Haven, CT: Yale University Press, Rev. ed. A. K. Peters.
- [Hołubiec, etc., 1997] Hołubiec J., Malkiewicz A., Mazurkiewicz M., Mercik J., Wagner D. Identification of ideological dimensions under fuzziness: the case of Poland. In: Consensus under fuzziness. Kluwer Academic Publishers, Boston/Londyn/Dordrecht.
- [Hołubiec, etc. 2007] Hołubiec J., Szkatuła G., Wagner D. Discovering electorate preferences in voting procedures. Homo Oeconomicus, vol 24, Nr 314, Accedo Verlagsgesellschaft, Munchen.
- [Hołubiec, etc., 2008] Hołubiec J., Szkatuła G., Wagner D. Machine learning approach for discovering electorate preferences during parliamentary election. Development in Fuzzy Sets, Intuitionistic fuzzy Sets, Generalized Nets and Related Topics, Applications, vol. II. Academic Publishing House EXIT, Warsaw.
- [Kacprzyk, etc., 2002] Kacprzyk J., Szkatuła G. An integer programming approach to inductive learning using genetic and greedy algorithms. W: L.C. Jain and J. Kacprzyk (eds.) *New learning paradigms in soft computing*. Studies in Fuzziness and Soft Computing. Physica-Verlag Heidelberg.
- [Szkatuła, etc., 2003] Szkatuła G., Wagner D. Programmes of parties versus their location on the political scene. Application of decision rules to describe the differences. In: Kacprzyk J., Wagner D. (eds.): *Group decisions and voting*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

---

## **Authors' Information**

---

**Jerzy Hołubiec** - professor, Systems Research Institute PAS, Warsaw, 01-447 Newelska 6, Catholic University J.P. II, Raclawicka 14, Lublin, e-mail: [jerzy.holubiec@ibspan.waw.pl](mailto:jerzy.holubiec@ibspan.waw.pl)

Major fields of Scientific Research: Artificial intelligence, voting procedures,

**Grażyna Szkatuła** - as. Professor, Systems Research Institute PAS, Warsaw, Newelska 6,

e-mail: [grazyna.szatula@ibspan.waw.pl](mailto:grazyna.szatula@ibspan.waw.pl)

Major fields of Scientific Research: Machine learning, computational intelligence,

**Dariusz Wagner** - as. Director, Systems Research Institute PAS, Warsaw, Newelska 6,

e-mail: [dariusz.wagner@ibspan.waw.pl](mailto:dariusz.wagner@ibspan.waw.pl)

Major fields of Scientific Research: Group decisions, voting procedures.

---

## **Appendix**

---

### 1) Attributes describing programmes of political parties

a1 - Way of governing the State

1. to strengthen the State and radical struggle against social ills - PiS, LPR
2. to try to reach a political compromise - PO, PSL
3. no proposals submitted - LiD, Sam.

a2 - Assessment of the two-year period 2005-2007

1. period of improving the condition of State -PiS, LPR
2. period of increasing the arrogance of authorities -PO, PSL, LiD
3. ambivalent attitude - Sam.

a3 - Attitude to the vetting

1. need of consistent vetting - PiS, PO, PSL, LPR
2. need to end the vetting - LiD, Sam.

a4 - Attitude to the Church and religion

1. need of close cooperation between the State and Church - PiS, LPR
2. to separate the Church from the State but to have respect for religious values - PO, PSL, LiD
3. no proposals submitted - Sam.

a5 - Unemployment

1. to decrease non - wage cost of works including the decrease of insurance fee - PO, PiS, Sam
2. insurance fee and income tax deduction - LiD, PSL
3. no proposals submitted - LPR

a6 - Education and science

1. consistent programme base for all the subjects with parallel increase of the school autonomy necessary for the adjustment to market demands - PO, PSL
2. to increase governmental funds for science and education - PiS, LPR, Sam
3. to decrease the school age - LiD

a7 - Economy

1. economical freedom based on private property - PO
2. development of small and medium enterprises including tax deduction and availability of credit; introducing the system of guarantees for enterprises - PiS, Lid, PSL
3. to stop privatization and control privatized enterprises - LPR, Sam.

a8 - Taxes

1. to simplify the tax system by means of introducing linear tax - PO
2. tax amnesty for Poles coming back to Poland - LiD
3. to decrease the income tax especially for the poorest ones; pro family policy including rent deduction - Sam., LPR, PiS, PSL

a9 - Health - care system

1. partition of the National Health Fund into several competing funds and to define the basket of services - PO
2. introducing the payment for medical services - LiD
3. unrestricted Access to the Basic health care system including formation of the charity fund and to transfer some Mount of Money from Work Fund to health care system - LPR, Sam., PSL, PiS

a10 - Internal safety

1. to form centralized and coordinated centre to counteract the most serious risk at the State level - PO, LPR

2. to improve the weaponry and equipment of uniformed services; to built the all - Poland communication system for rescue services - PiS, LiD
3. no proposals submitted - PSL, Sam.

a 11 - Social policy

1. Introduction of family benefits new benefits for children and to extend the maternity leave for both parents - PO, PiS, LPR
2. introduction of home allowances and pension reform; to increase benefits and assistance to the poorest - LiD, Sam.
3. Introduction of the family tax and common tax return for families - PSL

a12 – Agriculture

1. use of European funds for development of The Polish agriculture - PO, PSL, Sam.
2. To support low – production and low – profit farms and strengthen family farms -PiS, LPR
3. to reintroduce structural rents and make more realistic insurance fee for farmers - LiD

a13 – Attitude to authorities and self-governments

1. administrative decisions should be made in accordance with the letter of the law - LiD, PiS
2. to decentralize and strengthen the property base of self-governments — PO, PSL, Sam
3. to restrict the autonomy of self-governments - LPR

a14 – Introducing the Euro currency

1. accession of Poland to the Euro zone not later than in 2015 - PiS
2. accession of Poland to the Euro zone as soon as possible - PO, PSL, LiD
3. to remain separate currency - LPR, Sam

a15 – Foreign policy

1. to back up the accession of the Ukraine, Georgia and Moldova to the European Union; to strengthen the role of Poland In relations with the neighbors to cooperate with countries of so called Weimar Triangle - PiS PO
2. good relations with all the neighbors, especially with the Ukraine and Germany - LiD
3. to make relations with Russia warmer - Sam., LPR, PSL

a16 – Attitude to the UE

1. to cooperate more closely with the European Union and take responsibility for its growth - PO, PiS, PSL, Sam.
2. joint foreign policy and attitude to Russia - LiD
3. to restrict cooperation with the UE - LPR

a17 – Attitude to the USA

1. to preserve the partnership in the framework of safety and economy systems - PO, PiS, LPR, PSL
2. to loosen relations - Sam.
3. no proposals submitted - LiD

a 18 – Attitude to the war in Iraq

1. to curry out mission responsibilities; no prolongation of the presence of Polish troops; to promote economy and political relations - PO, LiD, Sam., LPR

2. to prolong participation of the Polish troops in the mission; to strengthen position and safety of Poland - PiS
3. no proposals submitted - PSL

a19 - Safety of energy supply

1. diversification of energy suppliers and use of own resources including use of clean energy - PO, PiS, PSL
2. to promote common policy of energy supply within the European Union - LiD
3. to normalize the cooperation with Russia - Sam., LPR

a20 – Fight against corruption

1. -transparent administrative procedures - PO
2. -to continue activities of the Central Anticorruption Bureau - PiS
3. -no proposals submitted - PSL, LiD, Sam., LPR

a21 – Attitude to the abortion law

1. abortion is allowed only when the pregnancy threatens mother's health - PO, LiD
2. total ban on abortion - PiS, LPR
3. no proposals submitted - PSL, Sam.

a22 – Character of the State

1. not expensive, decentralized and public - spirited - PO
2. social solidarity - PiS, Sam., LPR, PSL
3. lawfull and nonpartisan - LiD

II) Attributes describing political parties and declared watchwords and values.

a23 – Organization of a political party

1. party with widely extended structures - PSL, PO, PiS
2. party with narrowly extended structures - LiD, Sam., LPR

a24 – Experience in governmental activities

1. long period of activities - PSL
2. short period of activities - PO, PiS, LiD, LPR, Sam.

a25 – Participation in activities of the former Parliament

1. having strong position - PO, PiS
2. having weak position - LiD, PSL, LPR, Sam.

a26 – Party image

1. modern - PO, LiD
2. conservative - PSL, PiS
3. opportunistic - LPR, Sam.

a27 – Position of a leader

1. distinctive dominating over the party - PiS, LPR, Sam.
2. distinctive cooperating with party members - PO
3. not distinctive - LiD, PSL



a28 – Declared political views

1. liberal - PO
2. centre - PiS, PSL, LiD, Sam.
3. rightist - LPR

a29 – Declared watchwords and values

1. social-liberal - PO
2. social-national - PiS, PSL
3. social-leftist - LiD
4. social-liberal - Sam.
5. Christian-national - LPR

III) Attributes describing the electoral campaign

a30 – Organization of electoral committees

1. professional - PO, PiS
2. not carefully prepared - PSL, LiD
3. amateurish - LPR, Sam.

a31 – Forms of running electoral campaign

1. run with the use of modern media means - PO, PiS
2. run with emphasis put on the direct access to votes - PSL, LPR, Sam.
3. run with emphasis put on the use of Internet - LiD

a32 – Orientation of the electoral campaign

1. addressed to all the votes - PiS, PO, LPR
2. addressed to town and city dwellers - PSL, Sam.
3. addressed to young people and stable leftist electorate - LiD

a33 – Characteristic of the electoral campaign

1. aggressive campaign based on watchwords and values - PiS, LPR, Sam.
2. peaceful campaign based on watchwords and values - PO, PSL
3. peaceful campaign based on the appearance of personages - LiD

IV) Decision attribute

a34 – Results of the election

1. entering the Parliament with strong support of votes - PO, PiS
2. entering the Parliament with small support of votes - LiD, PSL
3. not entering the Parliament - LPR, Sam.

## TACIT KNOWLEDGE AS A RESOURCE FOR ORGANIZATIONS AND ITS INTENSITY IN VARIOUS VALUE CREATION MODELS

Sumeer Chakuu

**Abstract:** *Tacit knowledge continues to play a challenging role in any type of industry and is no doubt an everlasting resource. So it is important to realize the intensity and impact of tacit knowledge in an organization at various stages. Doing so will help us to achieve a conceptual consideration about the concentration of tacit knowledge capturing techniques, technologies and methodologies at a specific stage in the business. The main aim of this article is to explore the tacit knowledge as an inimitable resource and its intensity in all types of organization which are becoming more and more knowledge intensive by time.*

**Keywords:** *Knowledge Management, Tacit knowledge, Knowledge in Value configurations, Knowledge discovery, Knowledge as resource.*

**ACM Classification Keywords:** *K.6.3 Resource Allocation, I.2.4 Knowledge Representation, I.2.6 Knowledge Acquisition.*

---

### Introduction

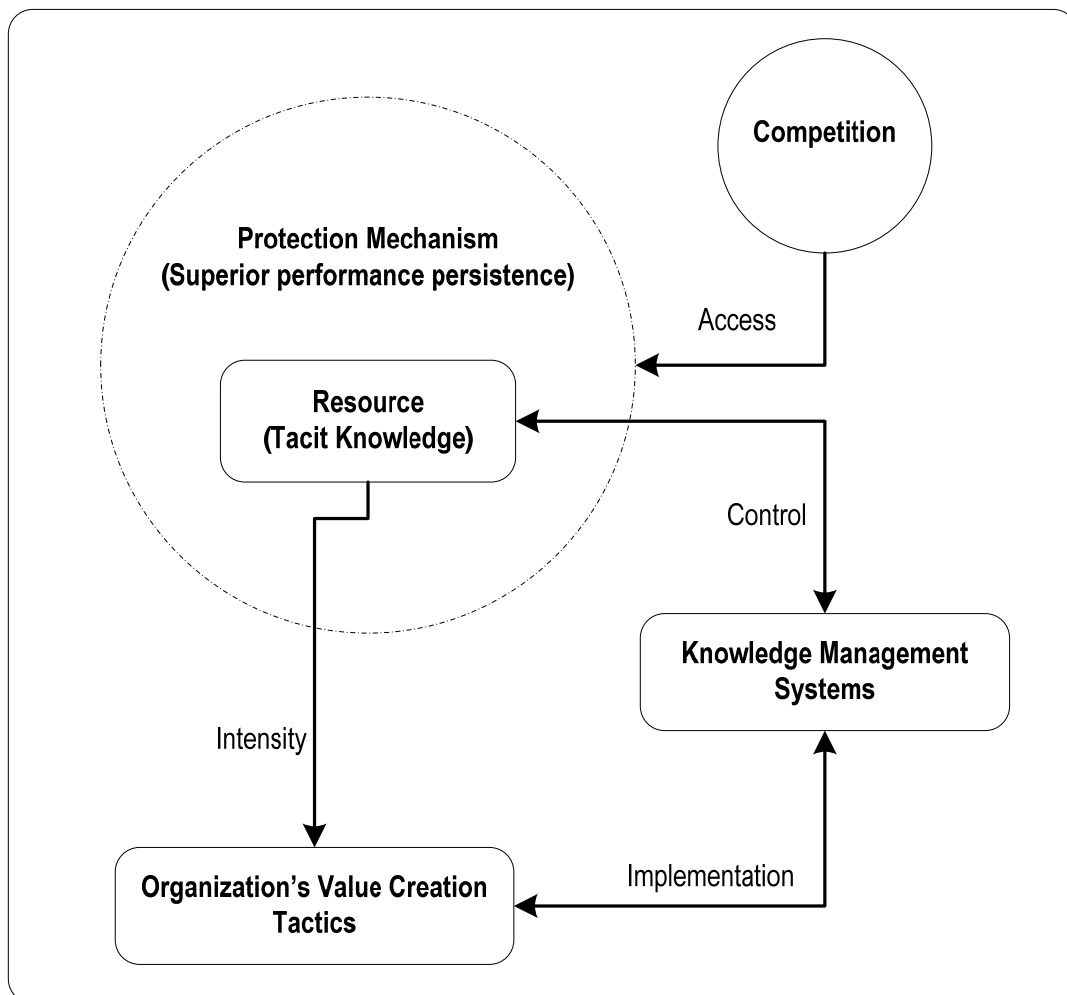
---

Tacit Knowledge Management is a young emerging discipline with plenty of topics to be researched, many theories and ideas yet to be tested, many issues yet to be resolved and much learning yet to be discovered. As the name suggests it implies managing tacit knowledge to be a leader in today's competitive age. In context to the today's leveraging knowledge economies its implementation area is vast encompassing big economies to small scale industries. Due to its importance and cohesiveness its now-a-days considered as the back bone of going to be globalised world. Being a borderless discipline now onwards I have narrowed my research only to its micro economic aspect. All the organizations in today's era are aware of the effectiveness of usage of tacit knowledge in their business processes but none is able to recognize its intensity at the base of their organizations. By effectiveness of usage I mean storage, transfer, retrieval, application, and visualization of knowledge. Though Knowledge is embedded in the organizations from the instant the industrial revolution started but its emergence as an independent subject was unknown due to lack of resources.

Advent of Information Technology had a great and impact on tacit Knowledge Management. The first and foremost level of impact of Information Technology was on the explicit knowledge at the point where work got done and transactions (e.g., orders, deposits, reservations) took place and were being used for future references. Later, the management information systems were used to aggregate data into useful information reports, often prescheduled, for the control level of the organization .Now-a-days Information Technology has started to facilitate the management of tacit knowledge. One of major issues which are in existence is the design of systems which can capture and disseminate tacit knowledge. As the scope is flourishing, strategic changes in Information technology for tacit knowledge management and redesign of Knowledge Management Systems has become a must be factor. In addition to this the transfer of tacit knowledge has always posed challenges to the knowledge workers and designers. There are many factors which tend to hurdle the tacit knowledge transfer especially cultural context in the organizations. There were large number of techniques employed in order to overcome it but all the techniques are theoretical with no or minimum usage of emerging technologies.

Tacit knowledge has been and will be always an imperishable resource for any type of organization. According to the resource-based theory of an organization tacit knowledge is considered as a valuable, unique, and difficult to imitate resource and in context with activity-based theory of an organization it is considered as a driver of all activities which provides basis for organization's performance and thereby providing organization a competitive advantage over others. Tacit knowledge is an intangible and dynamic asset of any organization. [Alavi, Leidner,2001] suggest that the long-term sustainable competitive advantage comes from the firm's ability to effectively apply the existing knowledge to create new knowledge and to take action that forms the basis for achieving competitive advantage from knowledge-based assets [Gottschalk,2004]. Tacit knowledge acts as an intangible resource or driver for an organization.

Based on discussion above the objective of this article is to determine the effectiveness of tacit knowledge and its intensity in company value analysis; main goal is to enhance the throughput of knowledge capturing systems through proper selection in accordance with the intensity of the tacit knowledge. Doing so will make it easier for knowledge workers to use their resources at par and generate the solutions which are economically and technically feasible. The conceptual map of the article is shown in figure 1.



*Fig. 1. Conceptual Map*

The conceptual map highlights the tacit knowledge as a resource for an organization and its interaction with the value creation tactics and knowledge management systems. Tacit knowledge should be always protected from the competition as it can be fatal for organization's operation if it will lose its tacit knowledge to competition. For avoiding such a scenario a protection mechanism should be provided. Protection Mechanism involves all the necessary actions which are supposed to be taken by an organization to keep its tacit knowledge inaccessible for competition. This is due to the fact that the resource is valuable only as long as it is unique from others. In other words the protection mechanism is for the persistence of superior performance in comparison to the competition. The required intensities of tacit knowledge (as a resource) at each stage of organization's value chain will enable knowledge workers and designers to use proper techniques and resources necessary to design knowledge management systems for managing tacit knowledge. This will enable them to increase the throughput of operations, solve the time complexity, proper budgeting of projects and skill requirements required at particular phase of value creations models. In addition to this, tacit knowledge directly controls the type of knowledge management system and technology which is needed according to the nature and complexity of tacit knowledge to be captured, stored and disseminated. Once we have all the above mentioned strategies accomplished we can design and implement knowledge management system which will serve the purpose of fulfillment of organization's operations tactics.

So, in accordance with the objective, the first part of this article justifies tacit knowledge as a imperishable and valuable resource for organizations and second part focuses on the impact of tacit knowledge in the various value creation models viz. Porter's value chain, Value shop and value network.

---

### **Tacit knowledge in organization performance**

---

The resource-based view of the firm has emerged as a major paradigm in the strategic management field [Barney, 1991]. At the basic level, the resource-based view of an organization is based on three straightforward propositions. The three propositions are as follows;

- Organizations differ on the basis of their resource endowments
- Resource heterogeneity gives rise to differential performance
- Superior performance persists as long as there are various mechanisms incorporated to protect the valuable and rare resources

As the theory has developed and research was performed to justify knowledge as a resource it was being concluded that tacit knowledge is an organization's intangible resource. Similar arguments concerning role of tacit knowledge as resource has diminished the valuability of human capital.

Human capital is no longer been argued as a critical resource in most organizations. Recent research suggests that human capital attributes such as tacit knowledge affect organization's outcomes and leverage their true capabilities [Hitt, Bieman, Shimizu, Kochhar, 2001]. Moreover it's the only resource having increasing returns as it is used. The more it is used, the more valuable it becomes, creating self-reinforcing cycle. To define tacit knowledge as a resource, we can correlate it with the three supplementary methods which exist to identify needs for knowledge. Figure 2 highlights the three methods viz. problem decision analysis, critical success factors and ends means analysis. If we closely refine their granularity we will come to the conclusion that definitely tacit knowledge is a resource which adds value to an organization.

Problem decision analysis involves identifying problems and taking appropriate decisions. Problem decision analysis is very critical for any type of organizations. Tacit knowledge is a valuable resource which enables appropriate, fruitful and precise identification of problems which underlie an optimal decision. Innovation and intuition are considered as the pillars in problem decision analysis which are no more than the tacit knowledge acquired by the decision makers over a time.

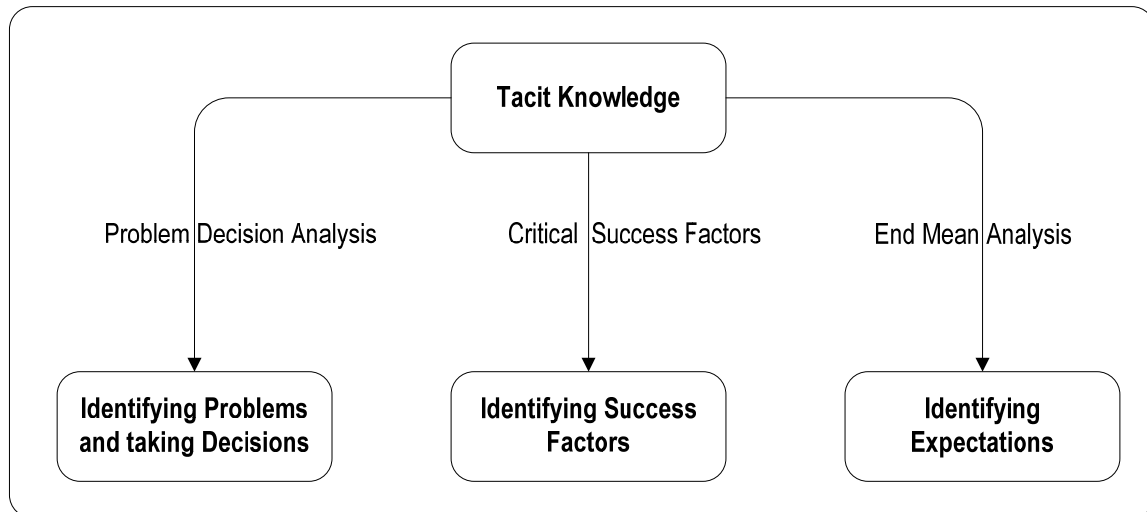


Fig. 2. Tacit knowledge as resource

Tacit knowledge as a resource for problem decision analysis is justified by its usage in:

- Exploring creative thinking in terms of what is required to be initiated and followed organizationally over time
- Looking at the various creative techniques that are useful to decision makers
- Exploring the use of problem finding from the standpoint of turning problems into opportunities
- Examining how problem finding can assist in expanding the wisdom of decision makers.

Critical Success factors play a vital role in means of planning, implementing and reviewing strategy to become industry leader. Tacit knowledge as a resource for identifying success factors is rationalized by various factors which can be applied for success of organizations. Tacit knowledge is used to

- Identify the critical success factors of any business by determining the challenges that may hinder organization's ability to grow and accomplish its target. These challenges can be internal viz. business politics or employee discontent or external viz. economic policies, political climate that affects business.
- Create a strategic plan that will wrap the challenges and help the organization to anticipate them.
- Understand targeted customers which involve learning more about the behavior of your targeted consumers and understanding the demographic of someone who is more likely to avail organization's product or service.
- Compare services with the competition and to assess how to serve the market niche in relation to the direct competitors to find out how product or service is competing among other brands.
- Examine the competition on the basis of how they operate. After looking at how the competitors serve targeted market. Examine quality control, performance and production cycles. Also comparison of these practices will lead to identify the shortcomings of the competition's strategy.
- Adjust the production or service providing strategies as needed. Looking at how competitors produce the same product or service will allow you to determine the weaknesses of organization's own strategy.

Tacit knowledge is a boost in identifying the expectations to be set for the successful operation of an organization. The tacit knowledge connects the goals and objectives attained to identify, clarify and agree to the identified expectations. Tacit knowledge of decision makers facilitates in

- Making employees and co workers understand the impact of their working culture in setting and identifying expectations
- Acknowledgement of the various perspectives set and achieved performance expectations
- Restating and clarifying expectations
- Being specific and descriptive
- Not too subjective expectation performance reviews

Let us consider a situation depicted in figure 3, which will demonstrate how tacit knowledge acts as a resource in correlation with above stated aspects.

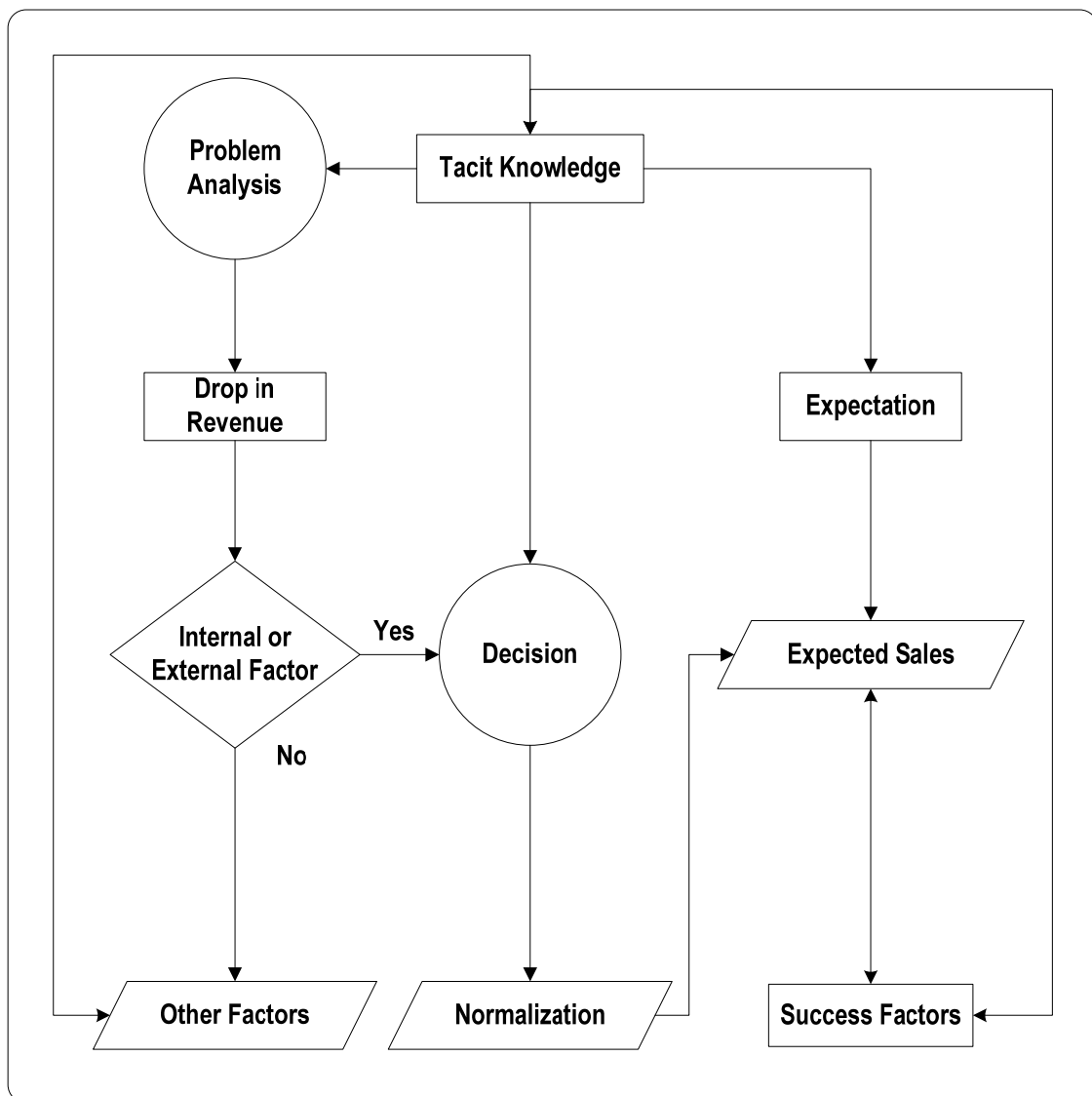


Fig. 3. Analogy Tacit Knowledge as resource

In this illustration I have considered a manufacturing organization which is producing a physical product. After a yearlong work on tactics and finalizing the goals and expectations the business outcome was not up to the mark. The financial statement highlighted that there is a decrease in revenue. Due to the decrease in revenue of an organization the managers will use embedded tacit knowledge and some figures to know the problem which led to such a situation. This is the problem analysis. After analysis it was found that the primarily reason for reduction

in revenue was due to drop in sales. Then tacit knowledge allows decision makers to think over a possible cause which involves factors viz. internal and external factors. Internal factors involve reduction in quality or rise in sales price and external factors involve the competition producing the same product. Depending on the type of factor involved in the drop of sales appropriate decision will be taken in accordance. After the decision is being taken the expected results will be analyzed, if positive, the solution will be incorporated in the list of success factors where tacit knowledge plays a significant role.

---

### **Tacit knowledge in Value Chain**

---

As stated and elaborated above that how tacit knowledge acts as an intangible resource or driver for an organization, now I have analyzed this resource as a part of value chain.

Porter's value chain framework [Porter, 1985] is presently the accepted language for both representing and analyzing the logic of organization-level value creation or performance builder. In spite of having some limitations which will be discussed in upcoming topic, it maintains its central role as a framework for the analysis of organization-level competitive strengths and performance. The value chain as described by Porter is a two-level generic taxonomy of value creation activities which in turn make organization's highly competitive by increasing their performance level. The framework is based on long-linked topology in which value is created by transforming input into products [Thompson, 1967]. The two levels in porter's value chain constitute primary and support activities of an organization. Primary activities consist of inbound logistics, operations, outbound logistics, marketing and sales, and service and support activities include Procurement, Technology development, human resource management and firm or organization infrastructure. If we speak about the usage or involvement of tacit knowledge in the primary activities, the high involvement can be seen in operations, marketing and sales, service and logistics have minimum involvement. In the same context the support activities all have highly embedded tacit knowledge.

Operations are associated with transformation of inputs into final products. So if we consider change in the market for any product, at that instant tacit knowledge will play a vital role in determination of any change in operations. Concerning marketing, sales and service it is necessary to determine the mindset of customers and market behavior and it is always important to have an exact view of all relating status and during this analysis tacit knowledge provides the intuition and zeal. Regarding support activities all except infrastructure require high intensity of tacit knowledge for proper management. All this is well supported by the above explained effect of tacit knowledge on organization's performance.

Figure 4 below demonstrates the intensity of tacit knowledge in porter's value chain model. The Tacit Knowledge Intensity (T.K.I.) is shown by low, moderate and high levels.

---

### **Tacit knowledge in Value shop and Value Network**

---

The topology described by Porter is well suited for traditional manufacturing organizations but this logic is less suitable to activities in a number of service industries like insurance companies, hospitals, educational institutes, banks, telephone companies and organization which are in similar genre of business. For these types of organizations value shop and value network configurations are desirable [Stabell, Fjeldstad, 1998].

The value shop models organizations where value is created by mobilizing resources and activities to resolve particular customer's problem. This model is based on intensive technology approach and these organizations are often called professional service firms or knowledge-intensive service firms [Gottschalk, 2004]. Knowledge is the most important resource, and reputation is critical to organization success. Instances of such organizations

are medicine, law, architecture and engineering. The ‘shop’ label captures that an organization so configured is directed at a unique and delineated class of problems [Stabell, Fjeldstad, 1998].

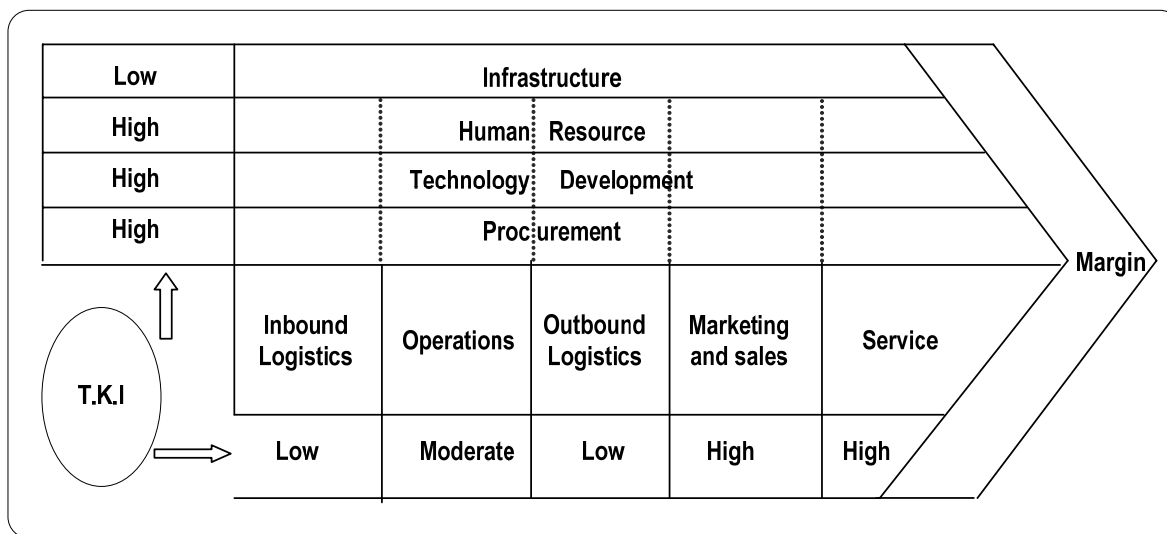


Fig. 4. Tacit Knowledge Intensity (T.K.I.) in porter's value chain

A value shop is characterized by five primary activities: problem finding and acquisition, problem-solving, choice, execution, and control and evaluation [Gottschalk, 2004]. In addition to this the support activities are similar as in value chain. As clear from the primary activities, we can say that they belong to problem-solving and decision-making genre, thereby making tacit knowledge an important aspect of the organization. Due to this reason we can typically find scientists and experts in these types of organizations. The intensity of tacit knowledge requirement is very high in all primary activities while the support activities have same level as described in porter's value chain. Figure 5 below shows the generic value shop diagram where post execution evaluation can be the problem-finding activity of a new problem-solving cycle and it also depicts Tacit Knowledge Intensity (T.K.I.) levels.

A value network is an organization that creates value by connecting clients and customers that are, or want to be, dependent on each other. These companies distribute information, money, products and services. While activities in both value chains and value shops are done sequentially, activities in value networks occur in parallel. The number and combination of customers and access points in the network are important value drivers in the value network. More customers and more connections create higher value to customers [Harrington, Voehl, 2007]. So a value network relies on mediating technology. Examples of organizations using mediating technology are telephone companies, retail banks, postal services and so on. The primary activity description in these organizations is inspired by that used in telecommunication because telecommunication is a rather generic form of mediation and because explicit activity decomposition models are well established both at the micro level of peer-to-peer communication and at the industry level in delineating industry actors [Murray, 2008]. The primary activities of value network are network promotion and contract management, service provisioning and network infrastructure operation and the support activities are the same as in porter's value chain and value shop. Talking about tacit knowledge intensity in value network, it is not as intensive as it is in value shop. Among all three activities, network promotion and contract management has a high intensity of tacit knowledge in it as it involves selling services, evaluating risk and monitoring contracts. As in my view as the activities in value network act in parallel so we can say that service provisioning as a direct effect on network promotion so has the same tacit knowledge intensity enveloping it.



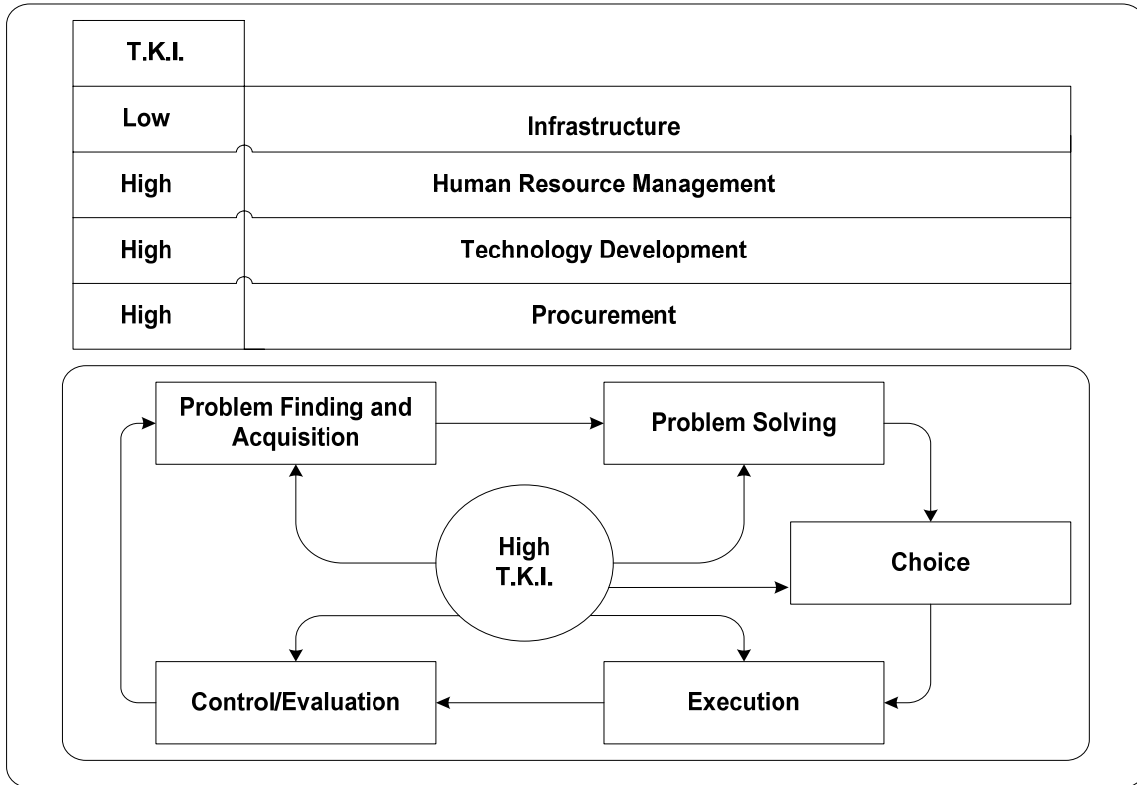


Fig. 5. Knowledge Intensity (T.K.I.) in Value shop

Regarding infrastructure operation the intensity is fairly less as these operations have proper procedures to follow. Figure 6 below depicts a value network [Stabell, Fjeldstad, 1998] and Tacit Knowledge Intensity (T.K.I.) in it.

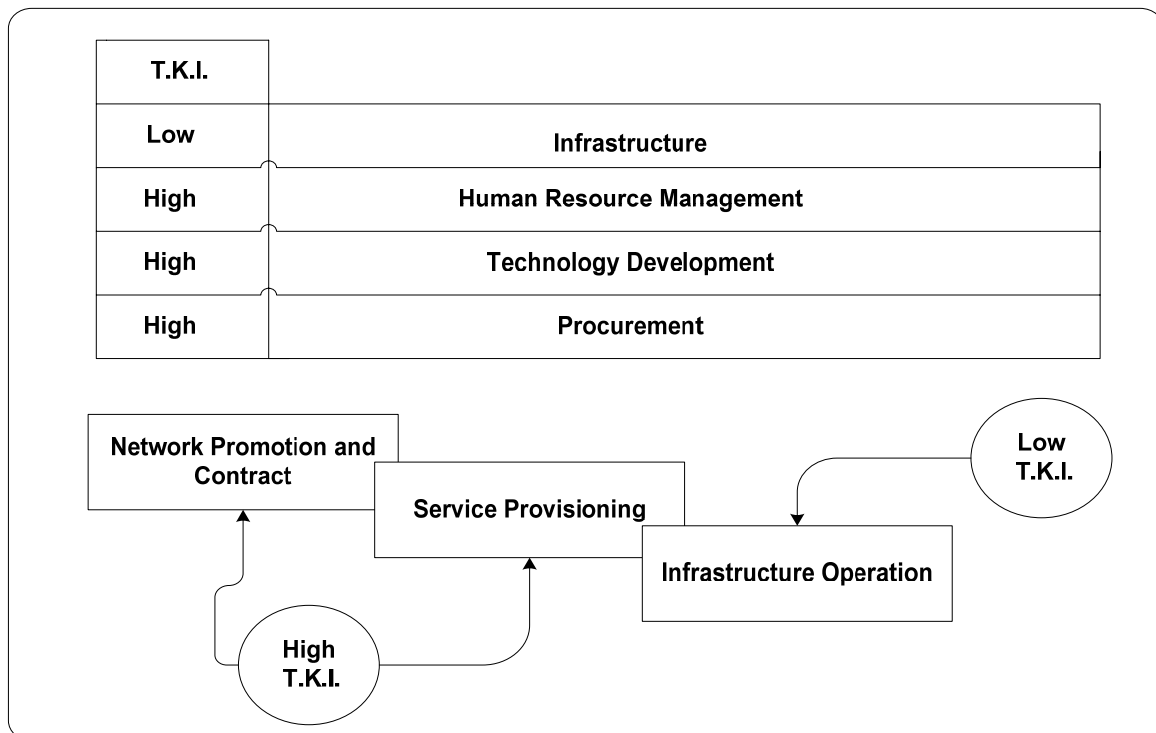


Fig. 6. Tacit Knowledge Intensity (T.K.I.) in Value Network

---



---

## Overall Tacit Knowledge Intensity

---

Value chain, value shop and value network are alternative value configurations that determine the intensity of tacit knowledge in an organization. Table 1 shows comparison between value configurations [Stabell, Fjeldstad, 1998] with an overall Tacit Knowledge Intensity (T.K.I.). Table 1 below shows the comparison between Value Configurations.

Table 1. Comparison between Value configurations

CHARACTERISTICS	VALUE CHAIN	VALUE SHOP	VALUE NETWORK
Value Creation Logic	Transformation of Inputs into Products	(Re)solving Customer or clients Problems	Linking Customers
Primary Technology	Long-Linked	Intensive	Mediating
Primary Activities Categories	Inbound Logistics Operations Outbound Logistics Marketing and Sales Service	Problem-finding and Acquisition Problem-solving Choice Execution Control/Evaluation	Network Promotion and Contract Management Service Provisioning Infrastructure Operation
Main Interactivity relationship Logic	Sequential	Cyclical, Spiraling	Simultaneously, parallel
Business Value System Structure	Interlinked Chains	Referred shops	Layered and interconnected Networks
<b>Overall T.K.I.</b>	<b>Moderate</b>	<b>High</b>	<b>Moderate</b>
Examples	Car Manufacturer	Law Firm	Telephone Company

---

## Conclusion

---

Tacit knowledge has always played a vital role in the successful business organizations and will continue to do so in future; only change will be the increase in intensity of its usage and applicability. Tacit knowledge as a significant and integral part of knowledge management has always been a challenge for an organization hence requires a special consideration. With the old workforce retiring and new ones highly fragile due to highly competitive markets it is important to take care of the tacit knowledge involved at various value creation stages in an organization so that knowledge and experience never expires and is readily available to the people requiring it at any moment of time as an imperishable resource. In this article, “**Tacit Knowledge Intensity**” (T.K.I.) is devised in order to highlight the significance of tacit knowledge in the form of its intensity at various value adding stages of an organization so that while incorporating mechanism to address tacit knowledge using knowledge management systems we are taking in account type of organization and intensity of tacit knowledge in it.

## Bibliography

---

- [Alavi, Leidner, 2001] Maryam Alavi, Dorothy E. Leidner, knowledge management and Knowledge management systems: Conceptual foundation and research issues, MIS Quarterly Vol. 25 No. 1/March 2001
- [Gottschalk, 2004] Dr. Gottschalk P., Strategic Knowledge Management Technology, Idea Group Publishing, UK 2004
- [Barney, 1991] Barney J., Firm resources and sustained competitive advantage, Journal of Management, 17:99-120, 1991
- [Hitt, Bieman, Shimizu, Kochhar, 2001] Michael A. Hitt, Leonard Bierman, Katsuhiko Shimizu, Rahul Kochhar, Direct and Moderate effects of human capital on strategy and performance in professional service firms: A resource based perspective, The academy of management journal, Vol. 44, No.1/ Feb. 2001
- [Porter, 1985] Michael E. Porter, The Value Chain and competitive Advantage: Creating and sustaining Superior performance: With a new introduction, Free Press, New York 1985
- [Thompson, 1967] Thompson J.D., Organization in action, McGraw- Hill, Irwin 1967
- [Stabell, Fjeldstad, 1998] Stabell C. B., Fjeldstad D., Strategic Management Journal, Vol.19, John Wiley and Sons Ltd., USA 1998
- [Harrington, Voehl, 2007] Harrington J., Voehl F., Knowledge Management Excellence, Paton Press LLC, USA 2007
- [Murray, 2008] Murray E. Jennex, Current Issues in Knowledge Management, Information Science Reference, USA 2008
- 

## Authors' Information

---



**Sumeer Chakuu, M.Phil.** –University of Information Technology and Management in Rzeszow, ul. Sucharskiego 2, 35-225, Rzeszow, Poland. ; e-mail: [schakuu@wsiz.rzeszow.pl](mailto:schakuu@wsiz.rzeszow.pl)

*Major Fields of Scientific Research: Knowledge Management systems, Methods to capture Tacit knowledge, Role of Knowledge management in future aviation.*

## KNOWLEDGE MANAGEMENT AS ACTIVE LABOUR MARKET POLICY DEVELOPMENT FACTOR

**Tatjana Bilevičienė, Eglė Bilevičiūtė**

**Abstract:** European Union gives priority to social policy, labor and employment, human resource development. The growing unemployment figures in most of European Union countries provide the need to overview possibilities to strengthen control and more efficiently regulate unemployment. Recently, active labor market policies (ALMP) are increasingly applied to the broader macro-economic, employment and social policy objectives. Active labor market policy measures include vocational training, employment promotion, direct job creation, business support for new entrepreneurs. Investing in ALMP measures is one of the most effective investments in labor market policy. Knowledge society and knowledge economy challenges change management models. Knowledge management could help to increase productiveness of employees, expanding sources of reachable for them knowledge. Human resource development and effective human resource management theory of mobility models for the coordination and efficiency of their practical application depends on the ability of organizations to integrate human resource management and knowledge management models. ALMP measures

---

---

*are directly linked to the learning process, innovation, ideas and competencies. Active labor market supports the management of individual or group learning. Research indicates that vocational training is closely linked to job skills and employment promotion subsidies. The principles of knowledge management (knowledge construction, knowledge embodiment, knowledge dissemination and use) disposed of the active labor market system. It can be argued that the only successful model of knowledge management can ensure the success of ALMP. In his article the author examines ALMP and knowledge management model, communication and applications.*

**Keywords:** *knowledge society, knowledge management, active social policy, labor market, active labor market policy.*

**ACM Classification Keywords:** *K.4.2. Social issues – Employment.*

---

## Introduction

---

Social policy is one of the economic adjustment measures, as a fundamental tool for the creation of the welfare state. Economics are open systems. They receive inflows of energy and materials. Economics use that incoming energy to develop and build new structures. Economy is primarily a social process, involving social factors: people, social groups, institutions and the state. These social entities: the first – an active force depends on economic growth, the second – a force that is closely connected with all areas of public life – politics, law, culture, ideology, family management. Social constraint reflects social aspect of system and added values that improve the quality of human life [Rudzkiene, Burinskiene, 2007]. European Union (EU) gives priority to social policy, labor and employment, human resource development. Social processes are not separated from economic change, they affect each other.

EU citizens' live and work is progressing rapidly, increasing the risk that the Social Security system is unsustainable. This process is linked to European and international economic integration, the new, particularly in information and communications technologies, the demographic aging of societies is still relatively low average level of employment. European Committee of Employment and Social Affairs acknowledged that the conditions of EU ground development should to be modified so that it reflects on today's political, economic and social realities [European Commission..., 2006]. In order to ensure prosperity and reduce the risks of social exclusion, it is necessary to modernize the social security system, more people to attract and keep active labor market policies.

Active labor market policies (ALMP) are very important in facilitating the most rapid employment of the unemployed and creating the right conditions for some economically inactive people back into the labor market, as well as addressing the problems of disadvantaged workers in the labor market. Major tasks of Small and Medium Enterprises (SME) promotion are to maintain the necessary jobs and create new jobs. This coincides with the active labor market policy objectives: promoting self-employment, structural unemployment and regional disparities in demand and supply deflection stop.

Knowledge management is optimal application of theoretical and practical knowledge in business processes – with purpose to reach durable advantage against rivals and bigger benefit of all shareholders of enterprise – investors, employees, managers, so common state benefit would be implemented [McGinn, 2001]. Transformation of modern society to knowledge society originates the absolutely new global social and economical contexts that require different management principles, skills, abilities and competences. The main factor of development of European economical space business organizations and economy would be the knowledge, generation of innovative products, perfection of production and management's methods.

---

### **Active labor market policies**

---

D. C. Vaughan-Whitehead [Vaughan-Whitehead, 2003] considers that the European social model can be defined as the EU and its Member States a set of legal rules and legal measures implemented to promote a coherent and comprehensive social policy in the EU. The key elements of the European social model are the current labor law, employment, equal opportunities, non-discrimination policy, employee participation, information and advice, and recognition of the social partners in decision-making process, social dialogue and collective bargaining, civil society, public services, fair earnings, social protection, social inclusion, ensure employment and social rights (workers and citizens in general), regional cohesion, social policy and international instruments [Melnikas, 2010]. Traditional universal welfare state hardly withstands globalization, liberalization and privatization influences and the different features of the traditional welfare states have a tendency to weaken.

Passive social protection is not effective in the sense that people returned back to the job market or at least maintain their ability to take care of themselves at home. The social reintegration of the results is important to show the real degree of effectiveness of social protection. Social reintegration is a key argument for the indisputable need for an active social policy and active social protection. Social policy, focused on increasing employment and reducing unemployment, wages and personal income growth is one of the most important investment and growth factors. In order to maintain an economically strong, stable and competitive position in the region, the European Union has formed a general policy, implemented in all EU member states. In 2010 the EU Commission has prepared the new ten-year economic recovery strategy. It provides innovative, sustainable and inclusive economic growth based on improved member states and the EU policy coordination vision [The Lisbon..., 2010].

Changing the system of unemployment insurance benefits and basic income support as well as the repertoire of active labor market policy instruments and making benefit receipt more conditional upon job search and acceptance of job offers was a major issue on the political agenda [Eichhorst et al., 2010]. *Passive* labor market policies are concerned with providing replacement income during periods of unemployment or job search, *active* policies emphasize labor market integration. Passive policies include unemployment insurance and early retirement measures; active measures include training, job creation measures, support for active job search, hiring subsidies and support for enterprise creation among the unemployed [Meager, 2009]. The European Commission recommends that the [European Commission, 2008] to ensure an integrated active inclusion policies work, in order to effectively address multiple dimensions of poverty and social exclusion.

Active labor market policies (ALMP) aim at enhancing labor market mobility and adjustment, facilitating the redeployment of workers to productive activities, and generally enabling people to seize new job opportunities as they arise [Armingeon, 2007].

---

### **Lithuania active labor market policies**

---

Welfare of state depends on its economic and labor market policies, the ability to per capital income. Employment is the most important way fully, actively and equally to participation in public life. The more people are participating in full-time or part-time labor market, the greater are their contribution to the availability of adequate social protection in that country. The expenditures for social welfare and unemployment regulation are every countries decision. S. Stoškus and D. Beržinskienė [Stoškus, Beržinskienė, 2002] define the employment model as the development of each individual and society concerning whole needs to find economic and social cohesion. Market economy, certain sections of the population in employment is becoming more vulnerable to failure to

adapt to changes in market relations. Growth of employment rates is the most efficient measure for economic growth and social inclusion, both to promote the economic protection for persons who can not work.

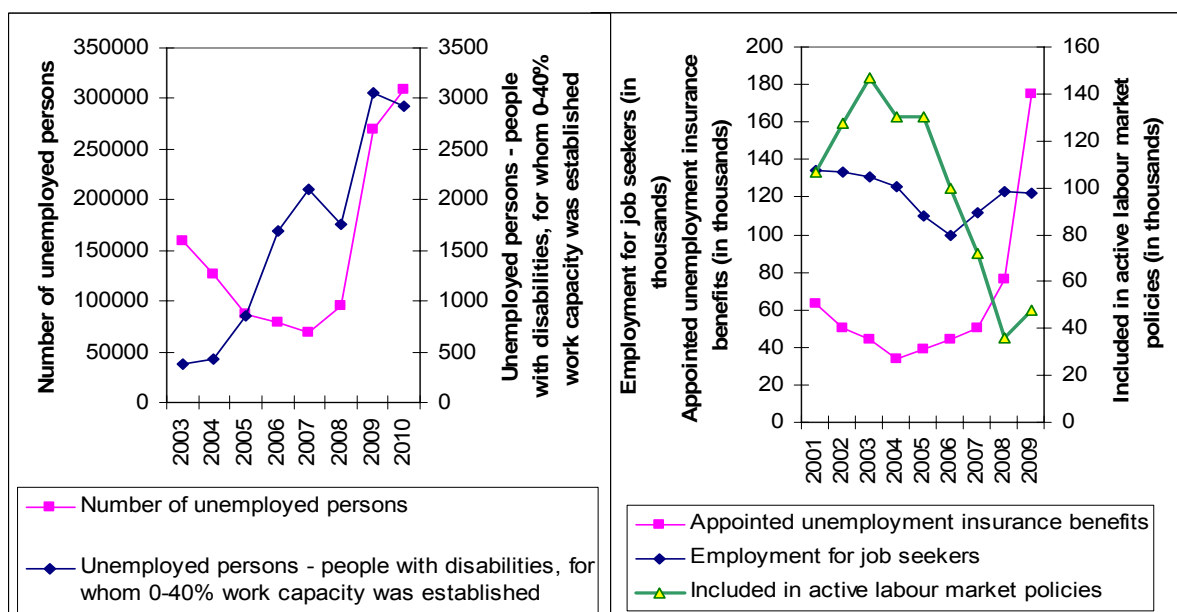


Fig. 1. Unemployment dynamics and dynamics of employment factors (Source: Statistics Lithuania [www.stat.gov.lt](http://www.stat.gov.lt))

EU employment policy is focused on socially vulnerable groups (e.g. long-term unemployed, people with disabilities), problem solving and quality of employment services and increasing the availability of closer cooperation with employers and social partners in development. All dimensions of sustainability should be considered in the process of strategic sustainable development planning. Strategic plan helps create management and planning systems of municipalities, based on the principles of sustainable development, democracy and market economy, and to assist for more rational use of limited resources of municipal budgets and for better coordination and implementation of programs in various sectors. *Lithuania for 2007–2013 EU Structural Assistance Strategy* proposes that it is important to the Lithuanian economy to create more new and better jobs. Such job growth would create more additional values to the Lithuanian economy and stop the drain of skilled manpower to foreign countries [Čiegis, Gineitienė, 2008].

However, the analysis of the Lithuanian Labor Exchange registered unemployed persons' metric (see Figure 1) we could see the continued growth in 2008. It is also a growing number of registered people with disabilities employment office, which determined 0 to 40 % work rate, the number of a relative decrease observed in 2008 and 2010. This indicates the need for more effective employment support measures.

The variety of tools used in different countries allows make a choice from the alternatives that could be useful in nowadays economic situation. Active labor market policies (ALMP) measures are an important tool in the implementation of the Lisbon Strategy. Lately, they are increasingly subject to the broader macro-economic, employment and social policy objectives. In the scientific literature noted that active labor market policies must cover all the objectives of diversity. These are: job creation, job reallocation of skills and human capital deepening, behavior (with) change, overcoming the timidity of job-seekers and the alienation of labor income increase, the broader macro-economic objectives, such as the potential labor supply, structural reduction of unemployment [Moskvina, 2008, Lapinskienė, Tvaronavičienė, 2009, Meager, 2009].

ALMP program of the Lithuanian Labor Exchange are subject, from its earliest beginnings in 1991. These measures serve job-seekers of employment growth, unemployment reduction and mitigation of negative consequences of a labor demand and supply-side alignment to maintain balance in the labor market and job

seekers through the working-age population in employment opportunities will be provided. Active labor market policy measures and procedures in terms of specification [Dėl aktyvios, 2009] provides the unemployed and those facing redundancy of working age employees in vocational training, supported employment, assistance for job creation and promotion of territorial mobility of the unemployed, the conditions and procedures. Active labor market policies include: unemployed and those facing redundancy of working age workers vocational training; supported employment (employment subsidies, job skills promotion, job rotation, public works); support for job creation (job creation subsidies, local employment initiatives, projects, self-supported employment); unemployed territorial mobility support.

Active unemployment regulation tools are more efficient and provide long-term effect. Vocational training, professional skills upgrading, consulting unemployed and employers can ensure work places in the future. The greatest attention should be paid towards entrepreneurship and support for small and medium business because that causes possibilities for some unemployed people to become self-employed and open some more new work places [Sakiene, 2010]. Lithuanian Labor Exchange, on behalf of *Active labor market policy effectiveness study* [Aktyvios..., 2007] showed ALMP measures of economic performance indicators of the fluctuations of 19.7% to 48.1%. This showed that investment in ALMP measures is one of the most effective investments in labor market policy.

In the most recent works on sustainable development, the social environment is looked upon as an absolutely equivalent factor, which influences social development to the same extent as economic growth or environmental sustainability [Misiūnas, Balsytė, 2008]. Equal opportunities are one of the key objectives of a democratic society. Additional choices include political, economic and social freedoms and opportunities to develop and manufacture, to live with respect for oneself and with human rights guarantees. While the significance of human resource development increases, the role of human resource development within the organization decreases, because a part of the work is transferred to specialized organizations, managers and colleagues participating more actively in the human resource development work [Kumpikaitė, 2008].

Dynamics of recruitment's factors in Lithuania (see Figure 1) indicates the necessity of active labor market policies' improvement. According data of Department of Statistics every year the number of employed persons differs just a little. Since 2005 designated by a growing unemployment insurance benefits, particularly noticeable in 2008 and 2009. At the same time, the steady decrease in active labor market policy number (the slight increase seen in 2009).

Analyzing the range of unemployment and total LMP measures dynamics (see Figure 2) we could see that unemployment range decreases (strong negative correlation  $r = -0,88$ ). Although the employment range depends on education level (see Figure 2). As higher education, so opportunity to be employment is higher. It confirms the necessity of ALMP application. The positive effects of ALMP vary by program type.

Raising the average educational level has economic and non-economic effects on society. A better educated population is not only able to perform higher-skilled jobs, but is also more likely to participate in the labor market. In this paper the potential benefits of investment in education are estimated. This estimation is based on three possible effects of an increase in the average educational level: improvement of the average earned salary, improvement of the average probability to find a job and positive non-economic effects, for instance on health and criminal behavior [Zandvliet at al., 2009]. The unemployed and those facing redundancy working life of employees training goal – to qualify, or (and) to acquire skills, needed employment.

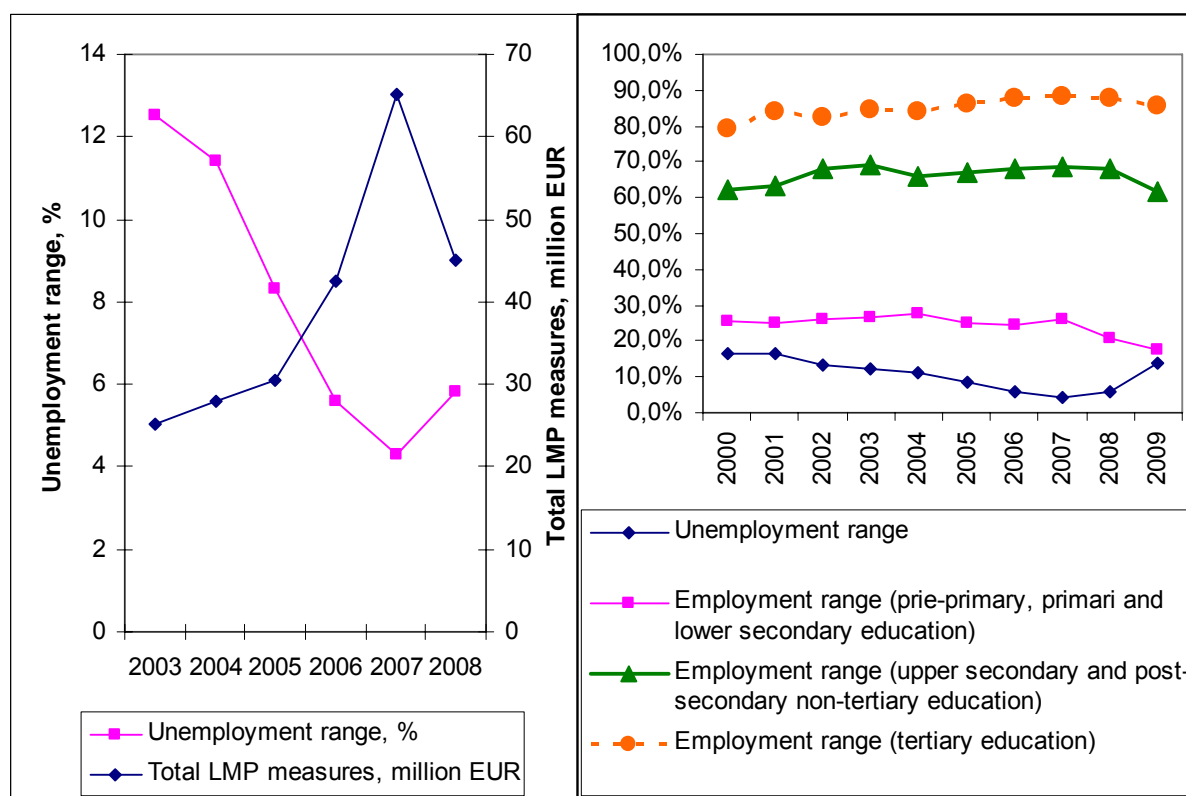


Fig. 2. Employment range and total LMP measures dynamics (Source: European Commission, 2010)

The examination of the Department of Statistics data shows that, people start working in the profession or occupation, number of dynamics coincides with the dynamics of vocational training (see Figure 3). It is the purpose of staff preparation is an important and useful.

Employment by subsidies aim – to help the labor market, further supported by the persons registered in the employment exchange, enter the labor market or temporary employment, and individuals who set up the level of 40% of capacity or severe disability rates, create special conditions in the labor force. Support for job skills to is to provide opportunities for job seekers to acquire skills gaps directly in the workplace. It can be argued (see Figure 3) that vocational training is closely linked to job skills and employment promotion subsidies, because of their similar behavior. It is important that the employee is prepared according to the program ordered by the employer, and that his employment would be based.

D. Gallie [Gallie, 2007] highlights the need to foster, through strong initial vocational training systems, specialized skills across the broad spectrum of the workforce (skilled manual workers, technicians, and engineers). Such skills should combine both industry-specific technological knowledge with company-specific knowledge of organization, processes, and products.

Training increases the expected productivity of the worker. The government can stimulate training by subsidizing training costs. We also take into account an alternative route to higher productivity. The government can stimulate the on-the-job training route by subsidizing the creation of vacancies. Simply because there are more vacancies, unemployed will flow more quickly into jobs and through learning by doing they flow from low productivity to high-productivity jobs (hence the transition from unemployment to high-productivity jobs happens more quickly) [Boome, Van Ours, 2009]. An increase in expenditures on labor market training causes unemployment to fall. The effect of expenditures on labor market training is larger the higher unemployment benefits are.



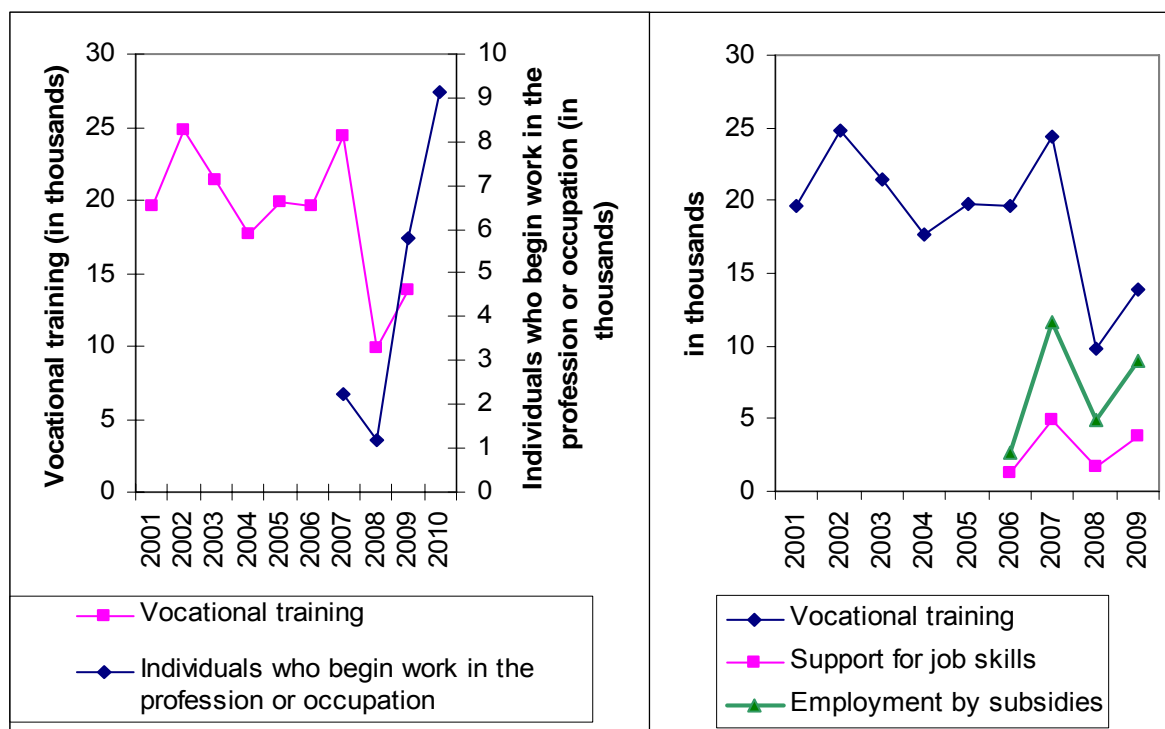


Fig. 3. Vocational training, job skills and employment promotion by subsidy dynamics (Source: Statistics Lithuania [www.stat.gov.lt](http://www.stat.gov.lt))

Training interventions are effective tools for integrating the unemployed and inactive into the labor market. The notion that skills-based measures might have an important role to play is intuitively plausible. At an aggregate level, there is a strong relationship between levels of initial education and continuing vocational training on the one hand and employment performance on the other [Meager, 2009]. Similarly, at the individual level, there is a strong relationship between training experience and the probability of being in work.

Job creation subsidy is organized labor of registered job seekers in employment support for an indefinite period. Subsidy paid to employers who employ people on permanent contracts for these belong to the labor market, further supported by the different categories of persons with disabilities of working age, employment office registered job seekers, who set up 25% of capacity or severe level of disability, level of working-age people, registered with the Labor Exchange unemployed people who set the level 30–55% of capacity or mild or moderate disability level. Support for job creation subsidies is provided to the employers creating new jobs or adapting the existing ones to the disability needs and hiring unemployed people under open-ended employment contracts, to former unemployed people within 36 months of the date of company registration or to entities implementing local employment initiatives to create jobs for unemployed registered with territorial labor exchanges.

If the problem of inefficient management and insufficient institutional quality is not properly assessed, and no attempts to solve it are made, it may lead to a situation when it is impossible to achieve the general country's sustainability either in the present or in the future [Misiūnas, Balsytė, 2009].

### The role of Knowledge Management in active labor market policies

Knowledge economy and knowledge society in terms of human resource development and effective human resource management theory of mobility models for the coordination and efficiency of their practical application depends on the ability of organizations to integrate human resource management and knowledge management

---

---

models. Conjunction of different knowledge parts to management of strategic intellectual capital brings people to new practice of management in information age. Knowledge of the different parts into a strategic management of intellectual capital brings people to the new boundaries of knowledge management practices in the information age. Knowledge management's essence becomes the management of individuals with particular skills and experience, with purpose to encourage particular behavioral models in organization and interaction of individual employees – socialization.

Knowledge Management (KM) comprises a range of strategies and practices that deal with how knowledge is acquired, transferred, and shared with all the members of the organization. Such strategies and practices seek to achieve the organization's objectives. Knowledge Management System (KMS) refers to a comprehensive information and communication technology platform used for managing knowledge in organizations for supporting creation, capture, storage and dissemination of information [Aktharsha, Anisa].

Knowledge management is a management discipline that seeks to increase aid effectiveness by adapting the business people, processes and technology synergies. Modern organization should comprehend knowledge management and implement it inside. Knowledge management is manage mental instrument supporting by different measures to create working environment in that seeking the best result they optimally create, spread and use their and others knowledge. The main result of knowledge management is environment stimulating employees to create, spread, keep and apply knowledge and consisting of all processes, roles, measures and structures that let to implement it.

In today's information society and turbulent environments all citizens have to be engaged in lifelong learning and self-development. Personal Knowledge Management as a concept based on wide range of individuals skills and competences undoubtedly can be a support of employability. Skills and competences are the crucial factors leading to success in self management including knowledge management, career management and employability management [Świgoń, 2011]. The ALMP analysis shows that there is the necessity for new programs, procedures, staff training, acquisition of new competencies, so there is the necessary of human resources management and knowledge management models for synthetic application.

In the knowledge society, a high level of economic performance and good living standards can only be achieved if an increasing share of the population attains a high level of education. Knowledge management is a complex process. Sandra Rodney McAdam and McCreedy [McAdam, McCreedy, 1999] proposed modified version of Demerest's knowledge management model. This model takes a balanced approach between scientific and socially constructed knowledge. Also the uses of KM are viewed as both emancipatory and as business oriented. Analysis of ALMP confirms that its main elements are closely connected with main KM model elements: knowledge construction, knowledge embodiment, knowledge dissemination and use. So, we can apply the McAdam and McCreedy KM model for realization of ALMP. Model suggested by authors is presented on Figure 4.

In this model we can distinguish *Vocational training* element. All ALMP measures are based on purchase of new knowledge and skills. Training programs are the most widely used active labor market measure in Europe. The assessment of their effectiveness shows rather mixed results; treatment effect estimates are negative in a few cases, and often insignificant or modestly positive. The training programs for unemployed may enable individuals to develop new skills and provide possibility increase their income in the future comparing to their current market wages and also facilitate labor market flexibility as the economies are being transferred to services and high technology [Sakiene, 2010, Bergemann at al., 2009].

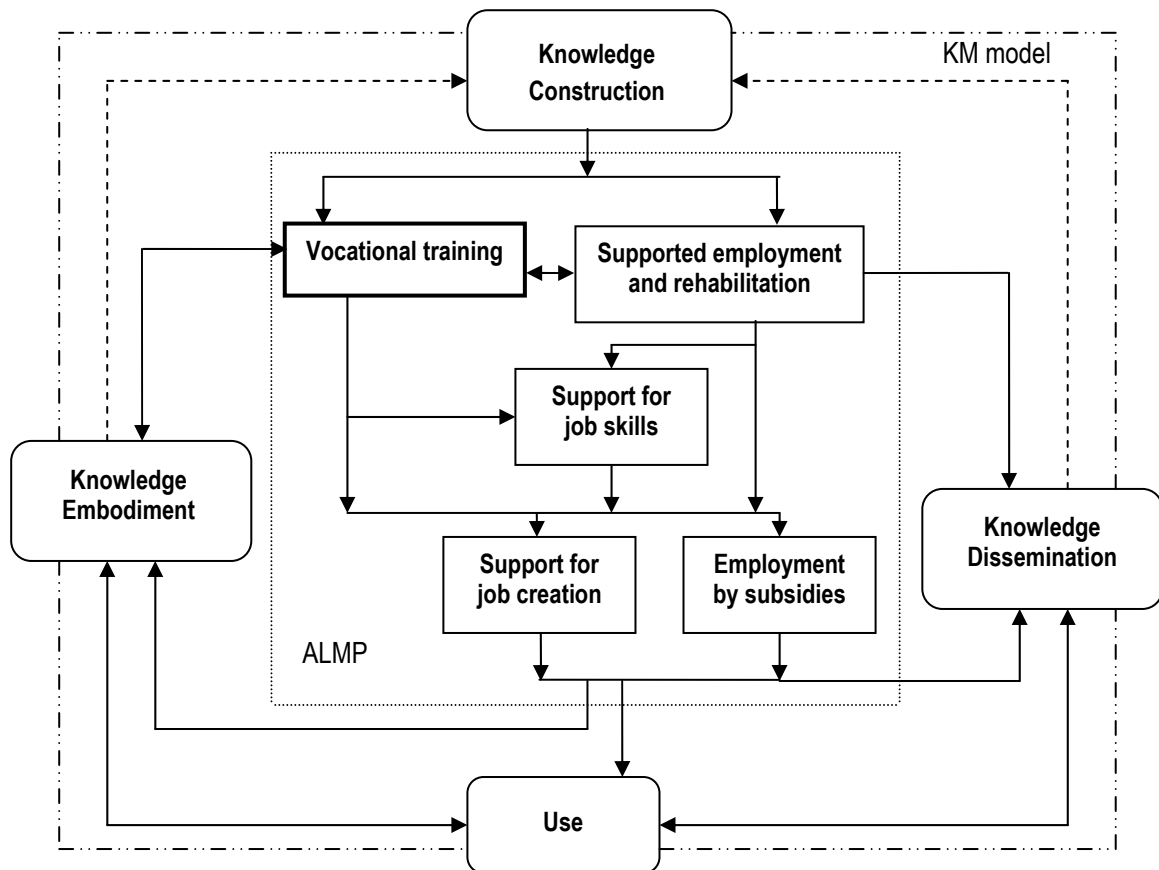


Fig. 4. ALMP and knowledge management model

Training programs involving private sector placements with on-the-job training are seen as having *strong* market orientation, while classroom-based schemes have *weak* market orientation. Similarly, on the demand side, traditional direct job-creation measures are weakly market oriented, while indirect measures subsidizing jobs in the private sector are strongly market oriented. In particular, programs with a stronger market orientation led to higher placement rates, longer job durations and higher earnings, than schemes with weak market linkages. Thus, the specific skills training program has far more positive impacts than the general training programs or the job-creation schemes. Their results also highlighted a need for targeting on the most disadvantaged groups [Meager, 2009]. The model provides continuous feedback for all stages.

## Conclusion

EU social policy is focused on improving the European social model. Social risk of social benefits is often the only one feasible mean of social protection, but active social policy helps to persons who want to work. European Employment Strategy provides to reach the full employment, quality and productivity and social cohesion of it. Active labor market policies could ensure the unemployed for people with disabilities to participate in the labor market. Create the necessary productive employment opportunities and ensure continued livelihoods is one of the most important and difficult task of every society. EU employment policy is focused on improving the European social model. Lithuania active labor market policies are described as the domestic legislation provides for measures to help job seekers improve their employment opportunities and improve job and prepare a balance between the ratios of skilled workers.

ALMP Department of Statistics data shows that people start working in the profession or occupation, number of dynamics coincides with the dynamics of training and preparation of the target employees is an important and

useful. However, training does not affect the activities of the dynamics of a business license, do not encourage entrepreneurship. Vocational training is closely linked to job skills and employment promotion subsidies, similar to their behavior. The employee is prepared according to the program ordered by the employer, and that his employment would be supported.

Knowledge management is connected with innovations, inter-connections, ideas, competences, structures. This management supports individual or groups' education, stimulates and enhances spread of experience, distribution of failures and good practises, choice of optimal solutions. Knowledge management technologies could be used for stimulation of dialogues, bargains, communication, but it is not essence of such management.

Analysis of ALMP confirms that its main elements are closely connected with main KM model elements: knowledge construction, knowledge embodiment, knowledge dissemination and use. So, we can apply the McAdam and McCreedy KM model for realization of ALMP. Model suggested by authors collects not only support of the feedback system, but after each cycle to assess the performance and result in a higher improved level of knowledge creation.

---

## Bibliography

---

- [Aktharsha, Anisa] U. S. Aktharsha, H. Anisa. Knowledge Management System and Learning Organization: An Empirical Study in an Engineering Organization.  
<[http://www.iupindia.in/411/Knowledge%20Management/Knowledge\\_Management\\_and\\_Learning\\_Organization\\_26.html](http://www.iupindia.in/411/Knowledge%20Management/Knowledge_Management_and_Learning_Organization_26.html)> [visited 2011 05 25].
- [Aktyvios..., 2007] Aktyvios darbo rinkos politikos priemonių efektyvumo tyrimas. IV-ojo mokslinio tyrimo etapo ataskaita. 2007. <<http://www.idb.lt/Informacija/Apie/Documents/ADRPP%20efektyvumo%20tyrimas.pdf>> [visited 2011 05 20].
- [Armingeon, 2007] K. Armingeon. Active labour market policy, international organizations and domestic politics. *Journal of European Public Policy*, 14:6, 2007, p. 905–932.
- [Bergemann at al., 2009] A. Bergemann, B. Fitzenberger, S. Speckesser. Evaluating the Dynamic Employment Effects of Training Programs in East Germany Using Conditional Difference-In-Differences. *Journal of Applied Econometrics*, 24 (2009), p. 797–823.
- [Boome, Van Ours, 2009] J. Boone, J. C. Van Ours. Bringing Unemployed Back to Work: Effective Active Labor Market Policies. *De Economist*, 157, No. 3, 2009, p. 293–313.
- [Čiegis, Gineitienė, 2008] R. Čiegis, D. Gineitienė, D. Participatory Aspects of Strategic Sustainable Development Planning in Local Communities: Experience of Lithuania, *Ūkio technologinis ir ekonominis vystymas – Technological and Economic Development of Economy* 14(2), 2008, p. 107–117.
- [Dėl aktyvios..., 2009] Dėl aktyvios darbo rinkos politikos priemonių įgyvendinimo sąlygų ir tvarkos aprašo patvirtinimo. Lietuvos Respublikos Socialinės apsaugos ir darbo ministro įsakymas Nr. A1-499, 2009.
- [Eichhorst at al., 2010] W. Eichhorst, M. Grienberger-Zingerle, R. Konle-Seidl. Activating Labor Market and Social Policies in Germany: From Status Protection to Basic Income Support. *German Policy Studies*, Vol. 6, No. 1, 2010, p. 65-106.
- [European Commission..., 2006] European Commission and Council of the Union. A European social model for the future. 2005/2248(INI). <[http://www.socialeconomy.eu.org/spip.php?article1296&debut\\_articles\\_rubrique=7](http://www.socialeconomy.eu.org/spip.php?article1296&debut_articles_rubrique=7)> [visited 2011 05 25].
- [European Commission, 2010] European Commission. The Social Situation in the European Union 2009. Directorate-General for Employment, Social Affairs and Equal Opportunities – Unit E.1. Eurostat – Unit F.4. Manuscript completed in February 2010.
- [Gallie, 2007] D. Gallie. *Employment Regimes and the Quality of Work*. University of Oxford, 2007.
- [Kumpikaitė, 2008] V. Kumpikaitė. Human Resource Development in Learning Organization, *Journal of business economics and management* 1(9), 2008, p. 25–31.
- [Lapinskienė, Tvaronavičienė, 2009] G. Lapinskienė, M. Tvaronavičienė. Darnusis vystymasis Centrinėje ir Rytų Europoje: pagrindiniai ekonominio augimo aspektai, *Verslas: teorija ir praktika* 3(10), 2009, p. 204–213.
- [McAdam, McCreedy, 1999] R. McAdam, S. McCreedy. A critical review of knowledge management models. *The Learning Organization*, Vol 6, Issue 3, 1999, p. 91-101.

- [McGinn, 2001] N.F.McGinn. Knowledge Management in the Corporate Sector: Implications for Education. 2001. <[http://www.norrag.org/db\\_read\\_article.php?id=734](http://www.norrag.org/db_read_article.php?id=734)> [visited 2009 04 05].
- [Meager , 2009] N. Meager. The role of training and skills development in active labor market policies. International Journal of Training and Development, 13:1, 2009, Blackwell Publishing Ltd, p. 1-18.
- [Melnikas, 2010] B. Melnikas. Sustainable Development and Creation of The Knowledge Economy: The New Theoretical Approach, Technological and Economic Development of Economy 16(3), 2010, p. 516–540.
- [Misiūnas, Balsytė, 2008] A. Misiūnas, I. Balsytė, I. The Essence of Sustainable Social Development and Possibilities for Measuring It, Intelektinė Ekonomika – Intellectual Economics 1(5), 2009, p. 61–71.
- [Moskvina, 2008] J. Moskvina, J. Aktyvios darbo rinkos politikos priemonių vertinimas. Probleminiai klausimai, Filosofija. Sociologija 19(4), 2008, p. 1–9.
- [Rudzkienė, Burinskienė, 2007] V. Rudzkienė, M Burinskienė. Assessment of Transformation Processes in The Complex Socio-Economic System of Transition Period, Intelektinė Ekonomika – Intellectual Economics 1, 2007, p. 74-81.
- [Sakiene, 2010] H. Sakiene. Analysis of Unemployment Regulation Tools in Lithuania. Economics and Management, 2010. 15, p. 219-225.
- [Stoškus, Beržinskienė, 2002] S. Stoškus, D. Beržinskienė. The Influence of Human on the Structural Labour Market Changes, Inžinerinė ekonomika – Engineering Economics 2(28), 2002, p. 57–60.
- [Świgoń, 2011] M. Świgoń. Personal Knowledge Management (PKM) and Personal Employability Management (PEM) – Concepts Based on Competences. Proceedings of The 3rd European Conference on Intellectual Capital held at University of Nicosia, Cyprus 18-19 April 2011, p. 432-438.
- [The Lisbon..., 2010] The Lisbon Review 2010 Towards a More Competitive Europe? World Economic Forum. <<http://www.weforum.org/en/initiatives/gcp/Lisbon%20Review/index.htm>> [visited 2011 04 05].
- [Vaughan-Whitehead, 2003] D. C. Vaughan–Whitehead, EU Enlargement versus Social Europe? The Uncertain Future of European Social Model. Cheltenham Northampton: Edward Elgar, USA, 2003.
- [Zandvliet et al., 2009] K. Zandvliet, T. Berretty, M. Collewet, O. Tanis. Activating Potential IC: Effective Addition to Active Labour Market Policy in Rotterdam? Proceedings of The European Conference on Intellectual Capital held at INHolland University of Applied Sciences, Haarlem, The Netherlands 28-29 April 2009, p. 581-588.

---

### Authors' Information

---



**Eglė Bilevičiūtė** – Mykolas Romeris University, Faculty of Law, Department of Administrative Law and Procedure, PhD, Professor, Ateities 20, LT-08303 Vilnius, Lithuania, e-mail: [eglek@mruni.eu](mailto:eglek@mruni.eu)

*Major Fields of Scientific Research: Processing of data and statistical methods for social scientific researches. Her current research interest includes law of research and studies, management of research, administrative law, forensic science, legal informatics, implementation of IT in law.*



**Tatjana Bilevičienė** – Mykolas Romeris University, Faculty of Economics and Finance Management, Department of Business Economics PhD, Assistant Professor, Ateities 20, LT-08303 Vilnius, Lithuania, e-mail: [tbilev@mruni.eu](mailto:tbilev@mruni.eu)

*Major Fields of Scientific Research: The e-inclusion problems of disabled persons is the main scientific research field of Tatjana Bilevičienė. In 2008 Tatjana Bilevičienė readied the equivalency doctoral dissertation that firstly in Lithuania presents the organisational model of telework of disabled persons and evaluation methodises of disabled persons' employment's quality.*

---

## Data Mining, Knowledge Acquisition

---

### DATABASE SERVER USAGE IN THE SOCIAL NETWORKS ANALYSIS

**Katarzyna Haręźlak**

**Abstract:** *In the field of computer science, topics regarding analysis of social networks grows more and more popular. It stems from the fact that the knowledge hidden in connections between people might enable us to answer many questions concerning various areas of life. When analyzing social networks, we are usually interested in such information as – network structure, how it is managed and what is the scheme of flow of goods or information. When starting a study of processes taking place in social networks it can be noticed that a lot of data subject to analysis is stored in structures of databases. Thus, database server's capabilities in terms of social network examining and assigning roles to its members are worth analysing. The aim of this paper is to present studies regarding using a database serve to gather data on social networks and mechanisms it provides to specify the type of connections between network elements.*

*One of the tasks of social network analysis is a graphic visualization of connections between various objects existing in the field of interest of a conducted research. The most convenient representation of such is a graph. Ensuring a possibility to gather data for a permanent storage and at the same time effectively operate on them required designing appropriate data structures. It was decided, that the relational database meets the demands. The basic structure of the proposed data model covers vital graph elements along with a description of their qualities.*

*Analysis of social networks presented as graphs poses a challenge when it consists of a huge number of nodes and connections between them. One of the most important functions for presenting a social network is to make it possible to locate and visualize only fragments of the whole graph in order to make analyses more effective. In the research, the mechanism for limiting set of nodes (creating sub-graphs, called paths for the purpose of this research) and for aggregating elements of a graph were prepared. Moreover, problem of graph search was also taken into consideration. The two algorithms, BFS and Dijkstra's, were implemented as a database procedures. Effectiveness of the operation of the first of algorithms was compared with its implementation in the. NET environment as well.*

**Keywords:** *social network analysis, database server, graphs, shortest paths.*

**ACM Classification Keywords:** *H. Information Systems H.3 INFORMATION STORAGE AND RETRIEVAL H.3.3 Information Search and Retrieval*

---

#### Introduction

---

In the field of computer science, topics regarding analysis of social networks grows more and more popular. It stems from the fact that the knowledge hidden in connections between people might enable us to answer many

questions concerning various areas of life [Chau, 2006, Giran, 2002, Razmerita, 2009]. When analyzing social networks, we are usually interested in such information as – network structure, how it is managed and what is the scheme of flow of goods or information [Newman, 2004]. Regardless of size and purpose of such network, there can always be made a distinction between the core and the periphery of the network, depicting the division of power, influence or status inside the network [Williams, 2001]. A more in-depth analysis of tasks performed in a social network leads to a more precise specification of roles and their assignment to particular person. It is used to describe the behaviour of a node in relation with the rest of a network as a whole. Among the roles, we can enumerate Organisers, Isolators, Communicators, Guards, Expanders, Supervisors, Connectors, Soldiers, Recruits, Neutral [Piekaj, 2007] and Ambassadors, Big Fish, Bridge [Scripps, 2007]. In order to specify a role of a particular node in a network many measurements have been created, the most popular of which are degree, closeness and betweenness [Guangming, 2009, Shaikh, 2006].

**Degree centrality** is a measure describing the number of connections of a particular node with other nodes in a network. High value of this measure may indicate that the person connected with such a node may be a leader of a group.

**Betweenness centrality** is described by a number of shortest paths between each pair of nodes crossing a particular node, which indicates that this node plays a vital role in transmitting information – they can thus be thought of as broker or gatekeeper.

**Closeness centrality** is a measure describing the distance between a particular node and other nodes. It is computed using a number of shortest paths connected with this node. High value of this measure also can indicate leaders of a group.

When starting a study of processes taking place in social networks it can be noticed that a lot of data subject to analysis is stored in structures of databases. Thus, database server's capabilities in terms of social network examining and assigning roles to its members are worth analysing. The aim of this paper is to present studies regarding using a database serve to gather data on social networks and mechanisms it provides to specify the type of connections between network elements.

---

### **The representation of a social network**

---

One of the tasks of social network analysis is a graphic visualization of connections between various objects existing in the field of interest of a conducted research. The most convenient representation of such is a graph  $G$ , with a weight function  $G=(V,E:w)$ , where  $V$  stands for a set of nodes,  $E$  is a set of connections,  $w$  is a function mapping each connection  $(u,v) \in E$ , to a weight  $w_{uv}$ , which shows the strength of connection between  $u$  and  $v$  [Yang, 2006]. Each node represents a person. Connection between nodes represents a type of a relation between two people. These graphs, due to the multiple analysis, should have a feature of durability, which means that a once-built graph should be kept for a renewed study.

#### **Graph data model**

Ensuring a possibility to gather data for a permanent storage and at the same time effectively operate on them required designing appropriate data structures. It was decided, that the relational database meets the demands. The basic structure of the proposed data model covers vital graph elements along with a description of their qualities. In this structure, the graph itself is an superior element. Graphs created by the user may concern various domains (billings, criminal network, connections between companies' managers), and within the domain various groups of objects represented by nodes (person, car, phone number). The last ones are usually described by a name, but, depending on the domain of the graph, their set of features may be different. For example, different figures are used to describe people and phone calls. The situation is very similar when

it comes to edges. There are two basic features, the direction of the connection and its weight, defining four types of graphs (undirected and unweighted, directed, weighted, directed and weighted). The other features are dependent on the type of graph. Taking this dependencies into account, a special model of data was elaborated, ensuring flexibility in determining a graph structure.

### Graph paths and their versions

Analysis of social networks presented as graphs poses a challenge when it consists of a huge number of nodes and connections between them [Buja, 2008]. The main problem faced by software visualising such networks is selecting appropriate methods of presenting them, ensuring proper deployment of nodes in a two or three-dimensional scope [Xu, 2005]. By proper, guaranteeing similar length of all edges and a minimal number of edges crossings is meant [Kulmar, 1994]. Furthermore, one of the most important functions for presenting a social network, is to make it possible to locate and visualize only fragments of the whole graph in order to make analyses more effective. Such fragments can be obtained by [Harezlak, 2010]:

- zooming in on a neighbourhood of a particular node,
- limiting set of nodes (creating sub-graphs, called paths for the purpose of this research),
- aggregating elements of a graph.

Some of these operations can be performed by an analytical application and some part of the functionality can be transferred to a database server. In the research, the mechanisms for the two last were prepared. Along with this work came the need to extend the graph with elements enabling keeping track of changes in graph structure resulting from an aggregation of its nodes. This operation entails an aggregation of edges connecting the aforementioned elements. This is why two tables **superiorNodes** and **superiorEdges** were added to the existing objects to enable collecting Master-Slave relationships. Their task is to gather all information on merged nodes or edges and new graph elements created on their basis. Thereby, in the object representing graph elements a special attribute, **AggregationIndex**, was added. It indicates on which level of graph development the element was created (positive value of the index) or aggregated (negative value of the index). Let us analyze the graph presented in the figure 1.

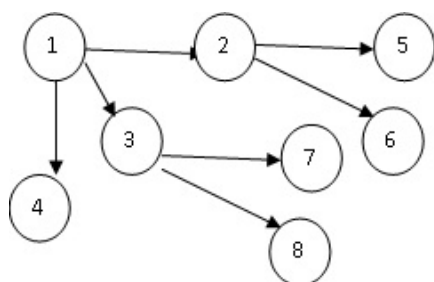


Fig. 1. A sample graph

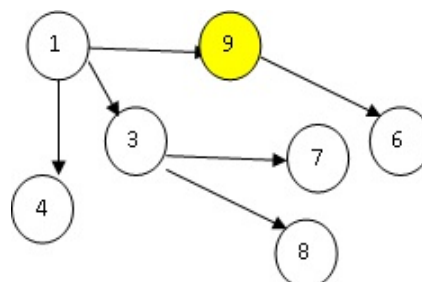


Fig. 2. A sample graph with integrated nodes

When a graph is being created, index attribute of all nodes is set to 1. In case when a user decides to aggregate nodes 2 and 5, a new node - number 9 - appears, and nodes 2 and 5 are no longer visible to the user, as presented in the figure 2.

The described operations causes a change of index values for nodes 2 and 5 - it is set to -2 (elements were deleted in the second version of the graph), and node number 9 is added to a database with the aggregation index set to 2 (the element was created in the second version of the graph). The actions being performed when deleting edges 1-2, 2-5 and 2-6, and creating 1-9 and 9-6 are very similar.

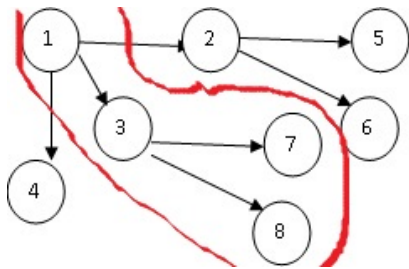


During the further research, a new functionality, allowing user for fragmentary analysis of a graph, was introduced as well. This way, idea of a graph **path** and its **version** was also proposed. A path is any sub-graph analyzed by a user at a given moment (figure 3).

With this subgraph a user can aggregate nodes and edges, creating consecutive versions of a path. Elements belonging to a version of a particular path have their own state, which equals the aggregation index from the above-mentioned solution. The state, if positive, corresponds to the version in which the element was created. If negative, it indicates that the element is deleted and which version of the path it was aggregated in.

In accordance with the data model proposed for performing this kind of operations, a full graph is at the same time its first path. Thus, the path presented in figure 3 is numbered 2, with a version number 1.

Described mechanisms are included in special database procedures, which input data consists of:



- graph identifier,
- path identifier,
- current version of the path,
- number and identifiers of aggregated nodes.

Fig. 3. A sample graph with a marked path

The effect of the procedure execution is creating another version of a chosen path of a particular graph and elements connected with it. It consists of a logical removal of the aggregated elements by changing their status and generating new objects instead, marking the version of their creation.

**Experiments**

The algorithms operating on a graph nodes implemented in the database server were tested using a sample data set. For the purpose of the test, data was imported to a model representing a graph. People became nodes, and connections between them edges. The import was performed taking directed connections into account. A database server chosen for the research was MS SQL Server 2008. The sample application developed during research was used for graphic presentation of the graph.

Proposed algorithms were tested in many variants of paths for the graph consisting of 1269744 nodes 5196852 edges. One of variants concerned five nodes with 107934, 110241, 142458, 142846, 132206 identifiers (marked in the figure 4 with graphic symbols OSO-01331206, OSO-01339508, OSO-01330596, OSO-01339610 OSO-01336904). They represent multi-level tree structure. Nodes on the last level of this structures (OSO-01331206, OSO-01339508) do not have any connections. The task of the node 132206 (OSO-01336904) is to connect a newly created node and the rest of the graph. Before aggregation it is connected to only one node from the aggregated structure. In such case, running the procedure resulted in (table 1):

Table 1. Nodes table

Node_ID	AggregationIndex
12231349	2
107934	-2
110241	-2
132206	1
142458	-2
142846	-2

- creation of one new node identified by 12231349 with aggregation index equal to 1,
- change of this index for aggregated nodes.

Table 2. SuperiorNodes table

SuperiorNode_ID	subordinateNode_ID
12231349	107934
12231349	110241
12231349	142458
12231349	142846

In the **SuperiorNodes** table four records can be found this time. Their aggregation brought to “life” a new node – 12231349 (table 2)

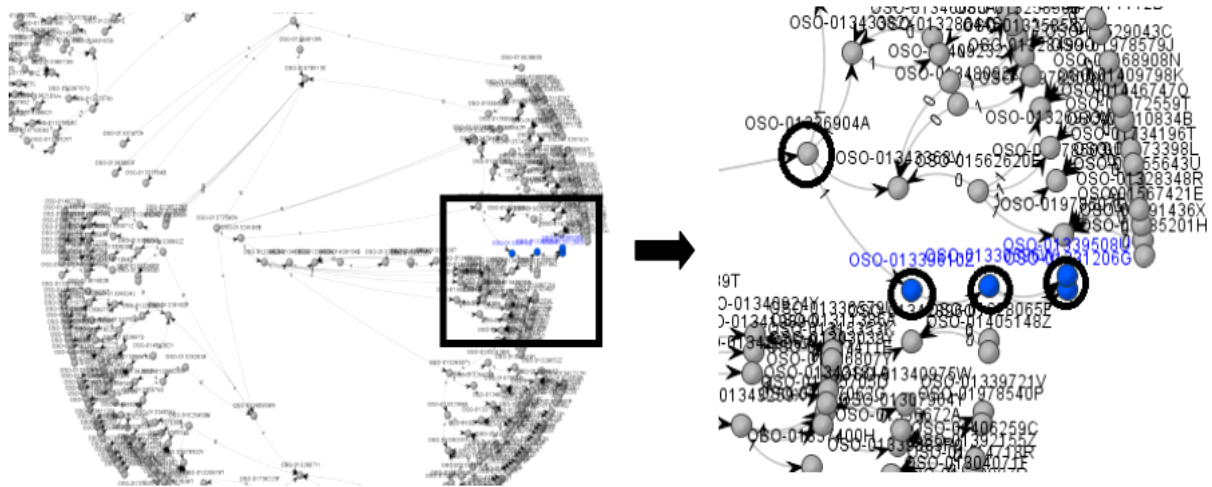


Fig. 4. Graphical representation of a network used in the first variant of tests

In the next presented step of testing the algorithms for nodes aggregation, a two-level sample of data was chosen, connected with the rest of the graph on both sides, characterized by various directions of connections. It consists of people with identifiers 131882, 132468, 156302 102566 (in the figure 5 OSO-01342822, OSO-01336979, OSO-01336825 OSO-01329263 respectively).

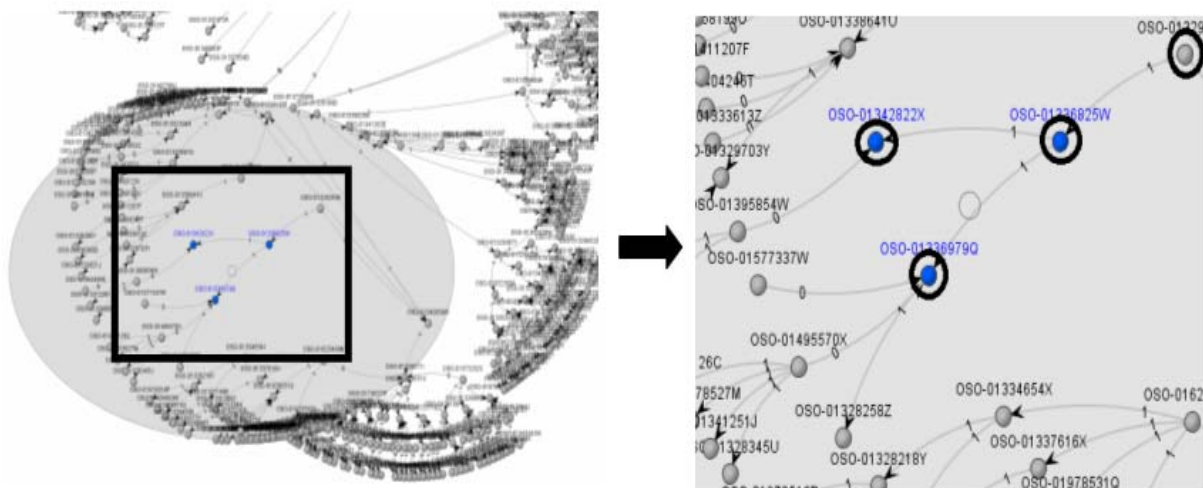


Figure 5. Graphical representation of a network used in the second variant of tests

Table 3. Nodes table

Node_ID	AggregationIndex
12231350	2
102566	1
131882	-2
132468	-2
156302	-2

Table 4. SuperiorNodes table

SuperiorNode_ID	SubordinateNode_ID
12231350	131882
12231350	132468
12231350	156302

The way in which the algorithm acted, in terms of operations on nodes, was identical with the cases described earlier. The aggregated nodes, 131882, 132468, 156302 are merged into one node 12231350 (table 4). Maintaining the index of aggregation is also conducted in a similar way, which is shown in the table containing data on nodes (table 3)

Table 5. SuperiorEdges table

SuperiorEdge_ID	subordinateEdge_ID
3052	2756
3053	2757
3054	2520
3055	2758
3056	2755

The aggregation of edges was conducted in a different way. The aggregated structure was connected with the rest of the graph by five edges, each ending in a different node. It caused the creation of five new edges, starting in these nodes and ending in the newly created one, preserving the primary binding direction of connection. The edges located in the middle of the aggregated structure had no influence on the number of new elements. Five new records were added to the table **SuperiorEdges** (table 5) and to the table representing data on graph edges (table 6). In the latter, aggregation index and weight of an edge were appropriately modified. In all the cases, their weight was inherited after removed edges.

Table 6. Edges table

EdgeID	Edge_A_ID	Edge_B_ID	Weight	AgregationIndex	Direction
2520	102566	131882	1	-2	1
2669	131882	156302	1	-2	1
2670	131882	132468	1	-2	1
2755	156302	401585	1	-2	0
2756	132468	98979	1	-2	1
2757	132468	1281444	1	-2	0
2758	132468	1863045	1	-2	0
3052	12231350	98979	1	2	1
3053	12231350	1281444	1	2	0
3054	102566	12231350	1	2	1
3055	12231350	1863045	1	2	0
3056	12231350	401585	1	2	0

### Operations on graphs representing a social network

A graph search or in other words passing through a graph is an activity consisting of visiting (in some systematic, regular way) all graph nodes in order to gather relevant information [Cormen, 2001]. There can be indicated two

most popular algorithms used for this purpose – BFS (breadth-first search) and DFS (depth-first search). Despite the fact that the analysis of database servers' mechanisms in terms of implementing both algorithms proved it to be theoretically possible, using the second one turned out to be impossible in practice. The cause of this was the necessity to use recursive queries, which were incapable of computing such a huge amount of data, encountered in social analysis. Not surprisingly, a decision was made to implement the BFS algorithm.

**BFS** algorithm is one of the simplest algorithms for searching graphs. The edges of the graph are systematically looked into, in order to visit every node accessible from the source node  $s$ . At the same time, distance (smallest number of edges) from source to each node is calculated. The effect of acting of the algorithm is also a breadth-first search tree from the  $s$  root to all accessible nodes. A path in the tree between  $s$  and node  $v$  is the shortest path between these two nodes in the graph. In order to make implementation of the algorithm, as a database procedure, possible, data model was extended with additional tables storing achieved tree and calculated paths.

In many cases of graph analysis searching the graph is performed in order to find the shortest path between two particular nodes. Among many available algorithms in this scope, Dijkstra's algorithms was chosen to be implemented [Cormen, 2001].

**Dijkstra's** algorithm determines the shortest paths starting in a particular node. It is performed for every node in the graph and requires weights of edges of a graph to be known beforehand. In the first step of the algorithm, distance between source node and each of the rest of the nodes is set to infinity. Next, during the process, the queue of neighbours of the currently analyzed node is created. For this purpose a special database table was created. Consecutive steps of the algorithm compare the distance between elements of the queue and the source node (which in the beginning equals infinity) with the weight of the edge connecting the current node and the analyzed one from the queue. The weight is obtained from the graph representing the social network. If the sum of this weight and distance from the current node to the source is smaller, it replaces the previous distance. The next current node is always selected acquisitively – meaning that it is always the unvisited node, to which the distance from the source is shortest. The algorithm effects in creating a set of shortest paths from the source node to each node in the graph. These paths are stored in the table of a graph data model. Thus, they can be easily accessed if necessary.

---

## Experiments

---

The chosen algorithms were tested with usage of the aforementioned graph. In the first step, BFS algorithm was used to check the consistency of the graph. Due to the fact that the graph consisted of huge number of elements, call of the database procedure was parameterized in terms of numbers of analyzed edges and neighbours.

Subgraphs, which consistency was proven, became the input data for the procedure implementing the second algorithm. The obtained result for sample procedure call is presented in the table 7.

*Table 7. Format of the shortest path from a sample node (11107)*

Node A ID	Node B ID	distance from the starting node
111107	113781	1
111107	262461	1
113781	101261	2
113781	113280	2
101261	129513	3
129513	148320	4
129513	158543	4
129513	679012	4
148320	102566	5
148320	108960	5
148320	321729	5

The weight of every edge was set to 1.

The elaborated mechanisms were also tested in terms of efficiency. For this purpose, BFS algorithm was implemented on the .NET platform using C# language. The same tests, which were run for database procedures, were repeated for the program constructed in this way. Both procedures were run in the same environment. Then, execution times of both procedures (with the same input data) were compared. Obtained results are presented in the figure 6. Their analysis clearly indicates that database server procedure was more efficient in checking consistency of a graph than the .NET one. It is worth to mention that .NET procedure connected to database server to achieve input data and meet the requirement concerning permanency of obtained results.

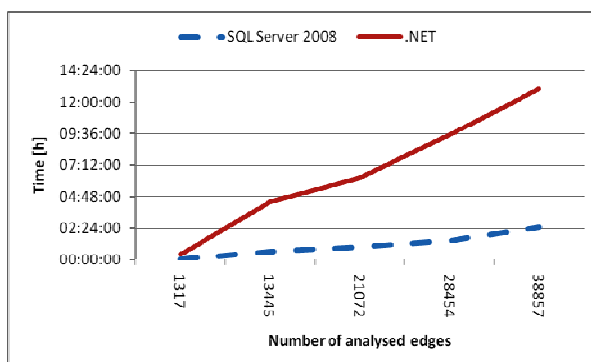


Fig. 6. Comparison of execution time of .NET and SQL Server 2008 procedures

## Conclusion

The aim of the research presented in the article was to assess the possibility of using database server resources to support social network analysis. There were two reasons of taking such approach. First of them was connected with the fact that information about objects constituting a social network are usually stored in database tables. The second reason resulted from the need for permanent storage of data on the previously computed characteristics concerning structure of such network. Adopting databases for such purpose seems to be the right concept. Within the research, a special data model was elaborated. Its architecture is universal enough to make usage of this model in analysis of social networks from various domains. The model consists of two groups of objects. In the first one, data on network elements is stored while the second consists of objects used to determine its characteristics. The second one is also exploited by database procedures implementing algorithms elaborated during the research, allowing for a change of social network structure (creating subgraphs, aggregation nodes and edges) and its searching (BFS and Dijkstra's algorithms). In order to check the correctness of the created mechanisms, tests for a sample social network were executed. They proved the preliminary assumption to be true. Furthermore, additional efficiency studies were performed, comparing the effectiveness of the mechanisms implemented on the database server with the corresponding ones created in the .NET platform. These tests also confirmed the rightness of the direction of the research.

The results obtained in the research are highly encouraging to further activity. In the next steps, elaborating solutions, allowing for usage of database servers in order to determine the roles of objects creating a social network, is planned. Moreover, what is also planned to be used for this purpose is multi-agent systems.

## Bibliography

- [Buja, 2008] A.Buja, D.F.Swayne, M.Littman, N.Dean, H.Hofmann, L.Chen. Interactive Data Visualization with Multidimensional Scaling, *Journal of Computational and Graphical Statistics*, 2008, Vol 17 (2): pp. 444–472.
- [Chau, 2006] M.Chau, J.Xu. A Framework for Locating and Analyzing Hate Groups in Blogs. *PACIS 2006 Proceedings*.

- 
- [Cormen, 2001] T.H.Cormen, C.E.Leiserson, R.L.Rivest, C.Stein. Introduction to Algorithms. The MIT Press, 2nd revised edition, September 2001.
- [Giran, 2002] M.Girvan, M.E.J.Newman. Community structure in social and biological networks, Proceedings of the National Academy of Sciences, Vol. 99, No. 12, pp. 7821-7826, 2002
- [Guangming, 2009] T.Guangming, T.Dengbiao, S.Ninghui. AParallel Algorithm for Computing Betweenness Centrality Parallel Processing, 2009. ICPP '09. International Conference on, 2009,pp. 340 - 347.
- [Hareźlak, 2010] K.Hareźlak, M.Kozielski. The methods of criminal network analysis (Metody analizy sieci kryminalnych). STUDIA INFORMATICA. Volume 31, Number 2A (89), 2010
- [Kulmar, 94] A.Kumar, R.H.Fowler. A spring modeling algorithm to position nodes of an undirected graph in three dimensions, Technical Report CS-94-7
- [Newman, 2004] M.E.J. Newman. Detecting community structure in networks, Eur. Phys. J. B 38, 321–330 (2004).
- [Piekaj, 2007] W.Piekaj , G.Skorek, A.Zygmunt, J.Koźlak Environment for identifying behavioral patterns using social networks (Środowisko do identyfikowania wzorców zachowań w oparciu o podejście sieci społecznych). Technologie Przetwarzania Danych, TPD 2007.
- [Razmerita, 2009] L.Razmerita, R.Firantas New Generation of Social Networks Based on Semantic Web Technologies: the Importance of Social Data Portability, <http://academic.research.microsoft.com/Publication/5968829/new-generation-of-social-networks-based-on-semantic-web-technologies-the-importance-of-social-data>, 2009.
- [Scripps, 2007] J.Scripps, P.N.Tan, and A.H.Esfahanian. Node roles and community structure in networks. In WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, ACM, 2007, pp 26:35..
- [Shaikh, 2006] M.A.Shaikh, J.Wang. Investigative Data Mining: Identifying Key Nodes In Terrorist Network. Multitopic Conference, 2006. INMIC '06. IEEE, pp 201 - 206.
- [Williams, 2001] P.Williams. Transnational Criminal Networks. [http://www.rand.org/pubs/monograph\\_reports/MR1382/MR1382.ch3.pdf](http://www.rand.org/pubs/monograph_reports/MR1382/MR1382.ch3.pdf), 2001.
- [Xu, 2005] J.Xu, H.Chen. CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery, ACM Transactions on Information Systems (TOIS), 23(2), pp. 201-226, 2005.
- [Yang, 2006] C.C.Yang, L.Nan, M.Sageman. Analyzing the Terrorist Social Networks with Visualization Tools, Intelligence and Security Informatics, Lecture Notes in Computer Science, Vol. 3975, 2006 pp.331-342.

---

## Acknowledgements

Project financed from the funds for learning in years 2010 - 2012 by a research and development grant number O R00 0113 12. I am grateful to the copartner WASKO S. A company for letting me exploit the screenshots of the sample application used in my figures

---

## Authors' Information



**Katarzyna Hareźlak** – Silesian University of Technology; e-mail: [katarzyna.harezlak@polsl.pl](mailto:katarzyna.harezlak@polsl.pl)

*Major Fields of Scientific Research: Distributed and Mobile Databases, Software Engineering, Agent Systems, Social Networks*

## ANALYSING AND VISUALIZING POLISH SCIENTIFIC COMMUNITY<sup>1</sup>

Piotr Gawrysiak, Dominik Ryżko

**Abstract:** *This paper describes the rationale and partial results of an ongoing experiment aiming to perform analysis and visualization of social and organizational structure of Polish scientific community. Work described in this paper concerns automated extraction of information (related to DSc theses) from Internet based sources and visualization of resulting data set. Primary database that was mined was a scientific information repository “Polish Science (“Nauka Polska” in Polish) maintained by the Information Processing Centre (OPI – “Ośrodek Przetwarzania Informacji” in Polish). The nature of this repository, specifically lack of any data export facility and web scrapping prevention provisions implemented (despite the fact, that database is supported by public funds), complicated data extraction process. However, when finally downloaded, verified and processed the data set proved to be very interesting and valuable, making possible statistical analysis and geospatial visualization. Specifically, this paper describes creation of a software tool able to create geographical maps depicting collaboration between various research institutions in Poland during the process of DSc theses review. It should be regarded as a report on a work in progress. It is a part of larger SYNAT project, being a government funded initiative to create an ICT infrastructure supporting scientific collaboration and exchange of research data in Poland. The results presented herein constitute just a proof of concept for a visualization approach that will be implemented when more detailed data, concerning Polish scientific community, will be collected during other SYNAT activities.*

**Keywords:** *bibliometrics, information visualization, social network analysis, web mining*

**ACM Classification Keywords:** *J.4 SOCIAL AND BEHAVIORAL SCIENCES*

**Conference topic:** *Data Mining, Knowledge Acquisition, Intelligent Web Mining and Applications*

---

### Introduction

---

Contemporary science is focused on collaboration. The nature of most important research problems - in engineering, but also in physics, biology, bioinformatics, medicine and many other fields - makes solitary research, carried out by an individual researcher or even by a single research center, impractical. One of the reasons is the size of the data and the complexity of the problems that are being tackled by modern science. Another one is the Internet that finally became a useful collaboration tool, linking together research centers across nations and even across continents.

Obviously the science itself was always about sharing results with fellow researchers. The more popular was a given scientific treatise, the more widely cited and commented, the more important it was for subsequent research. However, with the advent of professional bibliometrics in XX<sup>th</sup> century, the citation count became a driving factor in modern science [Walter, 2003]. One might even risk a statement that nowadays the merits of scientific research are much less important than a number of people knowing about this particular research and

---

<sup>1</sup> This work is supported by the National Centre for Research and Development (NCBiR) under Grant No. SP//I/177065/10 by the strategic scientific research and experimental development program: “Interdisciplinary System for Interactive Scientific and Scientific-Technical Information”

referring to it. This however, being nothing else than a *popularity* of a given idea or a researcher, is related not only to the quality of the research, but also to the quality of marketing (or public relations) strategy employed by the research center or by given scientists. This became especially evident in recent years, with exploding popularity of the Internet, which is used as a primary tool for disseminating scientific information. The teams that are able to use the Web more effectively are also able to promote their research better, which increases the chances of getting high citation count. This could be potentially disadvantageous for disciplines (e.g. in humanities) or teams (e.g. based in less developed countries, where access to computing and communication infrastructure might be problematic) which have not yet fully embraced the Internet as a research tool.

On the other hand, the popularity of the global network as a tool used by global scientific community presents an opportunity for better understanding the collaboration patterns that emerge within this community. Due to the largely open nature of the Internet, but also thanks to the adoption of “open access” [Antelman, 2004] as a one of the most important models of publishing scientific papers, gathering statistical information concerning collaboration patterns between researchers and teams became relatively easy. Of course lack of standardization of protocols and data formats (even for such seemingly simple information as bibliographic references) complicates this process, but it can be – for the first time in history of science – automated.

This is therefore obvious that the growing importance of the Internet presents itself both as an opportunity for a local scientific community and as a threat that needs to be addressed. In recent years most developed countries have launched various initiatives, targeting these two issues. Poland is no exception here. Last year a strategic government project was funded, named SYNAT. It is aiming to improve national and regional (i.e. in the area of Central and Eastern Europe) web based science infrastructure and to map and analyze the state of Polish scientific community. This paper presents information on some initial analysis experiments related to this task.

The paper is structured as follows. Second and third chapters contain an overview of the entire SYNAT project and provide background information concerning the nature of IT systems that the project is implementing. Next chapters describe a web mining experiment, in which data concerning DSc theses has been extracted from the database maintained by the government agency “Information Processing Centre” (“Ośrodek Przetwarzania Informacji” in Polish) and later visualized using geospatial approach. Finally, the last chapter contains concluding remarks and a vision for future work.

---

### **Building science infrastructure - the SYNAT project**

---

The research that is the subject of this paper has been carried out within a project, funded by the Polish Ministry of Higher Education and is being implemented by the consortium of Polish universities (see the project website [Synat, 2011] for a complete list). The project is part of a larger effort aiming to improve the state of scientific research in Poland and also in Central and Eastern Europe. One of the main obstacles hindering the growth of scientific research, and perhaps even more importantly, making the distribution of research results more difficult than necessary is lack of information exchange systems pertaining to these results.

In short, the lack of standards and systems supporting storage of scientific oriented information caused large distribution of repositories, storing data mostly in unstructured formats. Even information which is relatively easy to structure, such as bibliographical data, is not stored in an easy to process way. These problems are common both for large universities and research institutes. As a result, the visibility of Polish science in the Internet is very poor and particular centers of research are often not aware of the work conducted by others so intensive scientific collaboration between Polish institutions is relatively rarely implemented in practice. Much more important effect is however a distorted view of Polish science, because the apparent activity of local scientists is much lower than their real efforts and results of their work.



The system to be developed in the SYNAT project is planned to be a heterogeneous repository of data coming from various structured and unstructured sources. Hosting capabilities will be provided to address issues described above. This will be helpful especially with respect to low funded centers of research, which do not have capabilities to set up extensive repositories; additionally it is also required in order to acquire data from external sources. This includes large repositories of scientific papers and information about researchers and projects. In short, the project aims to create a backbone for scientific information storage in Poland. The envisaged system should also be able to retrieve potentially useful unstructured information from the Internet. Blogs, forums, project homepages, science funding schemes etc. constitute a large fraction of available knowledge scattered across the Web. Using such sources means that acquired data can contain missing information, errors or overlaps. The system should be able to: identify duplicates, merge partly overlapping objects, identify object versions, verify completeness of data objects (e.g. bibliography items), identify key words and proper names etc. It is required that the system will be able to perform search for new resources, especially in the areas heavily searched by the end users. The user should also be able to query the system repository but also start an off-line search process in order to discover resources according to specific requirements. Any new findings relevant to a particular user profile should be reported. Once discovered, sources of data have to be monitored in order to track any changes to their contents

Information storage capabilities drafted above constitute, however, only a part of the system's capabilities. Apart from just being able to store results of research in the form of publications and experimental data, the system should also support the research process itself, by providing tools for rapid dissemination of partial research efforts, for discovery of institutions or groups with similar research interests or even supporting some computationally intensive applications on the system infrastructure itself. The system is not being designed as a computational grid, however the distributed nature of storage subsystem means, that it can be also, for some specific use cases, utilized to perform some calculations (such as creating visualizations or statistical and/or knowledge discovery computations).

In a sense, the system's functionality in this context (i.e. not directly related to storage and distribution of bibliographical data) will be similar to the functionality of a social network system. Obviously calling such a system the scientific social network (or even "Facebook for scientists" as it has been unofficially dubbed) might be an overstatement. However, the similarities between Facebook and the system are also visible in a way in which its services are exposed to external parties. The system is structured as a set of modules with well-defined functionality and data types that can be interconnected by external institutions by using public system APIs. In this way it is possible to extend the functionality of the system or rather build other, specialized research support tools basing on the system infrastructure. Such tools could be both of commercial and open nature and possibly in the long run a larger ecosystem of services, supporting the scientific community, could be created.

---

### **System architecture overview**

---

The requirements described in the previous chapter indicate an explicit distributed nature of the problems to be addressed. On the data acquisition side, the Internet is a network of loosely connected sources, which can be processed more or less in parallel. On the end user side, each one of them can generate concurrent requests for information. At the same time these parallelisms do not forbid overlap or contradiction. All of the above calls for a highly distributed architecture, with autonomy of its components, yet efficient communication and synchronization of actions between them.

The envisaged approach is based on multi-agent paradigms, which introduce a concept of an intelligent, autonomous and proactive agent. Various agent roles have been identified while analyzing processes to be implemented in the system. Personal agents will be responsible for interaction with end users. They will receive

---

---

queries, preprocess them, pass to the knowledge layer and present results returned from the system. User feedback will also be collected here. Personal agents will store history of user queries and maintain a profile of interests to improve results and proactively inform the user about new relevant resources.

The main data acquisition process will be performed by specialized harvesting agents. Their task will be twofold. Firstly, they will perform a continuous search for new relevant resources. Secondly, they will perform special searches for specific queries or groups of queries. The main task of harvesting agents will be to manage a group of web crawlers to perform the physical acquisition of data.

Different users can generate queries, which return partially overlapping results. This means some search tasks should be merged. On the other hand the same results can be delivered from different sources and only one of them should be used. All this means that matchmaking and coordination between demand and supply of data generates some sophisticated problems. This can be mitigated by introduction of brokering agents (also called middle agents) [Klusch2001], whose purpose is to coordinate efficient matching between personal agents and harvesting agents.

Special agents should be dedicated to the process of managing data already incorporated into the system. They will be responsible for finding missing data, inconsistencies, duplicates etc. Finding such situations will result in appropriate action e.g. starting a new discovery process to find new information, deletion of some data, marking for review by administrator etc.

The bottom layer of the system will consist of a group of web crawlers. They will search the Internet for relevant resources and pass the data to appropriate agents responsible for its further processing. The crawlers will use various methods (heuristics, machine learning) to perform focused crawling for new documents based on classified examples.

A focused crawler is a program that traverses the Internet by choosing relevant pages to a predefined topic and neglecting those out of concern. The main purpose of such a program is to harvest more information on the topic that matches the expectation of the user while reducing the number of web pages indexed. A focused crawler has three main components: URL queue (container of unvisited pages), downloader (downloads resources from WWW), classifier (compound model which categorizes the type of information resource, and its domain).

The crawler's classifier includes two modules: extraction module and relevance analysis. The extraction module parses the web page, and identifies the main parts of it. Humans can easily distinguish the main content from navigational text, advertisements, related articles and other text portions. A number of approaches have been introduced to automate this distinction using a combination of heuristic segmentation and features.

In PASSIM we would deploy the solution proposed in [Kohlshutter, 2010]. In this work a combination of two features was used - number of words and link density. This approach leads to simply classification model that achieves high accuracy. Web pages are segmented into atomic text blocks using html tags. Found blocks are then annotated with features and on this basis classified into content or boilerplate. The features are called shallow text ones. They are higher, domain and language independent level (i.e. average word length, average sentence length, absolute number of words). Atomic text blocks are sequence of characters which are separated by one or more html tags, except for "A" tags. The presence of headline tag, paragraph, division text tag are used to split content of web page into set of structural elements. To train and test classifier for various feature combinations we would use well known scientific conference, journal page, home pages of scientists, research institutes. The labeled set is then split into training and a test set (using i.e. 10-fold cross validation) and fed into a classifier mode l(Support Vector Machine).

Relevance analysis uses the significant parts of web resource, which were detected by above described extraction module. It would use intelligent classifier to categorize resource as scientific or not. It is also possible to

do first domain classification, and label document to the field of science (i.e. biology, history, computers). The analysis of topic similarity is the most important stage for a topic-specific web crawling. The relevancy can be determined by various techniques like the cosine similarity between vectors, probabilistic classifiers, or BP neural network. During relevance analysis it may be useful to identify type of resource which was positive categorized. Machine learning classifier is trained with features, which includes i.e. hosting domain, non HTML markup words, URL of page, outgoing links. Such model could be enough relevant to classify resource as home-page, institute page, conference, or blog. Type of resource may be used as a filter during invoking searching process.

We have two types of focused crawlers. First is used in harvesting mode to detect all probably scientific resource, which are further processed by middle layer (specific agents). The second type of crawler is used to find resources relevant to natural language query (NL query). Natural query would be provided by user in similar way as we type queries in Google search engine interface. Next, this query would be analyzed in the context of ontology which describe meta-data of conferences, journals, articles, institutes, researchers. Ontology gives us ability to make user's query much more semantic and structured. This approach achieves advantage over Google like search engines. General purpose searchers could not be ontology-driven, because it is impossible to build ontology of whole world.

Web crawlers have URL queue which contains a list of unvisited web resources. In harvesting mode this queue would be initialized with seed URLs. Seed URLs may be built on data taken from e.g. DBLP [DBLP, 2011] and Citeseer. Those two sources have links to relevant sources, but also meta-data concerning author names, title, publication date. Found URL's within DBLP and Citeseer would build initial URL queue. The mentioned meta-data would be entered to Google Scholar service and returned URL's could enrich seed entries.

The classifier analyzes whether the content of parsed pages is related to topic or not. If the page is relevant, the URLs extracted from it will be added to queue, otherwise will be discarded.

The process described above results in discovery (hopefully) of several classes of resources for various fields of science. Therefore, the data has to be properly classified according to the type of information it represents (e.g. scientific paper, blog, conference homepage etc.). Once we have such classification we can use ontology to decompose the document into appropriate components. For example, if we deal with a scientific paper, we will expect to find title, authors, affiliation, abstract etc. In the case of a conference website, the ontology will tell us what are the roles related to a scientific conference (general chair, organizing chair, program committee member etc.), what is a special session, a paper and so on. Another dimension for classification of resources is the field of science which they belong to. Finally documents – classified and decomposed - will be stored in the system repository. From there they can be accessed by the system users. They will also undergo further processing in order to improve knowledge quality. Duplicates will be eliminated, missing information filled, inconsistencies resolved etc.

More detailed description of the SYNAT project or the infrastructure that will be implemented is out of the scope of this paper. For more information see e.g. [Bembenik, 2011] or [Synat, 2011].

---

### **Data extraction from Web resources**

---

Automated or semi-automated extraction of data from web resources, such as roughly outlined in previous chapter, is planned as one of the most important functionalities of SYNAT infrastructure. Because the nature and quality of the data that is currently available is highly inconsistent, various approaches need to be adopted in order to efficiently gather scientific information. From this perspective existing web resources can be roughly classified as follows:

- 
- 
- *structured data sources exposing consistent API* – this category includes scientific information databases (both bibliographic and containing experimental data) that have predetermined structure and that can be accessed not only via web interface (designed for human users) but also via an API, allowing querying the database and downloading contents in a structured format (e.g. in XML or JSON). Examples of such databases include DBLP Computer Science Bibliography [DBLP, 2011] or European Nucleotide Archive [ENA, 2011]. Such databases are obviously the most easy target for web mining and could be even directly linked to other parts of SYNAT infrastructure (a technique which – in the case of bibliographic databases – is most commonly referred to as “*harvesting*” [Sompel, 2004]);
  - *structured data sources without external API* – this includes resources, that have an internal controlled structure (e.g. store the data in relational database), but which expose the contents only through web interface. Examples include reference databases such as Polish Science database [OPI, 2011], some open access journals, or information repositories maintained by universities. Such resources are in most cases quite easy to mine, however an initial processing step is usually required where a parser is constructed which is able to extract important structured information from contents of web pages generated by the resource;
  - *semi-structured data sources* – this category includes these resources, that are created manually, but where an initial structure has been decided upon and is maintained, usually automatically by some kind of a content management system. Web newspapers and information portals such as CNN are probably most common examples of this resources; Wiki services (especially those using popular server backends such as MediaWiki) are another one – probably much more relevant in the context of scientific information retrieval. Information extraction difficulty is comparable to that described above, provided that some automated tools are available. For the SYNAT project several experimental systems, able to utilize machine learning techniques in order to automatically remove non-essential parts of webpages (such as navigational and static elements) as well as perform appropriate HTML stripping. Mentioned in previous chapter – see e.g. [Kolaczowski, 2011] and [Kohlschutter, 2010];
  - *handcrafted information repositories* – all other manually created data sources, such as personal and institutional web pages, project websites and even – in some cases – full scientific papers posted online. These resources usually require individual approach and in most cases are not susceptible to fully automated harvesting, however most valuable repositories will be imported to internal SYNAT database in later stage of the project.

Second category – the structured data sources without API – was selected as most important subject for initial experiments with web mining in SYNAT project. Most of the scientific databases available in Polish internet fall under this category, contrary to process of attaching external sources exposing well formed API to other IT systems is relatively trivial task, the project team expected some minor difficulties in this case. The aforementioned database Polish Science (“Nauka Polska” in Polish) was selected as one of the first systems that would be mined.

This database constitutes currently probably the most complete information repository regarding Polish scientists. It contains bibliographical records of over 121 000 people, who has been awarded a PhD in Poland, together with information about their professional affiliations, field of study and their professional career (including information about DSc theses and full professor titles awarded). The database is maintained by a government agency Data Processing Centre (“Ośrodek Przetwarzania Informacji” in Polish) [OPI, 2011] and quality of information contained there is generally considered to be high, mostly because of the fact that providing the agency with

relevant information by universities was (until recently) compulsory. The database provides also some information about scientific publications and events (i.e. conferences and symposia) but these parts are very incomplete.

The database is not equipped with any data export or external query provisions (albeit it is possible to formulate some simple queries via a web interface), but the generated HTML is quite consistent so no difficulties in data extraction were expected. However, it quickly turned out, that the database – despite the fact that it has been created and is maintained with public funds – is heavily protected from automated download attempts, blocking access as soon as after just 10 subsequent requests originating from a single IP number. Further investigation showed that one of the reasons is that the agency provides data processing services (such as performing specialized queries over database contents) for a fee, so naturally is not interested in facilitating access for other entities. Because the SYNAT project is also an official government initiative, negotiations concerning data access have been started, but a decision was also made to create a set of tools able to bypass the “data scrapping protection” mechanisms, as such tools might be useful also in future, when dealing with other systems.

The download framework has been implemented in Python and is able to use public anonymous proxy servers in order to “fake” access from the system that is being mined. The framework is able to automatically extract lists of proxies from various web sources (such as public proxy forums) and test their reliability. The accesses to a target database are also appropriately throttled down, with randomization applied both to the delays between consecutive HTTP requests, and to order of records that is being requested.

For test purposes a subset of the entire database was selected, concerning DSc theses defended during 2010. This dataset includes biographical information of authors and reviewers of theses (usually 4 independent reviews are required in the DSc process in Poland) thus allowing to analyze the relationship between various universities. Reviewers are usually invited to the events related to thesis defense by the author’s university, so such occasions are usually an occasion for discussion with fellow researchers and in many cases also make possible to start collaboration between various research groups. The extraction process took approximately one week, however during this time the database was offline for two days – later it turned out that such intermittent outages are quite common and might happen even as often as several times a month.

Basic statistical information concerning extracted data is as follows: 783 records describing DSc defended during 2010 has been downloaded. These theses have been reviewed by 3132 university professors working in 1836 institutions, of which only two are outside of Poland. The institutions employing reviewers are located in 76 cities. The quality of records was highly uneven – some records contained full biographical info including list of publications of a given scientists, some only cited the works he or she reviewed or defended, while in some cases only name and affiliation could be extracted. On the contrary, the quality of institutional records was very high, with regard especially to their address information, what turned out to be an important factor for further visualization experiment.

---

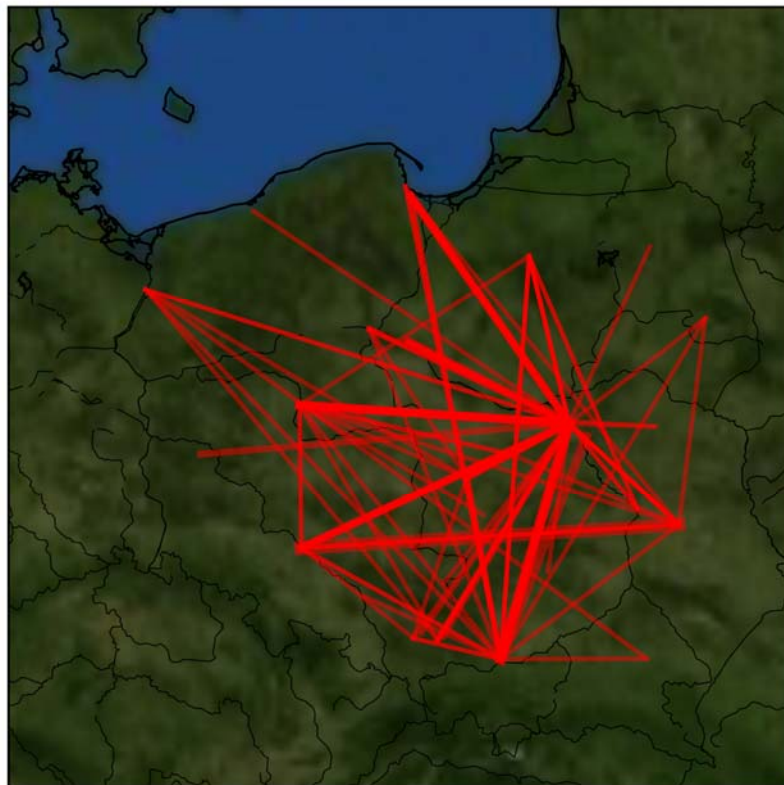
## **Data visualization**

---

As mentioned above the data set that was extracted contains information about social interactions between Polish scientists. Of course this is not a level of daily communication that is exhibited (and frequently visualized – see e.g. [Butler, 2010]) in social networks such as Facebook. However, due to the reasons mentioned above, it should allow at least to identify the associations between research institutions located in various parts of the country, that under normal circumstances, in day to day work, rarely collaborate together. For this purpose the data set was converted into a graph representation as follows: the bibliographic records of reviewers and authors have been parsed in order to extract information about their home institutions; the institutions records were analyzed in order to extract their addresses and the city portions of the address have been identified. Resulting

graph contains nodes corresponding to the cities, with vertices representing a “collaboration” events i.e. a meeting of two reviewers from two different cities during DSc defense.

Such graph can be obviously simply plotted e.g. via GraphViz and SFDP [Gansner, 2009], however because of the special nature of the data the most interesting way of visualizing it would involve placing it on a map. Of course in order to do it, a geographical coordinates of all the cities present need to be established, or “geocoded” and a specialized library was created that is able to do it for an arbitrary city name, by querying an external geospatial database. Google Maps [Maps, 2011] is probably one of the most popular such databases used currently, however it is not suitable for use in projects such as this due to processing limitations. Instead Geonames database [Geonames, 2011] (which is licensed under Creative Commons license) was used, and proved to be of high enough quality. For the actual drawing the Python Basemap toolkit was used [Basemap, 2011], allowing to incorporate a satellite ground map from the NASA Blue Marble project and draw the graph vertices not as straight lines but as parts of great circles (which will be of course more useful in future, when visualizations of collaboration with institutions in other countries will be performed; for the map of Poland parts of great circle are practically the same as lines). Various experiments with weight and coloring schemes of vertices were performed and finally a relatively simple method was adopted that associates the thickness and saturation of a line with logarithm of number of edges connecting two vertices. Resulting visualizations are presented below.



*Fig. 1 Collaboration visualization depicting only most connected institutions (at least 3 edges in collaboration graph).*



*Fig. 2 Alternative approach to representing intensity of collaboration, using only saturation without altering line thickness.*

---

## **Conclusions and future work**

Relatively small temporal scope of the data set (data from only one year i.e. 2010) that was used in this experiment does not allow drawing well founded conclusions about the state of scientific collaboration in Poland and indeed the experiment was thought mainly as a proof of concept and will be expanded in future. However, even within small data set that was a subject of this one can make some interesting observations. For example – it is quite evident that the Warsaw is the most important city as far as scientific research in Poland is concerned, the area most active in research corresponds roughly to central and greater Poland (i.e. the areas that have not been part of Poland before II World War are seem to be not entirely integrated as far as collaboration is concerned).

As it is mentioned above these observations should be taken with a grain of salt. However during next phases of the SYNAT project the entire database maintained by OPI will be downloaded – so this experiment will be repeated, but this time with information spanning several tens of years. Additionally other visualizations will be prepared, using software created during this experiment, but with different data sets (first candidate is obviously an experiment similar to DSc analysis described herein, but concerning PhD theses).

---

## **Acknowledgements**

This work is supported by the National Centre for Research and Development (NCBiR) under Grant No. SP//I/77065/10 by the strategic scientific research and experimental development program: “Interdisciplinary System for Interactive Scientific and Scientific-Technical Information”

---

## Bibliography

---

- [Antelman, 2004] K. Antelman, Do Open-Access Articles Have a Greater Research Impact?, *College & Research Libraries* vol. 65 no. 5 372-382, 2004
- [Basemap, 2011] Basemap library documentation, <http://matplotlib.sourceforge.net/basemap/doc/html/>, accessed on 30/05/2011
- [Bembenik, 2011] R. Bembenik et al. Retrieval and management of scientific information from heterogeneous sources. In: SYNAT Workshop 2011 Proceedings, *Studies in Computational Intelligence*, Springer Verlag, 2011
- [Butler, 2010] P. Butler, Visualizing Friendships, Facebook, USA, 2010, [http://www.facebook.com/note.php?note\\_id=46971639891](http://www.facebook.com/note.php?note_id=46971639891)
- [DBLP, 2011] DBLP Bibliography, <http://www.informatik.uni-trier.de/~ley/db/>, accessed on 30/05/2011
- [ENA, 2011] European Nucleotide Archive, <http://www.ebi.ac.uk/ena/>, accessed on 30/05/2011
- [Gansner, 2009] E. Gansner et al., Efficient Node Overlap Removal Using a Proximity Stress Model, *Lecture Notes in Computer Science*, 2009, Volume 5417/2009
- [Geonames, 2011] Geonames Database, <http://www.geonames.org>, accessed on 30/05/2011
- [Klusch, 2011] Klusch M., Sycara K. P. Brokering and matchmaking for coordination of agent societies: A survey. In *Coordination of Internet Agents: Models, Technologies, and Applications*, pp.197-224. Springer, 2001.
- [Kohlschutter, 2010] "Boilerplate Detection using Shallow Text Features" by Christian Kohlschütter et al., presented at WSDM 2010 -- The Third ACM International Conference on Web Search and Data Mining New York City, NY USA, 2010
- [Kolaczowski, 2011] P. Kolaczowski, P. Gawrysiak, "Extracting Product Descriptions from Polish E-Commerce Websites Using Classification and Clustering". In: 19th International Symposium on Methodologies for Intelligent Systems, *Lecture Notes in Computer Science*, Springer Verlag, 2011
- [Maps, 2011] Google Maps, <http://maps.google.com>, accessed on 30/05/2011
- [OPI, 2011] Nauka Polska, Ośrodek Przetwarzania Informacji, <http://nauka-polska.pl/>, accessed on 30/05/2011
- [Sompel, 2004] H. Sompel et al., Resource Harvesting within the OAI-PMH Framework, *D-Lib Magazine*, Volume 10 Number 12, 2004
- [Synat, 2011] SYNAT Project website. <http://www.synat.pl>, accessed on 30/05/2011
- [Walter, 2003] G. Walter et al., Counting on citations: a flawed way to measure quality, *Medical Journal of Australia*, Vol. 178, Nr. 6 (2003) , p. 280-281, 2003

---

## Authors' Information

---



**Piotr Gawrysiak** – Deputy Director for Scientific Research, Institute of Computer Science, Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland; e-mail: [P.Gawrysiak@ii.pw.edu.pl](mailto:P.Gawrysiak@ii.pw.edu.pl)

*Major Fields of Scientific Research: natural language processing, text, web and data mining, mobile computing, human computer interaction, information visualization.*



## ARSIMA MODEL

Vitalii Shchelkalin

**Abstract:** In presented work the further development of Box-Jenkins technique for models constructing and improvement of themselves ARIMA models is produced. A novel autoregressive – spectral integrated moving average (ARSIMA) model founded on joint use of the Box-Jenkins method (ARIMA models) and "Caterpillar"-SSA method with model trained on competitive base is developed.

**Keywords:** modeling, filtering, forecasting, control, "Caterpillar»-SSA method, ARIMA model, "Caterpillar»-SSA – ARIMA – SIGARCH method, ARSIMA model, ARSIMA – SIGARCH model, heteroskedasticity, Levenberg-Marquardt method.

---

### Introduction

---

The research progress constantly develops economic, technical, social, medical and other systems, complicating their structure and enlarging amount complex their internal intercoupling and external factor, from which they hang and majority from which take into account impossible. Therefore, it is actual to develop and use the universal mathematical models and techniques allowing to modeling, forecast and control the wide class of the processes since this will allow to raise efficiency of control and planning state of working complex systems, spare the significant facilities and resources not by development of new energy-saving resources, but by way of any mathematical subterfuges, as well as offloads the work of the personnel of various organizations and allows to anticipate emergencies and brings many other benefits.

The main requirements to the mathematical models construction in 70 - 90 of the last century, is an economical number of parameters, the velocity of the model determination and its resource-intensive for use on available then computers with low productivity. However, modern computer technology and mathematical modeling methods provide a great possibilities for analysis, modeling and forecasting time series of the different nature. Therefore, at present these requirements are not crucial and modern computing tools and systems allow to stand on the first plan the requirement of modeling accuracy, quality of the analysis and forecasting.

One of the most widely used models that corresponding to above requirements are the models ARIMA. ARMA method works only with pre-reduced to the stationary form time series. Nonstationary series are usually characterized by the presence of high power at low frequencies. However, in many practical applications of interest information may be concentrated at high frequencies. In such cases, all that was done - it is filtered out non-stationary low-frequency components and was used the remainder of the series for further analysis. At the same time as a filter to eliminate low-frequency component in the ARIMA model used a filter of the first differences or maximum second. Watching the gain of the filter can be seen that low frequency considerably weakened and, therefore, be less visible at the filter output. So the method of seasonal ARIMA model was satisfactorily predicted only with a relatively simple structure time series.

In the 80's years of last century Granger and Djoyo [Granger 1980] proposed a new class of ARFIMA models is convenient to describe the financial and economic time series with the effects of long and short memory.

2000's years are characterized by the using for a wide range of models for time series analyzing and forecasting, as well as ensembles of models with different structures. With the advent of high-speed computer was occurred the transition from ensembles of predictive models to its combination. The difference between the combined models and its ensembles lies in the simultaneous adjustment of model parameters.

---

The most important characteristics of models at analysis and choice of the most appropriate mathematical models of the following main important features are:

- method of modeling the trend component of the time series;
- method of nonlinear modeling of time series;
- method of modeling the random component of the time series;
- way of accounting for the influence of external factors on the process.

Therefore, the priority type of models are combined probability and deterministic prediction models with nonlinear complexity, because in this models simultaneously used as statistical and deterministic components that allows to reach the best quality of the forecasting [Седов, 2010]. Among the deterministic models the priority variant is a deterministic model of the spectral decomposition, which implements simulation-based expansion in the deterministic orthonormal basis different from that of harmonic functions. While the most priority among the probabilistic models is the model of auto regression - integrated moving average.

In the scientific literature have long been known combined probability and deterministic model presented in [Седов, 2010]. In given paper is offered next modification of the ARIMA and the GARCH models and at first proposed "Caterpillar"-SSA – ARIMA – SIGARCH method and combined probabilistic and deterministic model of auto regression – spectrally integrated moving average with spectrally integrated generalized autoregressive heteroskedasticity (ARSPSS – SIGARCH) and the method of its construction, which allows greater flexibility in analyzing, modeling and forecasting time series in comparison with the ARIMA – GARCH models.

---

### Summary of the main material of the study

---

The essence of the method was at first in a multi-dimensional decomposition of exogenous time series, and the propagated time series for basic latent components, including their extents and combinations, obtained by the principal (PCA), smooth (SCA) and independent components (ICA), in the selection of the basic components of design method of fast orthogonal search (FOS), one of the most effective and economical methods for time spent, and cutting off the destructive, thus forming the transfer function of a mathematical model of the process, to further identify the noise of a mathematical model of the process due to the seasonal autoregressive models — moving average, and the simultaneous parameter identification of model structure obtained by the Levenberg-Marquardt algorithm [Щелкалин, Тевяшев, 2010]. As it became known later, the proposed method is similar to the decomposition method of modeling (DMM) [Седов, 2010]. The method of "Caterpillar"-SSA also uses the decomposition of time series of singular values (SVD). Known publications using the "Caterpillar"-SSA method in various branches of science and technology as a method of fairly good description of non-stationary time series with linear, parabolic or exponential trend with not always stable oscillatory component, however studies have identified a number of significant shortcomings of the method, greatly limiting its applicability. Method for modeling uses suboptimal in terms of accuracy of some time series of orthogonal basis vectors of the trajectory matrix. Therefore, the main idea of the origin of the proposed method was first concluded in joint use the "Caterpillar"-SSA method and models of autoregressive - moving average, trained on a competitive base, with account generalized criterion of the accuracy and adequacy. Using such combination was dictated by the fact that individually, these approaches have several disadvantages, but their joint use brings synergy, increasing their efficiency, robustness and adequacy. However, the trend separation by "Caterpillar"-SSA method, as well as any other method, the residual component of the series in most cases is non-stationary, and therefore hereinafter "Caterpillar"-SSA method has been used in combination with the model of autoregressive - integrated moving average (ARIMA). In this case, joint use of the above methods imply that the parametric identification computes the parameter estimates ARIMA and nonlinear generalizations of principal components of the autoregression,

minimizing the sum of the squares error modeling, taking into account the time series, obtained by the "Caterpillar"-SSA, which during modeling does not have the model and parameters, respectively, and Levenberg-Marquardt method, calculating the parameters of the remaining parts of the additive model helps to define the exterior of a time series of "Caterpillar"-SSA method, acting as assistant to the definition of a deterministic (trend) component of the process.

The proposed approach is a variant of the priority to date combined probability and deterministic approaches, because herewith are simultaneously used as statistical and deterministic components that allows to reach the best quality of the forecasting. It also implemented the so-called trend approach, where the process is modeled as the deviation of actual values from the trend (which is presented here as time series obtained by the "Caterpillar"-SSA method), and which ensures the stability of the model and obtained the required accuracy of modeling, whereas previously the probabilistic model ARIMA tried to describe the entire process. Thus, a successful attempt made after more than thirty years after the establishment of the Box-Jenkins method and the "Caterpillar"-SSA method to combine them. However, for satisfaction of such requirements to models, as: learning rate, labor content, resource use, presentation models, ease of use and interpretability, time- and resource-consuming method "Caterpillar"-SSA was later offered to be used only for preliminary structural identification and rough parametric identification of the so-called integrating a polynomial of the operator of the delay  $L$  of proposed model as well as for a rough structural and parametric identification of a polynomial of the delay, whose presence is distinguishes more general polynomial model from the Box-Jenkins model, structure and whose coefficients are equal to those of recurrence prediction model of the "Caterpillar"-SSA method.

The "Caterpillar"-SSA method is also offered to use for preliminary generalized co-integration of multiply time series modeling processes, as well as for separation for a finite and separately for deadbeat regulators in the case of use the proposed model in control theory [Щелкалин, Тевяшев, 2011]. It is also possible nonlinear complication of the model transfer function of one of the ways: FOS, GMDH, RBF, LARS, built on the principal component of their degrees and combinations. So, first of all, the author proposed a model aimed at the automatic control theory, modeling and forecasting of technical systems and technological processes, due to the fact that their transfer functions are more determined and have a complex nonlinear structure.

### Description of the proposed mathematical models

A mathematical model of the processes that depend from several exogenous factors in the operator form can be presented as a model of the seasonal autoregressive - integrated moving average (SARIMA) [Евдокимов, Тевяшев, 1980]:

$$z_t^y = \sum_{i=1}^N \frac{b_{n_{b^i}}^i(L)}{a_{n_{a^i}}^i(L)} \cdot z_{t-m_i}^{x^i} + \frac{c_{n_c}^{\Pi}(L)}{d_{n_d}^{\nabla}(L)} \cdot e_t, \quad (1)$$

where  $L$  – the shift operator in time by one unit back, so that  $L^i x_t = x_{t-i}$ ,  $N$  – the number of exogenous variables;  $z_t^y$  – normalized from 0 to 1 according to the formula  $z_t^y = \frac{y_t - y_t^{\min}}{y_t^{\max} - y_t^{\min}}$  or some other way a time series  $y_t$  of simulated and predicted process subtracted the average value;  $z_{t-m_i}^{x^i}$  – normalized in the same way the  $i$ -th exogenous time series  $x_t^i$  subtracted from the average;  $m_i$  – the delay of the  $i$ -th exogenous time series  $x_t^i$  in time relative to the forecasted time series  $y_t$ ;  $a_{n_{a^i}}^i(L)$ ,  $b_{n_{b^i}}^i(L)$  – polynomials from  $L$  of  $n_{a^i}$  and

$n_{b^i}$  degrees respectively;  $c_{n_c^\Sigma}^\Pi(L) = c_{n_c^1}^1(L^{s_1}) \cdot c_{n_c^2}^2(L^{s_2}) \times \dots \times c_{n_c^{n_s}}^{n_s}(L^{s_{n_s}}) = \prod_{i=1}^{n_s} c_{n_c^i}^i(L^{s_i})$  – polynomial from

$L^{s_i}$  of  $n_c^i$  degree, defining component of the moving average of the periodic component with a periods  $s_i$ ,

$$n_c^\Sigma = \sum_{i=1}^{n_s} n_c^i \cdot s_i;$$

$$\begin{aligned} d_{n_d^\nabla}^\nabla(L) &= d_{n_d^\Pi}^\Pi(L) \nabla_{s_1}^{D_1} \nabla_{s_2}^{D_2} \dots \nabla_{s_{n_s}}^{D_{n_s}} = \\ &= d_{n_d^1}^1(L^{s_1}) \cdot d_{n_d^2}^2(L^{s_2}) \cdot \dots \cdot d_{n_d^{n_s}}^{n_s}(L^{s_{n_s}}) \nabla_{s_1}^{D_1} \nabla_{s_2}^{D_2} \dots \nabla_{s_{n_s}}^{D_{n_s}} = \prod_{i=1}^{n_s} d_{n_d^i}^i(L^{s_i}) \nabla_{s_1}^{D_1} \nabla_{s_2}^{D_2} \dots \nabla_{s_{n_s}}^{D_{n_s}}, \end{aligned}$$

$n_d^\nabla = n_d^\Sigma + \sum_{i=1}^n D_i \cdot s_i$ ,  $d_{n_d^\Sigma}^\Pi(L) = d_{n_d^1}^1(L^{s_1}) \cdot d_{n_d^2}^2(L^{s_2}) \times \dots \times d_{n_d^{n_s}}^{n_s}(L^{s_{n_s}}) = \prod_{i=1}^{n_s} d_{n_d^i}^i(L^{s_i})$  – polynomial from

$L^{s_i}$  of  $n_d^i$  degree, defining component of the autoregressive seasonal component with a period  $s_i$ ,

$$n_d^\Sigma = \sum_{i=1}^{n_s} n_d^i \cdot s_i; e_t - \text{residual errors model: } D_i - \text{the procedure of taking the differences } s_i; \nabla_{s_i} \text{ and } L^{s_i} -$$

simplify the operators such that  $\nabla_{s_i} y_t = (1 - L^{s_i}) \cdot y_t = y_t - y_{t-s_i}$ .

Equation (1) in more compact form can be represented as:

$$\tilde{a}(q) \cdot z_t^y = \sum_{i=1}^k \tilde{b}^i(q) \cdot z_{t-m_i}^{x^i} + \tilde{c}(q) \cdot e_t$$

$$\text{where } \tilde{a}(L) = d_{n_d^\nabla}^\nabla(L) \cdot \prod_{i=1}^N a_{n_a^i}^i(L), \quad \tilde{b}^i(L) = b_{n_b^i}^i(L) \cdot d_{n_d^\nabla}^\nabla(L) \cdot \prod_{j=1, j \neq i}^N a_{n_a^j}^j(L),$$

$$\tilde{c}(q) = c_{n_c^\Sigma}^\Pi(q) \cdot \prod_{i=1}^N a_{n_a^i}^i(q), \quad i = \overline{1, N}.$$

The expression for the prediction of pre-emption  $l$  using the proposed model of the joint use to models of the "Caterpillar"-SSA method and seasonal autoregressive model - integrated moving average with exogenous variables, after adduction it from rational form to the difference equation takes the type:

$$\begin{aligned} \hat{y}_t(l) &= \hat{y}_t^{SSA}(l) + \sum_{j=1}^{n_d^\Sigma + \sum_{a^i} n_a^i} \tilde{a}_j \cdot h_{t+l-j}^y + \\ &+ \sum_{i=1}^N \left( \tilde{b}_0^i \cdot x_{t+l-m_i}^i - \sum_{j=1}^{n_d^\Sigma + n_b^i + \sum_{a^p} n_a^p} \tilde{b}_j^i \cdot x_{t+l-m_i-j}^i \right) - \sum_{j=1}^{n_c^\Sigma + \sum_{a^i} n_a^i} \tilde{c}_j \cdot e_{t+l-j}; \quad (2) \\ \hat{x}_t^i(l) &= \hat{x}_t^{SSA}(l) + \sum_{j=1}^{n_{x^k}^k} \tilde{a}_{x^k j} \cdot z_{t+l-j}^{x^k} + \sum_{j=1}^{n_{x^k}^k} \tilde{c}_{x^k j} \cdot e_{x^k t+l-j}, \quad k = \overline{1, N}, \end{aligned}$$

where  $y_{t+j} = \begin{cases} y_{t+j}, j \leq 0, \\ \hat{y}_t(j), j > 0, \end{cases}$   $x_{t+j}^i = \begin{cases} x_{t+j}^i, j \leq 0, \\ \hat{x}_t^i(j), j > 0, \end{cases}$   $i = \overline{1, N}$ ;  $e_{t+j} = \begin{cases} e_{t+j}, j \leq 0, \\ 0, j > 0. \end{cases}$ ;  $h_i^y = y_i - \tilde{w}_i^{N+1}$ ;

$$\hat{y}_t^{SSA}(i) = \sum_{j=1}^{L-1} f_j^y \cdot \tilde{w}_{t+i-j}^{y, N+1}, \quad i = \overline{1, L}; \quad \tilde{w}_i^{y, N+1} = \begin{cases} \hat{y}_i, i > t, \\ \tilde{w}_i^{y, N+1}, i \leq t; \end{cases}$$

$$(f_{L-1}^y, f_{L-2}^y, \dots, f_1^y)^T = \frac{1}{1 - \sum_{i=1}^R (u_L^i)^2} \sum_{i=1}^R u_L^i \cdot (u_1^i \quad u_2^i \quad \dots \quad u_{L-1}^i)^T, \text{ where } (u_1^i \quad u_2^i \quad \dots \quad u_{L-1}^i)^T - \text{ a}$$

vector consisting of the first  $(L-1)$  elements of the singular value decomposition eigenvector  $U^i$  of the trajectory matrix of exogenous and simulated processes

$$X = (X_1 \quad X_2 \quad \dots \quad X_N \quad Y) = (X_{1,1} \quad X_{1,2} \quad \dots \quad X_{1,K} \quad X_{2,1} \quad X_{2,2} \quad \dots$$

$$\dots \quad X_{2,K} \quad \dots \quad X_{N,1} \quad X_{N,2} \quad \dots \quad X_{N,K} \quad Y_1 \quad Y_2 \quad \dots \quad Y_K);$$

$$Y_j = (y_{j-1} \quad y_{j-2} \quad \dots \quad y_{j+L-2})^T, \quad X_{i,j} = (x_{j-1}^i \quad x_{j-2}^i \quad \dots \quad x_{j+L-2}^i), \quad i = \overline{1, N}, \quad j = \overline{1, K} -$$

transformation of the  $(N+1)$ -dimensional time series ( $N$  - number of exogenous time series and one predictable) in the sequence  $L^y$ -dimensional vectors ( $L^y$  - width of the window), whose number is equal  $(N+1) \cdot K$ ,  $K = n - L^y + 1$ , where  $n$  - the length of time series;  $u_L^i$  - the last element of the vector  $U^i$ ;

$R$  - the number of elements of the singular value decomposition;  $\hat{y}_t^{SSA}(i) = \sum_{j=1}^{L-1} f_j^y \cdot \tilde{w}_{t+i-j}^{y, N+1}$  - model of

recurrent prediction "Caterpillar"-SSA method of time  $y_t$ ,  $t = \overline{0, n-1}$ ; which in turn can often be economically

recorded by seasonal ARIMA (ARI) or by Almon model of distributed lags, where  $\tilde{w}_0^{y, N+1}$ ,  $\tilde{w}_1^{y, N+1}$ , ...,  $\tilde{w}_{n-1}^{y, N+1}$  -

a time serie corresponding to the transformation of the predicted time serie  $y_t$ ; time series  $\tilde{w}_0^{x^i}$ ,  $\tilde{w}_1^{x^i}$ , ...,

$\tilde{w}_{n-1}^{x^i}$ ,  $i = \overline{1, N}$  correspond to the transformations  $i$ -th time series of the exogenous time series using singular

spectrum analysis at the stage of diagonal averaging which takes the matrix  $\tilde{Z}^i$ ,  $i = \overline{1, N+1}$  consisting of  $K$

columns from  $(i-1) \cdot K$ -th to  $i \cdot K - 1$ -th matrix  $Z$  to series  $\tilde{w}_0^{x^i}$ ,  $\tilde{w}_1^{x^i}$ , ...,  $\tilde{w}_{n-1}^{x^i}$ , according to the formula

$$\tilde{w}_i^{y, k} = \begin{cases} \frac{1}{i+1} \sum_{j=1}^{k+1} \tilde{z}_{j, i-j+2}^i, i = \overline{0, \min(L, K) - 2}; \\ \frac{1}{\min(L, K)} \sum_{j=1}^{\min(L, K)} \tilde{z}_{j, i-j+2}^i, i = \overline{\min(L, K) - 1, \max(L, K) - 1}; \\ \frac{1}{n-i} \sum_{j=k-\max(L, K)+2}^{n-\max(L, K)+1} \tilde{z}_{j, i-j+2}^i, i = \overline{\max(L, K), n-1}, \end{cases}$$

$k = \overline{1, N+1}$ ; where  $Z = \tilde{Z}^1 + \dots + \tilde{Z}^j$  - the amount of decomposition matrices

$$\tilde{Z}^i = \left( U^i \cdot (U^i)^T \cdot X_1 \quad U^i \cdot (U^i)^T \cdot X_2 \quad \dots \quad U^i \cdot (U^i)^T \cdot X_N \quad U^i \cdot (U^i)^T \cdot Y \right), \text{ selected by standard}$$

analysis of eigenvalues of the trajectory matrix in the "Caterpillar"-SSA method. There is also a modification of the

recursive method of SSA-forecasting – vector SSA-prediction [Голяндина, 2004], which in some cases provides more accurate forecasts.  $\hat{x}_t^{k, SSA}(i) = \sum_{j=1}^{L^y-1} f_j^{x^k} \cdot \tilde{w}^{x^k}_{t+i-j}$ ,  $k = \overline{1, N}$ , same thing, only for  $k$ -th exogenous time series, as well as all the variables and parameters with an index  $x^i$ ,  $i = \overline{1, N}$  similarly interpreted as for the time series  $y_t$ ,  $t = \overline{1, n}$ ;

With the nonlinear complexity of the transfer function the first equation (2) becomes:

$$\begin{aligned} \hat{z}_t^y(l) &= \hat{y}_t^{SSA}(l) + \sum_{i=1}^r g_i \cdot p_t^i + \sum_{j=1}^{n_d^{\Sigma} + \sum_{i=1}^N n_{a^i}} \tilde{a}_j \cdot h_{t+l-j}^y + \\ &+ \sum_{i=1}^N \left( \tilde{b}_0^i \cdot z_{t+l-m_i}^{x^i} - \sum_{j=1}^{n_d^{\Sigma} + n_{b^i} + \sum_{p=1}^N n_{a^p}} \tilde{b}_j^i \cdot z_{t+l-m_i-j}^{x^i} \right) - \sum_{j=1}^{n_c^{\Sigma} + \sum_{i=1}^N n_{a^i}} \tilde{c}_j \cdot e_{t+l-j}, \\ \hat{x}_t^i(l) &= \hat{x}_t^{k, SSA}(l) + \sum_{j=1}^{n_{x^k}^{\Sigma}} \tilde{a}_{x^k j} \cdot z_{t+l-j}^{x^k} + \sum_{j=1}^{n_{x^k}^{\Sigma}} \tilde{c}_{x^k j} \cdot e_{x^k t+l-j}, k = \overline{1, N}, \end{aligned}$$

where  $g_i \cdot p_t^i$  – the members of the Kolmogorov-Gabor polynomial:

$$\sum_{i=1}^r g_i \cdot p_t^i = FOS \left( \sum_{i=1}^M \tilde{g}_i \cdot \tilde{x}_t^i + \sum_{i=1}^M \sum_{j=1}^M \tilde{g}_{ij} \cdot \tilde{x}_t^i \cdot \tilde{x}_t^j + \sum_{i=1}^M \sum_{j=1}^M \sum_{k=1}^M \tilde{g}_{ijk} \cdot \tilde{x}_t^i \cdot \tilde{x}_t^j \cdot \tilde{x}_t^k + \dots \right),$$

$$g_n = \tilde{g}_{ij\dots k}, n = \overline{1, r} \quad \text{or} \quad \sum_{i=1}^r g_i \cdot p_t^i = FOS \left( \sum_{i=1}^{M_2} \tilde{g}_i \cdot \varphi_i(\tilde{x}_t) \right), \quad \text{where}$$

$$\varphi_i(\tilde{x}_t) = \frac{1}{(2\pi)^{-\frac{M}{2}} \cdot |\Sigma_i|^{-\frac{1}{2}}} \cdot e^{-\frac{1}{2}(\tilde{x}_t - \bar{c}_i) \Sigma_i^{-1} (\tilde{x}_t - \bar{c}_i)^T}, \quad \bar{c}_i - \text{the vector of mathematical expectation of time series,}$$

representing the principal components,  $\Sigma_i$  – the covariance matrixes,  $i = \overline{1, M}$ ; or

$$\sum_{i=1}^r g_i \cdot p_t^i = GMDH(\tilde{x}_t^1, \tilde{x}_t^2, \dots, \tilde{x}_t^M) - \text{structural identification of GMDH method, } p_t^i = \tilde{x}_t^i \cdot \tilde{x}_t^j \cdot \dots \cdot \tilde{x}_t^k,$$

consisting of the principal components identified as  $\tilde{x}_t^i$ ,  $i = \overline{1, M}$ , and their degrees and combinations, selected by FOS-algorithm, determining the nonlinear part of the proposed model; *FOS* – the function of the structural simplification of the model, written in her argument, given the nature of the behavior of time series with fast orthogonal search algorithm;  $h_i^y = z_i^y - \tilde{w}_i^{N+1} - \sum_{i=1}^r g_i \cdot p_t^i$ .

The process of finding such combined models (2) of the joint use of ARIMA and of the "Caterpillar"-SSA method can be prolonged due to the resource-intensive of the "Caterpillar"-SSA. Therefore, the "Caterpillar"-SSA method analysis is proposed to use only for pre-structural identification and rough parametric identification of integrating polynomial  $w(L)$  (hence the name of an autoregressive model - spectrally integrated moving average) of the delay operator  $L$  of model, which can also be interpreted as an transfer operator into the state space.

$$f(L) \cdot w(L) \cdot y_t = \frac{c(L)}{d(L)} \cdot e_t - \text{ARSIMA model,} \quad (3)$$

and the recursion SSA-forecasting method for gross-structural and parametric identification of a polynomial  $f(L)$ , whose structure and coefficients are equal to those of first recurrence prediction model of the "Caterpillar"-SSA method and the presence of which distinguishes the more general polynomial model from the Box-Jenkins model and in conjunction with  $w(L)(\frac{w_1(L)}{w_2(L)})$  which may have, generally, a rational view), determining the long-term memory model, describing a more wide class of processes of long-term memory than fractional integration in the ARFIMA model, which in turn was invented to overcome the lack of ARIMA models for modeling and forecasting processes in a long memory - loss (distortion) of long-term information in the when taking the incrementations. Polynomials  $d(L)$  and  $c(L)$ , in turn, determine the short-term depending of the process.

When analyzing and forecasting time series, depending on several other essential balance of the dynamic properties of the variables on the left-and right-hand sides of the equation of model. In this case the ideas of the "Caterpillar"-SSA method stand for pre-generalized cointegration of time series and model is divided as follows:

$$\begin{aligned} \hat{w}_t^y &= \frac{b_{n_{b^y}}^y(q)}{a_{n_{a^y}}^y(q)} \cdot z_t^y + \sum_{i=1}^N \frac{b_{n_{b^i}}^{w^y i}(q)}{a_{n_{a^i}}^{w^y i}(q)} \cdot z_{t-m_i}^{x^i} + \frac{c_{n_c}^{w^y \Pi}(q)}{d_{n_d}^{w^y \Pi}(q)} \cdot e_t^{w^y}; \\ z_t^y &= f^y(q) \cdot \hat{w}_t^y + \sum_{i=1}^N \frac{b_{n_{b^i}}^i(q)}{a_{n_{a^i}}^i(q)} \cdot z_{t-m_i}^{x^i} + \frac{c_{n_c}^{\Pi}(q)}{d_{n_d}^{\Pi}(q)} \cdot e_t; \\ f^{x_j}(q) \cdot \omega(q) \cdot z_t^{x^j} &= \frac{c_{n_c}^{x^j \Pi}(q)}{d_{n_d}^{x^j \Pi}(q)} \cdot e_t^{x^j}, j = \overline{1, N}, \end{aligned} \tag{4}$$

$\hat{w}_t^y$  – approximation by time series  $\hat{y}_t$  and by exogenous time series  $x_t^j$ , time series  $\tilde{w}_t^y$  with seasonal ARIMAX model (or ARIX - integrated auto regression with exogenous variables), which was originally obtained by the "Caterpillar"-SSA method and, subsequently, adjustable by the optimization method with competitive learning of model;  $\omega(L)$  – integrating polynomial that takes the time series  $x_t^j$  in time series  $\tilde{w}_t^{x^j}$  – an approximation of the time serie  $\tilde{w}_{x^k}^{N+1}$  by ARIMA model; the initial rough values of the polynomials coefficients  $f^y(L)$ ,  $f^{x_i}(L)$  and their number can be taken equal to the coefficients  $f_j^y$  и  $f_j^{x_i}$ ,  $j = \overline{1, N}$  of models SSA-

recursive prediction method  $\hat{y}_t^{SSA}(i) = \sum_{j=1}^{L^y-1} f_j^y \cdot \tilde{w}_{t+i-j}^{y, N+1}$  and  $\hat{x}_t^{k, SSA}(i) = \sum_{j=1}^{L^{x^k}-1} f_j^{x^k} \cdot \tilde{w}_{t+i-j}^{x^k, N+1}$  respectively;

$L^y$  and  $L^{x_i}$  – appropriate lengths of windows, and then iteratively tune with the rest of the coefficients of model (4) using of Levenberg-Marquardt method. Model (4) benefits significantly by the time training a combined model of sharing seasonal ARIMAX model and the method of "Caterpillar"-SSA (2), but slightly inferior to it by the statistical properties with regard to the manner of its construction is called as seasonal autoregressive model - spectrally integrated moving average model with exogenous variables ( ARSIMAX) and can be written as follows:

$$\hat{z}_t^y(l) = \sum_{j=1}^{L^y+n_{b^y}+\sum_{i=1}^N n_{a^i}^{w^y}+n_{a^d}^{w^y \nabla}+\sum_{i=1}^N n_{a^i}+n_d^{\nabla}} \tilde{a}_j \cdot z_{t+l-j}^y +$$

$$\begin{aligned}
& + \sum_{i=1}^N \left( \tilde{b}_0^i \cdot z_{t+l-m_i}^{x^i} - \sum_{j=1}^{n_b^y} \tilde{b}_j^i \cdot z_{t+l-m_i-j}^{x^i} \right) - \sum_{j=1}^{n_c^{\Sigma} + n_{a^y} + \sum_{i=1}^N n_{a^i}^{w^y} + n_{a^y} + \sum_{i=1}^N n_{a^i} + n_d^{\nabla}} \tilde{c}_j \cdot e_{t+l-j} - \\
& - \sum_{j=1}^{L^y + n_{a^y}^{\Sigma} + n_{a^y} + \sum_{i=1}^N n_{a^i}^{w^y} + \sum_{i=1}^N n_{a^i} + n_d^{\nabla}} \tilde{d}_j \cdot e_{t+l-j}; \\
& \hat{z}_t^{x^k}(l) = \frac{L^{x^k} + n_{a^d}^{\nabla}}{\sum_{j=1}^{L^{x^k} + n_{a^d}^{\nabla}} \tilde{a}_j^{x^k} \cdot z_{t+l-j}^{x^k} + \sum_{j=1}^{n_{a^c}^{\Sigma}} \tilde{c}_j^{x^k} \cdot e_{t+l-j}^{x^k}}, \quad k = \overline{1, N},
\end{aligned}$$

where  $\tilde{a}_j$ ,  $j = 1, L^y + n_{b^y} + \sum_{i=1}^N n_{a^i}^{w^y} + n_{a^y} + \sum_{i=1}^N n_{a^i} + n_d^{\nabla}$  – coefficients of the polynomial

$$\tilde{a}(L) = f^y(L) \cdot b_{n_{b^y}}^y(L) \cdot \prod_{i=1}^N a_{n_{a^i}^{w^y}}^{w^y}(L) \cdot d_{n_d^{\nabla}}^{\nabla}(L); \quad \tilde{b}_j^i, \quad i = \overline{1, N}, \quad j = 1, 2, \dots, n_b^y; \quad n_b^y =$$

$$= L^y +$$

$$+ \max \left( n_{b^i}^{w^y} + \sum_{j=1, j \neq i}^N n_{a^j}^{w^y} + n_{a^y} + n_{a^y}^{\nabla} + \sum_{j=1, j \neq i}^N n_{a^j} + n_d^{\nabla}, j = \overline{1, N}, n_{b^i} + n_{a^y} + \sum_{j=1, j \neq i}^N n_{a^j}^{w^y} + n_{a^y}^{\nabla} + \sum_{j=1, j \neq i}^N n_{a^j} + n_d^{\nabla} \right)$$

– coefficients of the polynomial

$$\begin{aligned}
\tilde{b}^i(L) &= f^y(L) \cdot \sum_{i=1}^N \left( b_{n_{b^i}^{w^y}}^{w^y}(L) \cdot \prod_{j=1, j \neq i}^N a_{n_{a^j}^{w^y}}^{w^y}(L) \cdot a_{n_{a^y}}^y(L) \cdot d_{n_{a^d}^{w^y} \nabla}^{w^y \nabla}(L) \cdot \prod_{j=1, j \neq i}^N a_{n_{a^j}}^j(L) \cdot d_{n_d^{\nabla}}^{\nabla}(L) \right) - \\
& - \sum_{i=1}^N \left( b_{n_{b^i}}^i(L) \cdot a_{n_{a^y}}^y(L) \cdot \prod_{j=1, j \neq i}^N a_{n_{a^j}^{w^y}}^{w^y}(L) \cdot d_{n_{a^d}^{w^y} \nabla}^{w^y \nabla}(L) \cdot \prod_{j=1, j \neq i}^N a_{n_{a^j}}^j(L) \cdot d_{n_{a^d}^{w^y} \nabla}^{w^y \nabla}(L) \right)
\end{aligned}$$

$\tilde{c}_j$ ,  $j = 1, n_c^{\Sigma} + n_{a^y} + \sum_{i=1}^N n_{a^i}^{w^y} + n_{a^y}^{\nabla} + \sum_{i=1}^N n_{a^i} + n_d^{\nabla}$  – coefficients of the polynomial

$$\tilde{c}(L) = c_{n_c^{\Sigma}}^{\Pi}(L) \cdot a_{n_{a^y}}^y(L) \cdot \prod_{i=1}^N a_{n_{a^i}^{w^y}}^{w^y}(L) \cdot d_{n_{a^d}^{w^y} \nabla}^{w^y \nabla}(L) \cdot \prod_{i=1}^N a_{n_{a^i}}^i(L) \cdot d_{n_d^{\nabla}}^{\nabla}(L), \quad i = \overline{1, N}; \quad \tilde{d}_j,$$

$j = 1, L^y + n_{a^y}^{\Sigma} + n_{a^y} + \sum_{i=1}^N n_{a^i}^{w^y} + \sum_{i=1}^N n_{a^i} + n_d^{\nabla}$  – coefficients of the polynomial

$$\tilde{d}(L) = f^y(L) \cdot c_{n_{a^c}^{\Sigma}}^{w^y \Pi}(L) \cdot a_{n_{a^y}}^y(L) \cdot \prod_{i=1}^N a_{n_{a^i}^{w^y}}^{w^y}(L) \cdot \prod_{i=1}^N a_{n_{a^i}}^i(L) \cdot d_{n_d^{\nabla}}^{\nabla}(L), \quad i = \overline{1, N}; \quad \tilde{a}_j^{x^i},$$

$j = 1, L^{x^k} + L^{\omega} + n_{a^d}^{x^k \nabla}$  – coefficients of the polynomial  $\tilde{a}^{x^i}(L) = f^{x^i}(L) \cdot \omega(L) \cdot d_{n_{a^d}^{x^i \nabla}}^{x^i \nabla}(L); \quad \tilde{c}_j^{x^i},$

$j = 1, n_{a^c}^{x^k \Sigma}$  – coefficients of the polynomial  $\tilde{c}^{x^i}(L) = c_{n_{a^c}^{x^i \Sigma}}^{x^i \Pi}(L).$

To account for heteroskedasticity of the process (changing the variance in time) applied the generalized model with autoregressive conditional heteroskedasticity GARCH( $m, r$ ), which has the form [Перцовский, 2003]:

$$\sigma_t^2 = w + \theta(L)\varepsilon_t^2 + \varphi(L)\sigma_t^2,$$



where  $\sigma_t^2$  – a time series of process dispersion changes  $y_t$ ,  $\theta(L) = \theta_1 L + \theta_2 L^2 + \dots + \theta_p L^m$ ,  $\varphi(L) = \varphi_1 L + \varphi_2 L^2 + \dots + \varphi_r L^r$ ,  $\varepsilon_t^2$  – the remainders of model. The GARCH( $m, r$ ) model can be expressed through the ARMA model as follows [Bollerslev, 1986]:

$$\varepsilon_t^2 = \frac{w + (1 - \varphi(L))}{(1 - \theta(L) - \varphi(L))} v_t,$$

where  $s = \max(r, m)$ ,  $v_t = \varepsilon_t^2 - \sigma_t^2$ .

Fractal integrated GARCH process can be written as follows:

$$(1 - L)^{2-H} \varepsilon_t^2 = \frac{w + (1 - \varphi(L))}{(1 - \theta(L) - \varphi(L)) \cdot (1 - L)^{-1}} v_t,$$

where  $H$  – the Hurst factor.

The proposed spectrally integrated generalized model with autoregressive conditional heteroskedasticity is as follows:

$$f^{\varepsilon^2}(L) \cdot \omega^{\varepsilon^2}(L) \cdot \varepsilon_t^2 = \frac{w + (1 - \varphi(L))}{(1 - \theta(L) - \varphi(L))} \cdot v_t,$$

where  $\omega^{\varepsilon^2}(L)$  – integrating polynomial of delay operator, with which approximated the time series variance of the noise  $\varepsilon_t^2$  into transformed smoothed by the “Caterpillar”-SSA method time series  $w_t^{\varepsilon^2}$ , and preliminary structural identification and a rough parametric identification of a polynomial of delay  $f^{\varepsilon^2}(L)$ , With which approximated itself series  $\varepsilon_t^2$  is made in determining the coefficients of the recursive prediction formula “Caterpillar”-SSA method;  $w$  – the average value or the level of time series  $\varepsilon_t^2$ .

Thus, the autoregressive - spectrally integrated moving average with the spectrally integrated generalized autoregressive conditional heteroskedasticity and exogenous variables model (ARSIMA – SIGARCH model) takes the form:

$$\begin{aligned} \hat{z}_t^y(l) = & \sum_{j=1}^{L^y + n_{b,y} + \sum_{i=1}^N n_{a_i}^{w,y} + n_{d,y}^{w,y} + \sum_{i=1}^N n_{a_i} + n_d} \tilde{a}_j \cdot z_{t+l-j}^y + \\ & + \sum_{i=1}^N \left( \tilde{b}_0^i \cdot z_{t+l-m_i}^{x^i} - \sum_{j=1}^{n_{b_j}^i} \tilde{b}_j^i \cdot z_{t+l-m_i-j}^{x^i} \right) - \sum_{j=1}^{n_c^{\Sigma} + n_{a,y} + \sum_{i=1}^N n_{a_i}^{w,y} + n_{d,y}^{w,y} + \sum_{i=1}^N n_{a_i} + n_d} \tilde{c}_j \cdot e_{t+l-j}^-, \\ & - \sum_{j=1}^{L^y + n_{c,y}^{w,y} + n_{a,y} + \sum_{i=1}^N n_{a_i}^{w,y} + \sum_{i=1}^N n_{a_i} + n_d} \tilde{d}_j \cdot e_{t+l-j}^{w,y}; \\ \hat{z}_t^{x^k}(l) = & \sum_{j=1}^{L^{x^k} + n_{x^k,d}^{\nabla}} \tilde{a}_j^{x^k} \cdot z_{t+l-j}^{x^k} + \sum_{j=1}^{n_{x^k,c}^{\Sigma}} \tilde{c}_j^{x^k} \cdot e_{t+l-j}^{x^k}, \quad k = \overline{1, N}, \\ & \varepsilon_t^2 = \sigma_t z_t, \end{aligned}$$

$$z_t \sim N(0,1),$$

$$f^{\varepsilon^2}(L) \cdot \omega^{\varepsilon^2}(L) \cdot \varepsilon_t^2 = \frac{w + (1 - \varphi(L))}{(1 - \theta(L) - \varphi(L))} \cdot v_t.$$

The model can be generalized to the multidimensional case.

## Results

Testing the proposed model (4) was carried out on real data of daily consumption of gas from air temperature changes over a three year period.

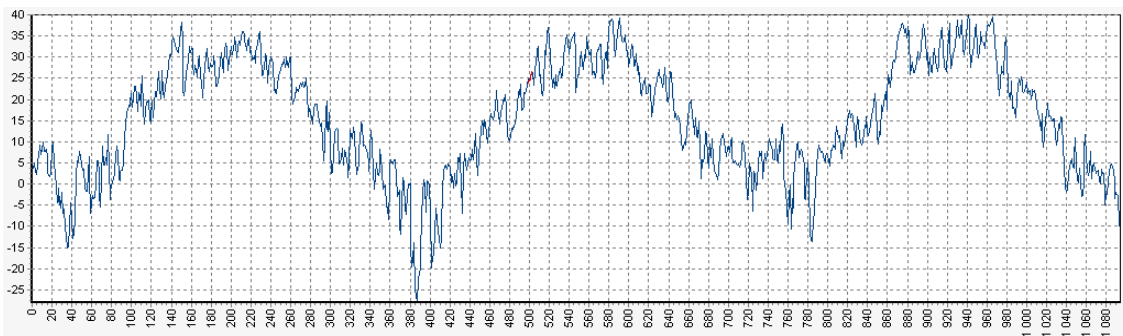


Fig. 1. The graph of daily changes in air temperature data

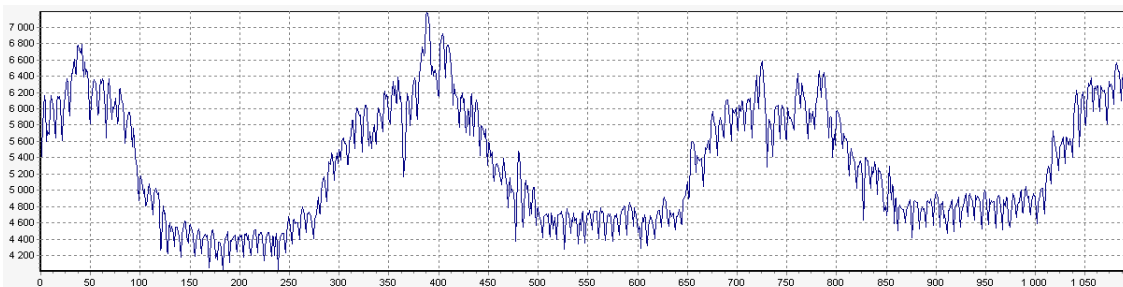


Fig..2. The graph of daily natural gas consumption data

Factoring time series corresponding to large and close-largest Eigen values of the information matrix of data can be judged weekly and yearly seasonal components of time series data.

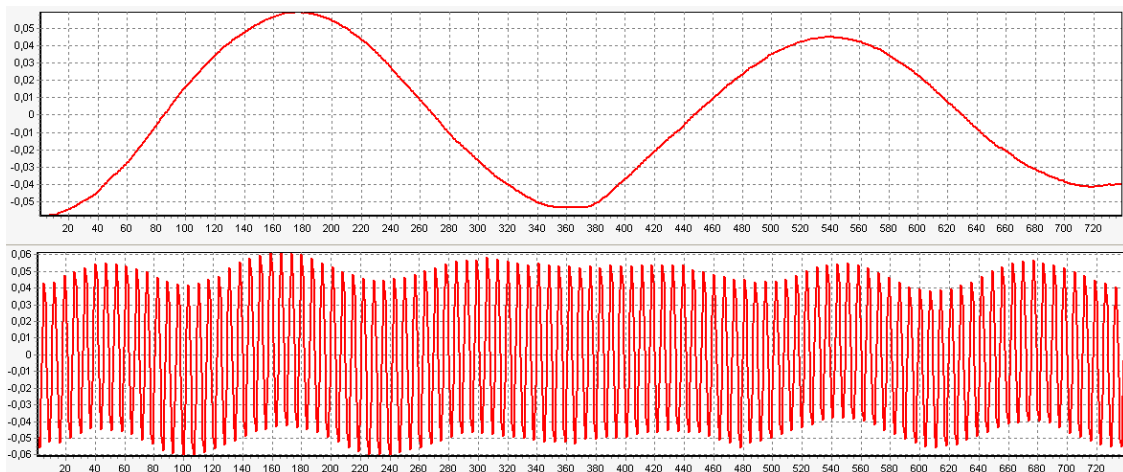


Fig. 3. Some of factor graphs of time series of given processes

Was obtained for the prediction SARIMAX model of natural gas consumption, taking into account changes in air temperature. The average percentage forecast errors was 1.87%.

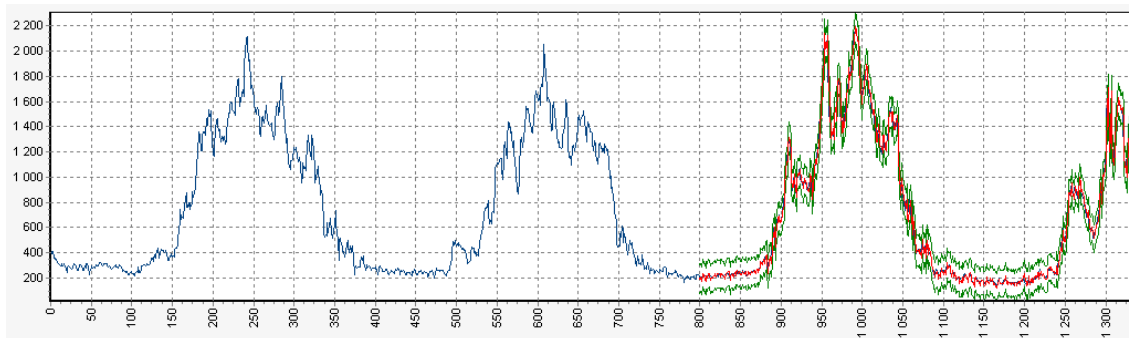


Fig. 4. Charts of forecasts of natural gas consumption by model SARIMAX and 95% confidence intervals

The average percentage error of forecasting natural gas consumption, taking into account changes in air temperature with proposed model was 1.12%. Together with a decrease in forecast errors and confidence intervals are narrowed.

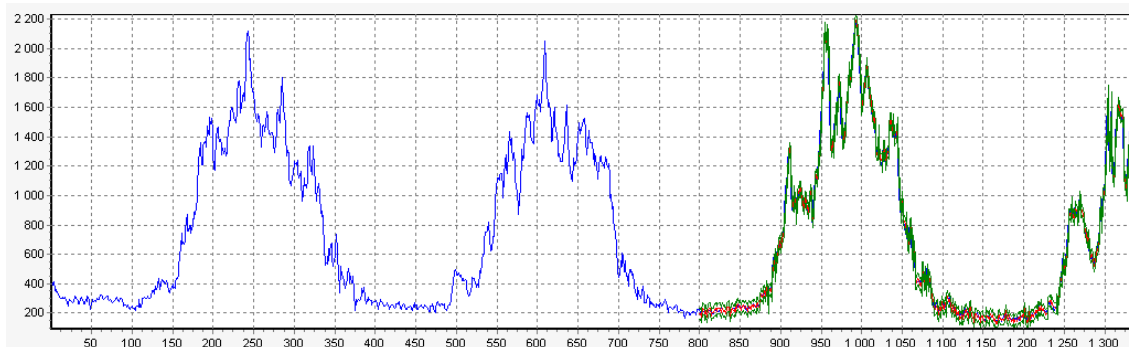


Fig. 5. Charts forecasts natural gas consumption of the proposed model and the 95% confidence intervals

In [Щелкалин, Тевяшев, 2011] presented the application of these models in various fields of science:

- operational forecasting of target products consumption processes in a housing and utilities infrastructure;
- simulation, prediction and control quasi-steady mode of gas-transport systems;
- automatized control for the construction of plants growing single crystals;
- for analysis and forecasting time series in economics;
- to describe and predict the physiological and psycho-physiological processes;
- to simulate the radio-processes and processes in the noise radar, etc;

## Conclusion

Thereby, to obtain adequate models of the complex processes, high-quality forecasts it is necessary to combine the models with miscellaneous structures, including nonlinear models, which are complementary in their competitive learning. The proposed method of "Caterpillar"-SSA – ARIMA – SIGARCH is a modification of the method of "Caterpillar"-SSA with automatic separation of short-term memory and periodic components and can be interpreted as the development of models in state space and the proposed model ARSIMA – SIGARCH – as a

model ARFIMA – FIGARCH is a next modification of the model ARIMA – GARCH, a method of constructing the proposed model is an extension of the method of Box-Jenkins, but to build a broader class of models.

The proposed model ARSIMA – SIGARCH and method of its construction some intermediate approach boundary classical regression and modern neural networks, but more formalized at the choice of structure, being herewith optimum in detail with provision for existing on the date of mathematical, human and machine as the strengths and achievements and shortcomings and limitations.

Summing up the above-described advantages of the proposed method, once again it should be noted that the basic idea is the effect of synergy, which arises from the combined use of two methods: the "Caterpillar"-SSA method and the Box-Jenkins method.

The main advantage of the proposed method of constructing an adequate model of the process under study is its rigorous formalization and, consequently, the ability to fully automate all phases of construction and use of the model.

---

## Bibliography

---

- [Granger, 1980] Granger C.W.J., Joyeux R. An Introduction to Long-Memory Time Series Models and Fractional Differencing // Journal of Time Series Analysis. 1980. N 1(1). P. 15-29.
- [Седов, 2010] Седов А.В. Моделирование объектов с дискретно-распределёнными параметрами: декомпозиционный подход / А.В. Седов; Южный научный центр РАН. – М. : Наука, 2010. – 438 с.
- [Щелкалин, Тевяшев, 2010] Щелкалин В.Н., Тевяшев А.Д. «Автоматизированная система анализа и оперативного прогнозирования процессов потребления целевых продуктов в жилищно-коммунальном хозяйстве». Международный конкурс инновационных проектов "Харьковские инициативы", 2010.
- [Щелкалин, Тевяшев, 2011] Щелкалин В.Н., Тевяшев А.Д. Модель авторегрессии – спектрально проинтегрированного скользящего среднего со спектрально проинтегрированной обобщенной авторегрессионной условной гетероскедастичностью для моделирования, фильтрации, прогнозирования и управления процессами в современных системах автоматизации. // Труды Международной научно-практической конференции «Передовые информационные технологии, средства и системы автоматизации и их внедрение на российских предприятиях» АИТА-2011. Москва, 4 – 8 апреля 2011 г. М.: Институт проблем управления им. В.А. Трапезникова РАН, 2011. с. 996 – 1022.
- [Евдокимов, Тевяшев, 1980] Евдокимов А. Г., Тевяшев А.Д. Оперативное управление потокораспределением в инженерных сетях. – Х. : Вища школа, 1980. – 144 с.
- [Голяндина, 2004] Голяндина Н. Э. Метод «Гусеница»-SSA: прогноз временных рядов: Уч. Пособие, СПб, 2004. – 52 с.
- [Перцовский, 2003] Перцовский О.Е. Моделирование валютных рынков на основе процессов с длинной памятью: Препринт WP2/2004/03 – М.: ГУ ВШЭ, 2003. – 52 с.
- [Bollerslev, 1986] Bollerslev T. Generalized autoregressive conditional heteroscedasticity // Journal of econometrics. – 1986. – V. 31. – PP. 307 – 327.

---

## Authors' Information

---



**Vitalii Shchelkalin** – graduate student, Department of Applied Mathematics, Kharkiv national university of radioelectronics, Lenina str., 14, Kharkiv, Ukraine; e-mail: [vitalii.shchelkalin@gmail.com](mailto:vitalii.shchelkalin@gmail.com)

Major Fields of Scientific Research: Mathematical modeling, prediction, control theory, data analysis, neural networks, data mining

## FUZZY SETS: MATH, APPLIED MATH, HEURISTICS? PROBLEMS AND INTERPRETATIONS

Volodymyr Donchenko

**Abstract:** *Number of Disciplines and Theories changed their status from status of Natural Science discipline to Mathematics. The Theory of Probability is the classical example of that kind. The main privilege of the new Math status is the conception of Math truth, which distinguishes Math from other theories. Some disciplines used in Applications pretended to be Math not really being it. It is entirely true of Fuzzy Subsets Theory with its pretension to be Math and to be exclusive tools in uncertainty handling. Fundamental pretensions of classical Fuzzy subset theory including pretension to be math as well as some gaps in the theory are discussed in the article.*

*Statistical interpretation of membership functions is proposed. It is proved that such interpretation takes place for practically all supporters with minimal constraints on it. Namely, a supporter must be a space with a measure. Proposed interpretation explains modification of classical fuzzy objects to fill the gaps. It is then possible to talk about observations of fuzzy subset within the conception of modification and to extend likelihood method to the new area. Fuzzy likelihood equation is adduced as an example of new possibilities within the proposed approach. One more interpretation for the Fuzzy subset theory is proposed for a discussion: multiset theory.*

**Keywords:** *f Uncertainty, Plural model of uncertainty, Fuzzy subsets Theory, statistical interpretation of the membership function, modification of Fuzzy subsets, Fuzzy likelihood equation, Multiset theory.*

**ACM Classification Keywords:** *G.2.m. Discrete mathematics: miscellaneous, G.2.1 Combinatorics. G.3 Probability and statistics, G.1.6. Numerical analysis I.5.1. Pattern Recognition: Models Fuzzy sets; H.1.m. Models and Principles: miscellaneous:*

---

### Introduction

Initially, the concept of fuzziness was intended to be the object of the proposed article. But later it became clear that the extent of the issue ought to be wider. First of all, the expansion must include a discussion about the role of fuzziness within the concept of uncertainty. What is the “uncertainty” itself? Is it mathematics? If not, where ought one to look for the origin of the concept? How does uncertainty sort with fuzziness? Which one of these two is primary? There are more pertinent questions related to the place of Math and Applied Math in definitions and applications of uncertainty and fuzziness as well as to the role of Heuristics in Applied researches. Uncertainty, surely, is the first in the discussion about priority within the mentioned pair. Pospelov, for instance, and his school [Поспелов, 2001] share this opinion. They consider fuzziness to be the means for handling uncertainty but not vice versa. As to Math, Applied Math, it is worth mentioning in connection with the Fuzzy subset theory (FzTh) coming into the world thanks to Lotfi Zadeh [Zadeh, 1965] (see also [Kaufmann, 1982]), that it was proclaimed to be mathematical panacea for uncertainty modeling.

---

### Mathematics

There is a principal consideration determining the relations between Math and Empiric experience.

As to Mathematics itself, Wikipedia, for example, [Wikipedia, Math] states the following: “Mathematics is the study of quantity, structure, space, and change. Mathematicians seek out patterns, formulate new conjectures, and

establish truth by rigorous deduction from appropriately chosen axioms and definitions.” Thus, specific objects (Math structures) and conception of Math truth (rigorous deduction) are the essence of Mathematics.

As to the Math structures (see for example [Донченко, 2009]). When saying “math structure” we ought to understand it as a set plus “bonds” or “relations” between the elements of the set. Correspondent “bonds” or “relations” in Math are specified by: 1) Math relations (for example “ $\leq$ ” in  $\mathbb{R}^1$ ); 2) functions; 3) operations (for example “+”, “ $\cdot$ ” in  $\mathbb{R}^1$ ); 4) collections of subsets (for example, collection of open subsets, collection of closed subsets or collection of neighbours in  $\mathbb{R}^1$ ); 5) combinations of the previous four. All Math structures were initially established for sets of numbers of different kinds: integers, real, complex. Then they were extended on abstract sets. Therefore, we now have, for example, a structure of metric space (an abstract set plus real valued non negative function of two arguments with certain properties), structure of a group, including an affine one (an abstract set plus binary operation with certain properties); structure of linear space (an abstract set with the structure of affine group plus product operations for each real number); structure of Euclidean and Hilbert space (structure of linear space plus non-negative real-valued function of two arguments: scalar product); topological space (an abstract set plus an appropriate collection of its subsets, with possibility to define the limit), measurable space (an abstract set plus a collection of its subspace, named by  $\sigma$ -algebra), linear topological space and so on. More detailed structure may be considered within the base structure: linear subspace or hyper plane within linear or Euclidean space, a subgroup within the group and so on.

---

### Math truth

---

The fundamental concept of Math truth is the concept of deducibility. It means that the status of truth (proven statement) is given to the statement which is terminal in the specially constructed sequence of statements called its proof. It is a peculiarity in sequence constructing that the next in it is produced by the previous one by special admissible rules (deduction rules) from initial, admissible statements (axioms and premises of a theorem). As a rule, corresponding admissible statements have a form of an equation with formulas on both its sides. So, each subsequent statement in the sequence-proof of the terminal statement is produced by the previous member of the sequence (equation) by changing a part of the formula on its left or right side to another one: from another side of equations-axioms or equations premises. The specification of restrictions on admissible statements and the deduction rules are the object of math logic.

---

### Applied Mathematics

---

The main aim of Applied Math (AppMa) is description of a real object by Math. This means that the object under consideration as “a structure” is represented by means of Math structure, i. e. main parts of the object under modeling and principal bonds, ties, relation between them are represented by means of Math structurization. It is a necessary condition for math modeling to have apt interpretation for correspondent Math objects and objects under observation. Such interpretations, for example for a function and its derivatives, are correspondingly path and speed. Integration and differentiation are the means to represent the relation between the speed and path. Likewise, frequencies of that group or those groups of results in a sequence of observations are interpreted as probabilities and vice versa. Surely, those interpretations can be applied under certain restrictions. So, we cannot investigate discrete systems by means of differential equations or apply probabilistic method out of fulfilling low-frequencies stability. The main aim of Math description, Math modeling of a real object, is to take advantage of establishing true statements for an apt Math object (target statement) which represents the real object: for its Math model, with following interpretation of correspondent statements. So, if the model of the real object is an equation, the target math statement is the statement on its decision and following interpretation of the decision for

the real object. So, the following three-step procedure is the essence of AppMa. 1. Math decrypting of an object on the base of available knowledge about the real object under consideration and with the help of an apt interpretation. 2. Establishing math truth for apt (target) statements within the Math model. 3) Interpretation of the target math statements for the real object. The first and the last steps are impossible without interpretation. The availability of interpretation is principal for applied math. Thus, interpretation plus math rigorous truth are the essence of applied math. So, for example, numerology is not Math and AppMa because it does not appeal to target (Math truth) statements.

---

### Heuristics

---

There are some more means of research which use Math at this stage of investigation but they do not follow the three-step procedure of AppMa. They may be designated by “intellectual calculus” as some investigators do it, but it is reasonable, to my mind, to use an old apt word “heuristic”. Indeed, [ by Wikipedia: heuristics] “heuristic or heuristics (from the Greek "Εὕρισκω" for "find" or "discover") refers to experience-based techniques for problem solving, learning, and discovery. Heuristic methods are used to speed up the process of finding a good enough solution, where an exhaustive search is impractical. Examples of this method include using a "rule of thumb", an educated guess, an intuitive judgment, or common sense”. Taking advantage of opportunity, we mention here the authors of the “heuristic” from Polya [Polya,1945] through A.Newell&J.C. Shaw& H.A.Simon[Newell& Shaw& Simon,1962] to D. Kahneman [Wikipedia: Kahneman].

---

### Uncertainty

---

It is common for investigators to say or use the expression “modeling under uncertainties”. It is also generally recognized that the theory of probability is the classical means for uncertainty handling when such uncertainty is shown as randomness. Determination of randomness appeals to the notion of experiment (observation, trail, test, sometimes – stochastic experiment). Thus, to understand what randomness is, it is necessary to look into the conception of “experiment”.

---

### Experiment

---

As analysis of numerous sources on Theory of Probability and Math Statistics [Донченко 2009] shows, the notion of experiment is associated with what is described as *conditions* (conditions of experiment) under which a phenomenon is investigated, and with what occurs under the conditions, described as *the results of experiment*.

Therefore, as in [Донченко 2009], it is proposed to consider “experiment” as the pair  $(c, y)$ :  $c$ - conditions of experiment (observation, trail, and test),  $y$  – result of experiment. Henceforth,  $Y_c$  for the fixed condition  $c$  will denote the set of all possible events that may occur in the experiment under conditions  $c \in C$ . Generally speaking  $Y_c$  is not singleton.

It is reasonable to mark out in a condition  $c$  variation, controlled, part  $x$ :  $x \in R^P$  as a rule and part  $f$ , which is invariable by default in a sequence of experiment. Condition  $c$  under such approach is denoted by the pair:  $c=(x, f), x \in X \subseteq R^P$ .

---

### Sequence of Experiments and their registration

---

If there are  $n$  experiments, then their registration is the sequence below:

$$(c_i, y_i), c_i \in C, y_i \in Y_{c_i}, i = \overline{1, n}. \quad (1)$$

Different variants of (1) can be implemented in practice:

$$((x_i, f), y_i), x_i \in X \subseteq R^p, y_i \in Y_{x_i}, i = \overline{1, n}, \quad (2)$$

$$(x_i, y_j), x_i \in X \subseteq R^p, y_j \in Y_{x_i}, i = \overline{1, n}, \quad (3)$$

$$y_i, y_i \in Y_{c_i}, i = \overline{1, n}, \quad (4)$$

$$y_i, y_i \in Y, i = \overline{1, n} \quad (5)$$

It is obvious that (5) is equivalent to (1) when all conditions are the same:

$$c_i \equiv c, i = \overline{1, n}.$$

But if otherwise, then

$$Y = \bigcup_{i=1}^n Y_{c_i} \neq \text{Singleton}$$

---

### Randomness as a classic example of uncertainty

---

Randomness in designation introduced above means, firstly, that the results of the experiment are not determined by conditions  $c \in C$  definitely, i. e.

$$Y_c \neq \text{Singleton}, c \in C. \quad (6)$$

And, secondly, that the observations satisfy low-frequency stability. This means: 1) in a sequence of experiment with fixed conditions  $c$ , frequency of each collection of possible results from  $Y_c$  turn to some limit value; 2) the limit value does not depend on the sequence of observations and characterizes the phenomenon under consideration. In the Theory of Probability  $Y_c$  is called Space of elementary events and is denoted by  $\Omega$ . Corresponding experiment is often called stochastic experiment.

---

### Plural model of uncertainty

---

As randomness is a special kind of uncertainty, so the definition of uncertainty one ought to look for is in the conception of experiment. Then, the natural definition of uncertainty coincides with the first part of randomness and is described by equity (5). Thus, uncertainty is defined on the basis of experiment and classifies certain relation between conditions of experiment  $c$  and its corresponding results  $y$ . This relation is stated in (5). We will name such conception of uncertainty Plural Model of Uncertainty (PluMoU).

---

### Mathematical means for uncertainty handling

---

There are comparatively few math tools for uncertainty handling. Having no possibility to discuss the issue in detail, we would like to at least mention that these are: 1) Theory of Probability; 2) Inverse Problem; 3) MaxMin method; 4) Hough Transform; 5) Multisets Theory; 6) Fuzzy Theory; 7) combination of 1)-6) issues. The last point



needs additional explanation in order to embed FzTh in PluMoU. Such embedding becomes feasible on the basis of two possible interpretation of FzTh: within Theory of Probability and Multisets theory.

### **Fuzzy Theory and statistical interpretation of membership function**

Fuzzy set  $\underline{A}$ , subset to be more precise (Kaufmann, 1982), as the object in mathematics is nothing more than a graphic image of real valued function  $\mu$  on an abstract crisp (usual) set  $E$  (henceforth - supporter of the Fuzzy subset). There is additional constraint on the value of this function, named membership function in Fuzzy theory: its values are bounded by the segment  $[0,1]$  :

$$\mu: E \rightarrow [0,1] , \underline{A} = \{(e, \mu_{\underline{A}}(e)) : e \in E\} .$$

There are no objections. The definition is perfect but trivial. There are great many functions in mathematics, there are great many graphics and there are no pretensions of the Fuzzy theory.

### **Limitations of Fuzzy Theory**

As it was mentioned above, there are several Math tools for uncertainty handling. All of them are well-grounded Math. So, FzTh is not exclusive in pretension to uncertainty handling. Also, the attention was drawn earlier to the importance of interpretation in Applied Math unlike in fundamental. As to FzTh, the lack of objective interpretation is a rather acute problem. The absence of its own set theory as well as a Fuzzy logic is the problem awaiting solutions. There are some steps relating to logic (see, for example, [Hajek, 1998F], [Hajek, 1998]), but the problem of interpretation in this case must also be solved. The importance of apt interpretation may be clearly demonstrated on history of modal logic.

There is no such a thing as axiomatic set theory in FzTh even in naive, Kantor's sense. Particularly, such axiom of paramount importance known as abstraction [Stoll, 1960] or separation principle [Kuratovski, Mostowski, 1967] is out of consideration. Implementation of a variant of this axiom in FzTh would help to overcome the "object" problem. Indeed, as is well known, the axiom under consideration establishes correspondence between classical (crisp) subsets and properties of the elements of a universal set – namely, predicates on the universal crisp set. So, a classical predicate has its object of characterization: the correspondent set, determined by abstraction axiom. In FzTh changing binary predicates by membership functions there is no defining of other elements in the pair (predicate, set). As a consequence, the object of fuzzy characterization is lost. Incidentally, Multiset theory (see below) with its technique could help solve the problem.

It is interesting that in obvious examples of membership functions out of the FzTh such objects are intrinsic to the definition of the correspondent objects. Namely, such examples are generalized variants of logit- and probit (GeLoPr) – regressions, transition matrix for Markov's chains and Bayesian nets are the examples mentioned.

### **Natural examples of membership function: generalized variants of logit and probit regression**

As to these examples, GeLoPr describes dependence of the frequencies (probabilities) of a certain event A on real valued vector under certain parameterization:

$$P\{A | H_x\} = G(\beta^T \begin{pmatrix} 1 \\ x \end{pmatrix}),$$

$$\beta \in \mathbb{R}^{n-1}, \beta^T = (\beta_0, \dots, \beta_{n-1}), x \in \mathbb{R}^n ,$$

Where  $G$  – distribution function  $F(z)$ ,  $x \in \mathbb{R}^1$  or correspondent tail:  $1-F(z)$  for the scalar distribution.

In this example,  $\text{GeLoPr } \mu(x) = P\{A | H_x\}$ ,  $x \in \mathbb{R}^{n-1} = E$  as a function of  $x \in \mathbb{R}^{n-1}$  is a membership function in classical FzTh, which corresponds to the certain object intrinsic to the theory: event  $A$ . It is important to remember that the event  $A$ , mentioned above, describes presence of a certain property in observation  $(x, y)$ ,  $y \in \{0, 1\}$ . The value 1 for  $y$  means fulfilling and 0 - not fulfilling the property in the observation.

---

### Natural examples of membership function: Markov chain

---

A transition matrix for the Markov's chain  $(\xi_n, n \in \mathbb{N})$ , with stated set  $\wp = \{S_1, \dots, S_M(\dots)\}$  is the  $M \times M$  matrix  $P = (p_{ij})$  of conditional probabilities:

$$p_{ij} = P\{\xi_{n+1} = S_j | \xi_n = S_i\}, i, j = \overline{1, M}$$

Each column with number  $j = \overline{1, M}$  of the matrix defines membership function  $\mu_j, j = \overline{1, M}$  on  $E = \wp$ :

$$\mu_j(S_i) = p_{ij} = P\{\xi_{n+1} = S_j | \xi_n = S_i\}, j = \overline{1, M}, \quad (6)$$

$$S_i \in \wp = E$$

In each of the  $M$  membership functions  $\mu_j(S), S \in \wp = E, j = \overline{1, M}$ , there are intrinsic objects of fuzzy characterization. Namely, these are correspondingly:  $\{\xi_{n+1} = S_j\}, j = \overline{1, M}$ .

It is interesting that it is natural to consider (6) to be a "full system" of membership functions: a collection of functions  $\mu_j, j = \overline{1, M}$  on  $E$ , for which, for any  $e \in E$ , there is:

$$\sum_{j=1}^M \mu_j(e) = 1, e \in E$$

---

### Natural examples of membership function: Bayesian nets

---

Any Bayesian net is, in the essence, a weighted directed graph associated with probabilistic objects. But, when in a classic probabilistic graph the weights are prescribed to the edges with one and the same head-nodes, in Bayesian – to the one with the same tail-nodes. Thus, the collection of probabilities is associated with each node: the probabilities which weigh the nodes of predecessors. So, correspondent probabilities (conditional by its nature) define a membership function.

---

### Probabilistic Interpretation of membership function

---

This subsection deals with probabilistic interpretation of the classical variant of FzTh (Donchenko, 1998, 3). Two variants of a supporter  $E$  are considered below: discrete and non-discrete. Discrete case is the one which fully illustrates the situation. Namely, each membership function of a fuzzy subset is represented by a system of

conditional probabilities of certain events, relatively complete collections of the sets  $H_e, e \in E$ . Saying "complete collection" we consider the collection  $H_e, e \in E$  to be the partition of the space of elementary events  $\Omega$  for a basic probability space.

---

### Probabilistic Interpretation of membership function: discrete supporter

---

The main point of the subsection is represented by theorem 1 [Donchenko, 1998, 3].

**Theorem 1.** For any classical Fuzzy Set  $(E, \mu_A(e))$  with discrete support  $E$ , there exist such discrete probability space

$$(\Omega, B_\Omega, P),$$

event

$$A \in B_\Omega$$

and complete collection of events

$$H_e : H_e \in B_\Omega, e \in E$$

within this probability space, such that membership function  $\mu_A(e)$  is represented by the system of conditional probabilities in the following form:

$$\mu_A(e) = P(A | H_e), e \in E. \quad (7)$$

**Theorem 2.** For any complete collection of Fuzzy subsets  $(E, \mu_{A_i}(e)), i = \overline{1, n}$ , with one and the same supporter  $E$  there exist:

discrete probability space  $(\Omega, B_\Omega, P)$ ;

collection of events  $A_i \in B_\Omega, i = \overline{1, n}$ ;

complete collection of events  $H_e : H_e \in B_\Omega, e \in E$  within the probability space  $(\Omega, B_\Omega, P)$ ,

such, that all of the membership functions  $\mu_{A_i}(e), e \in E, i = \overline{1, n}$  are simultaneously represented by the systems of conditional probabilities in the following way:

$$\mu_{A_i}(e) = P(A_i | H_e), e \in E, i = \overline{1, n}.$$

---

### Probabilistic Interpretation on membership function: non discrete supporter

---

The issue in the previous subsection may be extended noticeably to non-discrete case if the supporter  $E$  possesses certain structure, namely, if it is space with a measure [Donchenko, 1998, 3].

**Theorem 3.** Given that:

$(E, \mathfrak{F}, m)$  - is space with a measure;

$(E, \mu_{A_i}(e)), i = \overline{1, n}$ , is complete Fuzzy subsets collection with the same supporter  $E$ ;

all of the membership functions  $\mu^{(A_i)}(e), i = \overline{1, n}$ , are  $\mathfrak{S}, \mathfrak{L}$ , - measurable ( $\mathfrak{L}$  – Borel  $\sigma$ -algebra on  $R^1$ ),

then, there exist: probability space  $(\Omega, B_\Omega, P)$ ,

$\xi$  - discrete random  $S_p$  – valued random variable on  $(\Omega, B_\Omega, P)$ , where  $S_p$  is any  $n$  -element set with elements, say,  $S_i, i = \overline{1, n}$ ;

$\eta$  random  $E$  – valued random variable on  $(\Omega, B_\Omega, P)$  such that for any  $i = \overline{1, n}$

$$\mu^{(A_i)}(e) = P\{\xi = S_i \mid \eta = e\},$$

where

$$P\{\xi = S_i \mid \eta\}$$

– conditional distribution of random variable(r.v)  $\xi$  respectively r.v.  $\eta$ .

The conditional distribution is regular: for any  $e \in E$   $P\{B \mid \eta = e\}$  there is a probability  $B$  respectively.

Remark on the proof. The proof is the result of extended ideas of the previous theorems, but it embodies an application of another technique: the technique of conditional distribution. The proof, however, being technically complicated is omitted here.

**Remark 1.** There are obvious objects of uncertainty characterization within the theorems 1-3.

---

### Modified definition of fuzzy sets

---

Straight reference to the object or uncertainly described property may be, in the author's opinion, a way to solve the problem of constructing an analog of the separation principle. This reference ought to be reflected evidently in the definition of the membership function:

$$\mu^{(T)}(e), e \in E,$$

where  $T$  – correspondent property (predicate) on certain set  $U$ . The last is the set of “uncertain characterization”.

It may coincide with  $E$ . So  $\mu^{(T)}(e)$  would be “uncertain characterization” of property  $T$  or corresponding crisp subset  $P_T \subseteq U$ . The last transition is possible owing to the separation principle for crisp sets. Two membership functions  $\mu^{(T_1)}(e)$  and  $\mu^{(T_2)}(e)$  with  $T_1 \neq T_2$  would specify two different Fuzzy sets, even if they are equal as the function of  $e, e \in E$ .

**Definition.** The pair

$$(E, \mu^{(T)}(e))$$

or

$$(E, \mu^{(P_T)}(e))$$

is called modified Fuzzy subset (MoF) with  $E$  as a supporter, which uncertainly describes crisp  $T$  on  $U$  (or correspondent crisp subset,  $P_T \subseteq U$ , where  $U$ - the “universal” crisp set of “uncertain characterization”), if:

$E$  – is the abstract crisp set, which is referenced to as a supporter;

$T$  - is a crisp predicate on  $U$  correspondingly,  $P_T$  - crisp subset of  $U$ , which corresponds to  $T$ ;

$\mu^{(T)}(e) \in [0,1]$  - function of two arguments:  $e, e \in E$  and  $T$  from the set of all crisp predicates on universal crisp  $U$ .

The function

$$\mu^{(T)}(e), e \in E$$

just as in classical theory of Fuzzy sets, will be referenced to as membership function, with note that it uncertainly characterizes property  $T$  (or correspondent subset  $P_T$ ).

**Remark 2.** Obviously, statistical interpretation of the theorems 1-3 is applicable to MoF.

---

### Observations of the Modified Fuzzy Sets

---

The modification of Fuzzy set definition introduced earlier in the paper imparts objectivity to Fuzzy sets and allows for a discussion about observations of modified Fuzzy sets (Donchenko, 2004). It's a very important ontological aspect of mathematical modeling using Fuzzy sets. The observation of modified Fuzzy sets is the pair  $(e, T(e)) - e, e \in E$  - element from the supporter and  $T(e)$  is the predicate value on this element. Namely,  $e$  is the element, displayed in observation and  $T(e)$  is the fixed information about fulfilling the property  $T$  in the observation, specified by  $e \in E$ . It is just in such a way the observations are interpreted in the logit- and probit - regressions and in its generalizations.

So, the observation sample is  $(e_i, t_i), t_i = T(e_i), i = \overline{1, n}$  and we can talk about independent observation within statistical interpretation.

---

### Likelihood method for modified fuzzy sets

---

Statistical interpretation of a membership function grants the possibility to talk about an extension of statistical MLM for estimating fuzzy parameter just as it happens in the regressions mentioned above.

Indeed, let

$$\mu^{(T)}(e), e \in E$$

-MoF with membership function from parametric collection of membership functions

$$\mu^{(T)}(e) = \mu(e, \beta), \beta \in R^P.$$

Let  $(e_i, t_i), i = \overline{1, n}$  independent observation of MoF. We determine "Fuzzy Likelihood function"  $FL(\beta)$  by the relation

$$FL(\beta) = \prod_{i=1}^n \mu^{t_i}(e_i, \beta) (1 - \mu(e_i, \beta))^{1-t_i}.$$

Correspondingly, we denote by

$$fl(\beta) = \ln FL(\beta) = \sum_{i=1}^n t_i \ln \mu(e_i, \beta) + \sum_{i=1}^n (1 - t_i) \ln (1 - \mu(e_i, \beta))$$

- logarithmic "Fuzzy Likelihood function".

Just as it is in statistical likelihood estimation

$$\hat{\mu}^{(T)}(e) = \mu(e, \hat{\beta}),$$

where

$$\hat{\beta} = \arg \max_{\beta \in \mathbb{R}^p} \text{FL}(\beta).$$

Just as in Statistics if  $\mu^{(T)}(e) = \mu(e, \beta)$ , necessary condition is the

$$\frac{\partial \text{FL}(\beta)}{\partial \beta} = 0$$

or

$$\frac{\partial \text{fl}(\beta)}{\partial \beta} = 0.$$

The last equation is equivalent to the first one under additional restriction that the set of zeroes of  $\mu(e, \beta), \beta \in \mathbb{R}^p$ , respectively  $\beta \in \mathbb{R}^p$  is the same for all  $e \in E$ .

It is reasonable to refer to the equations of necessary conditions as “fuzzy likelihood equations”.

**Theorem 4.** Under all necessary restrictions “fuzzy likelihood equations” are of the following form:

$$\sum_{i=1}^n \frac{t_i - \mu(e_i, \beta)}{\mu(e_i, \beta)(1 - \mu(e_i, \beta))} \frac{\partial \mu(e_i, \beta)}{\partial \beta_j} = 0,$$

$$j = \overline{1, p}, \beta = \begin{pmatrix} \beta_1 \\ \dots \\ \beta_p \end{pmatrix} \in \mathbb{R}^p.$$

Expert estimating can also be used by combining LSM and MLM.

---

## Multisets Theory

---

Multisets (see, for example, reviews: [Blizard, 1989], [Буй, Богатирьова 2010]) are the Math's answer to the necessity of describing sets with elements which may “repeat”. Thus, originally the concept of multiset implements the idea of repetition  $\text{rep}(u)$  for elements  $u$  from subset  $D$  of a certain universal set  $U$ . Which are the sets  $D$  and  $U$ , and correspondingly  $\text{rep}(u)$ , depends on peculiarities of applied problem. So, for example,  $D$  can be a set of answers for this or that call in the Internet, answers and  $\text{rep}(u)$  – number of repetition for each record.

There is a natural way of implementing the idea of repetition: to provide each  $u \in D$  with a number or repetition  $n_u : n_u \in \{1, 2, \dots, n, \dots\} \equiv \mathbb{N}^+$ .

So, we get the first variant for determining multiset 1. We will call the multiset the set of pairs

$\bigcap_{u \in D} \{(u, n_u)\}, n_u \in \mathbb{N}^+, u \in D \subseteq U$  for any subset  $D$  of certain universal set  $U$ . Then,  $D$  will be the base of

the multiset and  $n_u$  -multiplicity or repetition factor. The terms will be used in all variants of multiset definitions below in an evident way. Multiset with base  $D$  will be denoted by  $D^{(\text{ms})}$ .

Thus, multiset  $D^{(\text{ms})}$  is the usual set  $D$  with “comments”  $n_u$  to its elements.

2. Within the frame of the second definition, multiset for any subset  $D$  of a certain universal set  $U$  is the transformation  $\alpha : D \rightarrow N^+$ , defined for any  $u \in D$  (see, for example, [Петровский, 2002], [Редько, 2001]). Equivalence of the first and second determination is evident:  $\alpha(u) = n_u, u \in U$ . One ought to remark that in the second variant the relation function substitutes the set.

3. Third variant:  $D^{(ms)}$  for  $D \subseteq U$  is the pair  $D^{(ms)} \equiv (D, \alpha) : \forall D \subseteq U, \forall \alpha : D \rightarrow N^+$ ,  $\alpha$  is defined on all elements of  $D$ . Thus, in this variant multiset is the pair: set  $D$  –“comment”  $\alpha$ .

When necessary, we will refer to the components of the multiset-pair  $D_{ms} = (D, \alpha)$  in an evident way, correspondingly, by  $D_\alpha$ , and  $\alpha_D$  as well as by  $D_{D^{(ms)}}, \alpha_{D^{(ms)}} : D = D_{D^{(ms)}}, \alpha_{D^{(ms)}}(u) = \alpha(u), u \in U$ .

Natural set terminology is applied for multisets: for standard operations (“ $\cup$ ”, “ $\cap$ ”) and for standard relation: “ $\subseteq$ ”. We will denote them for multisets correspondingly “ $\cup_{ms}$ ”, “ $\cap_{ms}$ ” “ $\subseteq_{ms}$ ”.

We will define them

$$\forall D_1^{(ms)} = (D_1, \alpha_1), D_2^{(ms)} = (D_2, \alpha_2) : D_i \subseteq U, i = 1, 2$$

by the relations, correspondingly:

1.  $D_1^{(ms)} \subseteq_{ms} D_2^{(ms)} \Leftrightarrow (D_1 \subseteq D_2 \ \& \ \alpha_1 \leq \alpha_2)$
2.  $D_1^{(ms)} \cup_{ms} D_2^{(ms)} \equiv (D_1 \cup D_2, \max(\alpha_1, \alpha_2))$
3.  $D_1^{(ms)} \cap_{ms} D_2^{(ms)} \equiv (D_1 \cap D_2, \min(\alpha_1, \alpha_2))$

As to operation “ $\bar{\quad}$ ”, it is necessary to “cut”  $N^+$  to  $N_M^+ = \{1, 2, \dots, M\}$  leaving the rest of the determinations unchangeable. Then  $\overline{D^{(ms)}} = \overline{(D, \alpha)}$  is determined by the relation

$$\overline{D^{(ms)}} = (D, M - \alpha)$$

Characteristic function  $\chi_{D^{(ms)}}(u)$  (see, for example, [Buy, Bogatyreva, 2010]) is convenient in multiset handling. It is determined by the relation

$$\chi_{D^{(ms)}}(u) = \begin{cases} \alpha(u), & u \in D \\ 0, & u \notin D \end{cases}$$

Namely, characteristic function is an extension of repetition factor or multiplicity on the universal set  $U$ .

The role of characteristic functions in multiset theory is fixed by the equivalency in determination of set operations and order described by the following relations:

1.  $(D_1^{(ms)} \subseteq_{ms} D_2^{(ms)}) \Leftrightarrow (\chi_{D_1^{(ms)}} \leq \chi_{D_2^{(ms)}})$
2.  $\chi_{D_1^{(ms)} \cup_{ms} D_2^{(ms)}} = \max(\chi_{D_1^{(ms)}}, \chi_{D_2^{(ms)}})$ ,
3.  $\chi_{D_1^{(ms)} \cap_{ms} D_2^{(ms)}} = \min(\chi_{D_1^{(ms)}}, \chi_{D_2^{(ms)}})$ .

## Multisets Theory and Fuzziness

It is evident that in multiset theory, repetition factor is the “absolute” variant of membership function. By this we mean absolute and relative frequency. Even more, in the variant of using  $N_M^+$  we get pure membership function by dividing repetition factor  $\alpha$  by  $M$ . However, there are essential differences between these two theories: all membership functions in FTh are referenced to one and the same  $E$  ( $U$  in the designations of the multiset theory) and are referenced to the particular  $D \subseteq U$  in multiset theory. Simple substitution: subsets  $D$  instead of one and the same universal set in Fth solves the problem of the object characterization:  $D$  is the object. All other limitations of the Fth are also immediately solved: 1) own set theory with corresponding set operations and order; 2) own logic: commonly used mathematical logic; 3) interpretation of  $(\alpha(u))$  as  $\text{rep}(u)$ ; 4) abstraction axiom: for each  $D \subseteq U$  there are many possible correspondent  $\alpha$ : any of them.

## Conclusion

General approach to describing uncertainty was expounded in the paper within conception of plurality in understanding uncertainty. The uncertainty is the quality of interaction between the researcher and phenomenon within an observation (experiment, trial, and test). Obviously, some formalization for the “observation” is proposed and discussed in the text. The proposed concept of uncertainty makes it possible for all to use math means for uncertainty handling. It is entirely true of Fuzzy approach after proving principal theorems on statistical interpretation of membership function. Some limitations of Fuzzy Theory were discussed and several examples and directions for overcoming them were demonstrated. Namely, these were modifications proposed for the membership function and Multiset Theory.

## Bibliography

- [Blizzard,1989] Blizzard W.D. The Development of Multiset Theory. In Notre Dame Journal of Formal Logic. - Vol.30, No.1. - 1989.-P. 36-66.
- [Buy, Bogatyreva, 2010] Buy D., Bogatyreva Ju. Multiset Bibliography. In Papers of 9th International Conference on Applied Mathematics, February 2-5,2010.– Bratislava.- 2010.- P.407 -413.
- [Donchenko, 1998,3] Donchenko V. Conditional distributions and Fuzzy sets. In Bulletin of Kiev University. Series: Physics and Mathematics, №3, 1998 (In Ukrainian).
- [Donchenko, 1998,4] Donchenko V. Probability and Fuzzy sets. In Bulletin of Kiev University. Series: Physics and Mathematics, №4,1998. (In Ukrainian)
- [Donchenko, 2004.] Donchenko V. Statistical models of observations and Fuzzy sets. In Bulletin of Kiev University. Series: Physics and Mathematics, №1, 2004 (In Ukrainian).
- [Донченко, 2009.] Донченко В.С., 2009. Неопределённость и математические структуры в прикладных исследованиях (Uncertainty and math structures in applied investigations) Human aspects of Artificial Intelligence International Book Series Information science & Computing.– Number 12.– Supplement to International Journal “Information technologies and Knowledge”. –Volume 3.–2009. – P. 9-18. (In Russian)
- [Hajek, 1998F] Hajek P. Ten questions and one problem on fuzzy logic. In Preprint submitted to Elsevier Science.- February 1998.- 10 p.
- [Hajek, 1998] Hajek P. Metamathematics of Fuzzy Logic.- Kluwer:- 1998.
- [-Wikipedia, Heuristics] Heuristics. [Electronic resource] -Wikipedia: <http://en.wikipedia.org/wiki/Heuristic>.
- [Wikipedia,Kahneman:] Kahneman.- [Electronic resource] -Wikipedia:



[http://en.wikipedia.org/wiki/Daniel\\_Kahneman](http://en.wikipedia.org/wiki/Daniel_Kahneman)

[Kaufmann, 1982] Kaufmann A. Introduction to the Theory of Fuzzy Sets, - Moscow.- 1982. (in Russian).

[Kuratovski, Mostowski,1967] Kuratovski K., Mostowski A. . Set theory – North Holland Publishing Company, Amsterdam.- 1967

[Newell et al, 1962] Newell A., Shaw J. C., Simon H. A. Empirical Explorations of the Logic Theory Machine: A Case Study in Heuristic. In J. Symbolic Logic.- 1962. -Volume 27, Issue 1- P. 102-103.

[Polya, 1945] Polya G. How To Solve It: A New Aspect of Mathematical Method.- Princeton, NJ: Princeton University Press. -1945.-ISBN 0-691-02356-5 ISBN 0-691-08097-6

[Петровский, 2002] Петровский А.Б. Основные понятия теории мультимножеств. – Москва: Едиториал УРСС. – 2002.- 80 с.

[Поспелов, 2001] Поспелов Д. Из истории развития нечетких множеств и мягких вычислений в России.- In Новости искусственного интеллекта. – 2001. – №2-3.

[Редько et al,2001] Редько В.Н., Брона Ю.Й., Буй Д. Б. , Поляков С.А. Реляційні бази даних: табличні алгебри та SQL-подібні мови. – Київ: Видавничий дім “Академперіодика”. – 2001- 198 с.

[Stoll, 1960] Stoll R. Sets, Logic and Axiomatic Theories. Freeman and Company, San Francisco.- 1960.

[Zadeh, 1965] Zadeh L. Fuzzy Sets. In Information and Control, 8(3). June 1965. pp. 338-53.

---

#### Authors' Information

---



**Volodymyr Donchenko** – Professor, National Taras Shevchenko University of Kyiv. Volodymyrs'ka street, Kyiv, 03680, Ukraine; e-mail: [voldon@bigmir.net](mailto:voldon@bigmir.net).

## ROUGH SET METHODS IN ANALYSIS OF CHRONOLOGICALLY ARRANGED DATA

**Piotr Romanowski**

**Abstract:** *The paper presents results of efforts of increasing predicting events accuracy by increasing a set of attributes describing the present moment by information included in past data. There are described two experiments verifying such an approach. The experiments were carried on by the use of the RSES system, which is based on the rough sets theory. The data analyzed in the first experiment, concerning the weather, were reported at the meteorological station in Jasionka near Rzeszów from 1 April 2004 to 30 september 2005. The second experiment deals with exchange rates based on the money.pl news bulletin data (<http://www.money.pl>).*

**Keywords:** *rough sets, prediction, temporal data.*

---

## Introduction

---

The intelligent analysis of data sets describing real-life problems becomes a very important issue of current research in computer science. Different kinds of data sets, as well as different types of problems that they describe, cause that there is no universal methodology nor algorithms to solve these problems. For instance, analysis of a given data set may be completely different, if we define a time order on a set of objects described by this data set, because the problem may be redefined to time dependencies. Moreover, the expectation of an analyst may be different for the same data set, up to the situation. Unknown objects classification on the basis of experience based on known objects is one of the essential issues of data exploration.

The rough set theory, presented by Z. Pawlak [Pawlak, 1981], is one of the most efficient tools in the data analysis. It is successfully used in many areas, such as expert systems, discovery and data mining or machine learning. In many cases, information is not only included in states of objects (attributes values), but in chronology of their occurrence as well. In such situations, we have to deal with temporal data. In the temporal series [Box, 1976], the time of object occurrence is treated literally, attributes' values are time functions, but it is not an essential condition for data to be treated as temporal ones. Information may be included in the sequence of object occurrence [Mannila, 1995], [Synak, 2005].

In the paper, it is assumed, that the size of the data, on the base of which the decision is predicted, does not exceed calculation capabilities of the computer system, and, there is a reserve of resources. Moreover, the analyzed data are chronological, that is why they may be treated similarly to time functions. Therefore, it seems reasonable to generate additional data, on the base of the previous objects and join them to the initial decision table, as new attributes. Such a case is analyzed in this paper, and two presented examples prove, that such an approach may increase efficiency of right decisions generating.

Next sections include a short description of basic concepts and algorithms used in the paper and the way of decision table extension. In the section Data and experiments, data, the methodology and experimental results are presented. Some suggestions for the further research are provided in section Conclusions.

---

## Basic Concepts and Algorithms

---

In the rough set theory, it is assumed, that the known world may be represented as a set of  $U$  objects [Skowron, 1993a], [Polkowski, 2002], [Bazan, 2000]. Each  $u_i$  object in  $U$  is described by a set of its attributes  $a_j$ . Object's attributes may have different meanings, and originally, they are described in different ways, but for practical purposes, they are recorded as values of certain attribute  $a$  from the finite set of values  $V_a$ , usually, the set of integer numbers (data discretization [Nguyen, 1997]). Consequently, the known world is represented by a table of numbers, rows of which are individual objects, and columns include values of corresponding attributes for individual objects. When one column is distinguished as a decision column  $d$ , the table of information is called a decision table  $\mathbf{A}=(U,A,d)$ , where  $U$  is a set of objects,  $A$  is a set of conditional attributes. The decision  $d$  divides a set of objects  $U$  into classes containing objects having the same value of the decision attribute. When a new object, with known values of conditional attributes and unknown value of decision appears, the decision table has to enable its classification to the certain class. Premise for such a classification is a possible similarity of the new object's conditional attribute set to conditional attributes of objects from one of the classes.

Due to big sizes of analyzed data and high complexity of deterministic algorithms there are used algorithms of lower accuracy, but of the lower complexity as well.

Decomposition trees are to divide data into fragments of the size defined earlier, which are represented as decomposition tree leaves. The global and local discretization can be used. More precise description of

classification according to decision trees and ways of discretization are presented in works [Bazan, Szczuka, 2000], [Nguyen, 1997].

Four methods of rules generating have been accepted in the paper.

In the exhaustive method the deterministic algorithm is used, that calculates all minimal rules (i.e. rules with a minimal number of descriptors on the left hand side). The discernibility matrix is generated, and on the base of it, there is build a logical formula, stating for attributes of each two objects if the objects are distinguishable. Such a formula is transformed into the simpler form by the means of absorption laws and new rules are created on the base of this formula. More precise description can be found in the work [Bazan, 2000]. The genetic method is modeled on the mechanism of genes' evolution in the nature. The initial set can be created even in random way. Then, there is a process of weaker elements elimination and better elements modification [Bazan, 2000]. The covering method is based on a set of rules generating on the basis of objects' subsets and than, creating their joint element [Bazan, 2000]. The LEM2 algorithm is based on local covering determination for individual objects from certain decision class, as presented in [Grzymala-Busse, 1997].

### Decision Table Extension

Additional attributes in the 'present' object can be created on the basis of attributes' values in previous objects in many ways. The right choice should result from the knowledge of phenomena peculiarities, described by certain data. In this paper, such additional information are not used. There are analyzed the results of extensions of the decision table by sums of values of the present and previous day:

$$a_i' = a_i + a_{i-1} \tag{1}$$

(such extensions of the decision table are special cases of the autoregression model usage for autoregression coefficients values accepted in advance [Box, 1976]),

and the results of extension data by coefficients A, B, C of parabola:

$$f(x)=Ax^2+Bx+C \tag{2}$$

which interpolates the data of three last days (see experiment 2).

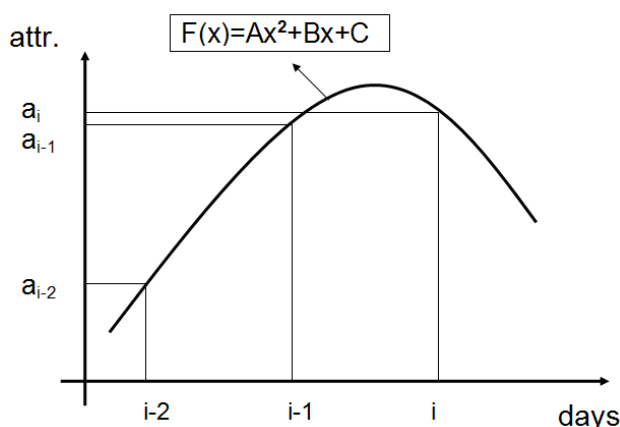


Fig. 1. Creation of new attributes A, B, C on the base of three days data

May be, it could be more profitable, to take into consideration longer preceding time, derivative equivalents or different ways of generating additional information from the past for each attribute. It could be the subject of further investigation.

## Data and experiments

In all calculations, there was used cross-validation method [Michie, 1994] for data classification. Objects (rows) of the decision table are randomly divided into two sets in the ratio, that is the calculation parameter (it was accepted 4:1). The first set is the base of decision tree or rules generating, the second one is for testing [Skowron, 1993]. The accuracy of decision predicting is expressed by the means of so called “confusion matrix” [RSES, 2006], Fig. 2.

		Predicted					No. of obj.	Accuracy
		0	2	5	1	4		
Actual	0	5.8	2.6	0.4	1.6	2.2	38.8	0.459
	2	2.4	2.8	0	0.4	1	33.6	0.438
	5	0.2	0.2	0	0.2	0	9	0
	1	1.4	0.2	0	0.8	0.2	14.6	0.217
	4	1.6	1.8	0	0	3	13	0.481
True positive rate		0.55	0.45	0	0.21	0.45		

Total number of tested objects: 109  
 Total accuracy: 0.433  
 Total coverage: 0.264

Fig. 2. The confusion matrix

Calculations efficiency was accepted as a product of “coverage” and “accuracy” (total coverage and total accuracy). There were accepted two kinds of classifiers: decomposition trees [Bazan, 2000], [Nguyen, 1997] and decision rules [Bazan, 1998].

Six sets of results were obtained from each experiment, which illustrate the effect of primary data extension.

The profit of efficiency, resulting from the information table broadening is called a proportion:

$$\text{profit} = \frac{\text{efficiency (extended data)} - \text{efficiency (primary data)}}{\text{efficiency (primary data)}} \quad (3)$$

### Experiment 1

There are analyzed the weather data, based on measurements made in the Meteorological Station in aeroplain station Jasionka near Rzeszów from 1 April 2004 to 30 september 2005.

In 548 days there were measured the following data: temperature, dew, humidity, pressure, visibility, wind and cloud cover. All of them were treated as parameters of the decision table. Moreover, there were recorded such events as : fog, rain, hail, snow or storm. They were treated as decisions. Events were presented in numerical way, by accepting the following denotations: 0 – none, 1 – fog, 2 – rain, 3 – hail, 4 – snow, 5 – storm. Table 1. illustrates fragment of the weather data.

Table 1. Fragment of the weather data

temp	dew	humidity	pressure	visibility	wind	clouds	events
[ F. deg. ]	[ 1 - 100 ]	[ % ]	[ hPa ]	[ 0 - 20 ]	[ m/s ]	[ 0 - 10 ]	[ 0 - 5 ]
43	31	58	1018.7	5	0	3	0
35	19	43	1018.6	20	10	0	0
38	15	40	1019	20	12	0	0
47	21	51	1019.2	7	10	4	2
50	38	75	1019.4	6	5	4	2

Figs 3 and 4 illustrate temperature and relative humidity (average of the day) and their changes.

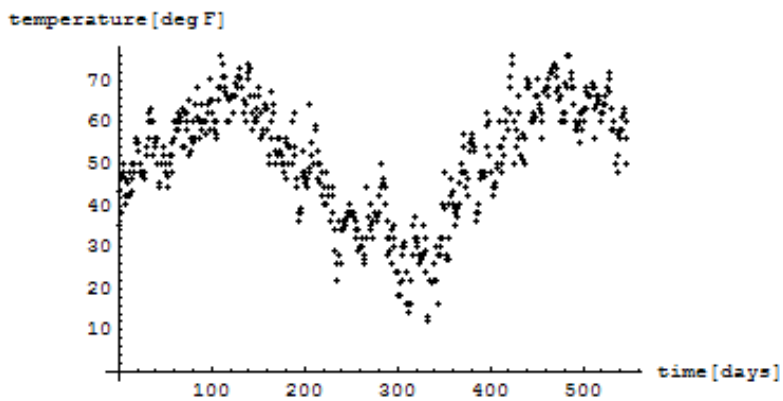


Fig. 3. The temperature (Fahrenheit's scale)

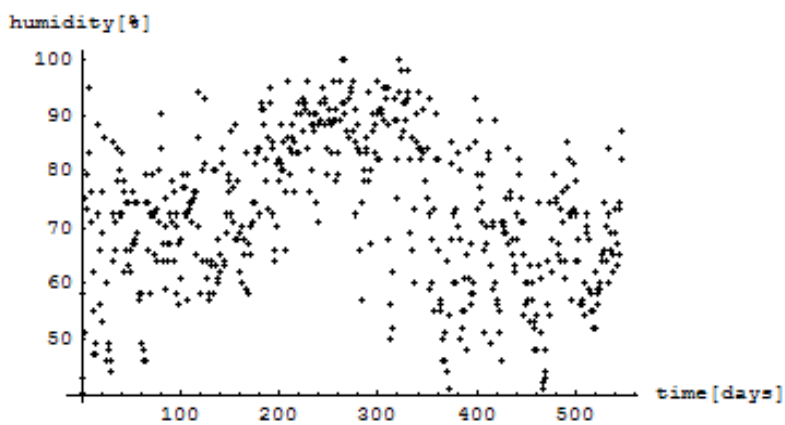


Fig. 4. Relative humidity [ % ]

Fig. 5 illustrates two days' sums of temperature.

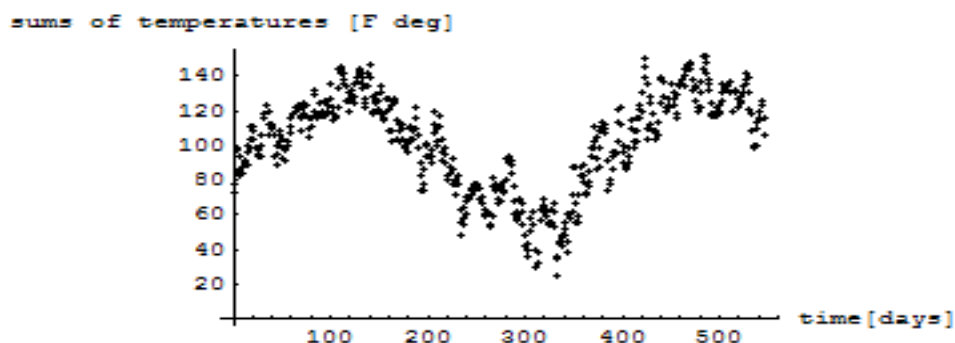


Fig. 5. Two days' sums.

The data analysis, which aim to predict phenomena – event of the preceding day, the decision 'n' is replaced with the following day event 'n+1'. Table 2 illustrates such a movement of the decision table.

Table 2. The decision table. A following day event is a decision

temp	dew	humidity	pressure	wizability	wind	clouds	events	following day events
[ F. deg. ]	[ 1- 100 ]	[ % ]	[ hPa ]	[ 0 - 20 ]	[ m/s ]	[ 0 – 10 ]	[ 0 - 5 ]	[ 0 - 5 ]
								0
43	31	58	1018.7	5	0	3	0	0
35	19	43	1018.6	20	10	0	0	0
38	15	40	1019	20	12	0	0	2
47	21	51	1019.2	7	10	4	2	2
50	38	75	1019.4	6	5	4	2	

For each of six possibilities there were performed four cycles of calculations (by the cross – validation method), the average calculation efficiency and the profit were calculated. The results are presented in Table 3.

Table 3. Experiment 1 – results

Exhaustive algorithm			Covering algorithm		
Data	primary	primary+ sums	Data	primary	primary+ sums
	0.347	0.404		0.130	0.249
	0.348	0.418		0.128	0.219
	0.376	0.400		0.115	0.209
	0.350	0.409		0.113	0.248
<b>average</b>	<b>0.355</b>	<b>0.408</b>	<b>average</b>	<b>0.121</b>	<b>0.231</b>
<b>profit</b>		<b>14.75%</b>	<b>profit</b>		<b>90.52%</b>

**LEM 2 algorithm**

Data	primary	primary+ sums
	0.113	0.125
	0.132	0.115
	0.113	0.102
	0.130	0.128
<b>average</b>	<b>0.122</b>	<b>0.117</b>
<b>profit</b>		<b>-3.68%</b>

**Genetic algorithm**

Data	primary	primary+ sums
	0.3520	0.4130
	0.3360	0.3433
	0.3470	0.387
	0.368	0.391
<b>average</b>	<b>0.351</b>	<b>0.384</b>
<b>profit</b>		<b>9.34%</b>

**Decomposition tree  
(global method)**

Data	primary	primary+ sums
	0.1547	0.1364
	0.1233	0.1324
	0.1547	0.114
	0.126	0.099
<b>average</b>	<b>0.140</b>	<b>0.121</b>
<b>profit</b>		<b>-13.69%</b>

**Decomposition tree  
(local method)**

Data	primary	primary+ sums
	0.1216	0.1207
	0.1483	0.1019
	0.1289	0.111
	0.149	0.118
<b>average</b>	<b>0.137</b>	<b>0.113</b>
<b>profit</b>		<b>-17.42%</b>

Although, the second experiment presents results obtained in the analysis of the data of different areas, results are quite similar.

**Experiment 2**

There are analyzed exchange rates of USD, Euro and CHF to PLN during 200 working days from 24 March 2007. Table 4. illustrates fragment of the Euro exchange rate data.

*Table 4. Fragment of the Euro exchange rate data*

USD	Euro	CHF	decision
[ PLN ]	[ PLN ]	[ PLN ]	(following day)
3.1553	3.9801	2.561	1
3.2012	4.0298	2.5948	-1
3.1936	4.0176	2.5888	1
3.186	4.035	2.5939	0
3.2104	4.0414	2.5945	1
3.2288	4.059	2.6027	1

Figure 6 presents Euro exchange rate.

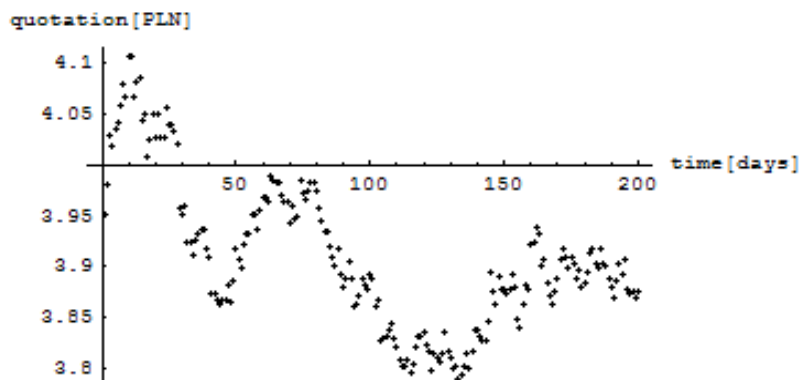


Fig. 6. Euro exchange rate

As a decision, there was accepted a EURO exchange rate in the following day in the form: decrease of more than 0.01 PLN – decision = -1, increase of more than 0.01 PLN – decision = 1 else – decision = 0. For the decision table built in such a way, the calculations were carried on in the same way as in experiment 1. The results are as follows:

Table 5. Experiment 2 – results.

#### Exhaustive algorithm

Data	primary	primary+ sums
	0,330	0,339
	0,333	0,313
	0,313	0,318
	0,327	0,369
<b>average</b>	<b>0,326</b>	<b>0,335</b>
<b>profit</b>		<b>2,84%</b>

#### Covering algorithm

Data	primary	primary+ sums
	0,114	0,275
	0,096	0,256
	0,125	0,282
	0,076	0,294
<b>average</b>	<b>0,103</b>	<b>0,277</b>
<b>profit</b>		<b>169,55%</b>

#### LEM 2 algorithm

Data	primary	primary+ sums
	0,126	0,144
	0,137	0,135
	0,159	0,137
	0,173	0,128
<b>average</b>	<b>0,149</b>	<b>0,136</b>
<b>profit</b>		<b>-8,51%</b>

#### Genetic algorithm

Data	Primary	primary+ sums
	0,301	0,349
	0,315	0,338
	0,305	0,385
	0,306	0,344
<b>average</b>	<b>0,307</b>	<b>0,354</b>
<b>profit</b>		<b>15,37%</b>



**Decomposition tree**

Data	primary	primary+ sums
	0,154	0,156
	0,161	0,118
	0,128	0,203
	0,142	0,133
<b>average</b>	<b>0,146</b>	<b>0,153</b>
<b>profit</b>		<b>4,30%</b>

**Decomposition tree**

Data	primary	primary+ sums
	0,136	0,187
	0,187	0,133
	0,118	0,171
	0,150	0,127
<b>average</b>	<b>0,148</b>	<b>0,155</b>
<b>profit</b>		<b>4,70%</b>

**Exhaustive algorithm**

Data	primary	primary+ ABC
	0,33	0,368
	0,333	0,368
	0,313	0,389
	0,327	0,384
<b>average</b>	<b>0,3258</b>	<b>0,3773</b>
<b>profit</b>		<b>15,81%</b>

**Genetic algorithm**

Data	Primary	primary+ ABC
	0,301	0,353
	0,315	0,358
	0,305	0,347
	0,306	0,358
<b>average</b>	<b>0,307</b>	<b>0,354</b>
<b>profit</b>		<b>15,31%</b>

**Conclusions**

Our main objective was to find the best method classifying unseen objects connected with two sets of chronologically arranged data (the weather data and the exchange rates data). The original data was extended by additional information. The experiments were executed by the use of six rough set algorithms implemented in the RSES system.

Classifications methods and theirs variants can be divided into 'time consuming' and 'economical' ones. The exhaustive and genetic methods are 'time consuming', whereas the others are 'economical'.

Generally, 'economical' methods and theirs extensions result in the loss of decision accuracy, except for the covering method of rules generating, but even this one, in spite of significant advantages in the used extension, does not give satisfying precision, as the precision is below the random choice. Extension of data by two days sums of attributes or coefficients of parabolas, using 'time consuming' methods are more satisfying, as they give significant profits and according to the author of this paper, encourage to further experiments.

**Bibliography**

[Bazan, 1998] Bazan, J.: A comparison of dynamic and not-dynamic rough set methods for extracting laws from decision table. In L. Polkowski, A. Skowron (eds.), Rough sets in knowledge discovery, Physica-Verlag, Heidelberg, pp. 321—365 (1998)

- 
- 
- [Bazan, 2000] Bazan, J., Nguyen, H., Nguyen, S., Synak, P. and Wróblewski, J.: Rough set algorithms In classification problem. In L. Polkowski, S. Tsumoto, and T. Lin, editors, Rough Set Method and Applications, Physica-Verlag, Heidelberg New York, , pp. 49--88 (2000)
- [Bazan, Szczuka, 2000] Bazan, J., Szczuka, M., RSES and RSESLib – A collection of tools for rough set computations. Lecture Notes in Artificial Intelligence 3066, Berlin, Heidelberg:Springer-Verlag, , pp. 592 – 601 (2000)
- [Box, 1976] Box, G., Jenkins, G.: Time series analysis: Forecasting and control. Holden-Day, San Francisco, CA, 2. edition, (1976)
- [Grzymała-Busse 1997] Grzymała-Busse, J.: A new version of the rule induction system LERS. Fundamenta Informaticae, vol. 31(1), pp. 27--39 (1997)
- [Mannila, 1995] Mannila, H., Toivonen, H., Verkamo, A.: Discovering frequent episodes in sequences, in U. Fayyad, R. Uthurusamy. First International Conference on Knowledge Discovery and Data Mining KDD. AAAI Press, Montreal, Canada, pp. 210--215 (1995)
- [Michie, 1994] Michie, D., Spiegelhalter, D, Taylor, C.: Machine learning, neural and statistical classification. Ellis Horwood, New York, (1994.)
- [Nguyen, 1997] Nguyen, H.: Discretization of Real Value attributes, Boolean reasoning approach. Ph. D. thesis, supervisor B. Chlebus, Warsaw University, (1997)
- [Pawlak, 1981] Pawlak, Z.: Information systems – theoretical foundations. Information systems, vol. 6, , pp. 205--218 (1981)
- [Polkowski, 2002] Polkowski, L.: Rough sets: Mathematical foundations. Advances in Soft Computing. Springer-Verlag, Heidelberg, Germany, (2002)
- [Skowron, 1993] Skowron, A.: Boolean reasoning for decision rules generation, in J. Komorowski, Z. Raś. Seventh International Symposium for Methodologies for Intelligent Systems ISMIS, vol. 689 Lecture Notes in Artificial Intelligence. Springer-Verlag, Trondheim, Norway, pp. 295--305. (1993)
- Skowron , A.: A synthesis of decision rules: applications of discernibility matrices, Proceedings of the Conference on Intelligent Information Systems, Practical Aspects of AI, Augustów, Poland, June 7-11, pp. 30--46 (1993)
- [Synak, 2005] Synak, P.: Temporal Templates and Analysis of Time Related Data, in W. Ziarko, Y. Yao (eds.), Second International Conference on Rough Sets and Current Trends in Computing, RSTC 2000, Banff, Canada, October 2000, Lecture Notes in Artificial Intelligence 2005, Springer, pp. 420--427 (2005)
- [RSES, 2006] RSES Homepage <http://logic.mimuw.edu.pl/~rses>

---

## Authors' Information

---

**Piotr Romanowski** -Chair of Computer Science University of Rzeszów, Poland 35-310 ul. Dekerta 2  
35 - 030 Rzeszów e-mail: [proman@univ.rzeszow.pl](mailto:proman@univ.rzeszow.pl)

## ABOUT MULTI-VARIANT CLUSTERING AND ANALYSIS HIGH-DIMENSIONAL DATA

Krassimira Ivanova, Vitalii Velychko, Krassimir Markov, Iliya Mitov

**Abstract:** *In this paper an example of multi-variant clustering is presented. The problems to be solved are described and multi-variant clustering based on pyramidal multi-layer multi-dimensional structures is outlined. The conclusion is that the multi-variant clustering combined with pyramidal generalization and pruning gives reliable results.*

**Keywords:** *Data mining, multi-variant clustering, pyramidal multi-layer multi-dimensional structures.*

**ACM Classification Keywords:** *H.2.8 Database Applications, Data mining; I.5.3 Clustering.*

---

### Introduction

---

Clustering is a fundamental problem that has numerous applications in many disciplines. Clustering techniques are used to discover natural groups in data sets and to identify abstract structures that might reside there without having any background knowledge of the characteristics of the data. They have been used in a variety of areas, including bioinformatics; computer vision; VLSI design; data mining; gene expression analysis; image segmentation; information retrieval; information theory; machine learning; object, character, and pattern recognition; signal compression; text mining; and Web page clustering [Kogan, 2007].

Clustering systems build a generalization hierarchy by partitioning the set of examples in such a way that similarity is maximized within a partition and minimized between them. At the lowest level of the hierarchy are the individual examples.

Clustering is especially suited to unsupervised learning, where the concepts to be learned are not known in advance, but it may also be applied to learning from examples. A new example is classified by considering adding it to each cluster, and determining which one it fits best. This process is repeated down the hierarchy until a cluster is reached that contains only examples of a single class. The new example adopts the class of this cluster. The main differences between different clustering methods are the similarity measure, and the method used to evaluate each cluster to determine the best fit for the new example. Approaches range from Euclidean distance to Bayesian statistics. Clustering is therefore the broad approach of concept formation by grouping similar examples. [Luo et al, 2009]

Clustering has attracted research attention for more than 50 years. A partial list of excellent publications on the subject is provided in [Kogan, 2007].

In this paper we present a simple example of multi-variant clustering and analysis high-dimensional data based on multi-dimensional pyramidal multi-layer structures in self-structured systems.

Let remember that the systems in which the perception of new information is accompanied by simultaneous structuring of the information stored in memory, are called **self-structured** [Gladun et al, 2008]. Self-structuring provides a possibility of changing the structure of stored in memory data during the process of the functioning because of interaction between the received and already stored information.

The building of self-structured artificial systems had been proposed to be realized on the basis of networks with hierarchical structures, named as "**growing pyramidal networks**" (GPN) [Gladun et al, 2008]. The theory as well

---

---

as practical application of GPN was expounded in a number of publications [Gladun, 1987, 1994, 2000; Gladun and Vashchenko, 2000].

**Pyramidal network** is a network memory, automatically tuned into the structure of incoming information. Unlike the neuron networks, the adaptation effect is attained without introduction of a priori network excess. Pyramidal networks are convenient for performing different operations of associative search. Hierarchical structure of the networks, which allows them to reflect the structure of composing objects and natural gender-species' bonds, is an important property of pyramidal networks. The concept of GPN is a generalized logical attributive model of objects' class, and represents the belonging of objects to the target class in accordance with some specific combinations of attributes. By classification manner, GPN is closest to the known methods of data mining as decision trees and propositional rule learning.

The growing pyramidal networks respond to the main requirements to memory structuring in the artificial intelligent systems [Gladun, 2003]:

- in artificial intelligent systems, the knowledge of different types should be united into net-like structure, designed according to principles common for all types of knowledge;
- the network should reflect hierarchic character of real media and in this connection should be convenient for representation of gender-type bonds and structures of composite objects;
- obligatory functions of the memory should be formation of association bonds by revealing intersections of attributive object representations, hierarchic structuring, classification, concept formation;
- within the network, there should be provided a two-way transition between convergent and displayed presentations of objects.

The research done on complex data of great scope showed high effectiveness of application of growing pyramidal networks for solving analytical problems. Such qualities as simplicity of change introduction the information; combining the processes of information introduction with processes of classification and generalization; high associability makes growing pyramidal networks an important component of forecasting and diagnosing systems. The applied problems, for solving of which GPN were used, are: forecasting new chemical compounds and materials with the indicated properties, forecasting in genetics, geology, medical and technical diagnostics, forecasting malfunction of complex machines and sun activity, etc.

The next step is using a new kind of memory structures for operating with growing network information structures. The new proposition is the multi-dimensional numbered information spaces [Markov, 2004]. They can be used as a memory structures in the intelligent systems, and in particular in the processes of data mining and knowledge discovery. Summarizing, the advantages of the multi-dimensional numbered information spaces are:

- possibility to build growing spaces hierarchies of information elements;
- easy building interconnections between information elements stored in the information base;
- practically unlimited number of dimensions - this is the main advantage of the numbered information spaces for well-structured tasks, where it is possible "to address, not to search";
- possibility to create effective and useful tools, in particular for clustering and association rules mining.

The further text of the paper is organized as follow. Firstly we describe the problems to be solved. In the next chapters we present an example of sparse high dimensional vectors and multi-variant clustering based on pyramidal multi-layer multi-dimensional structures. Finally, the conclusions are outlined.

---

## Basic problems to be solved

---

For a given set of instances  $\mathbf{R} = \{R^i, i \in 1, \dots, r\}$  and a query  $Q$  one often is concerned with the following basic problems:

1. Find instances in  $\mathbf{R}$  “related” to the query. If, for example, a “distance” between two instances  $R^i$  and  $R^j$  is given by the function  $d(R^i, R^j)$  and a threshold  $tol > 0$  is specified one may be interested in identifying the subset of instances  $\mathbf{R}_{tol} \subseteq \mathbf{R}$  defined by  $\mathbf{R}_{tol} = \{R : R \in \mathbf{R}, d(Q, R) < tol\}$ .
2. Partition the set  $\mathbf{R}$  into disjoint sub-collections  $\pi_1, \pi_2, \dots, \pi_k$  (called clusters) so that the instances in a cluster are more similar to each other than to instances in other clusters. The number of clusters  $k$  also has to be determined.

When “tight” clusters  $\pi_i, i = 1, \dots, k$  are available, “representatives”  $\mathbf{C}_i$  of the clusters can be used instead of instances to identify  $\mathbf{R}_{tol}$ . The substitution of instances by representatives reduces the set size and speeds up the search at the expense of accuracy. The “tighter” the clusters are the less accuracy is expected to be lost.

Building “high quality” clusters is, therefore, of paramount importance to the first problem. Applications of clustering are in particular motivated by *the Cluster Hypothesis* which states that “closely associated instances tend to be related to the same requests.”

Sets of instances are often changing with time (new instances may be added to the existing set and old instances may be discarded). It is, therefore, of interest to address the clustering problem under the assumption  $\mathbf{R} = \mathbf{R}(t)$  (i.e., the set of instances  $\mathbf{R}$  is time-dependent) [Kogan, 2007].

Natural steps to approach the two above-mentioned problems are:

*Step 1.* Embed the instances and the query into a metric space.

*Step 2.* Handle problems 1 and 2 above as problems concerning points in the metric space.

For instance, a vector space model may map instances into vectors in a finite dimensional Euclidean space, i.e., let the vector space is of dimension  $n = 17$ , and we will be building vectors in  $\mathbf{R}^{17}$ .

One can expect sparse high dimensional vectors (this is indeed the case in many applications) [Kogan, 2007].

---

## Input Data

---

One possible approach to handle the sparse high dimensional vectors is the Multi-layer Growing Pyramidal Networks (MPGN) realized in system INFOS and presented in [Mitov, 2011]. In this work we use this approach for multi-variant clustering high dimensional data. We will illustrate this by an implementation of MPGN for discovering regularities in data received by National Scientific Center “Institute of mechanization and electrification of agriculture” of Ukrainian Academy of Agriculture Sciences. The observations had collected high dimensional data about wheat crop, including data about fertilizing, weather, water reserves in the top layer of earth, temperature, wind, etc.

In our example we will use a small part of this data to illustrate the idea. In further work it may be extended to whole number of features. The extracted data set from main data collection contains data from 252 real observations of the fertilizing and the corresponded crop of the wheat provided in black earth regions Ukraine, which are rich of humus. Three kinds of fertilizers were chosen: nitric (N), phosphorus (P) and potassium (K) and four selected varieties of wheat – Caucasus, Mironov Jubilee, Mironov 808 and Kharkov 81 (Table 1).

Table 1. Observations of the fertilizing and the corresponded crop of the wheat

variants				Caucasus	Mironov Jubilee						Mironov 808		Kharkov 81						
n	N	P	K	1972	1971	1974	1975	1976	1977	1973	1978	1979	1980	1981	1982	1983	1984	1985	
1	0	0	0	24.6	35.0	31.5	24.9	48.0	27.8	24.6	28.8	23.3	33.4	25.1	15.2	21.6	7.1	25.3	
2	0.6	0.6	0	29.2	40.6	42.1	24.5	58.8	34.8	42.8	42.7	31.9	33.7	40.2	29.4	39.9	10.0	32.6	
3	1.2	1.2	0	-	-	-	-	58.0	36.6	-	50.6	31.2	32.7	47.1	38.5	41.5	11.9	49.0	
4	0	0.6	0.3	24.0	40.2	37.0	24.4	46.7	32.3	38.0	26.9	25.2	38.1	30.4	16.2	22.3	7.2	25.6	
5	0.6	0.6	0.3	26.5	43.8	32.2	29.5	57.7	32.9	42.6	42.2	32.4	35.5	42.3	29.9	36.7	9.9	31.0	
6	0.9	0.6	0.3	26.5	44.2	45.7	31.4	61.3	33.1	41.6	50.6	32.8	35.7	47.4	32.9	39.8	10.2	36.6	
7	1.2	0.6	0.3	26.5	40.4	44.2	30.3	57.9	34.9	40.6	50.6	33.1	34.9	46.8	36.4	43.3	12.4	42.6	
8	1.5	0.6	0.3	-	-	-	-	53.0	35.4	-	49.5	32.1	32.7	46.5	41.6	43.7	9.6	41.9	
9	0.6	1.2	0.3	29.2	46.2	42.8	28.3	58.6	38.0	43.2	44.5	31.8	37.1	39.4	28.5	35.7	10.9	33.2	
10	0.6	0.9	0.3	25.8	42.7	41.9	30.3	60.1	35.3	41.7	44.0	30.3	35.9	40.9	28.4	36.0	14.3	34.4	
11	0.6	0	0.3	25.8	32.6	34.4	26.5	46.1	32.1	40.6	40.8	29.7	35.5	36.5	20.5	30.7	8.1	26.4	
12	0.6	0.6	0.6	28.8	42.7	43.4	32.4	54.4	32.9	43.6	43.3	30.6	38.0	37.7	31.1	37.0	9.8	33.4	
13	0.9	0.9	0.6	-	-	-	-	56.0	40.5	-	49.4	34.1	34.7	46.7	36.1	40.1	12.6	38.0	
14	0.9	0.6	0.6	-	-	-	-	59.6	35.5	-	47.9	34.3	37.0	45.0	33.2	38.6	13.0	35.0	
15	1.2	1.2	0.6	28.8	-	48.1	27.6	56.6	40.2	43.3	48.3	33.1	32.2	50.5	39.6	44.0	13.7	41.2	
16	1.2	0	0.6	24.9	-	33.3	25.2	54.3	31.0	38.7	51.3	31.3	35.2	43.2	28.7	39.6	8.2	31.3	
17	0	1.2	0.6	28.0	-	38.3	35.3	44.5	32.2	41.3	27.0	25.0	39.7	28.0	16.1	23.2	7.6	26.2	
18	0.6	0.6	0.9	-	-	-	-	53.6	33.2	-	43.9	32.6	37.4	42.9	27.5	34.0	10.5	32.2	
19	1.2	1.2	0.9	-	-	-	-	60.4	36.8	-	51.4	34.7	34.3	49.8	36.7	42.6	13.0	43.8	

Usually, the research is concerned on the data of every variety separately without relationships with others. Here we will try to analyze all varieties in one data set.

Because of great distribution of the values of the crop for different varieties (shown in Figure 1) the normalization of data was provided. The distribution after normalization is shown on Figure 2. After normalization the values of the crop are in the interval [17.58, 49.41] (before it, the interval was [7.10, 61.30]).

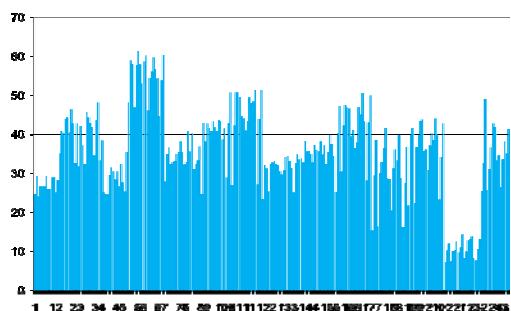


Figure 1. Values of the crop of the different varieties of the wheat before normalization – the vertical interval is [7.10, 61.30]

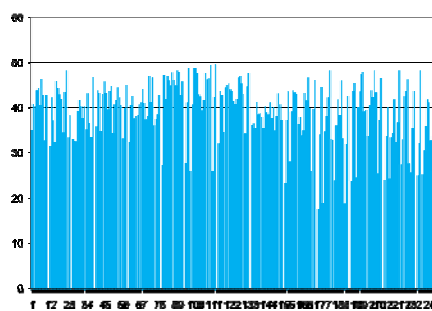


Figure 2. Values of the crop of the different varieties of the wheat after normalization – the vertical interval is [17.58, 49.41]

### Multi-variant clustering

We cluster the data using different kinds of distances between the values of the normalized crop. We provide four different types of clustering:

- Case A. One cluster – no distances are used. All instances are assumed to be in this cluster;
- Case B. Four clusters based on discretization based on human given intervals. The boundaries are respectively: 35, 40 and 45;
- Case C. Five clusters based on discretization realized in system PaGaNe [Mitov et al, 2009a] and especially – the Chi-merge discretization of the normalized crop values [Mitov et al, 2009b];
- Case D. Two clusters based on merged clusters from case C: (1+2+3) and (4+5)

The corresponded boundaries of the intervals are presented in Table 2.

Table 2. Boundaries of the intervals for different cases of discretization of the values of the normalized crop of the wheat

Class		Crop normalized	
		min	max
<b>A. One cluster</b>			
1		17.58	49.41
<b>B. Four clusters based on discretization based on human given intervals</b>			
1		17.58	34.99
2		35.00	39.99
3		40.00	44.99
4		45.00	49.41
<b>C. Five clusters based on Chi-merge discretization of the crop values</b>			
1		17.58	23.88
2		24.22	28.07
3		30.43	36.66
4		37.00	43.00
5		43.09	49.41
<b>D. Two clusters based on merging clusters from case C.: (1+2+3) and (4+5)</b>			
1		17.58	36.66
2		37.00	49.41

The results in Case A –one cluster, are not informative (Table 3). At the top of pyramids we receive practically all values used in the experiments. No conclusion may be made.

Table 3. Case A. One cluster – no distances are used. All instances are assumed to be in this cluster

N	P	K	variety
			Caucasus
			Kharkov 81
			Mironov 808
			Mironov jubilee
0N			
0.6N			
0.9N			
1.2N			
	0P		
	0.6P		
	0.9P		
	1.2P		
		0K	
		0.3K	
		0.6K	
		0.9K	

The Case B corresponds to the human common sense for clustering the data (5 points per interval). The intervals are chosen on the base of understanding that the interesting data are in the top intervals, which were chosen to be equal. The low intervals were merged in one big interval. This way four intervals were created: (17.58, 34.99), (35.00, 39.99), (40.00, 44.99) and (45.00, 49.41).

This case is more informative (see Table 4). The main conclusion from this case is that the variety “Mironov 808” gives most good crop if the fertilizing is in any of the combinations in class 4. “Mironov jubilee” and “Caucasus” as a rule have middle values of crop. The worst values belong to “Kharkov 81”.

Table 4. Case B. Four clusters based on discretization based on human given intervals. The boundaries are respectively: 35, 40 and 45

Class	N	P	K	variety
1	0.6N	0.9P	0.3K	Kharkov 81
1	0.6N	0.6P	0.9K	Kharkov 81
1	0.9N	0.6P	0.3K	Kharkov 81
1	0.9N	0.6P	0.6K	Kharkov 81
1	1.2N	0P	0.6K	Kharkov 81
1	1.2N	1.2P	0.6K	Kharkov 81
1	0N	0P	0K	Mironov 808
1	0N	0.6P	0.3K	Mironov 808
1	1.2N	0P	0.6K	Mironov jubilee

Class	N	P	K	variety
3	0N	1.2P	0.6K	Caucasus
3	0.6N	0.6P	0K	Caucasus
3	0.6N	0.6P	0.6K	Caucasus
3	0.6N	1.2P	0.3K	Caucasus
3	1.2N	1.2P	0.6K	Caucasus
3	0N	0.6P	0.3K	Mironov 808
3	0.6N	0.6P	0.9K	Mironov 808
3	0.6N	0.9P	0.3K	Mironov jubilee
3	0.9N	0.6P	0.6K	Mironov jubilee
3	1.2N	0.6P	0.3K	Mironov jubilee
3	1.2N	1.2P	0.9K	Mironov jubilee
3	1.2N	1.2P	0K	Mironov jubilee

Class	N	P	K	variety
2	0N	0P	0K	Caucasus
2	0N	0.6P	0.3K	Caucasus
2	0.6N	0P	0.3K	Caucasus
2	0.6N	0.6P	0.3K	Caucasus
2	0.6N	0.9P	0.3K	Caucasus
2	0.9N	0.6P	0.3K	Caucasus
2	1.2N	0.6P	0.3K	Caucasus
2	1.2N	0P	0.6K	Caucasus
2	0.6N	0.6P	0.9K	Mironov jubilee

Class	N	P	K	variety
4	0.9N	0.6P	0.3K	Mironov 808
4	0.9N	0.6P	0.6K	Mironov 808
4	0.9N	0.9P	0.6K	Mironov 808
4	1.2N	1.2P	0.6K	Mironov 808
4	1.2N	1.2P	0K	Mironov 808
4	1.2N	1.2P	0.9K	Mironov 808
4	1.5N	0.6P	0.3K	Mironov 808

In the same time, after the pruning, no generalized patterns exist and, maybe, some important regularity is not discovered. Because of this we continue the experiment with two other cases.

The Case C is based on discretization realized in the system PaGaNe [Mitov et al, 2009a] and especially – the Chi-merge discretization of the normalized crop values. In general, pyramidal classifier trained on data preprocessed by Chi-merge achieves lower classification error than those trained on data preprocessed by the other discretization methods. The main reason for this is that using Chi-square statistical measure as criterion for class dependency in adjacent intervals of a feature leads to forming good separating which is convenient for the pyramidal algorithms [Mitov et al, 2009b].

The crop values presented in Table 1 were discretized in five intervals based on the Chi-square statistical measure, respectively (17.58, 23.88), (24.22, 28.07), (30.43, 36.66), (37.00, 43.00), (43.09, 49.41).

In Table 5 the results of clustering in the Case C are presented.

Table 5. Results from Case C of clustering

Class	N	P	K	variety	Crop
1	0	0	0	Kharkov81 1982	17.58
1	0	1.2	0.6	Kharkov81 1982	18.62
1	0	0.6	0.3	Kharkov81 1982	18.73
1	0	0	0	Kharkov81 1981	23.18

4	0.6	0.6	0	Mironov jubilee 1977	40.36
4	1.2	0.6	0.3	Mironov jubilee 1971	40.40
4	1.2	0	0.6	Mironov jubilee 1976	40.41
4	1.2	1.2	0.6	Kharkov 81 1985	40.44
4	0.6	0.6	0.6	Kharkov 81 1983	40.45



1	0	0	0	Kharkov81 1983	23.61
1	0.6	0	0.3	Kharkov81 1982	23.70
1	0	0	0	Kharkov81 1984	23.88

Class	N	P	K	variety	Crop
2	0	0.6	0.3	Kharkov81 1984	24.22
2	0	0.6	0.3	Kharkov81 1983	24.38
2	0	0	0	Kharkov81 1985	24.84
2	0	0.6	0.3	Kharkov81 1985	25.13
2	0	1.2	0.6	Kharkov81 1983	25.36
2	0	1.2	0.6	Kharkov81 1984	25.56
2	0	1.2	0.6	Kharkov81 1985	25.72
2	0	1.2	0.6	Kharkov81 1981	25.85
2	0	0.6	0.3	Mironov 808 1978	25.86
2	0.6	0	0.3	Kharkov81 1985	25.92
2	0	1.2	0.6	Mironov 808 1978	25.95
2	0	0	0	Mironov 808 1973	27.14
2	0.6	0	0.3	Kharkov81 1984	27.25
2	1.2	0	0.6	Kharkov81 1984	27.58
2	0	0	0	Mironov 808 1978	27.69
2	0	0.6	0.3	Kharkov81 1981	28.07

Class	N	P	K	variety	Crop
3	0.6	0.6	0.3	Kharkov81 1985	30.43
3	1.2	0	0.6	Kharkov81 1985	30.73
3	0	0	0	Mironov jubilee 1974	31.50
3	0.6	0.6	0.9	Kharkov81 1985	31.61
3	0.6	0.6	0.9	Kharkov81 1982	31.80
3	0	0	0	Kharkov81 1979	31.89
3	0.6	0.6	0	Kharkov81 1985	32.00
3	0.6	0.6	0.3	Mironov jubilee 1974	32.20
3	0	0	0	Mironov jubilee 1977	32.24
3	1.5	0.6	0.3	Kharkov81 1984	32.29
3	0	0.6	0.3	Mironov jubilee 1975	32.35
3	0.6	0.6	0	Mironov jubilee 1975	32.48
3	0.6	1.2	0.3	Kharkov81 1985	32.59
3	0.6	0	0.3	Mironov jubilee 1971	32.60
3	0.6	0.6	0.6	Kharkov81 1985	32.79
3	0.6	0.9	0.3	Kharkov81 1982	32.84
3	0.6	1.2	0.3	Kharkov81 1982	32.95
3	0.6	0.6	0.6	Kharkov81 1984	32.96
3	0	0	0	Mironov jubilee 1975	33.01
3	0	1.2	0.6	Mironov jubilee 1976	33.12
3	1.2	0	0.6	Kharkov81 1982	33.18
3	1.2	0	0.6	Mironov jubilee 1974	33.30
3	0.6	0.6	0.3	Kharkov81 1984	33.30
3	1.2	0	0.6	Mironov jubilee 1975	33.41
3	0.6	0	0.3	Kharkov81 1983	33.56
3	0.6	0.6	0	Kharkov81 1984	33.64
3	0.6	0	0.3	Kharkov81 1981	33.70
3	0.6	0.9	0.3	Kharkov81 1985	33.77
3	0.6	0.6	0	Kharkov81 1982	33.99
3	0	1.2	0.6	Kharkov81 1979	34.22
3	0.9	0.6	0.3	Kharkov81 1984	34.31
3	0.6	0	0.3	Mironov jubilee 1976	34.31
3	0.9	0.6	0.6	Kharkov81 1985	34.36
3	0.6	0	0.3	Mironov jubilee 1974	34.40
3	0	0.6	0.3	Kharkov81 1979	34.49
3	0.6	0.6	0.3	Kharkov81 1982	34.57
3	0	0.6	0.3	Mironov jubilee 1976	34.76
3	0.6	0.6	0.6	Kharkov81 1981	34.81
3	1.2	1.2	0.6	Kharkov81 1980	34.86
3	0	0	0	Mironov jubilee 1971	35.00
3	0.6	0	0.3	Mironov jubilee 1975	35.13
3	0	0.6	0.3	Caucasus 1972	35.29
3	0.6	0.6	0.9	Kharkov81 1984	35.32

4	1.2	0.6	0.3	Mironov jubilee 1977	40.47
4	0.6	0.6	0.9	Kharkov 81 1980	40.49
4	0.6	0.6	0.6	Mironov jubilee 1976	40.49
4	0.6	0.6	0.3	Mironov 808 1978	40.57
4	0.6	0.6	0	Mironov jubilee 1971	40.60
4	0.6	0	0.3	Kharkov 81 1979	40.65
4	0.6	0.9	0.3	Mironov jubilee 1977	40.94
4	0.6	0.6	0	Mironov 808 1978	41.05
4	1.5	0.6	0.3	Mironov jubilee 1977	41.05
4	1.5	0.6	0.3	Kharkov 81 1985	41.13
4	0.6	0.6	0.6	Kharkov 81 1980	41.13
4	0	1.2	0.6	Caucasus 1972	41.17
4	0.9	0.6	0.6	Mironov jubilee 1977	41.17
4	0	0.6	0.3	Kharkov 81 1980	41.24
4	0.6	0.9	0.3	Kharkov 81 1979	41.48
4	0.9	0.6	0.6	Kharkov 81 1981	41.55
4	0.6	0.6	0.6	Mironov 808 1978	41.62
4	0.9	0.6	0.3	Mironov jubilee 1975	41.63
4	0.9	0.9	0.6	Mironov jubilee 1976	41.68
4	1.2	0.6	0.3	Kharkov 81 1984	41.71
4	0.9	0.9	0.6	Kharkov 81 1982	41.74
4	1.2	0.6	0.3	Kharkov 81 1985	41.82
4	0.6	0.6	0.6	Kharkov 81 1979	41.89
4	0.6	0.9	0.3	Mironov jubilee 1974	41.90
4	0	0.6	0.3	Mironov 808 1973	41.92
4	1.2	0.6	0.3	Kharkov 81 1982	42.09
4	0.6	0.6	0	Mironov jubilee 1974	42.10
4	1.2	1.2	0.6	Mironov jubilee 1976	42.13
4	0.9	0.6	0.6	Kharkov 81 1983	42.20
4	0.6	0.6	0.9	Mironov 808 1978	42.20
4	0.6	0.9	0.3	Mironov 808 1978	42.30
4	0.6	0.6	0.6	Caucasus 1972	42.34
4	1.2	1.2	0.6	Caucasus 1972	42.34
4	0.9	0.9	0.6	Kharkov 81 1984	42.38
4	1.2	1.2	0.9	Kharkov 81 1982	42.43
4	1.2	1.2	0	Mironov jubilee 1977	42.45
4	1.2	1.2	0.9	Mironov jubilee 1977	42.68
4	1.2	0	0.6	Mironov 808 1973	42.69
4	0.6	0.9	0.3	Mironov jubilee 1971	42.70
4	0.6	0.6	0.6	Mironov jubilee 1971	42.70
4	1.2	1.2	0	Kharkov 81 1979	42.71
4	0.6	1.2	0.3	Mironov 808 1978	42.78
4	0.6	1.2	0.3	Mironov jubilee 1974	42.80
4	1.2	0	0.6	Kharkov 81 1979	42.84
4	0.6	0.6	0	Caucasus 1972	42.93
4	0.6	1.2	0.3	Caucasus 1972	42.93
4	1.5	0.6	0.3	Kharkov 81 1981	42.94
4	0.6	0.6	0.3	Mironov jubilee 1976	42.95
4	0.6	0.6	0.6	Mironov jubilee 1975	42.96
4	0	1.2	0.6	Kharkov 81 1980	42.97
4	1.2	1.2	0.9	Kharkov 81 1985	43.00

Class	N	P	K	variety	Crop
5	1.2	0.6	0.3	Mironov jubilee 1976	43.09
5	0.9	0.9	0.6	Kharkov 81 1981	43.12
5	1.2	1.2	0	Mironov jubilee 1976	43.17
5	1.2	0.6	0.3	Kharkov 81 1981	43.21
5	1.2	0	0.6	Kharkov 81 1983	43.29
5	0.6	0.6	0.6	Mironov jubilee 1974	43.40
5	1.2	1.2	0	Kharkov 81 1981	43.49
5	0.9	0.6	0.3	Kharkov 81 1983	43.51
5	0.6	1.2	0.3	Kharkov 81 1979	43.53
5	0.6	1.2	0.3	Mironov jubilee 1976	43.62
5	0.6	0.6	0	Kharkov 81 1983	43.62
5	0.6	0.6	0	Kharkov 81 1979	43.67
5	0.9	0.6	0.6	Kharkov 81 1984	43.73

3	1.2	1.2	0	Kharkov81 1980	35.40
3	1.5	0.6	0.3	Kharkov81 1980	35.40
3	0	0	0	Mironov jubilee 1976	35.73
3	0.9	0.6	0.3	Kharkov81 1985	35.93
3	1.2	0	0.6	Mironov jubilee 1977	35.95
3	0.6	0.6	0.6	Kharkov81 1982	35.96
3	0	0	0	Kharkov81 1980	36.16
3	0	0	0	Caucasus 1972	36.17
3	0.6	1.2	0.3	Kharkov 81 1981	36.38
3	0.6	0.6	0	Kharkov 81 1980	36.48
3	1.2	1.2	0.6	Mironov jubilee 1975	36.59
3	1.2	0	0.6	Caucasus 1972	36.61
3	0.6	1.2	0.3	Kharkov 81 1984	36.66

Class	N	P	K	variety	Crop
4	0	0.6	0.3	Mironov jubilee 1974	37.00
4	0.6	0.6	0	Kharkov 81 1981	37.12
4	1.2	1.2	0.9	Kharkov 81 1980	37.13
4	0.6	0.6	0.9	Kharkov 81 1983	37.17
4	0.6	0	0.3	Mironov jubilee 1977	37.23
4	0.9	0.9	0.6	Kharkov 81 1985	37.30
4	0	1.2	0.6	Mironov jubilee 1977	37.34
4	0	0.6	0.3	Mironov jubilee 1977	37.46
4	0.6	1.2	0.3	Mironov jubilee 1975	37.52
4	0.9	0.9	0.6	Kharkov 81 1980	37.56
4	0.6	0.9	0.3	Kharkov 81 1981	37.77
4	1.2	0.6	0.3	Kharkov 81 1980	37.78
4	0.6	0.9	0.3	Caucasus 1972	37.93
4	0.6	0	0.3	Caucasus 1972	37.93
4	0.9	0.6	0.3	Kharkov 81 1982	38.04
4	1.2	0	0.6	Kharkov 81 1980	38.10
4	0.6	0.6	0.3	Mironov jubilee 1977	38.16
4	0.6	0.6	0.6	Mironov jubilee 1977	38.16
4	0	1.2	0.6	Mironov jubilee 1974	38.30
4	0.9	0.6	0.3	Mironov jubilee 1977	38.39
4	0.9	0.6	0.6	Kharkov 81 1982	38.39
4	0.6	0.6	0.3	Kharkov 81 1980	38.43
4	0.6	0	0.3	Kharkov 81 1980	38.43
4	0.6	0.6	0.9	Mironov jubilee 1977	38.50
4	0.9	0.6	0.3	Kharkov 81 1980	38.65
4	0.6	0.9	0.3	Kharkov 81 1980	38.86
4	0.6	0.6	0.3	Caucasus 1972	38.96
4	0.9	0.6	0.3	Caucasus 1972	38.96
4	1.2	0.6	0.3	Caucasus 1972	38.96
4	0.6	1.2	0.3	Kharkov 81 1983	39.03
4	0.6	0.6	0.3	Kharkov 81 1981	39.06
4	0.6	0.6	0.3	Mironov jubilee 1975	39.11
4	0.6	0	0.3	Mironov 808 1978	39.22
4	0.6	0.9	0.3	Kharkov 81 1983	39.35
4	1.5	0.6	0.3	Mironov jubilee 1976	39.45
4	0.6	0.6	0.9	Kharkov 81 1981	39.61
4	1.2	0	0.6	Kharkov 81 1981	39.89
4	0.6	0.6	0.9	Mironov jubilee 1976	39.89
4	1.2	1.2	0	Kharkov 81 1984	40.03
4	0.9	0.6	0.6	Kharkov 81 1980	40.05
4	0.6	0.6	0.3	Kharkov 81 1983	40.12
4	0.6	1.2	0.3	Kharkov 81 1980	40.16
4	1.2	0.6	0.3	Mironov jubilee 1975	40.17
4	0.6	0.9	0.3	Mironov jubilee 1975	40.17
4	0	0.6	0.3	Mironov jubilee 1971	40.20

5	1.2	1.2	0.9	Kharkov 81 1984	43.73
5	0.6	0.6	0	Mironov jubilee 1976	43.76
5	0.9	0.6	0.3	Kharkov 81 1981	43.77
5	0.6	0.6	0.3	Mironov jubilee 1971	43.80
5	0.9	0.9	0.6	Kharkov 81 1983	43.84
5	1.5	0.6	0.3	Kharkov 81 1979	43.94
5	0.6	1.2	0.3	Mironov jubilee 1977	44.07
5	0.9	0.6	0.3	Mironov jubilee 1971	44.20
5	1.2	0.6	0.3	Mironov jubilee 1974	44.20
5	0.6	0.6	0.3	Kharkov 81 1979	44.35
5	0.9	0.6	0.6	Mironov jubilee 1976	44.36
5	1.2	1.2	0	Kharkov 81 1982	44.52
5	0.6	0.6	0.9	Kharkov 81 1979	44.62
5	0.6	0.9	0.3	Mironov jubilee 1976	44.73
5	1.2	0.6	0.3	Mironov 808 1973	44.79
5	0.6	0	0.3	Mironov 808 1973	44.79
5	0.9	0.6	0.3	Kharkov 81 1979	44.90
5	1.2	1.2	0.9	Mironov jubilee 1976	44.96
5	1.2	0.6	0.3	Kharkov 81 1979	45.31
5	1.2	1.2	0.6	Kharkov 81 1979	45.31
5	1.2	1.2	0	Kharkov 81 1983	45.37
5	0	1.2	0.6	Mironov 808 1973	45.56
5	0.9	0.6	0.3	Mironov jubilee 1976	45.62
5	0.9	0.6	0.3	Mironov jubilee 1974	45.70
5	1.2	1.2	0.6	Kharkov 81 1982	45.79
5	0.9	0.6	0.3	Mironov 808 1973	45.89
5	1.2	1.2	0.9	Kharkov 81 1981	45.98
5	0.6	0.9	0.3	Mironov 808 1973	46.00
5	0.9	0.6	0.6	Mironov 808 1978	46.05
5	1.2	1.2	0.6	Kharkov 81 1984	46.08
5	0.6	1.2	0.3	Mironov jubilee 1971	46.20
5	1.2	1.2	0.6	Mironov 808 1978	46.43
5	1.2	1.2	0.9	Kharkov 81 1983	46.57
5	1.2	1.2	0.6	Mironov jubilee 1977	46.62
5	1.2	1.2	0.6	Kharkov 81 1981	46.63
5	0.9	0.9	0.6	Kharkov 81 1979	46.68
5	0	1.2	0.6	Mironov jubilee 1975	46.80
5	0.9	0.6	0.6	Kharkov 81 1979	46.95
5	0.9	0.9	0.6	Mironov jubilee 1977	46.97
5	0.6	0.6	0.3	Mironov 808 1973	47.00
5	0.6	0.6	0	Mironov 808 1973	47.22
5	1.2	0.6	0.3	Kharkov 81 1983	47.33
5	0.9	0.9	0.6	Mironov 808 1978	47.49
5	1.2	1.2	0.9	Kharkov 81 1979	47.50
5	1.5	0.6	0.3	Mironov 808 1978	47.58
5	0.6	1.2	0.3	Mironov 808 1973	47.66
5	1.2	1.2	0.6	Mironov 808 1973	47.77
5	1.5	0.6	0.3	Kharkov 81 1983	47.77
5	1.2	1.2	0.6	Mironov jubilee 1974	48.10
5	0.6	0.6	0.6	Mironov 808 1973	48.10
5	1.5	0.6	0.3	Kharkov 81 1982	48.10
5	1.2	1.2	0.6	Kharkov 81 1983	48.10
5	0.6	0.9	0.3	Kharkov 81 1984	48.10
5	1.2	1.2	0	Kharkov 81 1985	48.10
5	1.2	1.2	0	Mironov 808 1978	48.64
5	0.9	0.6	0.3	Mironov 808 1978	48.64
5	1.2	0.6	0.3	Mironov 808 1978	48.64
5	1.2	0	0.6	Mironov 808 1978	49.31
5	1.2	1.2	0.9	Mironov 808 1978	49.41

The results given in Table 5 show that the clustering is not enough to discover regularities in the data. The additional processing of clusters is needed. Using clusters as classes in MPGN, we have built corresponded pyramids for every case, and have made pruning for the cases B, C, and D. This way, in the corresponded cases we received a number of generalized patterns, which are not contradictory between classes.

Such discretization seems to be more informative but the received results are similar to Case B (Table 6). In the same time, the instances of the class 1 are contradictory to instances of class 2; and two instances from class 2 are contradictory to instances of class 4 and class 5. Because of this we have to remove them from the resulting Table 6; i.e. to make pruning of the instances by removing the contradictory ones. In Table 6, the contradictory instances are given in italic. After the final pruning there are no instances in class 1 (Table 7).

Table 6. Case C. Five clusters based on the Chi-merge discretization before the final pruning

Class	N	P	K	variety
1	<i>0N</i>	<i>0P</i>	<i>0K</i>	<i>Kharkov 81</i>
1	<i>0N</i>	<i>0.6P</i>	<i>0.3K</i>	<i>Kharkov 81</i>
1	<i>0N</i>	<i>1.2P</i>	<i>0.6K</i>	<i>Kharkov 81</i>
1	<i>0.6N</i>	<i>0P</i>	<i>0.3K</i>	<i>Kharkov 81</i>

Class	N	P	K	variety
2	<i>0N</i>	<i>0P</i>	<i>0K</i>	<i>Kharkov 81</i>
2	<i>0N</i>	<i>0.6P</i>	<i>0.3K</i>	<i>Kharkov 81</i>
2	<i>0N</i>	<i>1.2P</i>	<i>0.6K</i>	<i>Kharkov 81</i>
2	<i>0.6N</i>	<i>0P</i>	<i>0.3K</i>	<i>Kharkov 81</i>
2	<i>1.2N</i>	<i>0P</i>	<i>0.6K</i>	<i>Kharkov 81</i>
2	<i>0N</i>	<i>0P</i>	<i>0K</i>	<i>Mironov 808</i>
2	<i>0N</i>	<i>0.6P</i>	<i>0.3K</i>	<i>Mironov 808</i>
2	<i>0N</i>	<i>1.2P</i>	<i>0.6K</i>	<i>Mironov 808</i>

Class	N	P	K	variety
4	0N	1.2P	0.6K	Caucasus
4	0.6N	0P	0.3K	Caucasus
4	0.6N	0.6P	0K	Caucasus
4	0.6N	0.6P	0.3K	Caucasus
4	0.6N	0.6P	0.6K	Caucasus
4	0.6N	0.9P	0.3K	Caucasus
4	0.6N	1.2P	0.3K	Caucasus
4	0.9N	0.6P	0.3K	Caucasus
4	1.2N	0.6P	0.3K	Caucasus
4	1.2N	1.2P	0.6K	Caucasus
4	<i>0N</i>	<i>0.6P</i>	<i>0.3K</i>	<i>Mironov 808</i>
4	0.6N	0.6P	0.9K	Mironov 808
4	0.6N	0.6P	0.9K	Mironov jubilee
4	1.5N	0.6P	0.3K	Mironov jubilee

Class	N	P	K	variety
3	0N	0P	0K	Kharkov 81
3	0.6N	0.6P	0.9K	Kharkov 81
3	0.6N	0.9P	0.3K	Kharkov 81
3	0.9N	0.6P	0.3K	Kharkov 81
3	0.9N	0.6P	0.6K	Kharkov 81
3	0N	0P	0K	Caucasus
3	0N	0.6P	0.3K	Caucasus
3	1.2N	0P	0.6K	Caucasus
3	0N	0P	0K	Mironov jubilee

Class	N	P	K	variety
5	<i>0N</i>	<i>1.2P</i>	<i>0.6K</i>	<i>Mironov 808</i>
5	0.9N	0.6P	0.3K	Mironov 808
5	0.9N	0.6P	0.6K	Mironov 808
5	0.9N	0.9P	0.6K	Mironov 808
5	1.2N	0.6P	0.3K	Mironov 808
5	1.2N	1.2P	0K	Mironov 808
5	1.2N	1.2P	0.6K	Mironov 808
5	1.2N	1.2P	0.9K	Mironov 808
5	1.5N	0.6P	0.3K	Mironov 808

Table 7. Case C. Five clusters based on the Chi-merge discretization after the final pruning

Class	N	P	K	variety
1	-	-	-	-

Class	N	P	K	variety
2	1.2N	0P	0.6K	Kharkov 81
2	0N	0P	0K	Mironov 808

Class	N	P	K	variety
3	0.6N	0.6P	0.9K	Kharkov 81
3	0.6N	0.9P	0.3K	Kharkov 81
3	0.9N	0.6P	0.3K	Kharkov 81
3	0.9N	0.6P	0.6K	Kharkov 81
3	0N	0P	0K	Caucasus
3	0N	0.6P	0.3K	Caucasus
3	1.2N	0P	0.6K	Caucasus
3	0N	0P	0K	Mironov jubilee

Class	N	P	K	variety
4	0N	1.2P	0.6K	Caucasus
4	0.6N	0P	0.3K	Caucasus
4	0.6N	0.6P	0K	Caucasus
4	0.6N	0.6P	0.3K	Caucasus
4	0.6N	0.6P	0.6K	Caucasus
4	0.6N	0.9P	0.3K	Caucasus
4	0.6N	1.2P	0.3K	Caucasus
4	0.9N	0.6P	0.3K	Caucasus
4	1.2N	0.6P	0.3K	Caucasus

Class	N	P	K	variety
5	0.9N	0.6P	0.3K	Mironov 808
5	0.9N	0.6P	0.6K	Mironov 808
5	0.9N	0.9P	0.6K	Mironov 808
5	1.2N	0.6P	0.3K	Mironov 808
5	1.2N	1.2P	0K	Mironov 808
5	1.2N	1.2P	0.6K	Mironov 808
5	1.2N	1.2P	0.9K	Mironov 808
5	1.5N	0.6P	0.3K	Mironov 808

4	1.2N	1.2P	0.6K	Caucasus
4	0.6N	0.6P	0.9K	Mironov 808
4	0.6N	0.6P	0.9K	Mironov jubilee
4	1.5N	0.6P	0.3K	Mironov jubilee

For the Case D we create two clusters based on merging clusters from case C: classes (1+2+3) and classes (4+5) from Table 6. This is again based on “the human common sense”. The idea is that the last two classes (4+5) may contain the most of interesting for us regularities. Table 8 presents the result, which was received after the five steps of processing:

- normalization of data of the crop;
- discretization by the PaGaNe discretizer (Chi-merge)
- merging received intervals into two main (1+2+3) and (4+5)
- generalization into two classes separately
- pruning of the contradictory vertexes and instances between classes.

Table 8. Case D. Two clusters based on merged clusters from case C.:  
classes (1+2+3) and (4+5) from Table 6

Class 1				Class 2			
N	P	K	variety	N	P	K	variety
0N	0P	0K	Caucasus	0.6N	0.6P	0.3K	Caucasus
0N	0.6P	0.3K	Caucasus	0.6N	0.6P	0.6K	Caucasus
1.2N	0P	0.6K	Caucasus	0.6N	0.6P	0K	Caucasus
0N	0P	0K	Kharkov 81	0.6N	0P	0.3K	Caucasus
0.6N	0.6P	0.9K	Kharkov 81	0.6N	1.2P	0.3K	Caucasus
0.6N	0.9P	0.3K	Kharkov 81	0N	1.2P	0.6K	Caucasus
0.9N	0.6P	0.3K	Kharkov 81	1.2N	0.6P	0.3K	Caucasus
0.9N	0.6P	0.6K	Kharkov 81	1.2N	1.2P	0.6K	Caucasus
0N	0P	0K	Mironov 808	1.2N	0.6P	0.3K	Kharkov 81
0N	0.6P	0.3K	Mironov 808	-	-	-	Mironov 808
0N	1.2P	0.6K	Mironov 808	0.6N	0.6P	0.6K	Mironov jubilee
0N	0P	0K	Mironov jubilee	0.6N	1.2P	0.3K	Mironov jubilee
				1.2N	0.6P	0.3K	Mironov jubilee
				1.2N	1.2P	0K	Mironov jubilee
				1.5N	0.6P	0.3K	Mironov jubilee

The main conclusion from this case is that the variety “Mironov 808”, “Mironov jubilee”, and “Caucasus” are good with small exception (3 for the first variety, one for the second, and 3 for the third). The worst values belong to “Kharkov 81”.

Let mention the special instance in class 2 for variety Mironov 808 which contains dashes in all positions. This means that all instances of Mironov 808 in class 2 are not contradictory to ones in class 1. Because of this only one generalized instance is given as result. In the same time in class 1 there exist just three instances which have no contradictory to instances of the class 2 and they are shown in the Table 8.

Again, the information from this case (as well as the previous cases) is not enough to make decision. We need additional information, which may be taken from the previous cases or from the clusterization using another system. Such results will be outlined shortly below.

---

## Experiments with program complex CONFOR

---

Knowledge discovery methods based on pyramidal networks and using the results for decision making firstly were implemented in the program complex CONFOR (Abbreviation of CONcept FORMation) [Gladun, 1987, 1994, 2000; Gladun and Vashchenko, 2000]. The basic functions of program complex CONFOR are:

- discovery of regularities (knowledge) inherent to data;

- using of the retrieved regularities for object classification, diagnostics and prediction.

Original methods of knowledge discovery based on the pyramidal networks are taken as a principle in the CONFOR system. A successful long-term application of the methods in different fields of research and development has confirmed their decisive advantages as compared to other known methods. Chemists have done over a thousand of high-precise prognoses for new chemical compounds and materials [Kiselyova et al, 1998]. CONFOR is used for analysis of information technologies market. Application field for CONFOR is also medicine, economy, ecology, geology, technical diagnostics, sociology, etc.

It is important to compare the results received by system INFOS presented in previous chapters with the results received by the program complex CONFOR. This way we will have independent processing of the same data by the other program system and the new variants of clustering will improve our conclusions.

We provide experiments with the same data as in cases A, B C and D. We have received similar results which in this case were based on logical inference. The most interesting is the case D. The main conclusion from this case is that the varieties “Mironov 808” and “Mironov jubilee” are the best choice. The logical expression of this generated by Confor is as follow:

```
[17] - N_0_6 & variety_Mironov jubilee
      AND
      NOT{K_0_3 & P_0}
      AND
      NOT{P_0 & K_0_3}
      AND
      2 excluded
      OR
[13] N_0_6 & variety_Mironov 808)
```

It means that variety Mironov Jubilee presented by 17 instances and variety Mironov 808 presented by 13 instances are good with small exceptions. The worst values belong to “Kharkov 81” – the logical expression is:

```
[54] variety_Kharkov 81
```

In other words, 54 instances of variety Kharkov 81 were clusterized in the class 1 “worst”.

In details these experiments will be presented in further publication.

---

## **Conclusion**

---

In this paper we have used a small part of data to illustrate a possible clustering approach to handle the sparse high dimensional vectors. The extracted data set from main data collection contained data from 252 real observations of the fertilizing and the corresponded crop of the wheat provided in black earth regions Ukraine, which are rich of humus. Three kinds of fertilizers were chosen: nitric (N), phosphorus (P) and potassium (K) and four varieties of wheat – Caucasus, Mironov Jubilee, Mironov 808 and Kharkov 81.

Our main goal in this work was to illustrate using the approach for multi-variant clustering high dimensional data based on the Multi-layer Growing Pyramidal Networks (MPGN) and the system INFOS. We outlined an implementation of MPGN for discovering regularities in data received by National Scientific Center “Institute of mechanization and electrification of agriculture” of Ukrainian Academy of Agriculture Sciences. The observations had collected high dimensional data about wheat crop, including data about fertilizing, weather, water reserves in the top layer of earth, temperature, wind, etc.

The analysis of the results from different cases permits us to say that the [Heady and Dillon, 1961] advices are still actual (in our example, too). The main theirs advice is not to accept only one equation for characterizing the agricultural production in different conditions.

Taking in account all cases we may draw inference that the variety "Mironov 808" is stable in all observations. "Mironov jubilee" shows less stability but with proper fertilizing gives good crop. "Caucasus" and "Kharkov 81" could not be recommended to be used. Let remember that our example do not take in account many factors which were observed. In further work, data will be extended to whole number of features. The conclusion may differ when we will use great number of dimensions.

Similar results were received by parallel independent experiments with the same data provided by the program complex CONFOR which is based on pyramidal structures, too.

A possible extension of the investigated area is in direction of fuzzy clustering [Hoepfner et al, 1997]. As it is outlined in [Bodyanskiy et al, 2011] the problem of multidimensional data clusterization is an important part of exploratory data analysis [Tukey, 1977], [Höppner et al, 1999], with its goal of retrieval in the analyzed data sets of observations some groups (classes, clusters) that are homogeneous in some sense. Traditionally, the approach to this problem assumes that each observation may belong to only one cluster, although more natural is the situation where the processed vector of features could refer to several classes with different levels of membership (probability, possibility). This situation is the subject of fuzzy cluster analysis [Bezdek, 1981]; [Gath and Geva, 1989]; [Höppner et al, 1999], which is based on the assumption that the classes of homogeneous data are not separated, but overlap, and each observation can be attributed to a certain level of membership to each cluster, which lies in the range of zero to one [Höppner et al, 1999]. Initial information for this task is a sample of observations, formed from  $N$ -dimensional feature  $x(1), x(2), \dots, x(k), \dots, x(N)$ .

The result of clustering is segmentation of the original data set into  $m$  classes with some level of membership  $u_j(k)$  of  $k$ -th feature vector  $x(k)$  to  $j$ -th cluster,  $j=1, 2, \dots, m$ . [Bodyanskiy et al, 2011]

What we have seen from the experiments is that the multi-variant clustering combined with pyramidal generalization and pruning give reliable results. Using algorithms for fuzzy clustering will give new possibilities.

---

## Bibliography

---

- [Bezdek, 1981] Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms, N.Y.:Plenum Press., 1981.
- [Bodyanskiy et al, 2011] Bodyanskiy Y., Kolchygin B., Pliss I. Adaptive Neuro-fuzzy Kohonen Network with Variable Fuzzifier. International Journal "Information Theories and Applications", Vol. 18, Number 3, 2011, pp. 215 – 223
- [Gath and Geva, 1989] Gath I., Geva A.B. Unsupervised optimal fuzzy clustering In: Pattern Analysis and Machine Intelligence., 1989., 2., 7., P. 773-787
- [Gladun and Vashchenko, 2000] Gladun V.P., Vaschenko N.D. Analytical Processes in Pyramidal Networks. Int. Journal Information Theories and Applications, Vol.7, No.3, 2000, pp.103-109.
- [Gladun et al, 2008] Gladun V., Velichko V., Ivaskiv Y. Selfstructured Systems. International Journal Information Theories and Applications. FOI ITHEA, Sofia, Vol.15,N.1, 2008, pp.5-13.
- [Gladun, 1987] Gladun V.P. Planning of Solutions. Kiev, Naukova Dumka, 1987, 168 p, (in Russian).
- [Gladun, 1994] Gladun V.P. Processes of New Knowledge Formation. Sofia, SD Pedagog 6, 1994, 192 p, (in Russian).
- [Gladun, 2000] Gladun V.P. Partnership with Computers.. Man-Computer Task-oriented Systems. Kiev, Port-Royal, 2000, 120 p, (in Russian).
- [Gladun, 2003] Gladun V.P. Intelligent Systems Memory Structuring. Int. Journal Information Theories and Applications, Vol.10, No.1, 2003, pp.10-14.

- [Heady and Dillon, 1961] Heady E.O., Dillon J.L. Agricultural Production Functions. Ames, Iowa : Iowa State University Press, 1961. 667 p.
- [Hoepfner et al, 1997] Hoepfner F., Klawonn F., Kruse R. Fuzzy-Clusteranalysen. –Braunschweig:Vieweg, 1997. – 280S.
- [Höppner et al, 1999] Höppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester: John Willey & Sons., 1999,
- [Kiselyova et al, 1998] Kiselyova N.N., Gladun V.P., Vashchenko N.D., LeClair S.R., Jackson G.G. Prediction of Inorganic Compounds Perspective for Search of New Electrooptical Materials// Perspektivnie Materiali, 1998, N3. pp.28 -32. (in Russian).
- [Kogan, 2007] Jacob Kogan. Introduction to Clustering Large and High-Dimensional Data. Cambridge University Press, UK, 2007. 222 p.
- [Luo et al, 2009] Ping Luo, Hui Xiong, Guoxing Zhan, Junjie Wu, Zhongzhi Shi. Information-Theoretic Distance Measures for Clustering Validation: Generalization and Normalization. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009. Published by the IEEE Computer Society. pp. 1249-1262.
- [Markov, 2004] Markov, K.: Multi-domain information model. Int. J. Information Theories and Applications, 11/4, 2004, pp.303-308.
- [Mitov et al, 2009a] Mitov, I., Ivanova, K., Markov, K., Velychko, V., Vanhoof, K., Stanchev, P.: "PaGaNe" – A classification machine learning system based on the multidimensional numbered information spaces. In World Scientific Proc. Series on Computer Engineering and Information Science, No.2, pp.279-286.
- [Mitov et al, 2009b] Mitov, I., Ivanova, K., Markov, K., Velychko, V., Stanchev, P., Vanhoof, K.: Comparison of discretization methods for preprocessing data for pyramidal growing network classification method. In Int. Book Series Information Science & Computing – Book No: 14. New Trends in Intelligent Technologies, 2009, pp. 31-39.
- [Mitov, 2011] Iliya Mitov. Class Association Rule Mining Using Multi-Dimensional Numbered Information Spaces. PhD thesis. Promoters: K. Vanhoof, Kr. Markov. Hasselt University, 2011
- [Tukey, 1977] Tukey J. W. Exploratory Data Analysis. Reading, MA: Addison-Wesley Publ. Company, Inc., 1977.

---

## Authors' Information

---



**Krassimira Ivanova** – University of National and World Economy, Sofia, Bulgaria  
e-mail: [krasy78@mail.bg](mailto:krasy78@mail.bg)  
Major Fields of Scientific Research: Data Mining



**Vitalii Velychko** – Institute of Cybernetics, NASU, Kiev, Ukraine  
e-mail: [Velychko@rambler.ru](mailto:Velychko@rambler.ru)  
Major Fields of Scientific Research: Data Mining, Natural Language Processing



**Krassimir Markov** – Institute of Mathematics and Informatics at BAS, Sofia, Bulgaria;  
e-mail: [markov@foibg.com](mailto:markov@foibg.com)  
Major Fields of Scientific Research: Multi-dimensional information systems, Data Mining



**Iliya Mitov** - Institute of Mathematics and Informatics, BAS, Sofia, Bulgaria;  
e-mail: [mitov@foibg.com](mailto:mitov@foibg.com)  
Major Fields of Scientific Research: Business informatics, Software technologies, Data Mining, Multi-dimensional information systems

## Natural Language Processing

---

---

### SOCIAL CONTEXT AS MACHINE-PROCESSABLE KNOWLEDGE

**Alexander Trousov, John Judge, Mikhail Alexandrov, Eugene Levner**

**Abstract:** *In this paper, we show how to represent to our formal reasoning and to model social context as knowledge using network models to aggregate heterogeneous information. We show how social context can be efficiently used for well understood tasks in natural language processing (such as context-dependent automated, large scale semantic annotation, term disambiguation, search of similar documents), as well as for novel applications such as social recommender systems which aim to alleviate information overload for social media users by presenting the most attractive and relevant content. We present the algorithms and the architecture of a hybrid recommender system in the activity centric environment Nepomuk-Simple (EU 6th Framework Project NEPOMUK): recommendations are computed on the fly by network flow methods performing in the unified multidimensional network of concepts from the personal information management ontology augmented with concepts extracted from the documents pertaining to the activity in question.*

**Keywords:** *multidimensional networks, graph-based methods, network flow methods, data mining, natural language processing, recommender systems.*

**ACM Classification Keywords:** *H.3.4 [Information Storage and Retrieval]: Systems and Software – information networks; H.3.5 [Information Storage and Retrieval]: Online Information Services – data sharing.*

---

### Introduction

---

We live in an increasingly interconnected world of socio-technological systems, in which technological infrastructures composed of many layers are interoperating within a social context that drives their everyday use and development. Nowadays, most digital content is generated within public systems like Facebook, Delicious, Twitter, blogs and wiki systems, and also enterprise environments such as Microsoft SharePoint, and IBM Lotus Connections. These applications have transformed the Web from a mere document collection into a highly interconnected social space where documents are actively exchanged, filtered, organized, discussed and edited collaboratively.

The emergence of the Social Web opens up unforeseen opportunities for observing social behavior by tracing social interaction on the Web. In these socio-technological systems “everything is deeply intertwined” using the term coined by the pioneer of the information technologies Ted Nelson [Nelson, 1974]: people are connected to other people and to “non-human agents” such as documents, datasets, analytic tools, tags and concepts. These networks become increasingly multidimensional [Contractor, 2008] providing rich context for network mining and understanding the role of particular nodes representing both people and digital content.

In this paper we show how to represent to our formal reasoning and to model social context as knowledge using network models to aggregate heterogeneous information. We show how social context could be efficiently used for well understood tasks of natural language processing (such as context-dependent automated, large scale



semantic annotation, term disambiguation, search of similar documents ([Troussov et al., 2009a], [Judge et al., 2008]), as well as for novel applications such as social recommender systems which aim to alleviate information overload for social media users by presenting the most attractive and relevant content.

The log-files of social services and the links between human and non-human agents can be interpreted as triples: who did what on a social service, what node/entity is connected to another, etc. If aggregated, these triples can be treated as a single multidimensional network, which we will refer to as the social context.

The social context can be considered as knowledge in the same way as the semantic networks which are formed from concepts represented in ontologies. From the point of view of the traditional dichotomy between codification and collaboration approaches to knowledge management, the social context could be considered as bottom-up created social knowledge. As knowledge, the social context is a weaker type of knowledge when contrasted with ontologies and taxonomies in that it lacks proper conceptualisation, the links are usually typed and cannot be readily used for inferencing. Correspondingly, the potential of this knowledge can be fully revealed only by robust methods which are tolerant to errors and incompleteness of knowledge which is endemic in any user created, user centric knowledge system. Therefore instead of relying on the traditional logical methods of working with ontological semantic networks, we rely on graph-based methods which can be interpreted as methods of soft clustering and fuzzy inferencing. Graph-based methods provide clear intuition and elegant mathematics to mine networks. The applications described in this paper are based on the use of spreading activation methods [Troussov et al., 2009], which are more generic diffusion-based methods when compared to link analysis in Google's PageRank [Langville and Meyer, 2006], and in the FolkRank algorithms used for tag recommendation in Folksonomies [Jaschke et al., 2007].

As a processable knowledge for understanding the documents embedded into a socio-technological system, this social context has advantages over traditional ontologies. The social context is up-to-date knowledge about a subject area or community of users of a system which changes rapidly to reflect interests and developments in the area. It is populated with nodes representing the current state of the information and how it relates to processed texts and to the realities of a particular socio-technological system (people, projects, social groups).

The current trend in corpus linguistics is for bigger and bigger corpora in order to draw more general analyses. In order to provide the type of text analysis needed to drive the development of the social web, we need to look beyond the corpora and documents themselves and draw upon the individual context within which the documents exist. Instead considering how documents in the system relate to each other and also entities (people, tasks, ideas...) outside the scope of the traditional corpus but which have relevance when it comes to analysing the data in the text itself. In addition to word-level, paragraph-level and corpus-level text processing, text analytics on the socio-technological level yields a wealth of interesting and useful data and will play increasingly important role in future advances in this area.

We outline the use of spreading activation methods to navigate multidimensional social networks and to rank the nodes representing the actors ([Kinsella et al., 2008], [Troussov et al., 2008a], and [Troussov et al., 2008b]).

Finally we describe the algorithms and the architecture of the hybrid recommender system (partially covered in [Troussov et al., 2008b],[ Nepomuk project, 2008], and [Troussov et al., 2009b]) in the activity centric environment Nepomuk-Simple (EU 6th Framework Project NEPOMUK): recommendations are computed on the fly by graph-based methods performing in the unified multidimensional network of concepts from the personal information management ontology augmented with concepts extracted from the documents pertaining to the activity in question.



recommender systems. We analyse several applications and show that computational methods used in these applications are based on the network flow process, “that focuses on the outcomes for nodes in a network where something is flowing from node to node across the edges” ([Borgatti and Everett, 2006]). Following [Troussov et al., 2009] we interpret this “something” as a relevancy measure; for instance, the initial seed input value which shows nodes of interest in the network. Propagating the relevancy measure through outgoing links allows us to compute the relevancy measure for other network nodes and dynamically rank these nodes according to the relevancy measures. The same paradigm could be used to address the centrality measurements in social network analysis. Centralisation of the network can be achieved when we assume that all the nodes are equally important, and iteratively recompute the relevancy measure based on the connections between nodes. In addition to “global” centralisation, “local” centralisation could be performed if the initial seed values represent the nodes of interest.

The applications constituting previous art are monolithic software applications. In this paper we present a novel computational paradigm which breaks these applications into “atomic” components, where the computational methods for propagation are separated as distinct “atomic” network flow engines. This approach provides a unified view of previous applications. From the software engineering perspective the advantages of such an approach includes easy software maintenance, reuse and optimisation of network flow engines, and the guide for new applications.

It appears that the desiderata list for properties of propagation depends on the network properties and the task of propagation itself. Therefore in 3.1 we introduce the formal description of major types of propagation and their use when embedded in larger applications. We also provide a formal description of the “objects” used in these engines – nodes and fuzzy sets of nodes. In further sections we analyse the previous art to explain and justify our list. For each of the engines we indicate the efficient (near-linear with respect to the network size) implementation, however we do not assume that the described implementations are necessarily the best way to perform the task. Finally, we present the generic architecture of a software system which utilises the social context.

### **3.1. Network Flow Operations over Network Objects**

A formal description of the network flow methods applications requires the introduction of some notation. We assume that the social context is represented by a multidimensional network modeled as a directed graph, which is a pair  $G = (V, E)$  where

$V$  – is the set of vertices  $v_i$

$E$  – is the set of edges  $e_j$  (although in oriented graphs edges are referred to as arcs).

We also introduce the following terms and notations describing the set of “atomic” operations from formal point of view (operands, etc) and the purpose.

*Object* (Network object) – is a node or a (fuzzy) set of nodes on the network (see [Chen, 1996]). Fuzzy sets are characterized by a membership function  $M$  which shows the degree of belonging of an element to the set.

$M$  – the membership function for fuzzy sets which is a non-negative real-valued function.

*Activation* – the membership function when it is not interpreted in the fuzzy sets paradigm. We use the activation (the activation of nodes, or objects) as an abstract relevancy measure.

*Cloud* (cloud object) – is formally the same as the object, however we use the term cloud where we want to emphasize the fact that the membership function is non-negative real-valued function, not Boolean valued. As usual, we assume that a node  $e$  belongs to the fuzzy set  $C$ , or in mathematical notations  $e \in C$  - if  $M(e) > 0$ .

$|\dots|$  - cardinality of sets. In case of clouds we define  $|C| = \sum M(e)$  for all nodes  $e$  such that  $e \in C$ .

*Query* – an object used as a seed for local ranking (defined below)

*Expansion* – is a unary operation which transform a cloud into another cloud: *Expansion*:  $C_1 \rightarrow C_2$ . If  $C_1$  and  $C_2$  are crisp sets, we assume that  $C_1$  is a proper subset of  $C_2$ :  $C_1 \subset C_2$ . In general, we assume that this operation does not change significantly the values of the membership functions on the nodes in  $C_1$ , and that  $|C_1| \leq |C_2|$ .

*Smoothing* – is formally the same as expansion, however the interpretation of this operation can not be done in the framework of fuzzy sets, instead, it roots in the operations with functions in calculus. We assume that smoothing makes the difference between the values of the function  $M()$  on neighbour nodes smaller.

*Local ranking* - is formally the same as expansion. The purpose of this operation is to get the value of the activation which shows the proximity, or relevance, of objects to a query.

*Shrinking* – is a unary operation which transforms a cloud into another cloud: *Shrinking*:  $C_2 \rightarrow C_1$ . If  $C_1$  and  $C_2$  are crisp sets, we assume that  $C_1$  is a proper subset of  $C_2$ , i.e.  $C_1 \subset C_2$ . In general, we assume that this operation does not change significantly the values of the membership functions on the nodes in  $C_2$ , and that  $|C_1| \leq |C_2|$ . Shrinking is a kind of inverse operation to expansion, although we do not necessarily assume that for any pair of such operations  $C_1 \equiv \text{Shrinking}(\text{Expansion}(C_1))$  for each object  $C_1$ .

### 3.2. Network Flow Operations over Network Objects

Spreading activation (SA) (see [Troussov et al., 2009]) is one of the network flow methods to implement the operations described above. This is a wide class of algorithms which iteratively propagate the activation (relevancy measure) from the initial seed to other nodes.

### 3.3. Network Flow Methods for Natural Text Processing

Spreading activation algorithms were used for knowledge based natural language (text) processing ([Judge et al., 2008], [Troussov et al., 2009], [Troussov et al., 2008a], [Troussov et al., 2009b], and [Judge et al., 2007]). In this approach the text is modeled as a cloud of concepts (in a formal definition introduced in Section 3.1) in a semantic network (such as network of concepts from ontologies) and graph-based operations were used for mining of text models. The rationale and intended goals of graph-based methods described in these papers could be recounted as follows.

We assume that the source text is coherent and cohesive as opposed to random list of words. Therefore if some concept are relevant to the text, as indicated by the big value of  $M$ , the “neighbour” concepts are also somehow relevant to the text, since the neighbourhood of nodes is defined by links which represent semantic relations between concepts such as synonymy, “is-a”, “part-of” etc. We also assume that the keywords (subject terms, subject headings, descriptors), defined in information retrieval as terms that capture the essence of the topic of a document, should have a special position within the clustering structure of the text models (for instance, they hardly exist outside of strong clusters induced by the terms mentioned in the document). Term disambiguation, and other tasks of natural language understanding, are usually perceived as inference: “mentioning of *car* in a sentence increases our awareness that the term *Jaguar* mentioned in the same text refers to a *car*” [Troussov et al., 2009]. However, inferencing from one term is not quite reliable, while inferencing based on mentions of various terms is more reliable from a probabilistic point of view, which was confirmed by our numerical simulation showing sharp phase transition from uncertainty to certainty with the increase of number of the nodes in the initial seed.

Finding the key terms is done by spreading the activation from concepts mentioned in the text to other concepts in a semantic network. From this point of view the purpose of these operations could be classified as local

ranking. The proper ranking should be achieved by *Smoothing* to account for inferencing. The key concepts of a text are not necessarily mentioned in the text, so the operation is one of *Expansion*. So as an end to end solution the knowledge based semantic processing of a text is transforming the seed (concepts mentioned in a text) to a larger set of concepts and providing *Local ranking*.

### **3.4. Navigating Networks of People and Associated Objects**

Social spaces such as blogs, wikis and online social networking sites are enabling the formation of online communities where people are linked to each other through direct profile connections and also through the content items that they are creating, sharing, tagging, etc. The Semantic Web provides a platform for gathering diverse information from heterogeneous sources and aggregating such linked data into multidimensional network of nodes representing people, organisations, projects etc. Spreading activation methods were used in [Kinsella et al., 2008] to augment objects from social spaces, by highlighting related objects, recommending tags, and suggesting relevant sources of knowledge.

Recommendations are performed using a network flow engine which, in light of the current paper, provides the operation of *Local ranking* defined in Section 3.1 in the ego-centric network defined by the *Query*.

### **3.5. Hybrid Recommender System in the Activity Centric Environment Nepomuk-Simple**

The concept of navigation in the ego-centric networks [Kinsella et al., 2008] by queries which are single nodes of the network, was extended to navigation by queries which are *Clouds* in [Troussov et al., 2008b].

This paper presents the architecture of the hybrid recommender system in the activity centric environment Nepomuk-Simple (EU 6th Framework Project NEPOMUK).

“Real” desktops usually have piles of things on them where the users (consciously or unconsciously) group together items which are related to each other or to a task. The so called Pile UI, used in the Nepomuk-Simple imitates this type of data and metadata organisation which helps to avoid premature categorisation and reduces the retention of useless documents.

Metadata describing user data is stored in the Nepomuk Personal Information Management Ontology (PIMO). Proper recommendations, such as recommendations for additional items to add to the pile, apparently should be based on the textual content of the items in the pile. Although methods of natural language processing for information retrieval could be useful, the most important type of textual processing are those which allow us to relate concepts in PIMO to the processed texts. Since any given PIMO will change over time, this type of natural language processing cannot be performed as pre-processing of all textual context related to the user. Hybrid recommendation needs on-the-fly textual processing with the ability to aggregate the current instantiation of PIMO with the results of textual processing.

Modeling this ontology as a multidimensional network allows the augmentation of the ontology with new information, such as the “semantic” content of the textual information in user documents. Recommendations in Nepomuk-Simple are computed on the fly by graph-based methods performing in the unified multidimensional network of concepts from PIMO augmented with concepts extracted from the documents pertaining to the activity in question. In this paper, we classify Nepomuk-Simple recommendations into two major types. The first type of recommendations is the recommendation of additional items to the pile, when the user is working on an activity. The second type of recommendation arises, for instance, when the user is browsing the Web; Nepomuk-Simple can recommend that the current resource might be relevant to one or more activities performed by the user. In both cases there is a need to operate with *Clouds* (fuzzy sets of PIMO nodes): *Clouds* describe topicality of documents in terms of PIMO as we described in Section 3.3, the pile itself is a *Cloud*.

---

## Similarity of Sets of Nodes and Search for Similar Sets

---

Discussion on the distinction between similarity and proximity of any given network nodes is outside of the scope of this paper. In this Section we present an empirical approach to the computation of similarity based on a network flow process. The similarity of network nodes, or more generally the similarity of two network objects (like clouds which are fuzzy sets of network nodes) could be described in terms of their ability to affect various parts of the network (like in viral marketing applications [Kempe et al., 2003], [Kempe et al., 2005]). In other words, the similarity of two sets  $A_0$  and  $B_0$  should be defined as the similarity of two fuzzy sets  $A=Expanding(A_0)$  and  $B=Expanding(B_0)$ , where the operation *Expanding* is done by network flow methods compatible with the targeted applications. For instance, if the target application is in the area of “viral marketing”, than we expect that the *Expanding* is done by network flow methods which model “viral marketing”.

In section 4.1 we provide additional arguments to justify our approach to similarity and introduce the similarity of two fuzzy sets on a network. In Section 4.2 we describe efficient and scalable implementation of search for similar sets.

Operations with network objects introduced in Section 3.1 could be classified in terms of number of arguments or operands that the operation takes. In logic, mathematics, and computer science, this number is called arity, in linguistics arity is sometimes called valency. All operations introduced in Section 3.1 are unary. Computation of similarity of two nodes or two objects defined in Section 4.1. is a binary operation, search for objects similar to object of interest in a collection of  $n$  objects is  $(n+1)$ -ary operation.

### 4.1 Similarity of Two Sets of Nodes

Traditional measures for set similarity (such as Jaccard similarity coefficient) describe how much in common two crisp sets have in terms of the "exact" match. The similarity value is a number in the range 0 to 1, 0 – no common elements, 1 – the sets are equal:

$$\text{Similarity (Set1, Set2)} = \frac{\text{The number of elements in the common}}{\text{The number of elements in the union of two sets}}$$

or, using mathematical notations:

$$S(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The links between nodes must be taken into account when comparing the sets of nodes. Instead of the degree of exact match, we need to use a "fuzzy" matching technique. To illustrate this “fuzzy” matching, let us consider a geometrical example of four sets of nodes on a two dimensional grid shown on the Fig. 3 with shapes shown on the Fig 2.

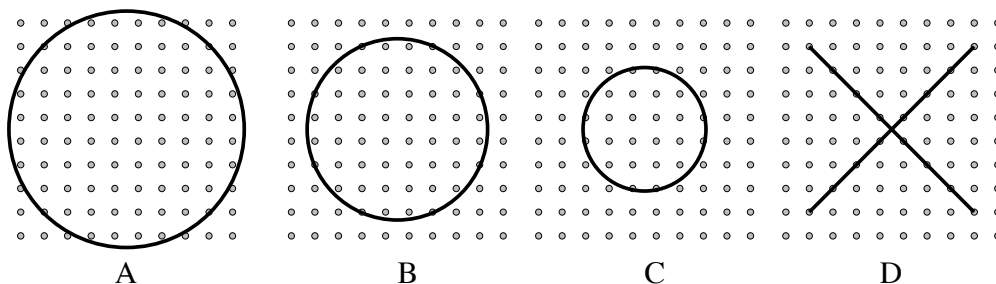


Fig. 2. Four sets of nodes on two dimensional grid

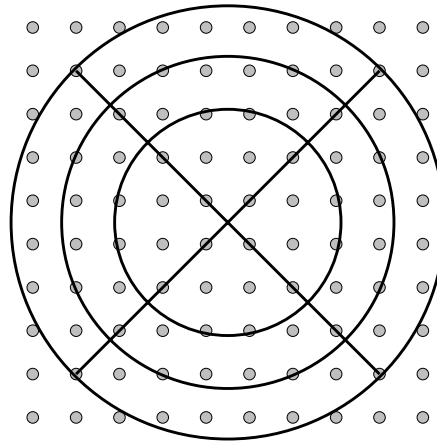


Fig. 3. Four sets of nodes on two dimensional grid with centres of symmetry placed at the origin of the grid

Which two sets of nodes are most similar? If our matching strategy is to look for an exact match, then the pair A and D would be most similar because they have the most nodes in common. However, intuitively, A and B are closest. How do we make a computation based on this intuition which will show us that A and B are very similar? Our approach for "fuzzy" matching is to expand all the sets by making their boundaries less well defined and more "fluffy" and as the measure of similarity between original pair we choose the exact match (i.e. overlap) of their expanded variants.

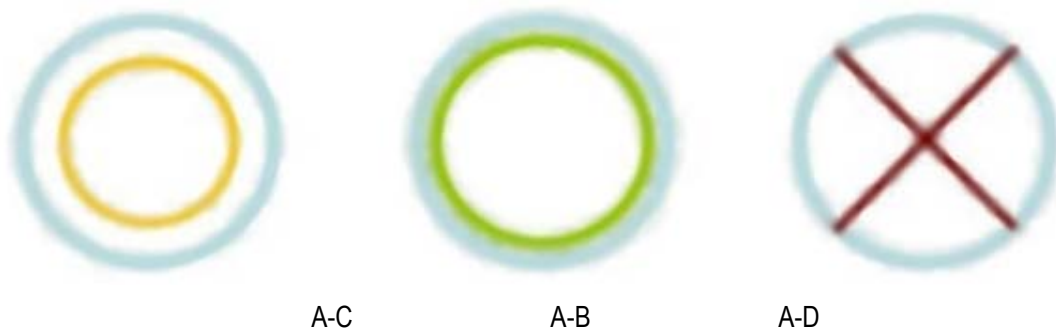


Fig. 4. To provide "fuzzy" comparison of original sets, we perform their "fuzzyfication" first by the operation defined as Expanding in Section 3.1

From the Fig. 4 one can see that expanded sets A and C still do not have common area, and hence the measure of their similarity is zero. Expanded sets A and B became practically indistinguishable, while sets A and D have common areas only in the four corners.

#### 4.2 Retrieval of Similar Sets in a Collection

Operations introduced in Section 3.1 are unary. Computing the similarity of two sets of nodes defined in Section 4.1. is a binary operation. Search for sets similar to the set of interest in a collection of  $n$  objects is  $(n+1)$ -ary operation. In the Nepomuk-Simple environment such an operation is used to find activities (piles) similar to user's current activity, or to provide the recommendation that the currently viewed web resource could be useful for particular activities of the user.

A scalable and linear wrt  $n$  implementation of this operation could be based on the algorithm suggested in [Troussov et al., 2009b] and the processing scheme on the Fig. 5.

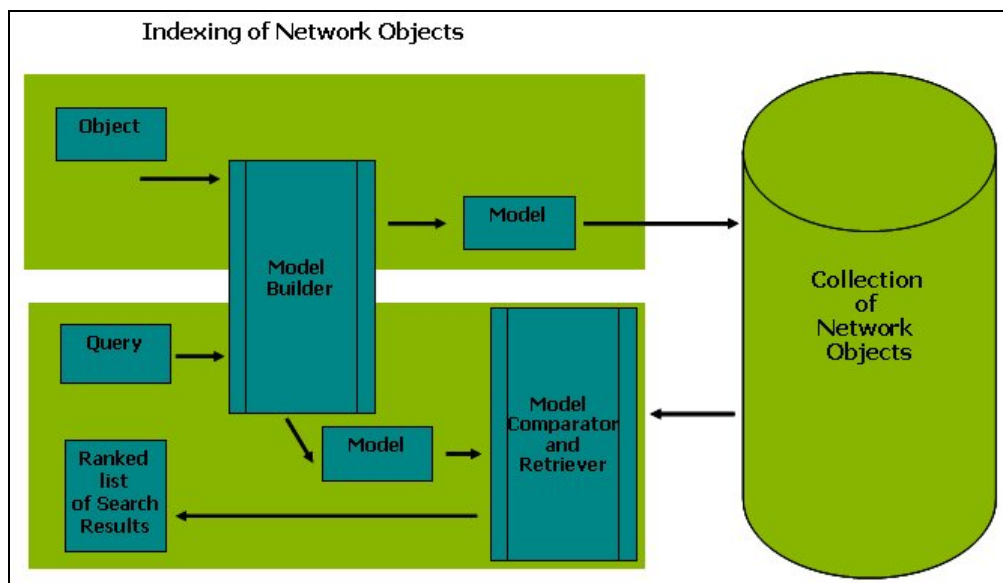


Fig. 5. Indexing and retrieval process to find similar fuzzy sets of nodes (or network objects defined in Section 3.1)

### Network-flow based Computational Systems for Mining and Use of the Social Context

In this Section we present a novel software framework for mining and use of network models of social context based on a set of atomic software engines implementing network flow operations described in Section 3.1. Arguments (operands) of these operations are network objects (fuzzy sets of network nodes).

This framework generalises the design of systems constituting the previous art without introducing new components which could potentially hamper performance and scalability. We show that efficient and scalable implementations for each of the atomic software engines already exist (although as part of monolithic software applications). For instance, the paper [Judge et al., 2007] described the system which performs atomic operations on networks with several hundred thousand nodes in 200msc on an ordinary PC. The paper [Troussov et al., 2008b] describes the large scale multifunctional application where various recommendations are done using hybrid methods including natural text processing. Therefore we conclude that the framework described in this Section could be successfully used to mine and exploit the social context.

Major steps involved in building the software application based on principles described in this paper are:

1. Modeling the social context (such as instantiations of socio-technological systems) using multidimensional networks.
2. Tasks are modeled as a *Cloud* - fuzzy set of nodes which perform the role of *Query*
3. Task dependent models of social context enhancement (the network is enhanced on the fly by new objects and new links between nodes, and augmented by new task dependent objects)
4. Local ranking using *Query* as the initial seed provides the ranked list of network objects relevant to the *Query*

The previous sections provide examples of these steps. For instance, in Nepomuk-Simple the underlying network is enhanced on the fly by concepts extracted from the textual content of pile items.



---

**Conclusions and Future Work**

---

We have revised the previous art in use of network models of weak knowledge, and we described the algorithms and the architecture of the hybrid recommender system in the activity centric environment Nepomuk-Simple (EU 6th Framework Project NEPOMUK):

The applications constituting previous art were monolithic software applications. In this paper we present a novel computational paradigm which breaks these applications into “atomic” components, where the computational methods for propagation are separated as distinct “atomic” network flow engines and described efficient scalable implementations of such operations with the performance on the subsecond level for networks with several hundred thousand nodes. This approach provides a unified view of previous applications. From the software engineering perspective the advantages of such an approach include easy software maintenance, reuse and optimization of network flow engines, and the guide for new applications.

Future work requires refining the set of atomic operations and selection of network flow methods for each of such operations. Evaluation of the results for each operation as well of the applications build from these operations is the next stage.

---

**Bibliography**

---

- [Nelson, 1974] Ted Nelson (1974): *Computer Lib/Dream Machines*. Self-published, 1974, sixth printing (May 1978), ISBN 0-89347-002-3, page DM45
- [Contractor, 2008] Contractor, N. (2008). “The Emergence of Multidimensional Networks.” Retrieved February 13, 2010, from [http://www.hctd.net/newsletters/fall2007/Noshir\\_Contractor.pdf](http://www.hctd.net/newsletters/fall2007/Noshir_Contractor.pdf)
- [Troussov et al., 2009a] Troussov, A, Judge, J., Sogrin, M., Akrouf, A., Davis, B., Handschuh, S. "A Linguistic Light Approach to Multilingualism in Lexical Layers for Ontologies", *SLT*, vol 12, Polish Phonetics Association, ed. G. Demanko, K. Jassem, S. Szpakowicz
- [Judge et al., 2008] Judge, J., Nakayama, A. , Sogrin, M., and Troussov, A. "Method and System for Finding a Focus of a Document". Patent Application US 20080/263038 Kind Code: A1. Filing Date: 02/26/2008
- [Troussov et al., 2009] Troussov, A., Levner, E., Bogdan, C., Judge, J., and Botvich, D. “Spreading Activation Methods.” In Shawkat A., Xiang, Y. (eds). *Dynamic and Advanced Data Mining for Progressing Technological Development*, IGI Global, USA, 2009.
- [Langville and Meyer, 2006] Langville, A.N. and Meyer, C. “Google's PageRank and Beyond: The Science of Search Engine Rankings.” Princeton University Press, 2006
- [Jaschke et al., 2007] Jaschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2007) “Tag Recommendations in Folksonomies.” *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD*, Warsaw, Poland.
- [Kinsella et al., 2008] Kinsella, S., Harth, A., Troussov, A., Sogrin, M., Judge, J., Hayes, C., & Breslin, J.G. (2008). “Navigating and Annotating Semantically-Enabled Networks of People and Associated Objects.” *Why Context Matters: Applications of Social Network Analysis* (T. Friemel, ed.), VS Verlag, ISBN 3531163280, 79-96
- [Troussov et al., 2008a] Troussov, A., Sogrin, A., Judge, J., Botvich, D. "Mining Socio-Semantic Networks Using Spreading Activation Technique". *Proceedings of I-KNOW '08 and I-MEDIA '08*, Graz, Austria, September 3-5, 2008
- [Troussov et al., 2008b] Troussov, A., Judge, J., Sogrin, M., Bogdan, C., Lannero, P., Edlund, H., Sundblad, Y. "Navigation Networked Data using Polycentric Fuzzy Queries and the Pile UI Metaphor". *Proceedings of the International SoNet Workshop*, pp. 5-12, 2008
- [Nepomuk project, 2008] Nepomuk PSEW Recommendation: Using the Recommendations View in PSEW. Retrieved June 30, 2011, from <http://dev.nepomuk.semanticdesktop.org/wiki/UsingPsewRecommendations>

- [Troussov et al., 2009b] Troussov, A., Sogrin, Judge, J., Botvich, D. "System and Method for Ontology-based Personalized Semantic Search", Disclosed by IBM. Loaded into the IP.com Prior Art Database on 2009-04-21 UTC, IPCOM000181971D
- [Borgatti and Everett, 2006] Borgatti, S. and Everett, M. (2006). A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484.
- [Chen, 1996] Chen, C.H., Ed. 1996 *Fuzzy Logic and Neural Network Handbook*. McGraw-Hill, Inc.
- [Kempe et al., 2003] Kempe, D., Kleinberg, J., and Tardos, E. Maximizing the Spread of Influence through a Social Network. *Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2003.
- [Kempe et al., 2005] Kempe, D., Kleinberg, J., and Tardos, E. Influential nodes in a diffusion model for social networks. In *Proc. 32nd Intl. Colloq. on Automata, Languages and Programming*, pages 1127–1138, 2005
- [Judge et al., 2007] Judge, J., Sogrin, M., and Troussov, A. "Galaxy: IBM Ontological Network Miner". *Proceedings of the 1st Conference on Social Semantic Web (CSSW)*, September 26-28, 2007, Leipzig, Germany.

---

## Authors' Information

---



**Alexander Troussov** – Ph.D., IBM Dublin Center for Advanced Studies Chief Scientist. Dublin Software Lab, Building 6, IBM Technology Campus, Damastown Ind. Est., Mulhuddart, Dublin 15, Ireland; e-mail: [troussov@gmail.com](mailto:troussov@gmail.com)

*Major Fields of Scientific Research: natural language processing, software technologies, network analysis*



**John Judge** – Ph.D., Centre for Next Generation Localisation Dublin City University Dublin 9, Ireland; e-mail: [jjudge@computing.dcu.ie](mailto:jjudge@computing.dcu.ie)

*Major Fields of Scientific Research: computational linguistics, natural language processing, social network analysis, semantic web applications*



**Mikhail Alexandrov** – Professor, Academy of national economy and civil service under the President of Russia; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, Bellaterra (Barcelona), 08193, Spain; e-mail: [MAlexandrov@mail.ru](mailto:MAlexandrov@mail.ru)

*Major Fields of Scientific Research: data mining, text mining, mathematical modeling*



**Eugene Levner** – Professor of Computer Science at Holon Institute of Technology and Bar-Ilan University, Israel. He is the full member of the International Academy of Information Sciences, a member of editorial boards of four international journals; 52, Golomb St, Holon 68102 Israel; e-mail: [levner@Hit.ac.il](mailto:levner@Hit.ac.il)

*Major Fields of Scientific Research: combinatorial optimization, operations research, design and analysis of computer algorithms, algorithm complexity and computability, scheduling theory, grid optimization, network analysis, and risk analysis*

## FOLKSONOMY - SUPPLEMENTING RICHE EXPERT BASED TAXONOMY BY TERMS FROM ONLINE DOCUMENTS (Pilot Study)

Aleš Bourek, Mikhail Alexandrov, Roque Lopez

**Abstract:** RICHE (Research Inventory of Child Health in Europe) is a platform developed and funded under the Health domain of 7<sup>th</sup> European Framework Program. The platform search engine is expected to use the multilingual taxonomy of terms for processing and classifying large volumes of documents of the RICHE repository. So far the experts participating in this project have produced the initial version of expert based taxonomy of terms relating to child health (based on existing taxonomies). In the paper we propose a simple man-machine technique for continuous support and development of the existing term list, which consists of three steps: 1) construction of various keyword lists extracted from a topic oriented document set using various levels of word specificity 2) selection of the most useful keyword lists using subjective criteria as a precision of selection and a number of new words 3) manual selection of new terms. Experimental material was represented by documents uploaded from three organizations active in child health improvement policies: World Bank, World Health Organization (WHO), and DG SANCO of European Commission (EC). The selection was performed in order to assess terms used in these documents that may be absent in the RICHE taxonomy. Presented work should be considered as a pilot (feasibility) study. The objective of the RICHE platform is to identify gaps in European child health research, so extensive mapping exercise has been started. Classification of identified studies is essential and cannot be based only on traditional terms of existing taxonomies. Emergent terms (such as for example “cyberbullying”) need to be identified and included into existing taxonomies. In our future work we focus on developing techniques related to multilevel and multiword term selection

**Keywords:** Child health, natural language processing, taxonomy, term selection

**ACM Classification Keywords:** I.2.7 Natural Language Processing

---

### Introduction

---

#### 1.1 RICHE project and its current taxonomy

The European Commission (EC) and other funding agencies make large investments in child health research. The health of our children is satisfactory, but there are serious concerns, for example obesity, mental health, alcohol abuse, and sexuality. The objective of EC efforts is to establish a sustainable network for researchers, funders, policy makers, advocates and young people in Europe to support collaboration in developing the future of child health research. In the RICHE project we are producing an inventory of research and reports on gaps in research and on roadmaps for future research [RICHE, [http](#)].

RICHE platform includes a search engine for efficient retrieval of information included in the platform and related to child health protection and child health and healthcare quality improvement. The effective function of this engine needs well prepared lexical resources based on specific terms reflecting the topic under consideration, for example, appropriate taxonomies [Taxonomy, [http](#)]. It was found that “in some cases, it was essential to look outside traditional health and social care search engines in order to fully understand and conduct a systematic review on subjects that are relevant and pertinent to public health, such as justice and police databases. As far as the taxonomy structures were concerned, the existing classification identified limited indexing used in some databases as a potential problem - “Where free text words are relied upon, variations in terminology used by different disciplines can create barriers and limit the value of the material retrieved in a search of the database/document repository” [RICHE, [http](#)].

To address this issue child health experts, mainly from the area of public health research, participating in RICHE project have prepared the initial version of taxonomy consisting of one-word and multiword terms distributed on 6 sub-topics (main axis of child health determinants): 1).Demographics 2).Population group 3).Agents, influences and settings 4).Health, disability, health issues and determinants 5).Language 6).Type of study. Table 1 shows some terms from the sub-topic 'Demography' as a part of the mentioned taxonomy

*Table 1. Example of terms of sub-topic 'Demography' taken from RICHE taxonomy*

Terms	Synonyms
Indeterminate / anomalous	unknown, unstipulated, uncertain, not determined at birth
Stillborn children	died before birth, died in utero, born dead
Genetic studies	heredity, inherited, chromosomal, inborn, genomic

It is obvious, that the RICHE taxonomy contains multiword terms located on two levels corresponding to given 6 sub-topics. Totally the current RICHE taxonomy contains 822 different one-word terms in stem-form. Currently RICHE experts continue to improve the taxonomy in two directions: modifying the term list and constructing a more detailed hierarchy

## 2.2 Problem settings

Goal of our contribution is to identify and add new terms to the existing "expert taxonomy" using appropriate NLP tools. On the given stage of our work we introduce two limitations:

- we deal with one-word terms and one-level term distribution
- we process limited number of documents ,

The first limitation is introduced by the fact that multiword and multilevel term list construction requires the use of sophisticated methods but as a feasibility study we decided to address the problem using as simple as possible tools. The second limitation is defined by our approach to analyze (when necessary) individual documents but not to work with descriptive statistics.

In the paper we propose a simple methodology for augmenting the existing term list constructed by RICHE experts, and to test this technique experimentally. The technique consists of 3 steps:

- construction of various keyword lists extracted from the above mentioned document set using various levels of word specificity
- selection of the most useful keyword lists using subjective criteria for the accuracy of selection and a number of new words to be analyzed
- manual selection of new terms by an expert

To demonstrate possibilities of the proposed methodology we performed experiments with different subsets of documents. The source of information for our experiments were 60 documents (7 Mb in plain textual format) downloaded from online resources of World Bank [World Bank, [http](http://)], World Health Organization [WHO, [http](http://)], and European Commission [EC, [http](http://)]. These organizations belong to main policy players in the area of child health in Europe.

The auxiliary problem we studied was the dependence of results on a concrete document set, which was used as a source of new terms. For this we considered various subsets of a given document corpus and compared their lexical resources from the point of view of our main goal - improvement of existing RICHE term list.

### *1.3 Related work*

Indicators of child health were introduced and studied in many projects, for example, [Rigby, 2002; Rigby, 2003]. These indicators need information reflecting current state of child health and RICHE platform is supposed to provide this information.

There are several works related with term selection focused on medical applications [Madden, 2007; Armstrong, 2009]. But our task is different: to supplement the existing term list by new terms from independent sources such as the Internet or the domain of "gray literature".

The key position in problem solution consists in constructing various keyword lists for consideration for further detailed analysis by a child health expert(s). The general approaches and algorithms for term selection are well presented in many publications, for example in the well-known monograph [Baeza-Yates, 1999]. An interesting approach to multilevel term selection is described in [Makagonov, 2005]. It is recognized that word collocations have a large informative and distinctive power. Just these collocations form so-called multiword terms [Yagunova, 2010]. But all these techniques are not simple. They often need complimentary information about word distribution in a corpus, correlation between words, etc. In this paper we deal with the simplest case: one-word and one-level term selection.

We apply the criterion of word specificity for extraction of keywords (candidates to be included in term list) from a given document set. This criterion was successfully used for constructing domain oriented vocabularies [Makagonov, 2000]. Recently free-share LexisTerm program was developed [Lopez, 2011] where both a traditional corpus based option and the new document based option are used [Lopez, 2011]. We use both of these options in our work.

In section 2 we describe the proposed methodology. In section 3 we demonstrate the results of experiments. Section 3 includes conclusions.

---

## **2. Methodology of term selection**

---

### *2.1 General description*

We use word 'keyword' instead 'term' on the stages of constructing initial keyword lists and selecting the best lists for further manual analysis. Here the selected keywords are only the "candidates to be" terms if an expert will select them.

As mentioned in introduction the proposed methodology consists in three stages:

- 1) Constructing several keyword lists on the basis of criterion of word specificity. We use the criterion of word specificity because the topic under consideration is not broad enough and we expect to obtain more or less useful keyword lists. But we do not know in advance what level of specificity and what option of selection will prove to be the most relevant to the existing expert list. For this reason we have to generate several keyword lists.
- 2) Selecting the most useful keyword lists. Here we compare each keyword list constructed on the previous stage with the expert list using indicator of precision and the number of keywords not included in the expert list (external keywords). We use indicator of precision to be more confident that not-common keywords are relevant to the expert list. But in general high precision refers to the case of very short keyword lists with very high level of word specificity. Such short lists can be un-useful. In this case an

expert must evaluate the number of non-common words and makes a decision whether the concrete keyword list is useful or not.

- 3) Extracting terms from the keyword lists selected on the previous stage. This is performed manually by an expert.

In our study we used stems instead of original word forms.

The auxiliary research concerned studying the dependence of results on concrete document sets. Here we performed two simple experiments with different document sets

- Comparison of keyword lists extracted from the half and from all documents with the expert list on the basis of indicators of precision and recall
- Comparison of keyword lists between themselves (without taking into account the expert list) constructed for a subset of 15% documents and for a different 15% subset of documents. The same procedure was implemented for 30% document subset and other 30% document subset, and finally for 50% documents and other 50% document subset. We use here only indicator of precision with respect to each document subset from the pair.

In these experiments we used the same fixed parameters for keyword extraction: level of keyword specificity and option of keyword selection.

### 2.2 Constructing keyword lists by the criterion of word specificity

To construct keyword lists we used the LexisTerm [Lopez, 2011] program. Following are some necessary definitions:

**Definition 1.** The general lexis is a frequency word list based on a given corpus of texts

The given corpus means here any standard document set reflecting the lexical richness of a given language. Generally such a corpus contains in a certain proportion the documents taken from newspapers, scientific publications related with various domains, novels and stories. For example, it could be the British National corpus.

**Definition 2.** The level of specificity of a given word  $\mathbf{w}$  in a given document corpus  $C$  is a number  $K \geq 1$ , which shows how much its frequency in the document corpus  $f_C(\mathbf{w})$  exceeds its frequency in the general lexis  $f_L(\mathbf{w})$ :

$$K = f_C(\mathbf{w}) / f_L(\mathbf{w})$$

**Definition 3.** The level of specificity of a given word  $\mathbf{w}$  in a given document  $D$  is a number  $K \geq 1$ , which shows how much its frequency in the document  $f_D(\mathbf{w})$  exceeds its frequency in the general lexis  $f_L(\mathbf{w})$ :

$$K = f_D(\mathbf{w}) / f_L(\mathbf{w})$$

Our research was done using keywords and terms presented in stem form. For this we had to transform both documents and general lexis based on British National corpus to their stem using the well-known Porter stemmer [Porter, 1980]

### 2.3 Measures for comparison of word lists

To select the most preferable lists of keywords we used two variables: precision of keyword selection and the number of keywords in the list not included in the expert list. Following is a description of these variables:

Let  $N_L$  is a number of terms in the expert list,  $N_W$  is a number of keywords in our list,  $N_{LW}$  is a number of common words in both lists. In this case a precision is calculated according the formula:

$$P = N_{LW} / N_W$$

That is the precision, it is a share of terms from the expert list in our keyword list. With the designation introduced above the number of new keywords in our list  $N$  is calculated by the formulae

$$N = N_W - N_{LW}$$

Additionally an expert can take into account the other two indicators of quality used in Information Retrieval: recall and so-called F-measure. They are calculated according the following formulae:

$$R = N_{LW} / N_L$$

$$F = 2(PR) / (P+R)$$

In the auxiliary experiments we needed to compare two keyword lists. Let  $N1_W$ ,  $N2_W$ ,  $N12_W$  be the number of keywords in the 1-st list, 2-nd list and the common keywords respectively. In this case one considers the precision with respect to each list and the average precision. They are calculated according the formulae:

$$P_1 = N12_W / N1_W$$

$$P_2 = N12_W / N2_W$$

$$P_{12} = (P_1 + P_2) / 2$$

All these indicators are used in the experiments described in the next section

## Experiments

### 3.1 Selection of preferable keyword lists

In this experiment we compared keywords selected from our full document set consisting of 60 documents with the full expert list. We constructed keyword list under different levels of word specificity ( $k=1,5,10,20,50,100$ ) and different options (C and D). The results are presented in Table 2. The designations in this table are described in the previous section. Figure 1 shows a graphical view of Table 2 for the precision

Table 2. Characteristics of different keyword lists, comparison with the complete expert list

	C, k=1	C, k=5	C, k=10	D, k=10	D, k=20	D, k=50	D, k=100
<b>Words</b>	1473	231	86	1807	1193	512	219
<b>P</b>	0.238	0.442	0.558	0.200	0.226	0.252	0.324
<b>R</b>	0.426	0.124	0.058	0.440	0.328	0.157	0.086
<b>F</b>	0.305	0.194	0.106	0.275	0.268	0.193	0.136
<b>N<sub>LW</sub></b>	350	102	48	362	270	129	71
<b>N</b>	1123	129	38	1445	923	383	148

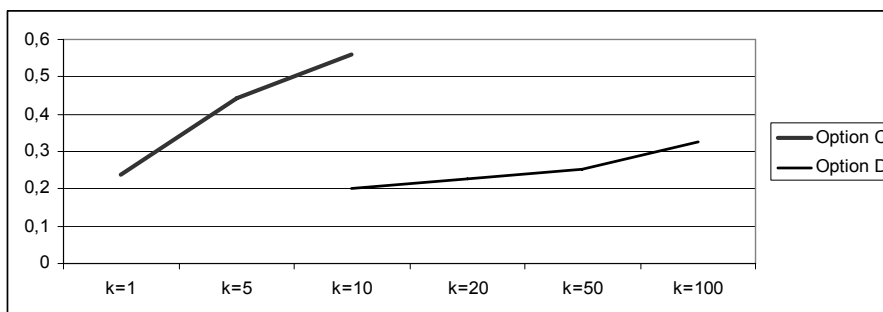


Fig.1 Graphical view of the Table 2 for the values of precision

One of the authors, an MD also engaged in medical informatics selected the two most useful/preferable lists with the parameters: option C,  $k=10$  and option D,  $k=100$ . With these parameters we obtained the highest precision

in the framework of given mode, and from the other hand the number of new non-common keywords is suitable for manual evaluation.

### 3.2 Contribution of sub-topics to keyword lists

In this experiment we compared our keyword lists with the terms of experts related with each category. The results are presented in the Table 3. Here we consider option C with the parameters  $k=5, 10$ . Figure 2 provides a graphed version of Table 2.

Table 3. Characteristics of different keyword lists, comparison with each category of the expert list

	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6
<b>Option C, k=5</b>	0.030	0.065	0.121	0.247	0.061	0.078
<b>Option C, k=10</b>	0.047	0.093	0.128	0.291	0.116	0.105

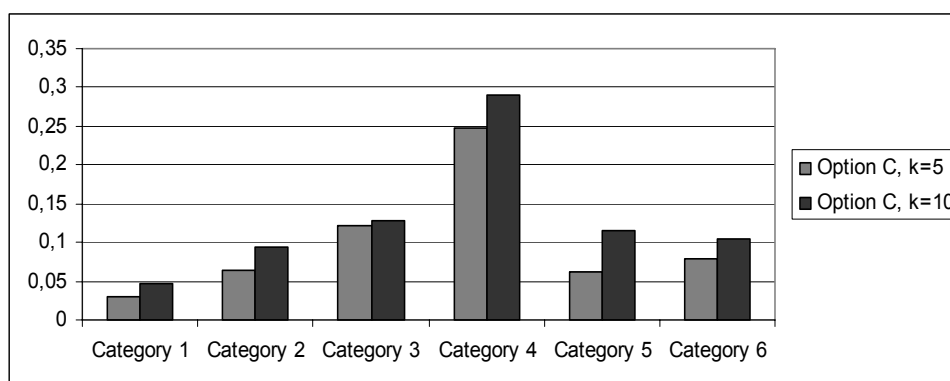


Fig.2 Graphed version of Table 2

It is easy to see that Category 4 (RICHE taxonomy axis 'Health Disability') is represented the best way in our keyword lists. Such result is not surprising, since this axis is the most comprehensive in included terms.

### 3.3 Lexical resources of subsets

In the first experiment we compare keyword lists extracted from the half and from all documents with the expert list on the basis of indicators of precision, recall and F-measure. The results are presented in the Table 4. In the second experiment we compared keyword lists for different pairs of document subsets using precisions with respect to each subset and the averaged precision. The results are presented in the Table 5. The designations of these tables are described in the previous section. In all experiments we used option C with the level of specificity  $k=1$ .

Table 4. Comparison of lexical resources of different document number with the expert list

	50% documents	100% documents
<b>Words</b>	1019	1473
<b>P</b>	0.277	0.238
<b>R</b>	0.343	0.426
<b>F</b>	0.306	0.305



Table 5. Cross-comparison of lexical resources of different document subsets

	10+10	20+20	30+30
<b>Words</b>	959/1171	1065/1234	1019/1333
<b><math>P_1</math></b>	0.766	0.783	0.838
<b><math>P_2</math></b>	0.628	0.676	0.656
<b><math>P_{12}</math></b>	0.697	0.729	0.747

Data of table 4 shows a naturally tendency: the more documents we consider, the more different words they have the less precision is. Table 5 demonstrates the relative stability of the results: any equal subsets of documents (having in view the number of documents) have close averaged precision. These circumstances inform about good quality of selected document corpus.

### 3.4 Term selection

As we have already mentioned in section 3.1 the preferable keyword lists for term selection prove to be those with the parameters: option C,  $k=10$  and option D,  $k=100$  – offering enough potential “candidate” terms for inclusion into the RICHE “expert taxonomy” but at the same time producing only a small number of false identified terms.

A quick “human” selection of potential terms (word stems) for classification of child health research documents identified by the machine-learning methodology presented from a volume of text relating to child health from websites of WHO, DG SANCO (European Commission) and World Bank (policy related documents on child health) was performed.

Following list shows some, not all, “candidate” term stems (word stems NOT included in the original RICHE child health experts version of the taxonomy) were identified:

- clinic* (possible candidate for classification of “clinical study, type of clinics”)
- HIV* (not included and exists only in a synonym classification option as “AIDS”)
- implement* (possible candidate for classification of “implementation research”)
- monitor* (possible candidate for classification of “monitoring research”)
- overview* (possible candidate for classification of “overview materials”)
- agenda* (possible candidate for classification of “agenda setting research”)
- cognit* (possible candidate for classification of “cognitive related research”)
- complex* (possible candidate for classification of “complexity research”)
- indicator* (possible candidate for classification of “indicator related research”)
- consensu* (possible candidate for classification of “consensus based documents”)
- emerg* (possible candidate for classification of “emergent issues related research”)
- fat* (not even the synonym *obes* (obesity) is included as a term in the RICHE expert taxonomy)
- framework* (possible candidate for classification of “framework setting research”)
- guideline* (possible candidate for classification of “clinical guidelines related research”)
- Mediterranean* (often used geographical term, NOT included in the Language/Geography axis)
- pregnan* (possible candidate for classification of “pregnancy related research”)
- priorit* (possible candidate for classification of “prioritization research, priority setting research”)
- protocol* (possible candidate for classification of “protocol setting research”)
- questionnaire* (possible candidate for classification of “questionnaire/survey related research”)
- satisfact* (possible candidate for classification of “health service satisfaction research”)
- vitamin* (possible candidate for classification of “vitamin related research”)
- facilit* /is included as a term in the RICHE expert taxonomy in the form of “Health care facility” BUT NOT

*for example as "facilitation study"/*

**feed** /is included as a term in the RICHE expert taxonomy BUT only in the form of "breastfeeding"/

**global** /is included as a term in the RICHE expert taxonomy BUT only in the form of "Global change and health (WHO-Europe)" BUT NOT as "globalization related research"/

On the other hand, four terms identified as "missing" by the machine-learning based methodology were already included in the original RICHE "expert taxonomy":

**analys** /is included as a term in the RICHE expert taxonomy/

**demograph** /is included as a term in the RICHE expert taxonomy/

**expenditur** /is included as a term in the RICHE expert taxonomy/

**outcom** /is included as a term in the RICHE expert taxonomy/

---

## Conclusion

We elaborated on and proposed the simplest way for supporting term list development experts of the RICHE project. Our methodology is based on criterion of specificity for keyword selection and characteristics of precision for keyword list selection having in view the possibilities of subsequent manual work of an expert. The results of our experiments may prove useful in evaluating how criterion parameters affect the list of selected terms.

Term stems expertly identified as possible classification term "candidates" have been correlated with terms of the RICHE\_expert\_taxonomy\_ver\_January\_2011. The four term stems followed by the remark "/is included as a term in the RICHE expert taxonomy/" represent false identified terms by means of our simple machine learning based approach. With the exception of these four terms all above listed stems have a potential for classifying child health related research documents of the RICHE repository, as commented in the brackets following the respective term. All of the "candidate" terms will be presented to the RICHE consortium group for expert evaluation and for inclusion of terms the experts will find consensus on into the most appropriate axis of the RICHE project taxonomy. Based on the presented small scale preliminary analysis of 60 documents, we demonstrate that the methodology has the strength and potential to identify terms possibly missed by the expert community, especially when a corpus of documents produced by experts focused on a different area of child health (policy issues rather than public health child research – which was the dominant area of expertise of the majority of RICHE project collaborators) is used. We conclude that even basic machine-aided document evaluation is a tool for consideration when addressing the issue of possible human bias of the taxonomy defining expert community.

---

## Bibliography

[Armstrong, 2009] Armstrong, R., Doyle, J., Waters, E. Cochrane Public Health Review Group Update: incorporating research generated outside of the health sector. *Journ. of Public Health*. Vol. 31, No. 1, pp. 187-189, 2009 (available at <http://jpubhealth.oxfordjournals.org/cgi/reprint/fdn116v1.pdf>)

[Baeza-Yates, 1999] Baeza-Yates, R., Ribero-Neto, B. *Modern Information Retrieval*. Addison Wesley, 1999.




[EC, http] [http:// ec.europa.eu](http://ec.europa.eu)

Lopez, R., Alexandrov, M., Barreda, D., Tejada, J. LexisTerm – the program for term selection by the criterion of specificity (this Proceedings)

[Madden, 2007] Madden, R., Sykes, C., Usten, T. World Health Organization Family of International Classifications, definitions, scope and purpose. 2007 (available at <http://www.who.int/classifications/en/FamilyDocument2007.pdf>)

- [Makagonov, 2000] Makagonov, P., Alexandrov, M., Sboychakov, K. A toolkit for development of the domain-oriented dictionaries for structuring document flows. In: Data Analysis, Classification, and Related Methods, Studies in classification, data analysis, and knowledge organization, Springer-Verlag, pp. 83–88, 2000
- [Makagonov, 2005] Makagonov, P., Figueroa, A., R., Sboychakov, K., Gelbukh, A. Learning a domain ontology from hierarchically structured texts. In: Proc. of Workshop “Learning and Extending Lexical Ontologies by using Machine Learning Methods” at 22-nd Intern. Conf. on Machine Learning (ICML 2005), Bonn, Germany, 2005.
- [Porter, 1980] Porter, M. An algorithm for suffix stripping. Program, 14, pp. 130–137, 1980.
- [RICHE, http] RICHE: <http://childhealthresearch.eu>
- [Rigby, 2002] Rigby, M., Kohler, L. (edit.) Child health indicators of life and development (Child): report to the European Commission, 2002 (available at <http://www.europa.eu.int/comm/health/ph/>)
- [Rigby, 2003] Rigby, M., Kohler, L., Blair, M, Metchler, R. A priority for a caring society. European Journ. on Public Health, vol. 13, pp. 38-46, 2003
- [Taxonomy, http] Taxonomy: [http://www.taxonomywarehouse.com/resultsbycat\\_include.asp?vCatUID=21&catcode=040100](http://www.taxonomywarehouse.com/resultsbycat_include.asp?vCatUID=21&catcode=040100)
- [WHO, http] World Health Organization: <http://www.who.int>
- [World Bank, http] World Bank: <http://www.worldbank.org>
- [Yagunova, 2010] Yagunova, E., Pivovarova, L., The Nature of collocations in the Russian language. The Experience of Automatic Extraction and Classification of the Material of News Texts // Automatic Documentation and Mathematical Linguistics, 2010, Vol. 44, No. 3, pp. 164–175. © Allerton Press, Inc., 2010.

### Authors' Information

	<p><b>Ales Bourek</b> – Senior lecturer, Masaryk University, Brno, Czech Republic; Head of Center for Healthcare Quality, Masaryk University. Kamenice 126/3, 62500 Brno, CZ. e-mail: <a href="mailto:bourek@med.muni.cz">bourek@med.muni.cz</a></p> <p>Major fields of interest: reproductive medicine – gynecology, health informatics, healthcare quality improvement, health systems</p>
	<p><b>Mikhail Alexandrov</b> – Professor, Academy of national economy and civil service under the President of Russia; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: <a href="mailto:MAlexandrov@mail.ru">MAlexandrov@mail.ru</a></p> <p>Major fields of scientific research: data mining, text mining, mathematical modeling</p>
	<p><b>Roque López</b> – Student of System Engineering at San Agustín National University, calle Santa Catalina N° 117 Arequipa, Peru; e-mail: <a href="mailto:rlopezc27@gmail.com">rlopezc27@gmail.com</a></p> <p>Major fields of scientific research: natural language processing, text mining, social network analysis</p>

## CLASSIFICATION OF FREE TEXT CLINICAL NARRATIVES (SHORT REVIEW)

**Olga Kaurova, Mikhail Alexandrov, Xavier Blanco**

**Abstract:** *The paper is a limited review of publications (1995-2010) related to the problem of classification of clinical records presented in a free text form. The techniques of indexing and methods of classification are considered. We also pay special attention to the description of document sets used in the mentioned research. Finally, we conclude about the perspective research directions related with the topic.*

**Keywords:** *Natural language processing, medical corpus, medical diagnostics, automatic medical classification*

**ACM Classification Keywords:** *I.2.7 Natural Language Processing*

---

### Introduction

---

For many years natural language processing (NLP) tools have been successfully used to process information in various applications areas. One of such areas is medicine, where the majority of documents are presented in free text form as primary care encounter notes, physical exams, radiology reports, progress notes, clinical histories, etc. The subject under consideration in this article is automatic classification of clinical data. This problem in effect is reduced to medical diagnostics and so, the topic we discuss can be titled as free text based medical diagnostics.

The automatic classification allows to detect incorrect diagnoses, to find particular clinical events in patient records, and to facilitate the exchange of clinical histories between hospitals. In recent years the problem of clinical text classification was under consideration in several publications.

The paper consists of 6 sections. In Section 2 we describe the problems related to application of NLP tools to clinical documents. We also consider statistical measures used in the reviewed articles. Section 3 refers to corpora used in the research. We describe features of the corpora and difficulties of document classification. NLP methods and tools are presented in Sections 4 and 5. Conclusions and short discussion are in Section 6. The paper contains two tables: the 1-st one with corpus data features and the 2-nd one with NLP methods and results of classifications; there is also an appendix with several corpus data samples.

---

### Problem description

---

Many researchers have attempted to use natural language processing tools, classification and text mining techniques for processing clinical narratives. But their works have centered mostly on clearly defined domains, such as detection of acute bacterial pneumonia from chest X-ray reports [Fizman, 2000], detection of breast cancer from mammogram reports [Jain, 1997], identification of episodes of asthma exacerbation [Aronow, 1995a], detection of pediatric respiratory and gastrointestinal outbreaks [Ivanov, 2003]. So, the NLP tools used in these researches were based on clearly defined vocabularies related to given domains.

Generally speaking, the main task of NLP in medical applications consists in automatic encoding clinical data from free text form to numerical form. Such a process requires semantic and syntactic information. Output data after such a transformation can be further automatically classified according to specific goals. For example, Aronow et al. in his work identifies episodes of asthma exacerbation. Basically, the problem is to sort a large collection of documents (medical record encounter notes) according to a specified set of characteristics. These

characteristics describe the presence of an acute exacerbation of asthma. Then an automatic system performs a three bin sort. One bin contains encounter notes classified as “very probably exacerbation”. The second bin contains “very probably not exacerbation”. And the third bin contains encounters which are uncertain from the point of view of the classifier [Aronow, 1995a].

Automatic assignment of medical codes is one of the NLP results. In healthcare, diagnostic codes are used to group and identify diseases, disorders, symptoms, human response patterns, and medical signs. We can name several coding systems: ICPC (International Classification of Primary Care), ICD (International Statistical Classification of Diseases and Related Health Problems), NANDA (North American Nursing Diagnosis Association) and others. Most of the works considered in the present review deal with ICPC codes. The ICPC is a method that allows for the classification of the patient’s reason for encounter, the problems/diagnosis managed and primary care interventions. These three elements make out the core constituent parts of an encounter in primary care. The ICPC was first published in 1987 by Oxford University Press and is now commonly used in the United Kingdom, the Netherlands, Norway and other countries.

A multi-class classification problem using ICPC is addressed in the work [Røst, 2006]. The authors classify encounter notes using ICPC codes as classification bins. However, they use not the codes themselves (because their number is very large) but their so called chapter values. The ICPC system contains 17 chapters. Thus, encounter notes are classified in 17 classes. Larkey et al. in their work [Larkey, 1996] operated with another coding system – ICD9. It is a more complex code than ICPC because it consists of two parts: a major category and a subcategory. This makes ICD9 more suited for specialized usage in hospitals. Larkey compared three different types of classifiers for automatic code assignment to dictated inpatient discharged summaries. Each possible code served as a category. The problem consisted in calculation of the probabilities that a document belonged to each category from a given list. As a result, a ranked list of codes (categories) for each document was built.

To evaluate the quality of binary classification of medical data the statistical measures “sensitivity” and “specificity” are used. Sensitivity measures the proportion of actual positives which are correctly identified as such, e.g. the percentage of sick people, who are correctly identified as having the condition. Specificity measures the proportion of negatives, which are correctly identified, e.g. the percentage of healthy people, who are correctly identified as not having the condition. In other words, in medical diagnostics sensitivity is the ability to correctly identify those with the disease, whereas specificity is the ability to correctly identify those without the disease. Chapman et al. [Chapman, 2005] used so-called “positive predictive value” besides sensitivity and specificity. It is the proportion of patients with positive test results, who are correctly diagnosed.

To evaluate the quality of categorization, statistical measures “precision” and “recall” are calculated. Just these characteristics are used in information retrieval. Precision is the proportion between relevant retrieved document set and all retrieved documents (the relevant and not relevant ones) and recall is the proportion between the same relevant retrieved document set and all relevant documents (the retrieved and not-retrieved ones).

However, there is a big difference between the typical information retrieval problem and classification of medical data. In information retrieval one aims to provide relevant documents at the top of a belief list. It gives high precision with low level of recall. In classification one aims to classify as many documents as possible, i.e. to achieve high recall, not the high precision [Aronow, 1995b]. The gold standard which is used for evaluation of a given classification procedure is normally prepared by a certified physician or several independent physicians [Chapman, 2005; Fiszman, 2000].

---

## Corpora and their features

---

Many difficulties in applying NLP methods to medical domain spring from the peculiar character of input data. A vast amount of patient data is available only in free text form: encounter notes, radiology reports, discharge summaries, admission histories, reports of physical examinations, etc. Below we consider some of these types of data in detail.

### 3.1. Primary care encounter notes

The characteristic features of a corpus made up of primary care encounter notes are: sparseness, brevity, heavy use of abbreviations, many spelling mistakes. This is due to the fact that the notes are normally written during the consultation with a patient when the time is limited. Another feature of such a corpus is that the data varies greatly in style and length. The texts are written by different physicians, each possessing their own manner of registering clinical information.

A dataset of free-text clinical encounter notes and their corresponding manually coded diagnoses is used in [Røst, 2006]. Totally, there are 482,902 unique encounters. The notes in the experimental dataset are coded according to the ICPC-2 coding system. Each encounter consists of a written note of highly variable length and zero or more accompanying codes. 287,868 of the available encounters have one or more ICPC codes. The final goal of the study was the classification of the notes according to the ICPC-2 code. So, in order to avoid ambiguity in the training data all encounters with more than one code were discarded. The final corpus consisted of 175,167 encounter notes. From these document set 2000 documents were selected randomly to be used as a test set, the remaining were used as a training set.

### 3.2. Medical texts of specific subdomain (e.g. containing specific illness)

A corpus of respiratory encounter notes is presented in the work of Aronow et al. [Aronow, 1995a]. The corpus is divided into two sets: a test and a training collection for automated identification of episodes of asthma exacerbation. The test collection consists of 965 encounter notes of 76 randomly selected asthmatic patients. The training collection consists of 1,368 encounters of other 100 random patients. The corpus is mostly made up of handwritten provider notes, manually entered letter-for-letter by trained inputters. The goal is to sort medical record encounter notes in two groups: those with the evidence of acute exacerbation of asthma and those without it. The corpus was filtered to reduce the number of irrelevant encounter notes. All the notes without the mention of code for asthma or asthma-like conditions (Acute Bronchitis, Bronchiolitis and Bronchospasm) were eliminated. Notes without a definite diagnosis were excluded as well. The resultant corpus numbered 231 texts in the testing collection and 357 in the training collection.

A different approach to creating a corpus was used by Fiszman et al. [Fiszman, 2000]. They dealt with about 15,000 chest x-ray reports produced during a six - month period. All reports were related to acute bacterial pneumonia. From this document set they selected 292 on the following basis: 217 were randomly selected from all the reports of the first three months, while the remaining 75 reports were randomly selected from the following three months from the list of patients with the diagnosis of bacterial pneumonia. Such an artificial way increased the prevalence of pneumonia-related reports in the sample, but at the same time it caused some doubts concerning validity of statistical assessments of classification quality.

### 3.3. Triage chief complaints, notational texts

Chapman et al. [Chapman, 2005] in their work created a collection of free-text triage chief complaints (TCC) - the earliest clinical data available on most hospital information systems. Triage chief complaints are used to describe the reason for a patient's visit to an emergency department. In their research the authors used 4700 complaints as a training set and 800 complaints as a test set in order to classify TCCs into 8 syndromic categories. Main

characteristics of the corpus data result from the purpose of triage chief complaints: to describe an emergency patient's condition using as short phrases as possible. So, the corpus is made up of short TCC strings that contain a lot of abbreviations, truncations, spelling and punctuation mistakes. This aggravates the problem of automatic medical classification as the data needs to pass through a complicated preprocessing stage in order to be expanded from abbreviated into a more complete form.

A different type of data but with similar "preprocessing" problems was analyzed in the work of Barrows et al. [Barrows, 2000]. The researchers tried to extract relevant diagnosis of glaucoma. A corpus they used was made up of 12,839 ophthalmology visit notes presented in the form of "notational text", a special kind of clinical documentation. Notational texts are typed by physicians during routine patient encounters. Therefore, the corpus data is terse, full of abbreviations and symbolic constructions, some of which may be specific to a medical sub-domain, to an institution, or even a clinician. Statements in notational text are poorly formed according to grammatical construction rules and are considerably lacking in punctuation.

#### *3.4. Discharge clinical notes*

Nowadays one of the problems to be solved is encoding medical documents using the ICD-9 code. The reason consists in wide computerization of hospitals. Franz et al. [Franz, 2000] in their work tested three different approaches to automatic disease coding using a German corpus of free-text discharge diagnoses. The corpus is made up of 120000 diagnosis records and the data covers the whole range of clinical medicine. The main problem the researchers faced in their work consisted in low quality texts with spelling errors, ambiguities and abbreviations. Sometimes one phrase included a combination of diagnoses.

In the framework of the automatic assignment of ICD-9 codes Larkey et al. [Larkey, 1996] operated with another corpus of discharge clinical data. Their corpus consisted of 11.599 discharge summaries divided into a training set of 10.902 documents, a test set of 187 documents, and a tuning set of 510 documents. Each summary included several ICD-9 codes (from 1 to 15). The characteristic feature of the corpus is that the summaries vary greatly in length, namely, from 100 to 3000 words per document. Moreover, the notes are written by different doctors, so the data also varies considerably in style.

A challenging problem of detecting possible vaccination reactions in clinical notes was addressed in the work of Hazlehurst et al. [Hazlehurst, 2005a]. The authors created a large corpus, the unusual feature of which consisted in the diverse nature of the data: telephone encounters, emergency department visits and outpatient office visits, including visits for immunization. Telephone encounters are included in the corpus because people often use the nurse advice line to inquire about possible reactions to immunizations converting it into a rich source of immunization-related events. The process of creating and processing the corpus was caused by the necessity to adapt the previously developed NLP-system (MediClass) to the particular goal of the given research. A manually coded training set of 248 patient records was created. Another set of 13657 visits was created, from which 1000 records were used to train the system. The efficiency of the system was finally evaluated on a test set of the remaining 12631 records (26 records were excluded due to corrupted text notes).

#### *3.5. Lexical resource*

In the reviewed works both one-word and multiword medical terms are used as keywords. Just these keywords form a parameter space for procedures of classification. However, there is no information on how multiword terms are constructed. In [Yagunova, Pivovarova, 2010] a statistical method for collocation construction is proposed. By a collocation the authors mean any nonrandom co-occurrence of two or more lexical units, specific for either language as a whole or particular genre of texts (corpus). Their method exploits two statistical measures: Mutual Information (MI) and t-score. MI is a coefficient of association strength, while t-score can be understood as a modification of the collocation frequency. The results showed that extracted MI-collocations consisted of such

multiword expressions as terminology and nominations; MI-measure proved to be useful for determining subject domain. T-score, in turn, picks out functional grammatical compounds and high-frequency constructions.

One should say that the potential possibilities of classification, any classification, are defined by relations between classes. The closer their characteristics are the lower level of quality we obtain while distributing objects between classes. In case of document classification the closeness between classes is mainly defined by the intersection of lexis related with each class. When classes, e.g. diseases, have absolutely different descriptor lists the quality of classification is the highest: we can avoid any errors. But when lexical resources of each class are similar then one can expect many errors. These extreme cases say about so-called wide domain and narrow domain with respect to classes, which compose this domain.

Unfortunately, the authors of the publications mentioned above did not consider their corpus of medical documents from the point of view of lexis used for disease description. Just for this reason we can say nothing about domains reflected in a corpus: whether they are wide or narrow ones. Taking this circumstance into account could essentially improve the results of classification. We could find only one work where the problem of clustering/classification of documents is considered from the point of view of width and narrowness of a given domain. It is the doctoral dissertation of David Pinto [Pinto, 2008].

In the Table 1 we present the features of the corpora used in the reviewed articles. Appendix contains some examples of clinical records from these corpora.

Table1. Features of the corpora

Study	Domain type	Total number of documents used in primary search	Training set	Test set	Typical features
Røst et al., 2006	Primary care encounter notes	175167	173167	2000	Sparseness, brevity, heavy use of abbreviations, spelling mistakes.
Barrows et al., 2000	Ophthalmology visit notes in the form of "notational text"	12,839			Terseness; abbreviations; specific symbolic constructions; ungrammatical statements; poor punctuation
Aronow et al., 1995	Encounter notes of respiratory care (acute asthma exacerbation)		1368 → 351	965 → 231	Handwritten provider notes, manually entered letter-for-letter as written by trained inputters => specific medical abbreviations and terminology.
Fiszman et al., 2000	Chest x-ray reports	15000	292		
Chapman et al., 2005	Triage chief complaints		4700	800	Very short, abbreviations, truncations, spelling and punctuation mistakes
Franz et al., 2000	Free-text discharge diagnoses	120000			Abbreviations; orthographic variations; combination of two diagnosis phrases within one diagnostic statement (typically, a noun phrase with a prepositional phrase)
Larkey et al., 1996	Discharge summaries	11599	10902	187	Vary in length; heterogeneous in linguistic style; much free form text irrelevant to the coding task
Hazlehurst et al., 2005	Post-immunization encounter records		248 + 1000	12631	



## Methods

---

The methods of classification used in the reviewed publications belong to methods of Machine Learning. So, one can find their descriptions in well-known books related with this area [Mitchell, 1997; Bishop, 2006]. The specificity of classification of clinical texts consists in

- preprocessing low quality data (abbreviations, ungrammatical statements, etc)
- necessity to take into account hidden information (relations between descriptors related with diseases)

### 4.1. Bayesian classifiers

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. Each node in the graph represents a random variable (observable quantities, latent variables, unknown parameters or hypotheses), while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node.

In the Table 2 we present a comparison of several NLP-tools. As one can see, the researchers mostly use Bayesian inference network, achieving good results [Aronow, 1995a, 1995b; Barrows, 2000; Fisman, 2000; Chapman, 2005; Larkey, 1996]. For example, Aronow [Aronow, 1995a] used the Bayesian inference network in order to make decision with respect to asthma and other adjacent diseases.

### 4.2. Support Vector Machine

Support vector machine (SVM) is very popular in naturally-scientific applications and nowadays it becomes widely used in the problems of medical document classification. Standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of. This makes the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other. A SVM model is a representation of the examples as points in space, mapped so that the examples of separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

SVM was used in the work of Røst [Røst, 2006]. His automatic system was trained on examples using SVM classifier. The achieved accuracy is 49,7%, but this approach is considered as quite promising.

### 4.3. K-nearest neighbors

k-nearest neighbors (k-NN) is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbors algorithm is amongst the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its  $k$  nearest neighbors.  $k$  is a positive integer, typically small.

An interesting method was proposed by Larkey et al. [Larkey, 1996]. The authors combined the described k-NN method with Bayesian classifiers and Relevance Feedback. Relevance Feedback has typically been used in information retrieval to improve existing queries. From the retrieved documents the user indicates a set of relevant documents. The original query and terms from the indicated documents are combined to produce a new query, which is better at ranking relevant documents over non-relevant documents. A small set of features is selected separately for each code, and a query is trained for each code. The comparison of different combinations of classifiers (k-NN, BC, RF) shows that the best result is obtained by combining them all.

#### 4.4. Decision trees

Decision tree learning is a commonly used data mining method. It uses a decision tree as a predictive model, which maps observations about an item to conclusions about the item's target value, i.e. predicts the value of a target variable based on several input variables. In tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree invented by Ross Quinlan. It can be summarized as follows: take all unused attributes and count their entropy concerning test samples → choose attribute for which entropy is minimum (or, equivalently, information gain is maximum) → make node containing that attribute.

Aronow et al. in their work considered an ID3 decision tree learning algorithm besides Bayesian network [Aronow, 1995b]. In their research the authors encoded each document as a list of feature-value pairs and passed it to an ID3 decision tree, which returned a probability that the document was a positive instance.

#### 4.5. Lexis based methods

In the work of Zhang et al. [Zhang, 2010] we find a description of an information retrieval method developed to automatically identify qualified patients for breast cancer clinical trials from free-text medical reports – Subtree match. It is an algorithm that finds structural patterns in patient report sentences that are consistent with given trial criteria. The sub-tree model is constructed in three steps. First, each training sentence is parsed by a syntax parser into a parse tree which describes sentence's syntax structure. Second, for each parse tree, all leaves that are keywords are located and from them, a program back-tracks up the tree for three levels to generate a set of subtrees. Finally, all subtrees found in this way are collected and represented as tree regular expressions. Tree regular expressions are used as search models for the given criterion [Zhang, 2010]. This algorithm proved to be very effective, yielding the results of 90%, 49%, 0.63 (Precision, Recall, F-score respectively) for the most complex model in which both manual and automatic keyword generation were used.

A quite novel approach in using keywords is presented in the paper of Catena et al. [Catena, 2008]. The authors elaborated a simple algorithm for automatic classification of clinical encounter notes based on general category descriptions and their comparison. Category description is a list of keyword frequencies reflecting documents of this category and the novelty of the method consists in using so-called positive, neutral and negative keywords. The sign of keywords (+1,0,-1) reflects its contribution to a given category and to its anti-category. Anti-category is presented by all documents, which do not belong to a given category. The problem of categorization is being solved basing on the rule that a keyword is taken into account only if its density in all texts of a given category exceeds its density in all texts of its anti-category. The results evaluated with Purity-measure and *F*-measure are 0.75 and 0.74 respectively.

Table 2 below presents NLP methods and achieved results of the reviewed works.

Table 2: NLP methods and reporting metrics

Study	Tool	Method	Accuracy	Recall	Precision	Specificity	Sensitivity	Positive predictive value
Røst et al., 2006	SVM-Light	Support Vector Machine	49,7%					
Barrows et al., 2000	MEDLEE	Bayesian inference network	95%	90%	100%			
Aronow et al., 1995	(1) INQUERY (2) FIGLEAF	(1) Bayesian inference network (2) ID3 decision tree algorithm			(1)71.7 % (2)80.8 %			

Fiszman et al., 2000	Symtext	Bayesian inference network		95%	78%	85%		
Chapman et al., 2005	Mplus	Bayesian inference network				98%	98%	91%
Franz et al., 2000	MS Access + Visual Basic	(1) Support Vector Machine (2) Heuristic approach	(1) 40% (2) 50,4%					
Larkey et al., 1996	INQUERY	k-NN, relevance feedback, Bayesian independence classifiers		77.6%	57.0%			
Hazlehurst et al., 2005	MediClass							57%
Zhang et al., 2010		Subtree match		49%	90%			

**Tools**

In this section we present tools used in the reviewed papers. We also inform about other tools, which were not mentioned in the publications but could be useful for automatic medical classification.

**SymText** (Symbolic Text Processor) is a NLP system, which uses syntactic and semantic knowledge to model the underlying concepts in a textual document [Koehler, 1998]. SymTexts syntactic component contains a parser that uses an augmented transition network grammar and a transformational grammar. SymTexts semantic component includes Bayesian networks that model the relevant medical domain. The system was created at LDS Hospital in Salt Lake City, Utah

**MedLEE** is a text processor that extracts and structures clinical information from textual reports and translates the information to terms in a controlled vocabulary [MedLEE, http]. Clinical information then can be accessed by further automated procedures. It has been used in radiology, discharge summaries, sign out notes, pathology reports, electrocardiogram reports, and echocardiogram reports, and can readily be ported to other clinical domains. MedLEE was created by Carol Friedman in collaboration with the Department of Biomedical Informatics at Columbia University, the Radiology Department at Columbia University, and the Department of Computer Science at Queens College of CUNY.

**INQUERY** and **FIGLEAF** were developed in the Center for Intelligent Information Retrieval (CIIR), University of Massachusetts, Amherst. The INQUERY is a text based information retrieval system that uses a probabilistic inference net model [Callan, 1992]. FIGLEAF (Fine Grained Lexical Analysis Facility) is a text classification system based on statistical analysis of semantic features. It uses decision trees derived from examples in a set of training documents [Lehnert, 1995].

**Mplus** (The Medical Probabilistic Language Understanding System) is a robust medical text analysis tool with a Bayesian network-based semantic model for extracting information from narrative patient records [Christensen, 2002]. The advantage of the system is that its semantic model can be trained in specific domains to adapt to new tasks.

**MediClass** (Medical Classifier) is a knowledge-based system that automatically classifies the content of a clinical encounter captured in the medical record [Hazlehurst, 2005b]. MediClass accomplishes this by applying a set of application-specific logical rules to the medical concepts that are automatically identified in both the free-text and precoded data elements.

**Weka** (Waikato Environment for Knowledge Analysis) is a machine learning software written in Java, developed at the University of Waikato, New Zealand [Weka, [http](#)]. Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. The system includes several standard data mining tools, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection, which turns it into a useful tool of NLP in medical domain.

**RapidMiner** (developed at the Artificial Intelligence Unit of the University of Dortmund, Germany) is an environment for machine learning, data mining, text mining, predictive analytics, and business analytics, written in Java [RapidMiner, [http](#)]. It provides data mining and machine learning procedures including: data loading and transformation, data preprocessing and visualization, modeling, evaluation, and deployment. It also integrates learning schemes of the Weka and statistical modeling schemes of the R-Project [R, [http](#)]

**CLUTO** is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of various clusters [CLUTO, [http](#)]. It was developed at the University of Minnesota, Minneapolis, USA. CLUTO consists of both a library and stand-alone programs, via which an application program can access directly the various clustering and analysis algorithms implemented in CLUTO. Among the key features of the system are: multiple classes of clustering algorithms (partitional, agglomerative, graph-partitioning based), multiple similarity/distance functions (Euclidean distance, cosine, correlation coefficient, extended Jaccard).

---

## Conclusion

---

The paper contains the review of a number of publications related with processing clinical narratives. By 'processing' we mean application of models and methods of computational linguistics for classification of short medical texts presented in free text form.

The review includes:

- analysis of medical corpora from the point of view of their style, volume and domains;
- consideration of methods of classification, which are part of machine learning methods;
- presentation of tools both mentioned in the publications and those could be useful for future applications.

Such a consideration allows to outline the following routes for improving the quality of classification:

- detailed pre-analysis of corpora for revealing linguistic properties of texts, in particular, whether we deal with wide- or narrow domain;
- application of advanced indexing techniques including word collocations, etc.;
- application of both classification methods and classification technologies including assembling, boosting, etc.

---

## Acknowledgements

---

As members of ACM's Special Interest Group on Health Informatics (SIGHIT) [<http://www.sighit.org>] we would like to thank all the group organizers and Dr. Ted Peterson in particular for providing this inexhaustible resource of publications. We also thank Dr. Elena Yagunova for sharing with us her experience and publications in the field of collocation construction.

---

**Bibliography**

---

- [Aronow, 1995a] D.B.Aronow, J.R.Cooley, S.Soderland. Automated identification of episodes of asthma exacerbation for quality measurement in a computer-based medical record. In: Proc. of Annual Symposium on Computer Applications in Medical Care. 309-13, USA, 1995.
- [Aronow, 1995b] D.B.Aronow, S.Soderland, J.M.Ponte, F.Feng, B.Croft, W.Lehnert. Automated classification of encounter notes in a computer based medical record. In: Proc. of MEDINFO '95 8th World Congress on Medical Informatics, Medinfo, Canada, p. 8-12, 1995.
- [Barrows, 2000] R.C.Barrows, M.Busuioc, C.Friedman. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. In: Proc. of American Medical Informatics Association Annual Symposium, 51-5, 2000.
- [Bishop, 2006] C. Bishop. Pattern Recognition and Machine Learning, Springer, 2006
- [Callan, 1992] J.P.Callan, W. B.Croft, S.M.Harding. The INQUERY Retrieval System. In: Proc. of DEXA-92, 3-rd International Conference on Database and Expert Systems Applications, pp. 78-83, 1992.
- [Catena, 2008] A.Catena, M.Alexandrov, B.Alexandrov, M.Demenkova. NLP-Tools Try To Make Medical Diagnosis. In: Proc. of the 1-st Intern. Workshop on Social Networking (SoNet-2008), Skalica, Slovakia, 2008.
- [Chapman, 2005] W.W.Chapman, L.M.Christensen, M.M.Wagner, P.J.Haug, O.Ivanov, J.N.Dowling, R.T.Olszewski. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artificial Intelligence in Medicine, 33(1), 31-40. 2005.
- [Christensen, 2002] L.M.Christensen, P.J.Haug, M.Fizman. MPLUS: a probabilistic medical language understanding system. In: Proc. of the ACL-02 workshop on Natural language processing in the biomedical domain (BioMed '02 ), Vol. 3, p. 29-36, Stroudsburg, USA, 2002. ( <http://acl.ldc.upenn.edu/W/W02/W02-0305.pdf>)
- [CLUTO, http] CLUTO: <http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview>
- [Fizman, 2000] M.Fizman, W.Chapman, D.Aronsky, R.S.Evans, P.J.Haug. Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports. American Medical Informatics Association Vol. 7 Num. 6, p. 593-604, 2000.
- [Franz, 2000] P.Franz, A.Zaiss, S.Schulz, U.Hahn, R.Klar. Automated coding of diagnoses--three methods compared. Proc. of AMIA Symp. 250-4, 2000.
- [Hazlehurst, 2005a] B.Hazlehurst, J.Mullooly, A.Naleway and B.Crane. Detecting Possible Vaccination Reactions in Clinical Notes. In: Proc. of AMIA Annu Symp Proc. 2005; 2005: 306-310.
- [Hazlehurst, 2005b] B.Hazlehurst, H.R.Frost, D.F.Sittig, V.J.Stevens. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc. 12(5):517-529, 2005.
- [Heinze, 2001] D.T.Heinze, M.L.Morsch, and J.Holbrook. Mining Free-Text Medical Records. In: Proc. AMIA Symp. 2001; 254-258, 2001.
- [Hofmans-Okkes, Lamberts, 1996] I.M.Hofmans-Okkes, H.Lamberts. The International Classification of Primary Care (ICPC): new applications in research and computer-based patient records in family practice. Family Practice; Vol. 13, No 3, p. 294-302, 1996.
- [Hripcsak, 2002] G.Hripcsak, J.Austin, P.Alderson & C.Friedman. Use of natural language processing to translate clinical information from database of 889,921 chest radiographic reports. Radiology (224), 157-163, 2002.
- [Ivanov, 2003] O.Ivanov, P.Gesteland, W.Hogan, M.B.Mundorff, M.Wagner. Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. In: Proc. of AMIA Annual Fall Symposium; p. 318-22, 2003.
- [Jain, Friedman, 1997] N.L.Jain, C.Friedman. Identification of Findings Suspicious for Breast Cancer Based on Natural Language Processing of Mammogram Reports. Proc. of AMIA Annu Fall Symp., p. 829-33, 1997.
- [Koehler, 1998] S.B. Koehler. Symtext: A Natural Language Understanding System For Encoding Free Text Medical Data. Doctoral Dissertation, Department of Medical Informatics, University of Utah, 1998
- [Larkey, Croft, 1996] L.S.Larkey and W.B.Croft. Combining classifiers in text categorization. In: Proc. of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96), pp. 289-97, Zurich, Switzerland. ACM Press, 1996.

- [Lehnert, 1995] W. Lehnert, S. Soderland, D. Aronow, F. Feng, A. Shmueli. Inductive text classification for medical applications. In: *Journal of Experimental and Theoretical Artificial Intelligence*, 7:49-80, 1995.
- [MedLEE, http] MedLEE: <http://lucid.cpmc.columbia.edu/medlee/>
- [Mitchell, 1997] T. Mitchell. *Machine Learning*, McGraw Hill, 1997.
- [Pinto, 2008] D. Pinto, *On Clustering and Evaluation of Narrow Domain Short-Text Corpora*. Doctoral Dissertation, Polytechnic University of Valencia, Spain, 2008
- [R, http] R-project: <http://www.r-project.org>
- [RapidMiner, http] RapidMiner: <http://rapid-i.com>
- [Røst, 2006] T.B. Røst, O. Nytro, A. Grimsmo. Classifying Encounter Notes in the Primary Care Patient Record. In: *Proc. of the 3-rd Intern. Workshop on Text-Based Information Retrieval (TIR-06)*. Univ. Press, p. 5-9, 2006.
- [Weka, http] Weka: <http://www.cs.waikato.ac.nz/ml/weka/>
- [Yagunova, Pivovarova, 2010] E. Yagunova, L. Pivovarova. The Nature of Collocations in the Russian Language. The Experience of Automatic Extraction and Classification of the Material of News Texts // *Automatic Documentation and Mathematical Linguistics*, 2010, Vol. 44, No. 3, pp. 164–175. © Allerton Press, Inc., 2010. Original Russian Text © E.V. Yagunova, L.M. Pivovarova, 2010, published in *Nauchno Tekhnicheskaya Informatsiya, Seriya 2*, 2010, No. 6, pp. 30–40.
- [Zhang, 2010] J. Zhang, Y. Gu, W. Liu, T. Zhao, X. Mu, W. Hu. Automatic Patient Search for Breast Cancer Clinical Trials Using Free-Text Medical Reports'. In: *Proc. of the 1-st ACM International Health Informatics Symposium*. New York, USA., 2010.
- [Zhou, 2006] X. Zhou, H. Han, I. Chankai, A.A. Prestrud, A. Brooks. Approaches to Text Mining for Clinical Medical Records. In: *Proc. of the 21-st Annual ACM Symposium on Applied Computing 2006, Technical tracks on Computer Applications in Health Care (CAHC 2006)*, Dijon, France. 2006.

---

## Appendix

---

Some samples of clinical corpus data

Sample 1: A typical encounter note from [Røst et al., 2006]:

“Inflamed wounds over the entire body. Was treated w/ apocillin and fucidin cream 1 mth. ago. Still using fucidin. Taking sample for bact. Beginning tmnt. with bactroban. Call in 1 week for test results”.

Sample 2: An encounter note for acute asthma exacerbation [Aronow et al., 1995]:

“G100 ASTHMA  
 COUGH & WHEEZE X1-2D HX PNEU 4/92 AFEB  
 ACTIVE RR=48 W/MOD RETRAX CHEST  
 DIFFUSE EXP WHEEZING & RHONCHI ONLY  
 SL. CLEARING AFTER 2 NEBS 02 SATS 93->  
 HOSP ER.”

Sample 3: Notational text from [Barrows et al., 2000]:

“3/1198 IPN  
 SOB & DOE\$  
 VSS, AF

CXR 3LLL ASD no A

WBC IIK

SIB Cx 69GPC c/W PC, no GNR

DIC Cef – PCNIV”

---

## Authors' Information

---



**Olga Kaurova** – Saint Petersburg State University (Department of Theoretical and Applied Linguistics - graduated in 2009); Autonomous University of Barcelona (International Master in “Natural Language Processing & Human Language Technology” - graduated in 2010; PhD program “Lenguas y Culturas Románicas” - current), 08193 Bellaterra (Barcelona), Spain;

e-mail: [kaurovskiy@gmail.com](mailto:kaurovskiy@gmail.com)

Major Fields of Scientific Research: automatic medical classification, sentiment analysis



**Mikhail Alexandrov** – Professor, Academy of national economy and civil service under the President of Russia; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;

e-mail: [MAlexandrov@mail.ru](mailto:MAlexandrov@mail.ru)

Major Fields of Scientific Research: data mining, text mining, mathematical modelling



**Xavier Blanco** – Cathedratric University Professor (Full Professor), fLexSem Research Laboratory (Fonètica, Lexicologia i Semàntica), Department of French and Romance Philology, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain; e-mail: [Xavier.Blanco@uab.cat](mailto:Xavier.Blanco@uab.cat)

Major Fields of Scientific Research: lexicology, lexicography, machine translation

---



---

## Automated transformation of algorithms

---

### MODELS OF THE PROCESS OF AN ANALYSIS OF XML-FORMATTED FORMULAE OF ALGORITHMS

Volodymyr Ovsyak, Krzysztof Latawiec, Aleksandr Ovsyak

**Abstract:** This paper describes two analytical models of the process of an analysis of XML-formatted formulae of algorithms represented in a special editor created for the algebra of algorithms. For identification and storage of types and orientation of selected operations and uniterms, two XML-formatted model algorithms are developed. The ability is shown for transformation of formulae of algorithms that results in 5-time reduction of a number of uniterms while maintaining the functionality of the algorithm.

**Keywords:** Algebra of algorithms, algorithm formulae editor, transformation of algorithms.

---

#### Introduction

Algebra of algorithms [Ovsyak et al, 2011] provides means to describe algorithms as mathematical formulae. Identity transformations are performed over formulae of algorithms just like those for mathematical expressions. The aim of these transformations is to reduce a number of uniterms and algorithm formulae, thus reducing both a time consumed to build and execute the code and a memory to store the code. An *xml* format is used to describe operations in the algebra of algorithms. Applications of the algebra of algorithms in synthesis and transformation of formulae of algorithms are illustrated by examples of two XML-formatted model algorithms, developed in a formulae editor whose GUI is shown.

Construction of the two model algorithms is supported with four theorems, including the one that establishes the functional equivalence of the two algorithms.

---

#### Elements of Editor of Algorithm Formulae

##### 2.1. Main window of editor

Fig.1 shows the main window of the editor of formulae of algorithms, including the menu of commands, operation menu and editing area.

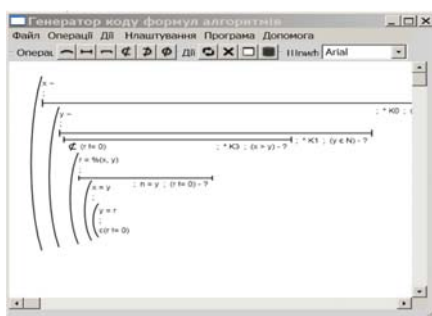


Fig.1. Main window of editor.



2.2. The encoding xml format for formulae of algorithms

Synthesized by the editor of the algorithm's formulas stored in computer memory as files with XML - similar format. For example, sequence of uniterms  $F(x, y)$  and  $S(z)$ , to separate them using a semicolon and arrangement - sign of operation of horizontal sequence has the following entries

$$\overbrace{F(x, y); S(z)}$$

and is recorded into computer's memory as follows:

```
<s sep="sem" ori="hor">
  <u>F(x, y)</u>
  <u>S(z)</u>
</s>
```

where the third line is an identifier of description of the sequence operation (s), uniterm's separator (sep =), a separator ("**sem**"), orientation identifier (ori =) and value ("**hor**"), into the following are recorded identifier of the uniterms (u) and their significance ( $F(x, y)$  oraz  $S(z)$ ) sequence operation with horizontal orientation of the sign sequence's operation

$$\left( F(x, y), S(z) \right)$$

and uniterms separator by comma has the following description:

```
<s sep="com" ori="ver">
  <u>F(x, y)</u>
  <u>S(z)</u>
</s>
```

XML - formats of the description of elimination operation (e) and parallelization (p) has a similar description. Beyond this operation of elimination has three uniterms elimination, separated by semicolons.

Description operation of cyclic sequence (cs) with horizontal (**hor**) and vertical (**ver**) orientation and condition of the cycle (**g-?**) and uniterm **H**, bound with operation of cycle is as follows:

```
<cs ori="hor">
  <u>g-?</u>
  <u>H</u>
</cs>
                                amd
<cs ori="ver">
  <u>g-?</u>
  <u>H</u>
<cs>
```

Operation of cycle elimination (ce) and parallelization (cp) have similar description.

**Formulas of Analysis Xml - Format of Algorithms Formulas**

For receiving in a computer's editor with XML - format formula algorithms necessary is to identify types of transactions, which are sequence, elimination, parallelization, cyclic sequence, elimination and parallelization and uniterms, data separator character and orientation of operations signs. With a view to optimizing the data from

XML - files formulas algorithms was synthesized algorithms formula (1), where  $\underline{pu} \underline{st}$  ( $w \in @T$ )  $Txml$  ( $t \in @T$ ,  $n \in @Nod$ ) - header of formula algorithm.

$$\underline{pu} \underline{st} (w \in @T) Txml(t \in @T, n \in @Nod) =$$

<pre> u ∈ @U; ; u.par = t ; u.val = n.IT ; * ; (n.IT.Le &gt; 0) - ? ; w = u. </pre>	<pre> s ∈ @S; A; (n.Na.Eq("s")) - ?; (n.Na.Eq("u")) - ?; Exc("\$xml"); (n ≠ \$) - ? ; s.par = t ; s.sep = @S.Sep.Sem ; s.sep = @S.Sep.Com ; Exc(„Błqd_s xml”). ; n.Attri["sep"].Val.Eq("com") - ? ; n.Attri["sep"].Val.Eq("sem") - ? ; s.ori = @S.Ori.Hor ; s.ori = @S.Ori.Ver ; Exc(„Błqd_s_O xml”) ; n.Attri["ori"].Val.Eq("ver") - ? ; n.Attri["ori"].Val.Eq("hor") - ? ; s.tA = Txml(s, n.CN[0]) ; s.tB = Txml(s, n.CN[1]) ; w = s. </pre>	(1)
---	--	-----

where  $\underline{pu}$  - identifier of access,  $\underline{st}$  - static property, ( $w \in @T$ ) - is the input parameter  $w$  belongs ( $\in$ ) to type  $T$  subsystem ( $@$ ) for uniterms processing,  $Txml$  - name of algorithm's formula,  $n \in @Nod$  - input parameters  $t$  and  $n$  - type subsystems  $T$  and  $Nod$  (abbreviated name of standard subsystem XmlNode [Petzold, 2002, MacDonald, 2008]);  $u \in @U$  - type variable  $U$  - subsystem "Uniterm" intended for uniterm processing;  $u.par = t$  - attributing to variable  $par$  value of input variable  $t$ ;  $u.val = n.IT$  - attributing variable  $val$  value of uniterm  $n.IT$ , selected by standard uniterm InnerText [MacDonald, 2008] from the input variable  $n$ ;  $w = u$  - attribution to input variable  $w$  variable value  $u$  subsystems  $U$  and the end ( $.$ ) of the algorithm implementation;  $((n.IT.Le > 0) - ?)$  - calculation  $(n.IT.Le)$ , using standard uniterm Length [3], the number of uniterm characters chosen by standard uniterm InnerText [MacDonald, 2008] to  $xml$  - file and verification this number of marks for the majority from zero;  $(n \neq \$) - ?$  - verification whether  $xml$  - file is not empty ( $\$$ );  $Exc("$xml")$  - view by standart uniterm Exception() [Petzold, 2002, MacDonald, 2008] report  $\$xml$  about empty ( $\$$ )  $xml$  - file and the end ( $.$ ) of the algorithm implementation;  $(N.Na.Eq("u") - ?)$  - in input variable  $n$  search of the name ( $Na$ ) keyword "u" string  $xml$  - file with using standard uniterm Equals() [MacDonald, 2008] compared  $Eq("u")$ ;  $(n.Na.Eq("s") - ?)$  - compared with the keyword  $s$ ;  $s \in @S$  - a variable of the subsystem  $S$ , appointed for working on sequence process;  $s.par = t$  - attributing value  $par$  to value of input variable  $t$ ;  $s.sep = @S.Sep.Sem$  - attributing variable  $sep$  separator  $Sem$ ;  $s.sep = @S.Sep.Com$  - attributing variable  $sep$  separator  $Com$ ;  $Exc("Błqd_s xml")$  - view of error message ("Błqd\_s xml") in the uniterm separator;  $n.Attri["sep"].(Val.Eq("com") - ?)$  - in  $n$  search ( $Attri["sep"]$ ) using standard unitermu Attributes ["" ] [Ptzold, 2002, MacDonald, 2008] keyword ( $sep$ ) and comparison ( $Eq("com")$ ), using

standard uniterm Equals(""), its value is recorded in Val - standard variable Value [Petzold, 2002, MacDonald, 2008], with "com"; (*n.Attri*["sep"].Val.Eq("sem") -?) - comparison with "sem"; *s.ori* = @S.Ori.Hor - attributing to variable *ori* identifier of horizontal orientation *Hor* sign of sequence operation, which is in division *Ori* subsystem *S*; *s.ori* = @S.Ori.Ver - attributing vertical (*Ver*) to orientation sign operation; *Exc*("Błąd\_s\_O xml") - view of error message in describing of orientation of the sign operation; (*n.Attri*["ori"].Val.Eq("ver") -?) - comparing the value orientation with *ver*; (*n.Attri*["ori"].Val.Eq ("hor") -?) - comparing the value orientation with *hor*; *s.tA* = *Txml*(*s*, *n.CN*[0]) - choice (CN) uniterm value from position [0] *xml* - description of the algorithm formula using standard uniterm *ChildNodes*[] [Petzold, 2002, MacDonald, 2008] and the algorithm *Txml*() attribution to variable *tA* first uniterm value; *s.tB* = *Txml*(*s*, *n.CN*[1]) - the choice from *xml* - description of algorithm of second uniterm and attributing its to variable *tB*; *w* = *s*. - attributing to input variable *w* value of variable *s* and the end (.) of the algorithm implementation.

Formula *A* differs from elimination by condition (*n.Na.Eq*("s ") -?) the elimination by condition (*n.Na.Eq* ("e ") -?) contains an identifier *e*; *e* ∈ @*E* - the creating variable *e* subsystem *E* working on elimination operation and using variable *e* instead of variable *s* and with missing uniterm processing separator and presence uniterm *cond* = *Txml*(*e*, *n.CN* [2]) - intended for attributing variable *cond* value of third uniterm, which is in *xml* - description of the algorithm formula.

$$A = \left( \begin{array}{l} e \in @E; B; (n.Na.Eq("e"))-? \\ ; \\ e.par=t \\ ; \\ e.ori=@E.Ori.Hor \\ ; \\ e.ori=@E.Ori.Ver \\ ; \\ Exc(„Błąd_e xml”) \\ ; \\ n.Attri["ori"].Val.Eq("ver")-? \\ ; \\ n.Attri["ori"].Val.Eq("hor")-? \\ ; \\ e.tA = Txml(e, n.CN[0]) \\ ; \\ e.tB = Txml(e, n.CN[1]) \\ ; \\ cond=Txml(e, n.CN[2]) \\ ; \\ w = e. \end{array} \right)$$

Formula *B* from elimination by condition (*n.Na.Eq*("s") -?) difference is in: in comparison by condition (*n.Na.Eq*("p") -?) identifier *p* is used; *p* ∈ @*P* - creating of variable *P* subsystem for working on parallelization and using variable *p* instead of the variable *s*.

$$\begin{aligned}
 B = & \left( p \in @P; D; (n.Na.Eq("p"))-? \right. \\
 & ; \\
 & \left. p.par = t \right. \\
 & ; \\
 & \left( p.sep = @P.Sep.Sem \right. \\
 & ; \\
 & \left. p.sep = @P.Sep.Com \right. \\
 & ; \\
 & \underline{Exc}(, Blqd\_p \text{ xml} ") \\
 & ; \\
 & \left. n.Attri["sep"].Val.Eq("com")-? \right. \\
 & ; \\
 & \left. n.Attri["sep"].Val.Eq("sem")-? \right. \\
 & ; \\
 & \left( p.ori = @P.Ori.Hor \right. \\
 & ; \\
 & \left. p.ori = @P.Ori.Ver \right. \\
 & ; \\
 & \underline{Exc}(, Blqd\_p\_1 \text{ xml} ") \\
 & ; \\
 & \left. n.Attri["ori"].Val.Eq("ver")-? \right. \\
 & ; \\
 & \left. n.Attri["ori"].Val.Eq("hor")-? \right. \\
 & ; \\
 & \left( p.tA = Txml(p, n.CN[0]) \right. \\
 & ; \\
 & \left. p.tB = Txml(p, n.CN[1]) \right. \\
 & ; \\
 & \left. w = p. \right.
 \end{aligned}$$

Formula *D* from elimination by condition  $(n.Na.Eq("e"))-?$  differs that: the elimination by condition  $(n.Na.Eq("cs"))-?$  identifier *cs* is used;  $cs \in @CS$  - the creating variable *cs* subsystem *CS*, appointed for working on process of cyclic sequence and using variable *cs* instead variable *e*.

$$\begin{aligned}
 D = & \left( cs \in @CS; I; (n.Na.Eq("cs"))-? \right. \\
 & ; \\
 & \left. cs.par = t \right. \\
 & ; \\
 & \left( cs.ori = @CS.Ori.Hor \right. \\
 & ; \\
 & \left. cs.ori = @CS.Ori.Ver \right. \\
 & ; \\
 & \underline{Exc}(, Blqd\_cs \text{ xml} ") \\
 & ; \\
 & \left. n.Attri["ori"].Val.Eq("ver")-? \right. \\
 & ; \\
 & \left. n.Attri["ori"].Val.Eq("hor")-? \right. \\
 & ; \\
 & \left( cs.tA = Txml(cs, n.CN[0]) \right. \\
 & ; \\
 & \left. cs.tB = Txml(cs, n.CN[1]) \right. \\
 & ; \\
 & \left. w = cs. \right.
 \end{aligned}$$

Formula *I* from elimination by condition  $(n.Na.Eq("cs"))-?$  differs that: in comparison by condition  $(n.Na.Eq("ce"))-?$  identifier *ce* is used;  $ce \in @CE$  - the creating variable *ce* subsystem *CE*, appointed for workion on cyclic elimination operation and using variable *ce* instead variable *cs*.

$$I = \left( \begin{array}{l} ce \in @CE; G; (n.Na.Eq("ce"))-? \\ ; \\ ce.par=t \\ ; \\ ce.ori=@CE.Ori.Hor \\ ; \\ ce.ori=@CE.Ori.Ver \\ ; \\ Exc(, Blqd_cs xml") \\ ; \\ n.Attri["ori"].Val.Eq("com")-? \\ ; \\ n.Attri["ori"].Val.Eq("sem")-? \\ ; \\ ce.tA = Txml(ce, n.CN[0]) \\ ; \\ ce.tB = Txml(ce, n.CN[1]) \\ ; \\ w = ce. \end{array} \right)$$

formula G from I differs by identifier and variable cp subsystem CP, appointed for working on operation of cyclic parallelization.

$$G = \left( \begin{array}{l} cp \in @CP; Exc(, Blqd"); (n.Na.Eq("cp"))-? \\ ; \\ cp.par=t \\ ; \\ cp.ori=@S.Ori.Hor \\ ; \\ cp.ori=@S.Ori.Ver \\ ; \\ Exc(, Blqd_cp xml") \\ ; \\ n.Attri["ori"].Val.Eq("com")-? \\ ; \\ n.Attri["ori"].Val.Eq("sem")-? \\ ; \\ ce.tA = Txml(cp, n.CN[0]) \\ ; \\ ce.tB = Txml(cp, n.CN[1]) \\ ; \\ w = cp. \end{array} \right)$$

**Theorem 1.** If F is XML - formula of algorithm described by format, then formula of algorithm (1) is described the identification in F operation algebra algorithms, their orientation, uniterms separators and uniterms selection and detection of errors in XML - formulas describing algorithms.

**Proof.** Elimination by condition  $(n \neq \$) -?$  checking whether the input variable n is not empty xml - description. If the variable is empty, the elimination by this condition is obtained by uniterm  $Exc("$ _xml")$ . with information  $(\$ _xml)$  about error in the input variable n. Otherwise elimination by condition  $n.Na.Eq("u") -?$  checking whether the line of the variable n contains uniterm identifier ("u"). If yes, then after the creating of variable u type subsystem for uniterms processing U and attributing for variable par value of abstract input variable t, checking whether the variable n is not empty  $((n.IT.Le > 0) -?)$ . Not empty value of the input variable is chosen  $(n.IT)$  and is attributed to variable val  $(u.val = n.IT)$ . Then the output variable (w) is attributing  $(w = u)$  value of variable u, which is the value of uniterm (abstract or recorded by xml - format). This is ending the algorithm implementation. In case the

condition is not executed  $n.Na.Eq("u")$  -? is the checking ( $n.Na.Eq("s")$ )-? wether string of variable n contains the name of identifier sequence transaction (s).

If the identifier xml (the description of sequence operation) is recognized - is creating variable sequence operations ( $s \in @S$ ). Uniterm ( $s.par = t$ ) is attributing variable *par* value of input variable t. Elimination by condition  $n.Attr["sep"].Val.Eq("sem")$  -? compared ( $Eq("sem")$ ) wether value of attribute sep ( $n.Attr["sep"]$ ) concordant with the name of the uniterms separator (*sem*). If convergence than variable sep is attributed ( $s.sep = @S.Sep.Sem$ ) *Sem* value. When it is not the convergence , then is checking ( $n.Attr["sep"].Val.Eq("com")$ )-? wether com is separator. If yes, then the variable sep is attributing ( $s.sep = @S.Sep.Com$ ) *Com* value. Otherwise, there is a view ( $Exc("Bład_s xml")$ ) of error in the description of the separator.

After identification and recording separator in elimination by conditions  $n.Attr["ori"].Val.Eq("hor")$  -? and  $n.Attr["ori"].Val.Eq("ver")$  -? performed the same identification and recording ( $s.ori = @S.Ori.Hor$  and  $s.ori = @S.Ori.Ver$ ) orientation for sequence operation  $s.ori=@S.Ori.Ver$ .

Next from xml - description is chosen line with uniterm ( $n.CN[0]$ ), getting the uniterm using algorithm ( $Txml()$ ) and attributing it to variable *tA*. Similarly is received second uniterm (*tB*) from xml - description. The output variable w is attributed value of variable s by the last uniterm ( $w = s$ ).

Thus it is proved that the algorithm (1) describes the identification in the XML - description sequence operation, its orientation, uniterm separator and also choice of uniterms and attributing this data to output variable.

Similarly we can prove description by the algorithm (1) identification and read data from XML - description data of elimination operation, parallelization, cyclic sequence, elimination and parallelization. Theorem proved

**Theorem 2.** If *F* is a description of the algorithm in XML - format, the formula of the algorithm (2)

$$\begin{array}{l}
 \overline{p u s t (w \in @T) T x m l (t \in @T, n \in @N o d) =} \\
 \left( \begin{array}{l}
 u \in @U; \\
 ; \\
 u . p a r = t \\
 ; \\
 u . v a l = n . I T \\
 ; \\
 * \\
 ; \\
 (n . I T . L e > 0) - ? \\
 ; \\
 w = u .
 \end{array} \right) \left( \begin{array}{l}
 \exists i \\
 i \in @b_i; \overline{Exc(„Bład”)}; (n . N a . E q (“i”)) - ? \\
 ; \\
 i . p a r = t \\
 ; \\
 *; j = 1; (i \in Q_3) - ?; \\
 \varnothing j \\
 (\exists (k \in Q_2)) \\
 \overline{i . x_j = @b_i . y_j . a_{k_j}; \overline{Exc(„Bład_i”)}; k - ?; (n . A t t r i f [“x_j”] . V a l . E q (“a_{k_j}”)) - ?} \\
 ; \\
 c_j \\
 ; \\
 \varnothing z \\
 \overline{i . x_z = T x m l (i, n . C N [z]); *; ((i = e) | (z \neq 2)) - ?} \\
 ; \\
 c_z \\
 ; \\
 w = i .
 \end{array} \right) , \quad (2)
 \end{array}$$

where

$$\begin{aligned}
 i \in Q_0 &= \overbrace{s; e; p; cs; ce; cp} ; b_i \in Q_1 = \overbrace{S; E; P; CS; CE; CP} ; j, k \in Q_2 = \overbrace{0; 1} ; \\
 Q_3 &= \overbrace{s; p} ; x_j \in Q_4 = \overbrace{sep; ori} ; y_j \in Q_5 = \overbrace{Sep; Ori} ; \\
 a_{k,j} \in Q_6 &= \overbrace{\left( \begin{array}{c} \overbrace{Hor; Ver} ; t_i = \\ ; \\ \overbrace{Sem; Com} \end{array} \right)} \begin{cases} tA, \text{ jeśli } i \in Q_7 = \overbrace{s; e; p}, \\ cond, \text{ jeśli } i \in Q_8 = \overbrace{cs; ce; cp}, \end{cases} r_i = \begin{cases} tB, \text{ jeśli } i \in Q_7, \\ t, \text{ jeśli } i \in Q_8, \end{cases} \\
 x_z &= \begin{cases} t_b, \text{ jeśli } z=0; & z \in Q_9 = \overbrace{0; 1; 2} \\ r_b, \text{ jeśli } z=1; \\ cond, \text{ jeśli } z=2, \end{cases}
 \end{aligned}$$

is describing an identification in  $F$  operations algebra algorithms, their orientation, uniterm separators and uniterm choice and view of errors in  $xml$  – description of algorithm formulas.

**Proof.** Elimination by condition ( $n \neq \$$ ) -? as well as formula (1) contains an elimination by condition  $n.Na.Eq$  (" $u$ ") -? and uniterm about error ( $Exc$  (" $\$\_xml$ ").). Execution of condition goes to attributing to the output variables, as well as in formula (1), an abstract value or the readout uniterm XML - format.

Let the variable  $i$  has value  $s$ . Then  $b_i$  has the value of  $S$ . Uniterms  $i \in @ b_i$ ,  $Exc$  (" $Bład$ "). and ( $n.Na.Eq$  (" $i$ ")-?), elimination by condition ( $n.Na.Eq$  (" $i$ ")-?), are as follows:  $s \in @ S$ ,  $Exc$  (" $Bład$ "). and ( $n.Na.Eq$  (" $s$ ")-?). All of them coincide with uniterm formula (1) for operation sequence.

Uniterm  $i.par = t$  after replacement  $i$  to  $s$  has the form  $(s.par = t)$  which is the same as in formula (1). Elimination by condition ( $i \in Q_3$ ) -? gives empty uniterm (\*), which can be left out because it cannot change the formula.

The variables  $j$  and  $k$  get the value 0. Variables  $x_0, y_0, a_{0,0}$  are getting values of  $sep, Sep, Sem$ , which gives elimination by condition  $n.Attr["x_j"].Val.Eq$  (" $a_{k,j}$ ")-? and formula:

$$\left[ \begin{array}{l} s.sep = @S.Sep.Sem \\ ; \\ c_k \\ ; \\ Exc("Bład\_s\_xml"). \\ ; \\ ((0+1) \in Q_2) -? \\ ; \\ (n.Attr["sep"].Val.Eq("Sem")) -? \end{array} \right. \quad (3)$$

Condition ( $0 \in Q_2$ ) -? cyclic elimination is performed, that's why the value of variable increases by 1 and becomes an returning to the cycle by variable  $k$ . We get expressions for variables uniterms  $x_0, y_0, a_{0,0}$  operation sequence  $sep, Sep, Com$ . Substituting them to formula (3) we obtain the following expression:

$$\begin{array}{l}
 \left[ \begin{array}{l}
 s.sep=@S.Sep.Sem \\
 ; \\
 s.sep=@S.Sep.Com \\
 ; \\
 c_k \\
 ; \\
 \underline{Exc}("Blqd_s"). \\
 ; \\
 ((1+1) \in Q_2) \text{-?} \\
 ; \\
 (n.Attri["sep"].Val.Eq("Com")) \text{-?} \\
 ; \\
 (n.Attri["sep"].Val.Eq("Sem")) \text{-?}
 \end{array} \right.
 \end{array}$$

In the last expression condition  $(1+1) \in Q_2$  -? not performed because elimination by this condition can be replaced on uniterm  $\underline{Exc}("Blqd_s")$ ., which gives the formula:

$$\begin{array}{l}
 \left[ \begin{array}{l}
 s.sep=@S.Sep.Sem \\
 ; \\
 s.sep=@S.Sep.Com \\
 ; \\
 \underline{Exc}("Blqd_s"). \\
 ; \\
 (n.Attri["sep"].Val.Eq("Com")) \text{-?} \\
 ; \\
 (n.Attri["sep"].Val.Eq("Sem")) \text{-?}
 \end{array} \right. \quad (4)
 \end{array}$$

Now variable of cyclical sequence  $j$  is increased by 1, and variable operation of cyclical elimination  $k$  becomes to the initial value 0. Then for variables  $x_1, y_1, a_0$ , we obtain such a value  $ori, Ori, Hor$  and for  $a_{1,1}$  we have a  $Ver$ . Similarly, as we get the expression (4), we get the formula:

$$\begin{array}{l}
 \left[ \begin{array}{l}
 s.ori=@S.Ori.Hor \\
 ; \\
 s.ori=@S.Ori.Ver \\
 ; \\
 \underline{Exc}("Blqd_s"). \\
 ; \\
 (n.Attri["ori"].Val.Eq("Ver")) \text{-?} \\
 ; \\
 (n.Attri["ori"].Val.Eq("Hor")) \text{-?}
 \end{array} \right.
 \end{array}$$

Now the variable  $z$  cycle takes the initial value of 0 and  $x_z = t_i = tA$ . Then condition  $((s = e) \mid (0 \neq 2))$  -? elimination is performed, which allows elimination from uniterm  $s.tA = Txml(i, n.CN[0])$  and makes returning to the cycle for the variable  $z$ , which takes the value 1. The second iteration  $x_z = r_i = tB$ , and the elimination condition  $((s = e) \mid (1 \neq 2))$  -? performed, giving the uniterm  $s.tB = Txml(s, n.CN[1])$ . On the third iteration  $z = 2$ . Condition  $((s = e) \mid (2 \neq 2))$  -? of elimination is not performed. Therefore there is receiving an empty uniterm, which can be left. From the last line of formula (2) we get  $w = s$ .. Thus obtained formula is concordant with the corresponding fragment of the formula (1):

$$\left( \begin{array}{l}
 s.tA = Txml(s, n.CN[0]) \\
 ; \\
 s.tB = Txml(s, n.CN[1]) \\
 ; \\
 w = s.
 \end{array} \right)$$



Thus it is shown that formula (2) in XML - description identifies data of sequence operation. Similarly, it can be proved also for the operation of elimination, and for operations of description cycles and parallelization. Theorem is proved.

---

### **Comparison Formulas Of Algorithms**

---

As it was proven formulas (1) and (2) describe the same process of analysis XML - the description formulas of algorithms. Lets compare formulas of the algorithms (1) and (2) due to the number of uniterms. Formula (1) contains 90 uniterms, while formula (2) has only 26 uniterms. Thus formula (2) has 3.5 times uniterms less.

**Theorem 3.** From formula (2) is derived formula (1).

The proof of the theorem is similar to the proof of Theorem 2.

**Theorem 4.** From formula (1) is derived formula (2).

The proof is based on the axiom operations of elimination descriptions cycles of algebra algorithms.

---

### **5. Conclusion**

---

Description of algorithms as formulas algorithms provides performance identical transformations of algorithms, which reduced expenses needed for implement algorithms.

---

### **Bibliography**

---

[Ovsyak et al., 2011] V. Ovsyak. Computation models and algebra of algorithms. Submitted to the Conference.

[Petzold, 2002] C. Petzold. Programming Microsoft Windows with C#, 2002.

[MacDonald, 2008] M. MacDonald. Pro WPF in C# 2008 Windows Presentation Foundation with .NET 3.5.

---

### **Authors' Information**

---



**Volodymyr Ovsyak** – Full Professor, Department of Electrical, Control and Computer Engineering, University of Technology, Box: 31, Sosnkowskiego, Opole 45-272, Poland, e-mail: [ovsyak@rambler.ru](mailto:ovsyak@rambler.ru)

*He specializes in theoretical and applied computer science, theory of algorithms, programming, information systems, mathematical modeling*



**Krzysztof Latawiec** – Full Professor, Department of Electrical, Control and Computer Engineering, University of Technology, Box: 31, Sosnkowskiego, Opole 45-272, Poland, e-mail: [lata@po.opole.pl](mailto:lata@po.opole.pl)

*His research interests concentrate on system identification, multivariable control, adaptive and robust control (also in networks) and fractional systems.*



**Aleksandr Ovsyak** – Phd. National University of Culture and Arts, The L'viv Campus, Box: 5, Kuszewicza, L'viv, Ukraine; e-mail: [ovsjak@ukr.net](mailto:ovsjak@ukr.net)

*He specializes in theoretical and applied computer science, theory of algorithms, programming, information systems, mathematical modeling of systems, computer simulation and mathematical modeling.*

---

---

## MODELLING AND CONTROL OF COMPUTATIONAL PROCESSES USING MAX-PLUS ALGEBRA

Jerzy Raszka, Lech Jamroż

**Abstract:** *In this paper we propose a modelling technique for the control of computational processes. The Petri Net model, particularly a Timed Event Graph (TEG) can be used for analyzing. The proposed model enables the determination of state equations. The max-plus algebra represents linear algebraic form of discrete systems and supplies new tools to their modelling. We develop a linear mathematical model under constraints in the Max-plus algebra. When using max-plus algebra with TEG, the arc weights are kept equal to one in order to be able to resolve the state equations. Structure of max-plus algebra is equipped with maximization and addition operations over of the real numbers and minus infinity. It can be used appropriately to determine marking times within a given Petri net and a vector filled with marking state at the beginning. Tools of max-plus algebra are useful to investigate properties of network models. Finally, numerical examples show the use of this model.*

**Keywords:** *Max-Plus-Linear Systems, Petri Nets, Discrete Systems, Modelling Technique*

**ACM Classification Keywords:** *H. Information Systems, H.1 MODELS AND PRINCIPLES, H.1.1 Systems and Information Theory; D. Software D.4, OPERATING SYSTEMS, D.4.1 Process Management, Multiprocessing/multiprogramming/multitasking, Synchronization.*

---

### Introduction

Recent technological achievements require advances beyond the existing computational models in order to be used effectively. For example the Internet has progressed from a simple store-and-forward network to a more complex communication infrastructure. In order to meet demands on security, flexibility and performance, network traffic not only needs to be forwarded, but also processed on routers. Pragmatic aspects of current and future computer systems will be modelled so that realistic estimates of efficiency can be given for algorithms and controlling of computations in these new settings. Proposed methods deals with the performance evaluation of a communication infrastructure system in terms of waiting times of data and starting points of computing in various connections. The behaviour of a system is studied in the framework of discrete systems

Discrete systems and specially discrete-event dynamical systems often arise in the context of parallel computing, manufacturing systems [Jamroż, Raszka 1997], for project management, railway networks [Goverde, Rob 2007], telecommunication networks, etc. In the last years there has been a growing quantity of research on discrete systems that can be modelled as max-plus linear systems. Most of the earlier literature on this class of systems set included modelling, performance and properties analysis, rather than control [Bacceli, Cohen and alt. 1992], [Jamroż, Raszka 1997], [Nait-Sidi-Moh, Manier 2009] and many others e.g. J. Bernd Heidergott, Geert Jan Olsder, Didier Dubois, Jean-Pierre Quadrat and Jacob van der Woude. Lately there are articles on control for max plus systems [Maia, Andrade 2011] [Nait-Sidi-Moh, Manier 2009], [Schutter, Boom 2001].

Two possible operating modes of discrete systems can be observed at each operating: periodic and no periodic mode. Two complementary tools, Petri nets and  $(\max, +)$  algebra, are used to describe the network by a linear state model. This one can be solved after solving the structural conflicts associated to the graphical

representation. From the characteristic matrix of the mathematical model, we may determine eigenvalues and eigenvectors that we use to evaluate the operations.

Petri nets (PN) [Murata 1989] are very popular formalism for the analysis and representation of parallel and distributed computing in concurrent systems that has draw much attention to modelling and verification of this type of systems. P systems, also referred to as membrane systems [Gutuleac 2006], are a class of parallel and distributed computing models [6]. The interest of relating P systems with the PN model of computation lead to several important results on simulation and decidability issues. Some efforts have been made to simulate P systems with Petri nets to verifying the many useful behavioural properties such as reachability, boundedness, liveness, terminating, etc.

When considering processes from manufacturing or chemical engineering, their behaviour can often be adequately represented by a discrete event model accounting for the typically discrete sensor and actuator equipment of such processes. In addition, the behaviour of these processes is often adequately described by a sequence of transitions between discrete process states. The focus of this contribution is on a particular class of such discrete event systems where synchronization and controlling of computational processes occurs. This system class has gained significant attention in recent years due to the fact that the sequences of event times for such processes can be described by equations which are linear in a particular algebra, the so called max-plus algebra [Baceli Cohen 1992]. The resulting equations exhibit a structural equivalence to system descriptions from conventional control engineering such as transfer functions or state space models.

Thus, a system theory for these max-plus-linear systems has been developed, and various concepts well known from control engineering have been adapted to this system class in control design and diagnosis.

Sometimes for modelling Timed event graphs (TEG's) are a subclass of timed Petri nets which can be used for modelling Discrete Event Dynamic Systems (DEDS) subject to synchronisation phenomena as manufacturing systems, multiprocessor systems and especially transportation.

In this paper, we propose a deterministic Petri net model for the computational system that can be considered as a discrete event system. Moreover, such DES can be easily modelled with a subclass of Petri net for evaluation purposes, we suggest then a TEG approach to model, analyse and control a computational process From this TEG model, we formulate a mathematical model based on the max-plus algebra. The behaviour of this DES can be described easily by a linear system in this algebra. In the second part, we introduce model of computational process. The third part presents an overview of max algebra theory and the specific model will be analysed. In the last part we present the simulations results.

---

## **Computational Processes**

---

When considering computational processes that allows event-driven applications to take advantage of multiprocessors by running code for event handlers in parallel. To obtain high performance, servers must overlap computation with I/O. Programs typically achieve this overlap using threads or events. Threaded programs typically process each request in a separate thread; when one thread blocks waiting for I/O, other threads can run. Event-based programs are structured as a collection of call-back functions which a main loop calls when I/O events occur. Threads provide an intuitive programming model, but require coordination of accesses by different threads to shared state, even on a uniprocessor. Event-based programs execute call-backs serially, so the programmer need not worry about concurrency control; however, event-based programs until now have been unable to take advantage of multiprocessors. Much of the effort required to make existing event-driven programs take advantage of multiprocessors is in specifying which events may be handled in parallel.

As example in this paper is consider the simple problem of designing control of system where the cost is chosen such that it provides a trade off between minimizing delays of time of end of operations of computational process (real time of complete of all tasks in computational periodic process, times of final results of one cycle) and periodicity of desired output (wished needed time) to complete process

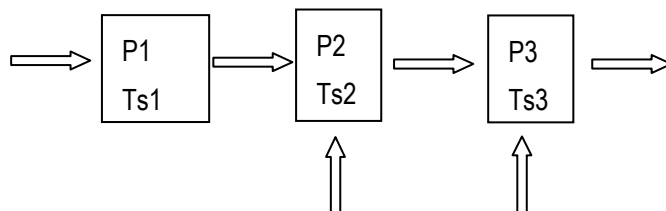


Fig. 1. Structure of the process

Simple computational process consist several tasks which linked by the waiting for I/O data (Figure 1). To illustrate our approach, we consider here a process that is constituted by three tasks: Ts1, Ts2, Ts3 which run on three processors: P1, P2 and P3. Every one of these tasks is operated on the dedicated processor. Within this process circulate information flows in several directions; these can be processing data input/output and signal data. Outer input data are processed as a first task on the P1 and its output data have to be saved in the memory waiting to be processed. The second and third processors operate in the same way but they input data are as output results from the first and second processor respectively. Moreover tasks Ts1 and Ts2 need an extra outer data.

The aim of this modelling is to evaluate command for the process according to pre-established criteria. For instance, to have a continuous computation (on processor P3, we have to prepare time for input data to other processors P1 and P2 also enough of memory to save an input, output and temporary data. A good schedule will allow maintaining a minimal costs and optimising the size of memory required.

---

## Petri Net Model

---

Petri nets, a graph-oriented formalism, allow to model and analyze systems, which comprise properties such as concurrency and synchronization.

A Petri net model of a dynamic system consists of two parts: net structure and marking. A net structure is a weighted-bipartite directed graph that represents the static part of the system. A marking is representing a distributed overall state on the structure. This separation allows one to reason on net based model at two levels - structural and behavioral.

Net structure is built on two disjoint sets of objects: places and transitions, which are connected by arcs. In the graphical representation, places are drawn as circles, transitions are drawn as thin bars and arcs are drawn as arrows. Places may contain tokens, which are drawn as dots. The vector representing in every place the number of tokens is the state of the Petri net and is referred to as its marking. This marking can be changed by the firing of the transitions. A Petri nets do not include any notion of time are aimed to model only the logical behavior of systems. The introduction of a timing specification is essential if we want to use this class of model to consider performance problem.

More formally timed Petri nets (TPN) are 5-tuples [Murata 1989] : $TPN = (P, T, F, M_0, \tau)$ , where  $P = \{p_1, p_2, \dots, p_n\}$ ,  $|P| \neq 0$ ;  $T = \{t_1, t_2, \dots, t_m\}$ ,  $|T| \neq 0$  is a finite disjunct set of suitable places and transitions;  $M_0: P \rightarrow N$  is the initial marking function which defines the initial number of tokens for every place. ( $N = \{0, 1, \dots\}$ );  $\tau: T \rightarrow R^+$  is

the firing time function, and  $F \subset (PxT) \cup (TxP)$  is the set of arcs. The problem evoked above can be modelled by a Petri Net (Figure 2).

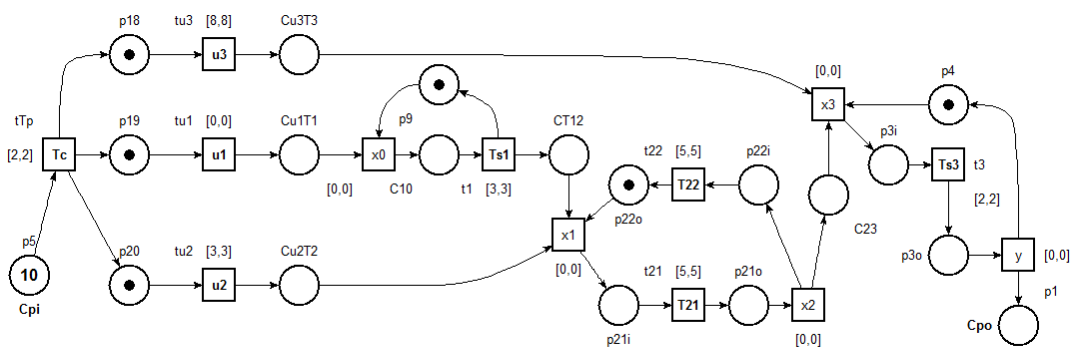


Fig. 2. Timed Petri net for simulation of the system process

The net has three inputs  $u_1$ ,  $u_2$ , and  $u_3$  and one output  $y$ . The firing time  $u_1$ ,  $u_2$ , and  $u_3$  are the start times of task  $Ts_1$ ,  $Ts_2$ ,  $Ts_3$  respectively. Finally, the ending time of the process is represented by the firing time of the transition  $y$ .

### Max-Plus Algebra

#### Define

The basic operations of the max-plus algebra [Bacceli Cohen 1992] are maximization and addition, which will be represented by  $\oplus$  and  $\otimes$  respectively:  $x \oplus y = \max(x, y)$  and  $x \otimes y = x + y$  for  $x, y \in \square_\varepsilon = \text{def } \square \cup \{-\infty\}$ . The reason for using these symbols is that there is a remarkable analogy between  $\oplus$  and conventional addition, and between  $\otimes$  and conventional multiplication: many concepts and properties from linear algebra (such as the Cayley-Hamilton theorem, eigenvectors and eigenvalues, Cramer's rule, ...) can be translated to the max-plus algebra by replacing  $+$  by  $\oplus$  and  $\times$  by  $\otimes$ . Therefore, we also call  $\oplus$  the max-plus-algebraic addition, and  $\otimes$  the max-plus-algebraic multiplication. Note however that one of the major differences between conventional algebra and max-plus algebra is that in general there do not exist inverse elements w.r.t.  $\oplus$  in  $\square_\varepsilon$ . The zero element for  $\oplus$  is  $\varepsilon = \text{def } -\infty$  we have  $a \oplus \varepsilon = a = \varepsilon \oplus a$  for all  $a \in \square_\varepsilon$ . The structure  $(\square_\varepsilon, \oplus, \otimes)$  is called the max-plus algebra.

Let  $r \in \mathbb{Z}$ . The  $r$ -th max-plus-algebraic power of  $x \in \square_\varepsilon$  is denoted by  $x^{\otimes r}$  and corresponds to  $rx$  in conventional algebra. If  $r \in \mathbb{Z}$  then  $x^{\otimes 0} = 0$  and the inverse element of  $x$  w.r.t.  $\otimes$  is  $x^{\otimes -1} = -x$ . There is no inverse element for  $\varepsilon$ , since  $\varepsilon$  is absorbing for  $\otimes$ . If  $r > 0$  then  $\varepsilon^{\otimes r} = \varepsilon$ . If  $r < 0$  then  $\varepsilon^{\otimes r}$  is not defined. In this paper we have  $\varepsilon^{\otimes 0} = 0$  by definition.

The rules for the order of evaluation of the max-plus algebraic operators correspond to those of conventional algebra. So max-plus-algebraic power has the highest priority, and max-plus-algebraic multiplication has a higher priority than max-plus-algebraic addition.

#### Max-plus-algebraic matrix operations

The basic max-plus-algebraic operations are extended to matrices as follows. If  $A, B \in \square_\varepsilon^{m \times n}$  and  $C \in \square_\varepsilon^{m \times p}$  then:

$$(A \oplus B)_{ij} = a_{ij} \oplus b_{ij} = \max(a_{ij}, b_{ij})$$

$$(A \otimes C)_{ij} = \bigoplus_{k=1}^n a_{ik} \otimes b_{kj} = \max_k(a_{ik}, b_{ki})$$

for all  $i, j$ . Note the analogy with the definitions of matrix sum and product in conventional linear algebra.

The matrix  $\mathcal{E}_{m \times n}$  is the  $m \times n$  max-plus-algebraic zero matrix:  $(\mathcal{E}_{m \times n})_{ij} = \varepsilon$  for all  $i, j$ ; and the matrix  $E_n$  is the  $n \times n$  max-plus-algebraic identity matrix:  $(E_n)_{ii} = 0$  for all  $i$  and  $(E_n)_{ij} = \varepsilon$   $i, j$  with  $i \neq j$ . If the size of the max-plus-algebraic identity matrix or the max-plus-algebraic zero matrix is not specified, it should be clear from the context. The max-plus-algebraic matrix power of  $A \in \square_{\varepsilon}^{n \times n}$  is defined as follows:  $A^{\otimes 0} = E_n$  and  $A^{\otimes k} = A \otimes A^{\otimes k-1}$  for  $k = 1, 2, \dots$

## Model of processes

### Investigation

The intend of this study is to show that, and discuss how, the process satisfying the above assumptions can be modelled in max-plus algebra, to determine the input vector  $u(k)$  for knowing values of  $y(k)$ , evaluate the error between real and desired output and estimate cost of the resources.

Let  $u(k) = [u_1, u_2, u_3]^T$  the input vector,  $x(k) = [x_1, x_2, x_3]^T$  the state vector and  $y(k)$  the model output. For each transition  $x_i$ , and  $u_i$  is associated an indicator  $x_i(k)$  and  $u_i(k)$  responsibly, which corresponds to the steps of the  $k$ 'th firing of transition  $x_j$  (resp.  $u_j$ ), and in the same way we have  $y(k)$ . The state system in the max plus algebra is follow:

$$\begin{aligned} x_1(k) &= t_1 \otimes u_1(k) \oplus u_2(k) \oplus t_{22} \otimes x_2(k-1) \oplus t_1 \otimes x_1(k-1) \\ x_2(k) &= t_{21} \otimes x_1(k) \\ x_3(k) &= x_2(k) \oplus u_3(k) \oplus y(k-1) \end{aligned}$$

For example the  $k$ 'th firing of transition  $x_1$  (when start Ts1), must wait of  $t_1$  units of time until that the  $k$ 'th input data  $u_1$  for task Ts1 is ready and  $k$ 'th input data  $u_2$  is given. Then, the linear equation of evolution in  $\square_{\varepsilon}$  of this discrete event dynamic system is as follows:

$$x(k) = A_0 \otimes x(k) \oplus A_1 \otimes x(k-1) \oplus B_0 \otimes u(k) \oplus D_0 \otimes y(k-1) \quad (1)$$

Where

$$A_0 = \begin{bmatrix} \varepsilon & \varepsilon & \varepsilon \\ t_{21} & \varepsilon & \varepsilon \\ \varepsilon & \mathbf{e} & \varepsilon \end{bmatrix}, \quad A_1 = \begin{bmatrix} t_1 & t_{22} & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon \end{bmatrix}, \quad B_0 = \begin{bmatrix} t_1 & \mathbf{e} & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \mathbf{e} \end{bmatrix}, \quad D_0 = \begin{bmatrix} \varepsilon \\ \varepsilon \\ \mathbf{e} \end{bmatrix}.$$

A solution of (1), is

$$x(k) = A_0^* \otimes (A_1 \otimes x(k-1) \oplus B_0 \otimes u(k) \oplus D_0 \otimes y(k-1))$$

Where

$$A_0^* = E \oplus A_0 \oplus A_0^2 \oplus A_0^3 \dots$$

$$A_0^* = \begin{bmatrix} e & \varepsilon & \varepsilon \\ t_{21} & e & \varepsilon \\ t_{21} & e & e \end{bmatrix}, \quad A = A_0^* \otimes A_1 = \begin{bmatrix} \varepsilon & t_{22} & \varepsilon \\ \varepsilon & t_{22} \otimes t_{21} & \varepsilon \\ \varepsilon & t_{22} \otimes t_{21} & \varepsilon \end{bmatrix}.$$

$$B = A_0^* \otimes B_0 = \begin{bmatrix} t_1 & e & \varepsilon \\ t_1 \otimes t_{21} & t_{21} & \varepsilon \\ t_1 \otimes t_{21} & t_{21} & e \end{bmatrix}, \quad D = A_0^* \otimes D_0 = \begin{bmatrix} \varepsilon \\ \varepsilon \\ e \end{bmatrix}.$$

Then, the model becomes:

$$x(k) = A \otimes x(k-1) \oplus B \otimes u(k) \oplus D \otimes y(k-1) \quad (2)$$

By recurrence we obtain the expression:

$$x(k) = A_0^{k-1} \otimes x(1) \oplus \sum_{i=2}^k A_0^{k-i} \otimes B \otimes u(i) \oplus \sum_{i=1}^{k-1} A_0^{k-i-1} \otimes D \otimes y(i) \quad (3)$$

To determine the command for the process, we have to define at first the whole order-1 model, which describes its global behaviour:

$$\begin{cases} x(k) = A \otimes x(k-1) \oplus B \otimes u(k) \oplus D \otimes y(k-1) \\ y(k) = C \otimes x(k) \\ u(1) = [u_1(1) \quad u_2(1) \quad u_3(1)]^T \\ x(1) = [x_1(1) \quad x_2(1) \quad x_3(1)]^T \end{cases} \quad (4)$$

Where  $C = [\varepsilon \quad \varepsilon \quad t_3]$ ,  $u(1)$  and  $x(1)$  are the initial conditions that we are going to determine below.

### Initials conditions

To respect the previous assumptions, we determine the initial values of  $u(1)$  and  $x(1)$  according to the end process  $y(1)$ . We suppose that initially there is at least one freight vehicle ready for departure:

$$y(1) = t_1 \otimes t_{21} \otimes t_3 \otimes u_1(1) = t_{21} \otimes t_3 \otimes u_2(1) = t_3 \otimes u_3(1)$$

Then the initial control is:

$$\begin{bmatrix} u_1(1) \\ u_2(1) \\ u_3(1) \end{bmatrix} = \begin{bmatrix} y(1)\phi(t_1 \otimes t_{22} \otimes t_3) \\ y(1)\phi(t_{22} \otimes t_3) \\ y(1)\phi(t_1 \otimes t_{22} \otimes t_3) \end{bmatrix}. \quad (5)$$

Where " $\phi$ " means "-", in  $\square_\varepsilon$ .

Consequently, the initial value of the state vector is

$$\begin{bmatrix} x_1(1) \\ x_2(1) \\ x_3(1) \end{bmatrix} = \begin{bmatrix} t_1 \otimes u_1(1) \\ t_{21} \otimes u_2(1) \\ u_3(1) \end{bmatrix}. \quad (6)$$

These two initial vectors mean that, if the task Ts1 begins for instance at  $t = 0$ ,  $t_1$  units time later, the data for task Ts2 ( $u_2$ ) is need be prepared and this task begins. This ensures a good starting up without delay for the evolution of the model.

### Procedure

As indicated first, the network operates under a schedule defined for final result; according to this schedule we find the suitable inputs of the model. We formulate the model output more explicitly as:

$$y(k) = C \otimes A_0^{k-1} \otimes x(1) \oplus \sum_{i=2}^k C \otimes A_0^{k-1} \otimes B \otimes u(i) \oplus \sum_{i=1}^{k-1} C \otimes A_0^{k-i-1} \otimes D \otimes y(i) \quad (7)$$

or

$$y(k) \geq \max \{ C \otimes A_0^{k-1} \otimes x(1), \sum_{i=2}^k C \otimes A_0^{k-1} \otimes B \otimes u(i), \sum_{i=3}^{k-1} C \otimes A_0^{k-i-1} \otimes D \otimes y(i) \}$$

which is equivalent to:

$$y(k) \geq C \otimes A_0^{k-1} \otimes x(1) \quad (8)$$

$$y(k) \geq \sum_{i=2}^k C \otimes A_0^{k-1} \otimes B \otimes u(i) \quad (9)$$

$$y(k) \geq \sum_{i=3}^{k-1} C \otimes A_0^{k-i-1} \otimes D \otimes y(i) \quad (10)$$

We are interested rather in the second in equation (9) that we transform on its equation form:

$$y(k) = \sum_{i=2}^k C \otimes A_0^{k-1} \otimes B \otimes u(i) \quad (11)$$

It is straightforward now that from (11) we can formulate the command  $u(k)$  for process if the values of the output  $y(k)$  are known. For all the rest  $y(k)$  will be the desired of the final result.

For  $k = 2, 3, 4, \dots$ :

$$y(2) = C \otimes B \otimes u(2)$$

$$y(3) = C \otimes A^1 \otimes B \otimes u(2) \oplus C \otimes B \otimes u(3) \quad (12)$$

$$y(4) = C \otimes A^2 \otimes B \otimes u(2) \oplus C \otimes A^1 \otimes B \otimes u(3) \oplus C \otimes B \otimes u(4)$$

...

Since our aim is to compute  $u(k)$  for specified  $y(k)$ , we have to solve the following equation:

$$y(k) = C \otimes B \otimes u(k) \quad (13)$$

For example, to calculate  $u(2)$ , a solution of  $y(2) \geq C \otimes B \otimes u(2)$ , we resolve its equation form (12) and we keep its smallest solution to be sure that it is verifies also the in equation. Note here that we proceed by a simplification of the terms such " $C \otimes A^2 \otimes B \otimes u(2) \dots C \otimes A^1 \otimes B \otimes u(k-1)$ " in the expression of  $y(k)$ . Indeed, these terms constitute the first condition of the desired output  $y(k)$ . More explicitly, if we consider that all expression  $y(k)$  are equal to  $C \otimes B \otimes u(k)$ , we must then assure that:

$$C \otimes A^2 \otimes B \otimes u(2) \oplus \dots \oplus C \otimes A^1 \otimes B \otimes u(k-1) \leq C \otimes B \otimes u(k)$$

Which means that we must have?

$$y(k) \geq t_1 \otimes t_{21} \otimes t_3 \otimes u_1(k-1) \oplus t_{21} \otimes t_3 \otimes u_2(k-1) \oplus t_3 \otimes u_3(k-1)$$



### Periodic processing

We assumed that we wish the output of final results of the process be once every time interval  $T_c$ . We shall see later how the network reacts according of various value of  $T_c$ .

Let  $y(k) \geq T_c^k \otimes y(0)$  where  $T_c$  is the periodicity of desired output. We use in our computations this condition in the following form:

$$T_c \geq \max(t_1 + t_{21} + t_3 + u_1(k-1), t_{21} + t_3 + u_2(k-1), t_3 + u_3(k-1)) / k$$

Solve the equations (11) and determine the control vector as follows:

$$u_j(k) = y(k) \phi(C \otimes B)_{i,j}; \quad k = 2, 3, \dots; \quad i = 1 \text{ and } j = 1, 2, 3.$$

Where:

$$C \otimes B = [\varepsilon \quad \varepsilon \quad t_3] \otimes \begin{bmatrix} t_1 & e & \varepsilon \\ t_1 \otimes t_{21} & t_{21} & \varepsilon \\ t_1 \otimes t_{21} & t_{21} & e \end{bmatrix} = [t_1 \otimes t_{21} \otimes t_3 \quad t_{21} \otimes t_3 \quad t_3]$$

More explicitly, for every  $k \in N / \{1\}$ , the general solutions are:

$$\begin{aligned} u_1(k) &= y(k) \phi(C \otimes B)_{1,1} \\ u_2(k) &= y(k) \phi(C \otimes B)_{1,2} \\ u_3(k) &= y(k) \phi(C \otimes B)_{1,3} \end{aligned} \tag{14}$$

These equations determine the appropriate control of the modelled process. On the other hand, (8) and (10) contain two constraints for the desired outputs of the system. While  $T_c$  is the periodicity of these outputs,  $y(k) \geq T_c^k \otimes y(1) = T_c^k$  the first constraint which ensues from (8) is:

$$y(k) = T_c^k \geq C \otimes A^{k-1} \otimes x(1) \quad \text{or} \quad T_c^k \geq t_3 \otimes (t_1 \otimes t_{21})^{k-1} \otimes x_2(1)$$

In practice,  $T_c \geq (t_2 + (t_1 + t_{21})(k-1) + x_2(1)) / k$  means that periodicity must be superior at certain value in order to have a good control.

The second constraint of  $T_c$  becomes from (10):

$$y(k) = T_c^k \geq \sum_{i=3}^{k-1} C \otimes A^{k-i-1} \otimes D \otimes y(i), \text{ this constraint is verified all time since the product } C \otimes A^{k-i-1} \otimes D$$

is null. To conclude this section we describe the preceding steps in the following algorithm:

- 1- Determine the state equations in max plus algebra as (4).
- 2- Calculate the global recurrence equation of the linear evolution of the system.
- 3- Determine initials conditions.
- 4- Calculate constraints of desired model output.
- 5- Calculate control vector using max plus algebra operations.

### Simulation and Results

For the simulation of the model of process we use a fixed interval  $T_c$  for desired outputs, we have then:  $y(k) = y(k-1) \otimes T_c$ ;  $k = 2, 3, \dots$ . Using (14), we can compute the vectors  $u(k)$  and we consider the initial values. Interested to the state of the places C12, C23 (Figure 2) which represents intermediate buffered data and

cost of its maximal length and calculate the error between wished outputs and real outputs. Figures 3 - 6 show time evolution interpolated values of count of processes - particularly finished last tasks of following processes and generated input tasks appointed by signs blue  $\blacktriangledown$  and red  $\blacklozenge$  respectively.

All the following examples results are obtained for various values of  $T_C$  and  $t_1$  (time of operations of task Ts1). In first example, computations have periodicity equal to 10. This value is enough large to contain all tasks of process and no error between wished and real outputs occur. Some differences in the following examples, can be reduced only by increasing resources - processors and / or memory units. Figures 8 and 9 show the time dependence of markers in places that represent changes in the allocation of memory (buffers).

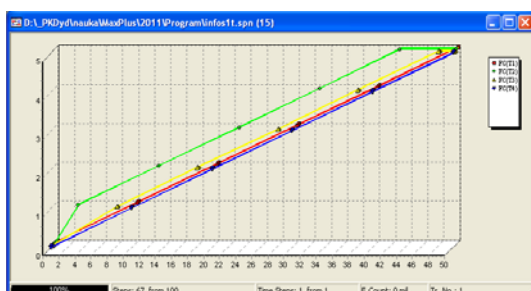


Fig. 3.  $T_C=10$ ,  $t_1=3$



Fig. 4.  $T_C=15$  and  $t_1=3$

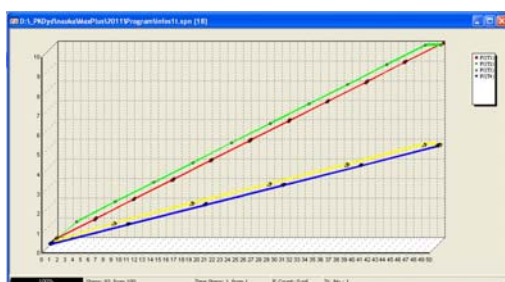


Fig. 5.  $T_C=5$  and  $t_1=3$



Fig. 6.  $T_C=10$  and  $t_1=13$

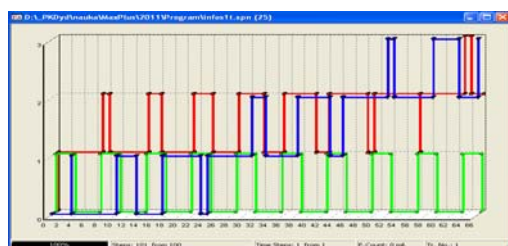


Fig. 7.  $T_C=7$ , Places P7,  $\blacktriangledown$  P8,  $\blacklozenge$

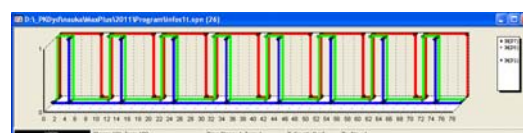


Fig. 8.  $T_C=10$  and  $t_1=3$ , Places P7,  $\blacktriangledown$  P8,  $\blacklozenge$

## Conclusion

Engineers who build discrete event systems have to confront dynamical problems as a matter of course. For the most part, they have had little mathematical support to do this, despite the considerable understanding of dynamical systems arising from classical methods. This article proposed and introduced max plus algebra - a new methodology which used to modelling and simulating discrete event processes. A control of tasks of process on simple multi-processors computational system is using as an example. This is only a part of the carried out studies, which require additional testing and even wider range of experience, especially practical applications. Further more, have to develop a way to control this processes and studied conditions for periodicity of the required results. An other research will include the application to larger models and improvements of the

optimisation procedure with respect to its efficiency. Specially topic for future research include all over methods of designing, synthesis controls of processes with output and/or state feedback and using models of predictive and adaptive control.

---

**Bibliography**

---

- [Baceli Cohen 1992] Baceli F.L.,Cohen G., Olsder G. J., Quadrat J.P.: Synchronization and Linearity. An Algebra for Discrete Event Systems, London, John Wiley & Sons Ltd, 1992.
- [Balduzzi 2000] F. Balduzzi, Has. Giua and G. Menga. "First-order hybrid Petri net: In model for optimisation and control" IEEE trans. On Rob. And Aut. 16 (4):382-399,2000.
- [Cassandras 2007] Cassandras Ch.G., Lafortune St.: Introduction to Discrete Event Systems Springer 2008, Kluwer Academic Publishers 2007
- [Elmahi 2004] Elmahi I., Grunder O., Elmoudni A.: A max plus algebra approach for modeling and control of lots delivery. Industrial Technology, 2004. IEEE ICIT '04. 8-10 Dec. 2004 p. 926- 931 Vol. 2
- [Goverde, Rob 2007] Goverde Rob M.P.: Railway timetable stability analysis using max-plus system theory, Elsevier, Transportation Research Part B 41, 2007, p. 179–201
- [Jamroż, Raszka 1997] Jamroż L., Raszka J.: Simulation method for the performance evaluation of system of discrete cyclic processes. Proceedings of the 16-th IASTED International Conference on Modelling, Identification and Control, Innsbruck, Austria, 17-19.02.97, p. 190-193
- [Maia, Andrade 2011] C.A. Maia, C.R. Andrade and L. Hardouin On the control of max-plus linear system subject to state restriction Automatica, Volume 47, Issue 5, May 2011, Pages 988-992 C.A. Maia, C.R. Andrade and L. Hardouin
- [Murata 1989] Murata T.: Petri nets: properties, analysis and applications. Proceedings of the IEEE, vol. 77, no. 4, p.541-580, 1989.
- [Gutuleac 2006] Emilian Gu\_tuleac Descriptive Timed Membrane Petri Nets for Modelling of Parallel Computing International Journal of Computers, Communications & Control, Vol. I (2006), No. 3, pp. 33-39
- [Nait-Sidi-Moh, Manier 2009] A Nait-Sidi-Moh, M -A Manier, A El Moudni; Spectral analysis for performance evaluation in a bus network European Journal of Operational Research Volume 193: Issue 1. Pages 289-302 16 February 2009
- [Schutter, Boom 2001] De Bart De Schutter, Ton van den Boom Model predictive control for max-plus-linear discrete event systems. Automatica 37(7):1049-1056, July. 2001

---

**Authors' Information**

---

**Jerzy Raszka** *PhD, Institute of Computer Science, Faculty of Physics, Mathematics and Computer Science, Tadeusz Kościuszko Cracow University of Technology Warszawska 24 St., 31-155 Kraków*

tel.:+48 12 628 27 08, e-mail: [jraszka@mail.pk.edu.pl](mailto:jraszka@mail.pk.edu.pl)

**Lech Jamroż** – *PhD, Institute of Computer Science, Faculty of Physics, Mathematics and Computer Science, Tadeusz Kościuszko Cracow University of Technology Warszawska 24 St., 31-155 Kraków*

tel.:+48 12 628 27 08, e-mail: [ljamroz@mail.pk.edu.pl](mailto:ljamroz@mail.pk.edu.pl)

---

## Intelligent Agents and Multi-Agent Systems

---

### TACTICAL MANAGEMENT OF SUPPLY CHAIN WITH AGENT BASED MODELING AND SIMULATION

Jacek Jakiela, Paweł Litwin, Marcin Olech

**Abstract:** *Excellence in Supply Chain Management depends on timely and effective translation of market demand into material and products control decisions across Supply Chain. As has been observed many times, this task may be complicated by several technical and business related constraints. Effective decision making processes are essential to execute Supply Chain in such a way that products will be made at lowest possible cost and delivered on time. In most cases, software support has to be used. Supply Chain related decisions are usually poorly structured and therefore only reasonable way to support them is to use simulation tools. The main aim of the paper is to show how Supply Chain Management process can be supported with Agent Oriented Simulation Platform. The case study presented describes in detail the decision situation that is analyzed from scratch, by simulating different scenarios for suggested problem solutions. Finally the paper shows how the analysis of simulation results may lead to final decision solving the problems encountered.*

**Keywords:** *Agent-based Models, Agent oriented Supply Chain Management, Agent-Based Simulation and Modeling, Agent-orientation as Modeling Paradigm*

**ACM Classification Keywords:** *1. Computing Methodologies; 1.2 Artificial Intelligence; 1.2.11 Distributed Artificial Intelligence; Multi-Agent Systems*

---

### Introduction

---

Nowadays business is conducted in the networked world. Increasing global and competitive marketplace forces enterprises to work together to achieve individual as well as collective goals in more effective and efficient ways. Business partners are not isolated. They operate as nodes in a network of suppliers, products warehouses and specialized service functions. Such network, called Extended Enterprise, Virtual Enterprise or Supply Chain has to be agile enough to rebuild and adjust plans and make decisions in real time to take care of unexpected events. The agility is related to effective Supply Chain Management which crucial part is decision making process. The general areas where the decisions are made address demand planning, master planning, procurement, manufacturing and transportation management. For many companies it has become a matter of survival to improve their decision making processes. What is more firms are trying to apply software support to decision situations. This support may take form of full-blown ERP system but, what is done more often, especially for semi-structured and unstructured decisions, it is used in the form of simulation workbench. Simulation tools may aid human decision maker to make right decision by providing information in proper context and related to whole Supply Chain. Thanks to well-designed simulation experiment, decision makers are able to understand the overall Supply Chain business logic and characteristics, capture system dynamics related to unexpected events as well

as their influence on Supply Chain, and conduct what-if analysis enabling to minimize risk associated with changes in the planning process.

As has been shown in [Nfaoui *et al.*, 2006] and [Kimbrough *et al.*, 2002] such business structures as Supply Chains require particular approach. In order to have all important characteristics of Supply Chain fully captured and included in the simulation model, agent-based approach is usually taken [Paolucci *et al.*, 2005], [North *et al.*, 2007]. This way of problem domain conceptualization has also been adopted in this paper, which main aim is to show how Agent Based Modeling and Simulation approach can be used for supporting decision making processes related to Supply Chain Management. Presented case study is based on simulation experiment executed on Agent Oriented Platform designed and developed by authors of this paper and described in several articles. In [Jakiela, 2006] the basic assumptions related to using agent orientation as a modeling paradigm for contemporary organizations were shown. The rationale for the research currently conducted has been presented in [Jakiela *et al.*, 2009]. Papers [Jakiela, Litwin, Olech, 2010a], [Jakiela, Litwin, Olech, 2010b], [Jakiela, Litwin, Olech, 2011b] describe the partial results of the research in the form of reference model for the simulation platform and its applications to bullwhip effect analysis. Finally, the paper [Jakiela, Litwin, Olech, 2011a] shows how the simulation platform developed may be used as a Supply Chain analysis workbench.

---

### **The Concept of Supply Chain Management**

---

According to Muckstadt *et al.*, Supply Chain term is defined as the set of firms acting to design, engineer, market, manufacture and distribute products and services to end-consumers. This set of firms is structured as a network with every firm operating as a node [Muckstadt *et al.*, 2001]. Such business structure is quite complex and its behavior often unpredictable. Therefore the properly tuned management process should be applied. This process is called Supply Chain Management and as a popular term has started to be used in early 1990s.

According to Chang *et al.* it is a process of coordinating activities of suppliers, manufacturers, warehouses and retailers in order to minimize costs and satisfy customer requirements [Chang *et al.*, 2001]. Supply Chain Management may also be defined as a set of approaches utilized to efficiently integrate suppliers, manufacturers, warehouses and stores, so that merchandise is produced and distributed in the right quantities, to the right locations, in the right time, in order to minimize system wide costs, while satisfying service level requirements.

In Supply Chain Management decisions may be classified as strategic, tactical and operational. As their name suggests, strategic decisions are related to firm's strategy, are usually long-term and involve most partners in Supply Chain. Tactical decisions are mid-term and made in individual area of Supply Chain – by specific business partner. The main problems addressed by these types of decisions are related to demand, procurement, production warehousing and distribution. These decisions are usually semi-structured, what means that they do not have predefined procedures used for making them. Operational decisions are related to day-to-day operations [Chang *et al.*, 2001].

Although all these decisions categories are equally important, the paper focuses on tactical level decisions related to retailers' problem called "out of stock". This choice should be regarded as next "prove of concept" for simulation platform developed by authors of this paper. The running research is supposed to take into consideration most of the decisions made in Supply Chain Management process, show how to support them with simulation platform and collect all the solutions in the form of decisions patterns library.

---

### **Simulation as a tool for Supply Chain Management**

---

As we mentioned above, Supply Chains are complex systems. Simulation offers an effective analytical tool for organizations that need to understand their behavior and measure the performance in the Supply Chain

---

---

environment. There are several reasons why simulation should be used for analysis and understanding the behavior of Supply Chains. North and Macal put them forward in the following way [North *et al.*, 2007]:

- No one is able to understand how all parts of the Supply Chain interact and add up to the whole.
- No one is able to imagine all the possibilities that the real Supply Chain could exhibit.
- No one is able to foresee the full effects of events with limited mental models.
- No one is able to foresee novel events outside of their mental models.
- Decision makers want to get insights into key variables and their causes and effects.
- Decision makers want to make predictions of how Supply Chain will behave. Thanks to simulation they can get educated guesses and be provided with the range of possible futures.

Simulation addresses all these motivations and seems to be only reasonable analysis method for understanding existing Supply Chains as well as designing new ones.

According to Banks *et al.* simulation as a tool used in Supply Chain Management may be applied to different areas and problems in the Supply Chain lifecycle. In design phase simulation may help in evaluation of possible configurations of Supply Chain, concerning different business partners and locations of manufacturing as well as distribution facilities. What is more simulation may be used for analysis of different product configurations, and establishing the inventory levels that allow to achieve service level goals. Operational phase may be supported in the areas of Supply Chain planning and execution, in the processes of establishing production and logistics plans and schedules to meet long and short term demands. During the termination phase simulation can be applied in the area of analysis and selecting emptying the pipeline plans as well as shut down manufacturing and distribution facilities plans [Banks *et al.*, 2006].

---

### **Supply Chain Modeling, Model Implementation and Simulation**

---

Modeling Supply Chain differs from modeling manufacturing systems, where models regard mainly material flow through machines and material handling systems, and are used for analysis of machine utilization, cycle times and bottlenecks. In case of Supply Chain, modeling the material flow is not enough and should be enriched by information flows related to business processes triggering and controlling flow of material, orders, products and money transactions. All these elements have been taken into consideration in the model used in simulation experiment presented in this paper.

The modeling process has been driven by agent orientation principles. As was shown many times, in case of such business models as Supply Chains using agents as a basic modeling constructs is reasonable solution because of the following characteristics of problem domain [North *et al.*, 2007]:

- Representation may be conceptualized as consisting of interacting agents.
- Decisions and behaviors can be defined discretely, that is, with well-defined boundaries.
- It is important that agents change their behavior and adapt.
- It is important that agents engage in dynamic strategic behavior.
- It is important that agents have dynamic relationships with other agents, and agent relationships form and dissolve.
- It is important that agents form organizations. What is more adaptation as well as learning are important at organizational level.
- The past may be a poor predictor of the future.
- Scaling up is important, and scaling up consists of adding more agents and agent interactions.
- Process structural change needs to be a result of the model, rather than an input to the model.

The agent oriented model of Supply Chain, used in the simulation experiment, has been based on the work of Veira *et al.* [Veira *et al.*, 2005]. Agents that constitute the building blocks of the model belong to the following categories:

- Supplier Agent,
- Manufacturing Agent,
- Retailer Agent,
- Market Agent.

Every agent plays the role related to the position in Supply Chain. Agents' responsibilities have been formalized in the behavioral rules driving their behavior during the simulation process. Sample behavioral rule for retailer agent is presented in Figure 1.

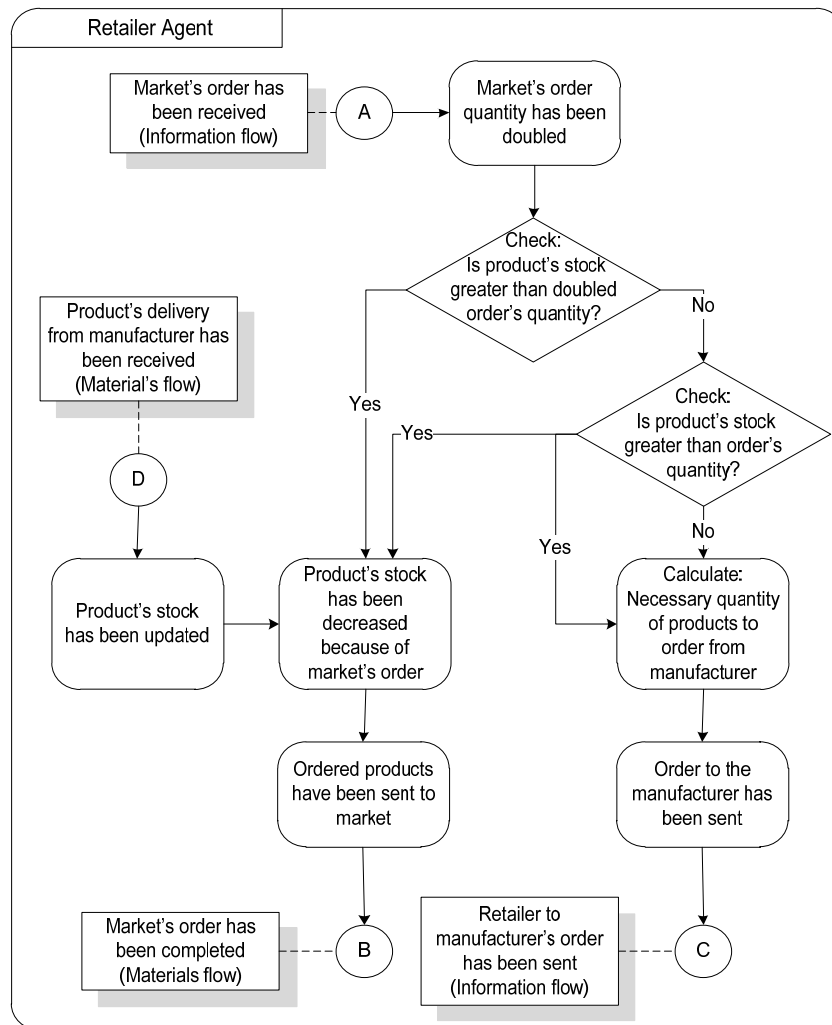


Fig. 1. Retailer Agent's Behavioral Rules

The starting point of Retailer Agent is the moment when an order is received, but in this case its source is the market. After the order is received and the number of order items is doubled, the retailer compares calculated value with its stock level. If the stock level is greater than the doubled number of order items, Market Agent is provided with products and process ends. In case the stock level is less than calculated value, but is enough to fulfill the order, the products are delivered and an order is placed to replenish the stock to the planned level. If the Retailer Agent does not have requested number of products it sends an order to Producer Agent and waits for delivery. When products are delivered, Retailer Agent fulfills the market demand and replenishes its stock level.

---

---

Detailed description of the reference model for simulation environment used in this paper may be found in [Jakiela, Litwin, Olech, 2010a].

Because the easiest and most effective way of conducting agent-based simulation is to use the package dedicated to this purpose, the model and the simulation experiment have been implemented in NetLogo environment, which is considered to be one of the most popular simulation environment. It includes several facilities such as tools for building user interface or system dynamics modeler. The environment is free of charges for educational and research purposes [Gilbert, 2007].

---

### **Case Study – Agent Based Supply Chain Management**

---

As was mentioned above, Supply Chain Management is deeply entrenched in decision making processes taking place in different nodes of Supply Chain. Even if the decision is made individually by specific Supply Chain partner the characteristics related to whole Supply Chain have to be taken into consideration. Therefore there was assumed that it is possible to show that Agent Based Modeling and Simulation may be used as a powerful tool for Supply Chain Management by showing how to support sample decision making processes.

The presented case study includes the decision situation analysis that has been characterized according to the pattern as follows. At first the problem to be solved is articulated. Then possible solutions to the problem are suggested. Next, metrics are carefully selected, formalized and the simulation is run. Finally simulation results are interpreted and applied to the decision making process. In the case study sample decision will be made with the support of agent based simulation and modeling platform.

#### **Decision situation – out of stock problems**

##### Problem to be solved and suggested solutions

This kind of problems is usually experienced by retailer. It is related to the situations when the customer's demand cannot be fulfilled, because there is not enough merchandise in stock. If the situation is frequent it may cause the decrease in customers' loyalty level and retailer's financial losses. Decisions made in case of such problems may be classified as a semi-structured and are regarded to be a tactical one.

The problem may be solved in the following ways:

1. Solution 1 – change the merchandise delivery parameters such as: different roads, routes and vehicles' loads. Every parameter setting may have specific influence on the problem solution – positive as well as negative. In order to find the optimal one, several simulations will be run with different parameters values. Simulations' results will be then analyzed and compared to the results used as a reference.
2. Solution 2 – find more appropriate safety level of the merchandise and improve current stock policy. This may increase the availability and eliminate out of stock situations but at the same time increase the costs of warehousing. Simulation experiment will take into consideration different values of safety stock levels. The experiment's results show how these values affect the problem solution.
3. Solution 3 – combine the solutions of types (1) and (2).

All the suggested solutions will be tested with the use of agent based simulation and finally the best solution will be selected and implemented.

##### Metrics Selection and Formalization

In order to test different possible problem's solutions, the following metrics have been selected:

1. OTIF (*On Time in Full*) – indicator that shows how often customers' demand is fulfilled directly from retailer's stock. This metric is calculated according to the formula below



$$OTIF = \frac{q_f}{q} \cdot 100 [\%], \tag{1}$$

where:  $q_f$  – quantity of orders fulfilled directly from retailer’s stock,  $q$  – total number of orders taken by retailer.

2. Delivery time – time measured between the moment the order has been taken and the moment the order has been fulfilled. With regard to this metric such statistics as maximum, minimum, average and standard deviation have been calculated. It’s important to mention that only deliveries related to customers’ orders have been included.
3. Total retailer’s profit – profit generated by retailer during the simulation run.
4. Total delivery cost – total cost of delivery between retailer and manufacturer.
5. Average unit delivery cost – average unit delivery cost calculated for merchandise sold.

All these metrics have been used in the simulation experiment; however different configurations of them were selected for verification of problem solutions.

Simulation Runs

The first run is treated as a source of reference data used during the analysis of experiment results. The parameters of reference run, which will not be changed during all simulation runs are presented in table 1.

*Table 1. Parameters of simulation runs*

<b>Parameters</b>	<b>Values</b>
Time of the simulation experiment	650
Initial part’s price	3.8
Demand distribution	Normal
Mean of demand distribution	5
Standard deviation of demand distribution	0.8
Safety part’s stock at supplier	40
Safety part’s stock at manufacturer	30

For simulation runs and suggested solutions the following settings have been used:

1. Solution 1 – change the merchandise delivery parameters.
  - a. Highway type road – the road selected for delivery is longer, faster and safer than the road type used in reference simulation run.
  - b. Mountainous type road – the road selected for delivery is shorter but slower and more risky in comparison with reference run road type.
  - c. Regular type road – alternative to reference run road.
2. Solution 2 – find proper safety level
  - a. Merchandise safety stock level is equal 1 item
  - b. Merchandise safety stock level is equal 2 items
  - c. Merchandise safety stock level is equal 3 items
  - d. Merchandise safety stock level is equal 5 items
3. Solution 3 – combination of best option from solution 1 and best option from solution 2.

Next section presents the results of simulation experiment executed. Every simulation run takes into consideration different settings as showed above. The results will provide us with the answer to the question: *Which solution is the optimal one?*

### Simulation Results Analysis

The values of metrics calculated in reference run are presented in table 2. As the value of OTIF indicator (which is not very high) and standard deviation of delivery time (which should be smaller) show, the customers complaints related to “out of stock” situations are well-founded. Solving the problem will require to go through all the alternatives defined and find the best solution possible.

Table 2. Metrics values used for reference run

Metric	Value
OTIF [%]	65,28%
Delivery time [h]:	
- minimum	2,60
- maximum	16,30
- average	8,78
- standard deviation	2,15
Total profit	7913,33
Delivery cost:	
- total	4818,00
- average cost per unit	1,59

#### Solution 1 – use different delivery parameters

As was mentioned before, the first solution is to check how the situation will change if the type of road will be different. The results of the simulation with the road type set to highway are presented in table 3.

Table 3. Metrics values calculated for “highway” road type

Metric	Value
OTIF [%]	67,38%
Delivery time [h]:	
- minimum	2,30
- maximum	15,60
- average	8,39
- standard deviation	1,92
Total profit	7531,27
Delivery cost:	
- total	4193,00
- average cost per unit	1,41

The OTIF indicator value is almost the same as in reference run. The conclusion is obvious – the road type does not influence the merchandise availability. What is also well visible, and may be inferred from data in table 2 and table 3, road type selection will strongly affect the delivery time. All metrics related to delivery time have been improved. As can be seen selecting different road type reduced delivery costs (total as well as unit cost). Data related to other road types (mountainous and regular) are gathered and presented in table 4.

*Table 4. Simulation results for mountainous and regular type roads*

<b>Metric</b>	<b>Mountainous type road</b>	<b>Regular type road</b>
OTIF [%]	66,78%	67,56%
Delivery time [h]		
- minimum	2,20	2,00
- maximum	15,70	13,40
- average	9,76	8,35
- standard deviation	2,50	1,89
Total profit	7661,68	7449,62
Delivery cost		
- total	8409,00	5776,75
- average cost per unit	2,79	1,95

The comparison of simulation results for mountainous type road with regular and highway type roads reveals that the delivery time is the longest for the former one. Using shorter route not always provides improvements in delivery time (route is shorter but more risky and therefore slower). What is more, the stability of deliveries made along this route, determined by standard deviation, in this case is higher than in others. The final factor that settles getting rid this option of is delivery cost, which for this road type has the highest value.

In case of regular type road, such metrics values as delivery time and standard deviation are smaller than values for highway type road. Unfortunately the value of delivery cost is higher. Relative differences among all the metrics' values may be compared with the help of summary of all data presented in table 5.

*Table 5. Metrics values summary for Solution 1*

<b>Metric</b>	<b>Used road type</b>						
	<b>Normal (reference)</b>	<b>Highway</b>		<b>Mountainous</b>		<b>Regular</b>	
		<b>Absolute</b>	<b>Relative</b>	<b>Absolute</b>	<b>Relative</b>	<b>Absolute</b>	<b>Relative</b>
Average delivery time	8,78	8,39	-4,44%	9,76	11,16%	8,35	-4,90%
Standard deviation	2,15	1,92	-10,70%	2,5	16,28%	1,89	-12,09%
Delivery cost	4818	4193	-12,97%	8409	74,53%	7449,62	54,62%
Average delivery cost	1,59	1,41	-11,32%	2,79	75,47%	1,95	22,64%

According to simulation results, the best possible choice is to use the highway road type. The improvements can be achieved in all metrics values – delivery time, delivery cost and standard deviation. At first sight regular road type may seem the optimal solution; however after we take a closer look at delivery cost, it becomes clear that it is not an appropriate choice. Obviously, the worst option is to select mountainous road type.

Solution 2 – find the right safety stock level

What has already been observed, changing delivery parameters may reduce the delivery costs and time but unfortunately does not have an impact on the availability of merchandise described by OTIF indicator. The availability mainly depends on the stock level of merchandise. Therefore setting the proper safety stock level may sort the problem out. In the first step the maximum, minimum and average customer's order lot has been calculated. Then the OTIF indicator value has been determined in order to check how often the merchandise was available offhand. It important to remember that safety level will influence the profit company generates, because profit will be decreased by storage costs. Storage costs are related to the number of items we have in the warehouse. The above mentioned metrics' values calculated in reference run are presented in table 6.

Table 6. Metrics values related to merchandise safety level

Metric	Value
Minimum order lot	3
Maximum order lot	8
Average order lot	5,038
OTIF	65,28%
Profit	7913,33

Based on the metrics values presented in table 6, four following alternatives of merchandise safety stock level have been taken into consideration in simulation experiment:

1. Safety stock level is equal 1 item.
2. Safety stock level is equal 2 items.
3. Safety stock level is equal 3 items.
4. Safety stock level is equal 5 items.

The simulation has been run four times, once for every alternative presented above. The main metrics calculated during every run are OTIF indicator and total profit. The results are presented in table 7.

Table 7. Metrics calculated for different safety stock levels

Option 1		Relative difference
OTIF	89,28%	24,00%
Profit	6088,97	-23,05%
Option 2		Relative difference
OTIF	91,81%	26,53%
Profit	4675,71	-40,91%
Option 3		Relative difference
OTIF	89,60%	24,32%
Profit	3027,80	-61,74%
Option 4		Relative difference
OTIF	94,96%	29,68%
Profit	1970,33	-75,10%

As analysis reveals, the safety stock level which is equal or greater than 1 item provides better availability of merchandise about at least 24%. Unfortunately at the same time, the profit decreases rapidly because of storage costs related to number of items that constitute the merchandise safety level. Therefore, only one item which is set aside as a reserve is economically viable.

Solution 3 – combination of best alternatives from solution 1 and 2

The final step is to check if the combination of two best options of suggested solutions will provide better results than each option separately or putting this differently to check if synergy effect will take place. The table 8 shows simulation results. In this case the parameters have been set up in the following way: safety stock level = 1 item; type of road = highway.

*Table 8. Metrics values for solution 3*

Reference run		Solution 3		Difference	
Metric	Value	Metric	Value	Absolute	Relative
OTIF [%]	65,28%	OTIF [%]	87,58%	22,30%	N/A
Delivery time [h]:		Delivery time [h]:			
- minimum	2,60	- minimum	2,30	-0,30	-11,54%
- maximum	16,30	- maximum	13,90	-2,40	-14,72%
- average	8,78	- average	8,70	-0,08	-0,89%
- standard deviation	2,15	- standard deviation	2,32	0,17	7,79%
Total profit	7913,33	Total profit	6026,14	-1887,19	-23,85%
Delivery cost:		Delivery cost:			
- total	4818,00	- total	4053,00	-765,00	-15,88%
- average cost per unit	1,59	- average cost per unit	1,36	-0,23	-14,47%

In solution 3, OTIF indicator's value has been improved by 22,3%. There are also better values of delivery time and delivery costs. Unfortunately total profit has decreased by 24%. What we have already mentioned this is related to the storage costs which increase when the number of items grows.

Final decision

After careful analysis of the “out of the stock” situation, supported with Agent Based Modeling and Simulation, the final decision can be made. According to simulation results, the solution 1(a) has turned out to be the optimal one. The reason of such choice is that it gives the shortest delivery time what can boost our customers' satisfaction level, however in order to improve customer service with regard to “out of stock” problem, the warehousing system should also be improved. As the case study showed, in the current form of warehousing system, the process of increasing number of items significantly affects storage costs in the negative way.

**Conclusions and Further Research**

The paper shows that simulation done with the use of Agent Based Platform may be very powerful tool for supporting decision making processes related to Supply Chain Management. Although only one decision situation was analyzed it's clearly visible how sensitive this kind of modeling is and how well decision maker may be supported. The approach presented is especially well suited for semi-structured and unstructured decisions,

where predefined solutions' procedures are not available and analysis has to include risk factor. The further research will be devoted to collecting decision situations patterns that take place in Supply Chain Management process and showing how these kinds of decisions may be supported by simulation experiments planned and executed on the developed platform. The vision of the research is to have library of simulation experiments patterns dedicated to the most common decision situations taking place in the area of Supply Chain Management. This library is supposed to be some kind of "prove of concept" for the reference model for Agent Based Simulation of Extended Enterprises which is the final product of the research.

---

## References

---

- [Anciaux et al., 2004] Anciaux, D., Monteiro, T., Ouzizi, L., Roy, D.: Multi-Agent Architecture for Supply Chain Management. In: Journal of Manufacturing technology and management. Logistics and Supply Chain Management with Artificial Intelligence Techniques – Part one. Vol 15, no 8, (2004)
- [Banks et al., 2006] Banks, J., Buckley, S., Jain, S., Ledermann, P.: Panel Session: Opportunities for Simulation in Supply Chain Management. In: Proceedings of the 2006 Winter Simulation Conference. Yucesan, E., Chen, C., Snowdon, J., L., Charnes, J., M., (2006)
- [Chang et al., 2001] Chang, Y., Makatsoris, H.: Supply chain modeling using simulation. International Journal of Simulation, 2(1), 24-30, (2001)
- [Gilbert, 2007] Gilbert N.: Agent-Based Models, Sage Publications, (2007)
- [Jakiela, 2006] Jakiela, J.: AROMA – AgentowożoRientowana metOdologia Modelowania orgAnizacji. WAEil, Politechnika Slaska, Gliwice (2006)
- [Jakiela et al., 2009] Jakiela J., Pomianek B.: Agent Orientation as a Toolbox for Organizational Modeling and Performance Improvement. International Book Series "Information Science and Computing", Book 13, Intelligent Information and Engineering Systems, INFOS 2009, pp. 113-124, (2009).
- [Jakiela, Litwin, Olech, 2010a] Jakiela J., Litwin P., Olech M.: Toward the Reference Model for Agent-based Simulation of Extended Enterprises. In: Setlak, G., Markov, K.: Methods and Instruments of Artificial Intelligence, pp. 34-66, (2010)
- [Jakiela, Litwin, Olech, 2010b] Jakiela, J., Litwin, P., Olech, M.: MAS Approach to Business Models Simulations: Supply Chain Management Case Study. In: KES AMSTA-2010, Jędrzejowicz, P., Nguyen, N., T., Howlett, R., Lakhmi, C. J., (Eds.), Part II, LNAI 6071, pp. 32-41, Springer-Verlag, Berlin Heidelberg, (2010)
- [Jakiela, Litwin, Olech, 2011a] Jakiela J., Litwin P., Olech M.: Multi Agent Based Simulation as a Supply Chain Analysis Workbench. Transactions on Computational Collective Intelligence, Springer-Verlag, Berlin Heidelberg, paper accepted for publication (2011)
- [Jakiela, Litwin, Olech, 2011b] Jakiela J., Litwin P., Olech M.: Prototyp platformy symulacji wieloagentowej rozszerzonych przedsiębiorstw. Studia Informatica, vol. 32, Number 2B (97), pp. 9-23, Gliwice (2011)
- [Kimbrough et al., 2002] Kimbrough, S., O., Wu, D., Zhong, F.: Computers play the Beer Game: Can artificial agents manage Supply Chains? Decision Support Systems 33. pp. 323–333, (2002)
- [Muckstadt et al., 2001] Muckstadt, J., Murray, D., Rappold, J., Collins, D.: Guidelines for collaborative supply chain system design and operation. Information Systems Frontiers 3, pp. 427–435, (2001)
- [Nfaoui et al., 2006] Nfaoui, E., H., Ouzrout, Y., El Beqqali, O.: An approach of agent-based distributed simulation for supply chains: Negotiation protocols between collaborative agents. In Proceedings of the 20th annual European Simulation and Modeling Conference, EUROSIS, Toulouse, France, pp. 290–295, (2006)
- [North et al., 2007] North, M.J., Macal, C.M.: Managing Business Complexity. Discovering Strategic Solutions with Agent-Based Modeling and Simulation. Oxford University Press (2007)
- [Paolucciet al., 2005] Paolucci, M., Sacile, R.: Agent-Based Manufacturing and Control Systems. New Agile Manufacturing Solutions for Achieving Peak Performance, CRC Press (2005)

[Vieira et al., 2005] Vieira, G.E., Cesar, O. Jr.: A conceptual model for the creation of supply chains models. Proceedings of the 37th conference on Winter simulation, pp. 2619 – 2627, Orlando, Florida, (2005)

---

### Authors' Information

---



**Jacek Jakiela, Ph.D., Eng.** – Department of Computer Science FMEA RUT; W. Pola 2, 35-959 Rzeszow, Poland; e-mail: [jjakiela@prz.edu.pl](mailto:jjakiela@prz.edu.pl)

Major Fields of Scientific Research: Software Development Methodologies, Agent and Object-Oriented Business Modeling, Internet Enterprises Models, Computational Organization Theory and Multi-Agent Simulation of Business Architectures.



**Paweł Litwin, Ph.D., Eng.** – Department of Computer Science FMEA RUT; W. Pola 2, 35-959 Rzeszow, Poland; e-mail: [plitwin@prz.edu.pl](mailto:plitwin@prz.edu.pl)

Major Fields of Scientific Research: Applications of Neural Networks in Mechanics, Computer Simulations, Finite Element Method.



**Marcin Olech, M.Phil., Eng.** – Department of Computer Science FMEA RUT; W. Pola 2, 35-959 Rzeszow, Poland; e-mail: [molech@prz.edu.pl](mailto:molech@prz.edu.pl)

Major Fields of Scientific Research: Multi-agent Simulation, Application of Artificial Intelligence in Industry.

## SEMANTICALLY RICH EDUCATIONAL WORD GAMES ENHANCED BY SOFTWARE AGENTS

**Boyan Bontchev, Sergey Varbanov, Dessislava Vassileva**

**Abstract:** *With steadily evolving new paradigms for technology enhanced learning, educational word games such as quizzes, puzzles and quests raise new appeal and motivation for students in following game based educational processes. Traditional word games may be applied more successfully to game based learning in given scientific domain provided they are highly oriented to the content and problems of that domain. Such games may be more efficient if they include artificial agents simulating opponents, advisors or collaborators of the player. Authors present a semantic structuring model of learning content for logic and word games serving for educational purposes and, next, show the place of artificial agents within the game construction and possible ways of agent's realization. There are given results from practical experiments with playing a memory game using the semantic content model, without and with agents integrated into the game.*

**Keywords:** Agents, quiz, word game, e-learning, game based learning.

---

## Introduction

---

The last developments of the Internet technologies and their usage for technology enhanced learning require blending traditional teaching and training with online processes. New trends in evolution of modern e-learning comprise various approaches of applying educational games in a complement to the traditional instructional learning design [Dempsey et al, 1996]. Word and logic games are considered as a rather effective mean for retaining interest of learners by attracting their attention for much more time than other approaches. Educational games appeared to be not only entertaining but also cognitive and educative for today learners, as far as they have precise definition of goals, constraints, built-in rules and consequences [Prensky, 2006]. Such logic games help learners to improve their knowledge in given domain by challenging them to solve various problems or to use practices within the domain area and, thus, to improve creative thinking [Salen and Zimmerman, 2003]. Mark Prensky explains in [Prensky, 2006] why and how games deliver pleasure and emotions while the learner plays a game. Games impose conflicts and competitions by using rules and establishing goals in order to facilitate evolution of creativity, motivate learners and, as well, to satisfy their ego in an interactive and pleasant way.

Nowadays, the mostly applied educational games are represented by quizzes, puzzles, quests and problem solving staging [Batson and Feinberg, 2006], [Ferreira et al, 2008]. Such logic word games make use of textual and even multimedia content of a proper domain and apply predefined rules for solving some problems, sometimes supported by counters or dices according given educational purposes. They may be represented by board games [Bontchev and Vassileva, 2011] which may be played by a single player alone or against a simple or intelligent software agent replacing the real opponent.

The paper describes a principal model for semantically enriched educational word games such as quizzes, puzzles and quests and, next, its realization by using an agent based software architecture. It explains in brief a model of semantic structuring of domain content by means of UML class diagrams. The model is used as a basic paradigm for the development of educational word games. The paper presents how it facilitates the implementation of a positional memory quest board game for matching symbols with minimum number of mouse clicks. Next, it shows how a software agent is used within the memory game as opponent of real player. There are given practical results from an experimental field trial aimed at evaluation of student appreciation of the single user memory game, with and without software agent.

---

## Motivational Background

---

Educational games used to support e-learning may vary from simple single user word games to complex multi-user collaboration games using augmented reality. The paper is focused on word games like quizzes, puzzles and quests because they suppose a rather simple construction process and, at the same time, are very useful for exercises and self-assessment.

---

## Educational Logic Games

---

Quizzes are popular as word games with educational goal thanks to their easy implementation. They can be used by students and teachers for self-assessment tests and control exams. Besides assessment of knowledge, they are purposed for producing more fun and increasing motivation for learning. For this reason, some of the educational logic and word games are implemented as combinations between board games and quizzes. This type of games usually uses board rules for navigation within a quiz. Thus depending on his/her position on the board, the player gets a question with a certain complexity [Bontchev and Vassileva, 2010]. Educational quizzes could be developed and managed using different existing tools such as Quiz Center and Quiz Builder [Bontchev and Vassileva, 2011]. These tools have opportunities for automatically evaluating students' answers, some of



them automatically generate questions from curriculum [Guettl et al, 2005] and other use authoring tools to create quizzes [Retalis, 2008]. In some authoring tools there can be created not only simple questions, but also parameterized ones [Feng, 2005].

In recent years, games are often used in e-learning as means of students' attention to being actively involved in learning process and as means of achieving a higher motivation for studying [Aleksieva-Petrova and Petrov, 2011]. Games are also successfully used in adaptive e-learning systems, e.g. like adaptive quizzes such as QuizGuide and QuizJet [Hsiao et al, 2009] and adaptive board games such as ELG [Retalis, 2008]. In the ELG different learning scenarios could be presented in the form of board games, through which students are expected to improve their performance and extend their knowledge. Games created with the ELG can be customized on different parameters such as learner's level of knowledge, preferences and educational goals. Games such as QuizGuide and QuizJet [Hsiao et al, 2009] support students to select self-assessment quizzes most suitable for learner's goals and preferences. These systems use adaptive navigation to increase learners' knowledge by selecting most important topic of training.

Generally educational word and board games use rule-based approaches for navigation through its board and require detailed knowledge of the subject concerned [Retalis, 2008]. It could be used different strategies for selecting a question, its level of difficulties and area of knowledge, depending on the learner/player profile (goals and preferences, learning style and performance) [Bontchev and Vassileva, 2010].

---

### ***Types of Software Agents for Educational Games***

---

Educational games may have three main different modes of play. These modes are multiple users mode, single user mode and single user against a simulated player mode. The last of these three modes of play appeared to be one of the most preferred by end users. Usually the simulated player is implemented by means of an intelligent software agent. Usage of software agents is an established and effective approach for realizations of various types of simulations. According [Varbanov et al, 2007] intelligent agents may be applied in several aspects and cases of serious games. Main cases of them are as follows:

- The case where a part of the game is implemented as an intelligent agent. The software agent performs a simulation process and its behavior unlike other cases is similar in each game and each player. In this case, the participation of the intelligent agent is hidden for gamers and it is realized as an autonomous system.
- The case in which the intelligent agents acts as opponents of the player. The aim is to simulate the competition and to encourage gamers to achieve better results. The software agents which participate in a game have different behavior and their numbers may vary depending on the level of a user.
- The case where an intelligent agent participates as business partner or/and task collaborator. The player communicates and performs various activities together with this type agent.
- The case where the intelligent agents are included in a game as assistants to players. These software agents are responsible for providing necessary information, guidance, comments and recommendations.

The agent may have various levels of implementation complexity, e.g. from a simple search procedure to controlling multidimensional state space of a virtual world model. Moreover, complex agents may simulate social behavior with knowledge about the game and adaptation to the physical player. As well, agents may cooperate within a community of social agents.

## Semantically Rich Educational Games

Logic and board games like word and problem-oriented games need a special organization of course content in order to integrate it to the game. A semantic structuring of the content will enable game engine to extract specific terms and their inter-relationships dynamically during the play process. It follows a brief representation of such a model and a sample content organization based on it.

### Semantically-Rich Content Model for Educational Word Games

The model for semantic content organization for educational games was proposed in (Bontchev and Vassileva 2011) and is based on UML class diagrams. It models visually content relationships by class hierarchies, instances, class attributes, metadata and relationships plus axioms. For any class, there may be presented its possible super-classes by means of *IS\_A* link. A class instance is linked to its class by an *instance Of* reference. A class may contain a content description text and image (again with public visibility and classifier scope) and a list of class properties and resources. Names and values of both properties and resources may be of private, protected, public or package scope and may be annotated within metadata. Any concept may be related to another term by UML relationships of type dependency, association, aggregation, and composition. As a minimum, a relationship may have name, direction, and roles and cardinality of the inter-related entities.

Figure 1 shows a UML class diagram of some terms from a bachelor course in XML technologies. Terms are semantically structured by their inter-relationships and, possibly, with some instances. For example, the terms “ANY”, “EMPTY” and “PCDATA” are subtypes of “Element type DTD declaration” having with it *IS\_A* relations. “Element type DTD declaration” may compose “Sequence list”, “Choice list” and “Element quantifier”. There is shown only two instances – one of “Choice list” and another of “PCDATA”.

In fact, fig. 1 illustrates only a small part of the whole UML diagram of XML terms plus their relationships and instances. Such semantically-rich content models are comprised by a plenty of specific terms (such as DTD, XML Schema, DOM, SAX, etc.) and their interrelationships. These terms and relationships are to be extracted and integrated (preferably, automatically) in any board games suitable for educational purposes. The section below represents an educational board game using this content and appropriate for agent-based implementation.

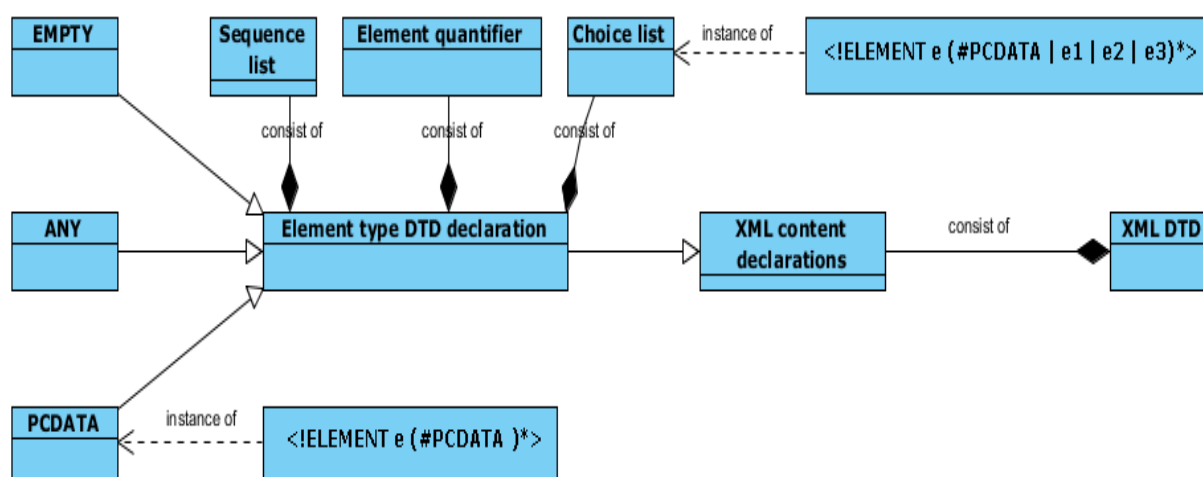


Fig. 1. UML diagram of some XML terms and their relationships and instances

**A Memory Game Using the Semantic Content Model**

Authors have investigated several educational logic games (both word and board games) extracting content from the semantic model. The simplest case of such games makes use of a single entity (class or instance) without utilizing its possible inheritance and/or association relations to other entities. Simple word games serve as good examples here, e.g. the classical hangman game and anagram games. More attracting games may be constructed by using type information about the entity, its properties and instances.

For a mass invasion of games usage in education, teachers need of simple and rapid process of constructing educational games. Classical word puzzles may be easily constructed as board mini-games (Bontchev and Vassileva, 2010). The building parts of a word puzzle being partial images or letters may be represented as objects, which are be moved to the correct positions while wrong moves are possible. More complex logical quizzes and quest may require not only simple moves but also some other player action types like single and double clicks onto objects and object relations. Fig. 2 provides a snapshot of a positional memory board game for matching XML terms to their instances – both linked into couples by means of *instanceOf* relationships in fig. 1.

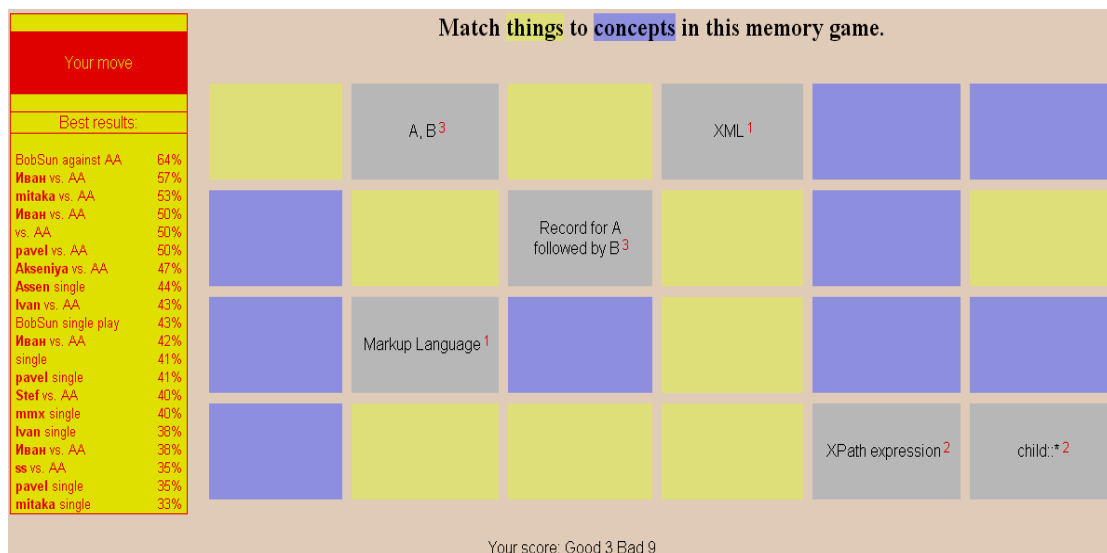


Fig. 2. Positional memory quest board game for matching symbols with minimum number of mouse clicks

The player has to click first onto a blue cell in order to see its hidden term. Next, he/she should click onto a yellow cell supposed to contain the hidden instance of this term. The yellow cell shows its contents and, if the instance matches the term, the player receives a point and both the term and instance are appended by a red suffix showing the number of that match, and continue being displayed. Otherwise, the player loses a point and both the term and instance disappear. The game is over when all the terms of the blue cells are matched to their instances.

Left on fig. 2, there is given a list of players' results until the present moment. Some of them are reached within single user games as explained above, while others are resulted in playing against an artificial agent called AA. In the last case, game play is controlled in turns – the player tries to match a term to its concept and next, the agent does the same. Thus, player is supposed to learn not only from his/her mistakes but from wrong moves of the agent, too.

## The Software Agent Architecture

In one of our previous works [Varbanov et al, 2007], we have used a three-layered architecture of the software agent, shown in fig. 3. The first layer contains methods for interfacing the virtual world in which the agent acts. The second layer represents the model of the virtual world, which the agent builds and uses. The third layer contains the agent's decision-making functionality, which applied to the virtual world model implements the decision-making process. For the implementation of those three layers, different technologies were used, accordingly:

- Java methods to implement agent's sensors and effectors, communicating with the game server via CORBA;
- Protégé [Drummond et al, 2005], as knowledge base management system to implement the virtual world model;
- Algernon [Algernon, 2005], as forward-backward rule based inference engine to implement the agent's decision-making functionality.

That approach, although fruitful, left the impression that in many cases the full power of Java, Protégé and Algernon would not be needed. Therefore, this three-layered architecture could be implemented by more "lightweight" means. The ideal solution would consist of maximum lightweight tools for the trivial cases, and enough openness to provide for easy inclusion of heavyweight tools, such as Protégé and Algernon, if needed.

In [Bontchev et al, 2010], we've described architecture based on Persevere Server [Persevere, 2011] - an object storage engine and application server (running on Java/Rhino) that provides persistent data storage of dynamic JSON data in an interactive server side JavaScript environment (fig. 4).

Placing Persevere Server at the base of our intelligent software agent's implementation, the three layers architecture mapping to technologies evolves to:

- JavaScript methods to implement agent's sensors and effectors, communicating with the game server through a standard JSON HTTP/REST Web interface;
- Persistent JavaScript/JSON classes, methods and instances to implement the virtual world model;
- JavaScript methods to implement the agent's decision-making functionality, when the decision-making process is not too complex.

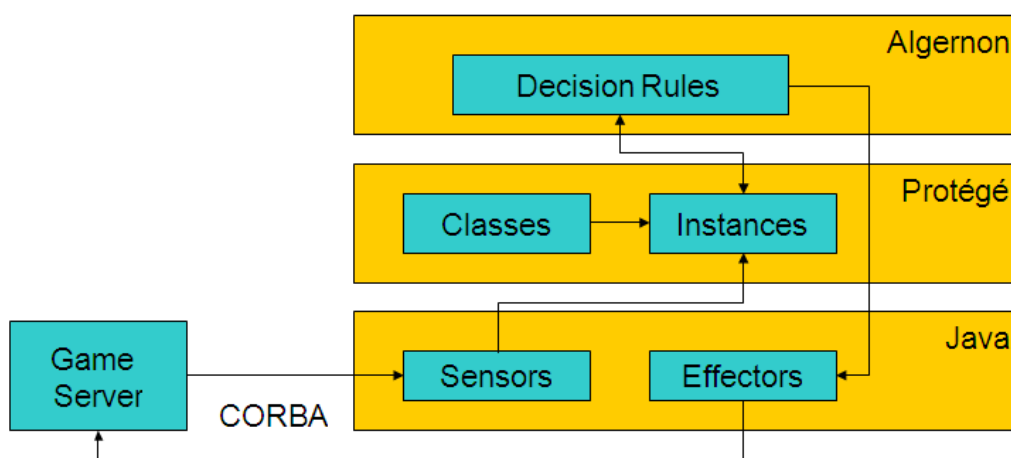


Fig. 3. Three layer agent's architecture used in PRIME project [Bontchev et al, 2010]

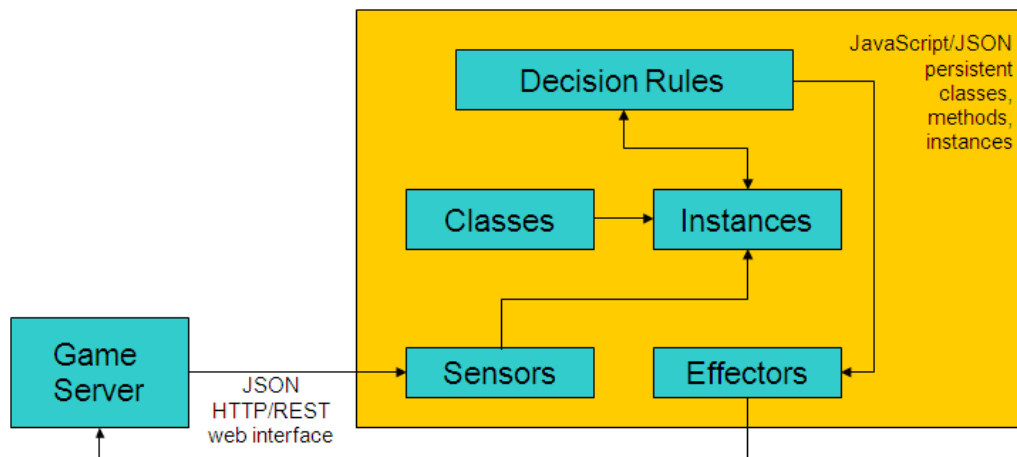


Fig. 4. A lightweight implementation based on Persevere Server [Bontchev et al, 2010]

If the case at hand imposes complex intelligent software agents, Persevere Server provides for easy inclusion of additional Java libraries. In fact, any tool written in Java and having well-documented Java API could be easily included and used.

This architecture is used in the presented here experiments.

### Implementation of the Memory Game Using Agent as Opponent

The memory game does not impose high complexity on the participating software agent, comparing with other types of games. Thus the agent's resources are concentrated on adapting a chosen strategy to the user's evolving model, built and exploited by the agent.

Let us consider several specific opponent agents, demonstrating, although quite in a basic way, how the agent's functionality responds to the strategic goal(s) we pursue:

1. Goal: Mimic human player's features
  - STM agent – a simple agent, which knows the proper associations, but simulates short-term memory (STM) capacity, i.e. memorizes only the last  $N$  moves. Comparing the user success to the results of agents initialized with different  $N$  values, some conclusions about the users STM capacity might be drawn.
  - PK agent – a simple agent, which models partial knowledge of the proper associations, by using parameter  $0 < P \leq 1$ , representing the probability to guess properly the correct association.
  - Note: Both described agents might be adaptable – they can dynamically change their  $N$  or  $P$  parameters, to keep their success rate close to the human player's.
2. Goal: Stimulate human's performance
  - Average Adaptive agent. The agent "knows" all tile's values from the start, but deliberately makes "wrong" moves to keep in pace with the human's rate of success. Sometimes, to help the user, the agent may make "wrong" moves by turning a tile, most appropriate for the human's next move. This agent might function in tolerant mode (by keeping its success rate a little bit below human's average) or in aggressive mode (by keeping its success rate a little bit above human's average), thus stimulating the user to perform better.

---

---

Those and more agent's types can be used in experimenting with different user target groups in order to specify more precisely the relationship between user classes and appropriate agent models and parameters, implementing strategies.

The memory game presented in 3.2 is implemented including an option to play against software agent of the Average Adaptive type.

---

## Practical Experiments and Results

---

The experimental field trial presented here aimed at evaluation of student appreciation of single user memory games with XML content - without and with software agents. In executed practical experiments participated 30 four-year students of the bachelor program in Software engineering at Sofia University, Bulgaria. After the game play, students filled a questionnaire about the efficiency of such game based learning approach. The questions had answers in Likert scale (Strongly disagree, Disagree, Neutral (neither agree nor disagree), Agree, Strongly agree) as given below:

- Q1: The presence of an agent motivates me to play better and more.
- Q2: The presence of the agent suppresses me and prevents me from playing well.
- Q3: The agent behavior designed within the game simulates successfully another real player.
- Q4: Educational games representing problems to be solved using course content are very useful for self-learning and self-test.
- Q5: I feel educational games will have a positive contribution to University education and technology enhanced learning at all.

As well, the questionnaire contained a question with answers in non-Likert scale:

- Q6: What type of game construction is most suitable for e-learning purposes? – possible answers here are as follows:
  1. Single user games with no agents
  2. Single user games with simple software agents
  3. Single user games with artificial intelligence (AI) agents
  4. Multi-user games with no agents
  5. Multi-user games with simple software agents
  6. Multi-user games with AI agents

Figure 5 provides a view of answers for questions from Q1 to Q5. It is obvious that the majority of learners find software agents do not prevent them from playing well even more – a presence of such an agent does stimulate them within the game play. However, students cannot agree on successful simulation of real player behavior on behalf of a software agent in the course of an educational game (question Q3) – in fact, their opinions are divided into two quasi equal groups. At the same time, all of them share the opinion that educational games representing problems to be solved using course content are very useful for self-learning and self-assessment (Q4). In general, they think games will have a positive contribution to University education and technology enhanced learning at all.

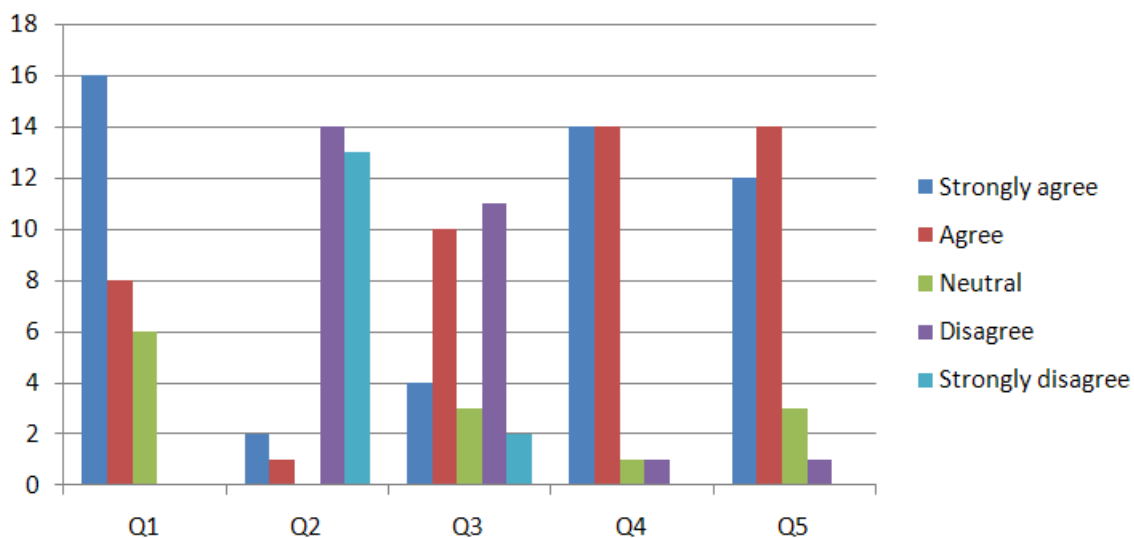


Fig. 5. Results for questions Q1 ÷ Q5

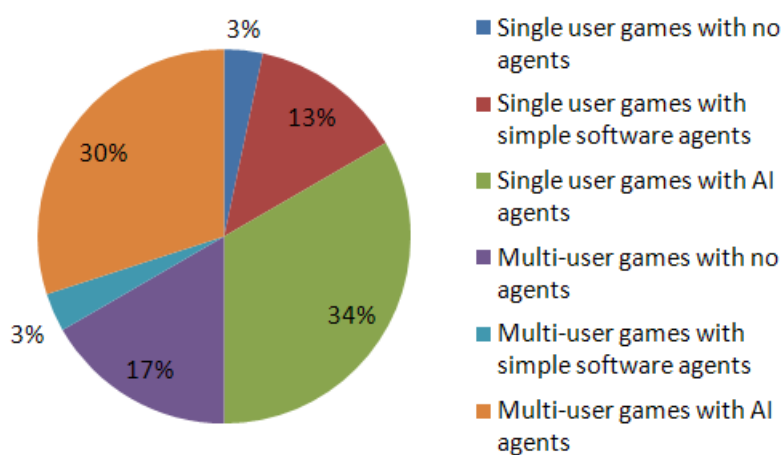


Fig. 6. Answers for question Q6

Figure 6 represents distribution of the answers of question Q6. The majority of students find both single and multi-user games with AI agents most suitable for e-learning purposes. As well, they agree on the assumption that even games with simple software agents will have a positive impact on educational process, especially single user games. The found results sound very encouraging in terms of new development of intelligent agents for educational games.

### Conclusions and Future Works

The present paper proposed a way of using semantically structured courseware content in simple educational logic games such as word or board games. Word games as simple or enhanced quizzes, mazes and quests for solving other logical problems are of great importance for modern e-learning as far as they offer a plenty of benefits and are very useful for self-learning and self-testing. They may be combined with some principles of board games in order to represent more complex logical problems which are to be solved by player actions under execution of some rules upon context conditions [Bontchev and Vassileva, 2010]. Thus, the game context could be based on semantic relationships among domain concepts and terms and, as well, may enable inclusion of a

more complex set of e-learning instructions. Moreover, the game context may include some learning activities such as scenarios typical for a given virtual world, for obtaining new knowledge and/or for maintaining cognitive skills. The presented example of a board game for training memory in semantic relationships within the XML domain is only one of the many examples of such games. It relies on extraction of semantically structured content which may be done by hand or automatically, by extraction of domain concepts and their relationships from an UML description file presented as XMI (XML Metadata Interchange) document.

The second focus of the paper was directed on construction and usage of software agents for educational board games relying on open architectures and agent programming. The software application architecture was chosen specially for shortening both the design and implementation phases. Authors plan to develop further the game to evolve in a more complex dynamic model. As well, the applied software agent may be rather primitive in the beginning and, next, to be replaced with an enhanced intelligent agent.

The proposed combination of semantic model, technologies and tools provides a good ground for further development of other word and board games and, on other side, for creation of new, more intelligent software agents. The conducted survey shows that presence of an agent motivates students to play better and more. As well, the authors are highly encouraged by the students' appreciation of the agent behavior designed within the game which may simulate successfully another real player. They plan to develop more intelligent agents in order to incorporate them into new single- or multi-player educational games.

---

### Acknowledgments.

The work reported in this paper is supported by the ADOPTA project funded by the Bulgarian National Science Fund under agreement no. D002/155.

---

### Bibliography

- [Algernon, 2005], Algernon Tutorial, <http://algernon-j.sourceforge.net/>
- [Aleksieva-Petrova and Petrov, 2011] Aleksieva-Petrova A., Petrov, M. ADOPTA Model of Learner and Educational Game Structure, Int. Conf. CompSysTech'11, 16-17. June 2011, Viena, Austria (in print)
- [Batson and Feinberg, 2006] Batson, L., Feinberg, S.: Game Designs that Enhance Motivation and Learning for Teenagers, Electronic Journal for the Integration of Technology in Education, Vol. 5, pp. 34-43, 2006.
- [Bontchev and Vassileva, 2010] Bontchev, B., Vassileva, D.: Modeling Educational Quizzes as Board Games, Proc. of IADIS Int. Conf. e-Society'2010, ISBN: 978-972-8939-07-6, March 18-21, Porto, Portugal, pp.20-27, 2010.
- [Bontchev and Vassileva, 2011] Bontchev, B., Vassileva, D.: A Light-Weight Semantic Content Model for Mobile Game Based Learning, Proc. of IADIS Mobile Learning 2011, Avila, Spain, ISBN:978-972-8939-45-8, pp.177-182.
- [Bontchev et al, 2010] Bontchev, B., Varbanov S., Vassileva, D.: Software Agents in Educational Board Games, LATEST TRENDS on ENGINEERING EDUCATION, 7th WSEAS International Conference on ENGINEERING EDUCATION (EDUCATION '10), Corfu Island, Greece, July 22-24, 2010, ISSN: 1792-426X, ISBN: 978-960-474-202-8, pp.387-392.
- [Dempsey et al, 1996] Dempsey, J. et al.: Instructional applications of computer games". American Educational Research Association, New York, 1996, pp. 8-12, 1996.
- [Drummond et al, 2005] Drummond N., M. Horridge, Knoblauch, H.: Protégé-OWL Tutorial, Proc. of 8th Int. Protégé Conf., Madrid, July 2005 (also at <http://protege.stanford.edu/>), 2005.
- [Feng, 2005] Feng, K.: Joyce: A Multi-Player Game on One-on-one Digital Classroom Environment for Practicing Fractions, Proc. of the Fifth IEEE Int. Conf. on Advanced Learning Technologies (ICALT'05), pp. 543-544, 2005.
- [Ferreira et al, 2008] Ferreira, A. et al.: The common sense-based educational quiz game framework "What is it?", ACM Int. Conf. Proceeding Series. Vol. 378, pp. 338-339, 2008.



- [Guetl et al, 2005] Guetl, C. et al.: Game-based E-Learning Applications of E-Tester. In P. Kommers & G. Richards (Eds.), Proc. of World Conf. on Educational Multimedia, Hypermedia and Telecommunications 2005, pp. 4912-4917, 2005.
- [Hsiao et al, 2009] Hsiao, I-Han, Sosnovsky S., Brusilovsky, B.: Adaptive Navigation Support for Parameterized Questions in Object-Oriented Programming, LNCS, Vol. 5794/2009, ISBN 978-3-642-04635-3, pp. 88-98, 2009.
- [Persevere, 2011], Persevere Documentation, <http://docs.persvr.org/>
- [Prensky, 2006] Prensky, M.: Don't bother me, mom, I'm learning!, St. Paul, Minn. MN: Paragon House, 2006.
- [Retalis, 2008] Retalis, S.: Creating Adaptive e-Learning Board Games for School Settings Using the ELG Environment, J. of Universal Computer Science, vol. 14, no. 17 (2008), pp. 2897-2908, 2008.
- [Salen and Zimmerman, 2003] Salen, K., Zimmerman, E.: Rules of play: Game design fundamentals, Cambridge, MA, USA, October, MIT Press, 2003.
- [Varbanov et al, 2007] Varbanov S., Stoykov S., Bontchev B., Software Agents in a Serious Business Game – the PRIME Project Case, LG2007 Learning With Games, eds. Marco Taisch and Jacopo Cassina, Sophia Antipolis, France, 24 – 26 September 2007, pp. 301 – 306, ISBN 978-88-901168-0-3, 2007.
- [White et al, 2004] White, M. et al.: ARCO - an architecture for digitization, management and presentation of virtual exhibitions, Proc. of Computer Graphics Int. Conf., pp. 622-625, 2004.

---

### Authors' Information

---



**Boyan Bontchev** – Assoc. Prof., PhD, Sofia University “St. Kl. Ohridski”, 125, Tzarigradsko shoes, bl. 2, Sofia 1113, Bulgaria; e-mail: [bbontchev@fmi.uni-sofia.bg](mailto:bbontchev@fmi.uni-sofia.bg)

*Major Fields of Scientific Research: Software architectures, Technology enhanced learning, Educational games, Adaptive systems*



**Sergey Varbanov** – Assist. Prof., Institute of Mathematics and Informatics, Acad. Georgi Bonchev Str., Block 8, Sofia 1113, Bulgaria; e-mail: [varbanov@fmi.uni-sofia.bg](mailto:varbanov@fmi.uni-sofia.bg)

*Major Fields of Scientific Research: Artificial Intelligence, Artificial Intelligence Applications, Intelligent Agents, Neural Networks, Genetic Algorithms, Educational Games, Serious Games.*



**Dessislava Vassileva** – Research Assoc., PhD, NIS at Sofia University “St. Kl. Ohridski”, 125, Tzarigradsko shoes, bl. 2, Sofia 1113, Bulgaria; e-mail: [ddessy@fmi.uni-sofia.bg](mailto:ddessy@fmi.uni-sofia.bg)

*Major Fields of Scientific Research: Software technologies, Technology enhanced learning, Educational games, Adaptive hypermedia systems*

---

## Engineering Applications of Artificial Intelligence

---

### MACROMODELING FOR VLSI PHYSICAL DESIGN AUTOMATION PROBLEMS

Roman Bazylevych, Marek Pałasiński, Lubov Bazylevych

**Abstract:** *The paper summarizes the authors methodology for solving the intractable combinatorial problems in physical design of electronic devices: VLSI, SOC, PCB and other. The Optimal Circuit Reduction (OCR) method has proved to be an efficient and effective tool to identify the hierarchical clusters' circuit structure. The authors review the applicability of this method for solving of some problems, including hierarchical clustering, partitioning, packaging, and placement. Developed approach based on multilevel decomposition with the recursive use of global and local optimization algorithms at it every level for unique, not very large size subproblems. At every step we receive some initial solutions which are improved by optimization algorithms. Experiments confirm the efficiency of developed approaches. For some well-known test cases the optimal results were achieved for the first time, while for many other cases improved results were obtained.*

**Keywords:** VLSI, SOC and PCB physical design, hierarchical clustering, partitioning, packaging, placement

**ACM Classification Keywords:** B.7.2 Design Aids

---

#### Introduction

---

Many of the most difficult in Design Automation are intractable combinatorial problems. They appear in physical design – partitioning, packaging, placement, routing as well as testing and other areas. Optimization are especially important for VLSI, SoC and PCB design. Rapid growth of electronic circuit complexity requires a further search for new robust, efficient and effective approaches to solve them with high quality. From mathematic point of view they belong to the very large-scale intractable combinatorial NP-class problems – nowadays chips have a few billions of transistors.

Many of these problems have identical input data. The ideas to solve the large-scale problems are to transfer the full mathematical model to the aggregate mathematical notation that could significantly decrease the number of arguments and to operate by the not very large number of hierarchically built macromodels instead of original elements, the number of which is extraordinarily high. This enables us, to decrease the size of the problem in every step of decision making, to reduce the calculation consumption, to improve the quality of solution, as well as to easier trap into the zone of the global optimum. Basic approaches and algorithms were developed for

- hierarchical circuit clustering:
  - free clustering,
  - partially enforced clustering,
  - enforced clustering;
- partitioning and decomposition :

- serial, parallel-serial and dichotomy partitioning,
- initial partitioning,
- partitioning optimization;
- packaging:
  - serial and parallel-serial packaging,
  - initial packaging,
  - packaging optimization;
- placement:
  - hierarchical initial placement by multilevel macromodels,
  - multilevel placement optimization by scanning area method with macromodels.

For these problems also were developed special algorithms for escaping from the local extreme.

The proposed algorithms have such properties:

- can be efficient for choosing the appropriate number of partitions to divide the circuit;
- arbitrary division ratio can be chosen for partitioning;
- many same procedures can be used for initial solution and their optimization;
- close to linear computational complexity;
- provide good quality of solutions;
- are appropriate for large and very large-scale problems.

Most likely, the first proposal to use the free hierarchical clustering for partitioning was in [Bazylevych, 1975]. It was further developed in [Bazylevych, 1981] and used for packaging and placement [Bazylevych, 2000, 2002, 2007] with good results. More lately hierarchical clustering, especially enforced, was used for hyper graph partitioning [Garbers, 1990], [Cong, 1993], [Dutt, 1996], [Karypis, 1997], [Saab, 2000] and for others problems.

For all test cases investigated, the results are not worse, and in many cases they are better comparatively with obtained by other known methods. For some cases, the optimal results were received for the first time.

---

### **Main stages for solving the problems by hierarchical clustering**

---

For solving the large-scale intractable combinatorial problems at the first step we must perform aggregation. We divide large problems into the set of small ones that are simulated by macromodels. Every macromodel include the fixed number of initial circuit elements. The number of macromodels in aggregated circuit is also very important. It is possible to create multilevel model in such way that the number of macromodels and numbers of their elements at very level of decomposition must be not very large to receive good quality solution. We build multilevel system of hierarchically built macromodels. In such system every subproblem could be solved with the high quality for not large CPU time. The main decisions that we must make in such approach are:

- how to chose the number of elements of basic subproblems that we can solve with high quality in a reasonable running time,
- what must be the number of level in macromodeling,
- what method is desirable to use for solving the basic subproblems,
- how merge the partial solutions of subproblems into one solution of whole problem,
- do we need to use additional optimization (refinement) algorithms or not,
- how to escape from the local extreme at every level of macromodeling ?

One of the first problems that appear here is to create the hierarchical macromodel of initial system. Thereto we must receive multilevel aggregation of circuit. One way is to reveal hierarchical built clusters. For this reason it is

possible in electrical system to use the Optimal Circuit Reduction method [Bazylevych, 1975, 1981]. By this method the problem solving is divided into the following steps (Figure 1):

- the bottom-up free hierarchical circuit clustering;
- the mathematical description of clusters by macromodels;
- the top-down multilevel solving with receiving global initial solutions and theirs local optimizations with macromodels at every level of decomposition.

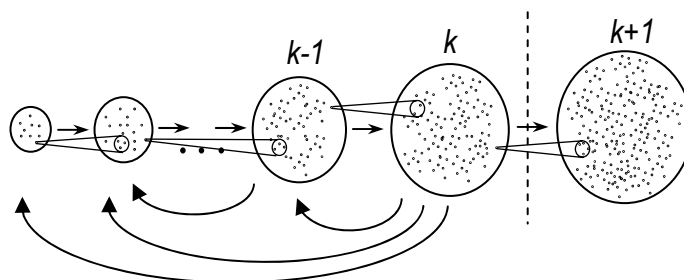


Fig. 1. Bottom-up (left arrows) hierarchical clustering and top-down (right arrows) problem solving

The main features of developed approach are:

- the problem size as a whole (the number of variables) increases step by step during the solution process from substantially reduced, initially (level 1) to real (level k);
- the number of tasks which are to be solved increases on each recursive level. However, all of them would be not large, and are properly hierarchically inserted one into the other, and thus can be solved by the same basic procedure with high quality.

The idea was to operate not by original elements, the number of which is extraordinarily high, but by the hierarchically built clusters of arbitrary sizes (not large) that could be mathematically described by macromodels. The  $(k+1)$ -level (Figure 1) shows that simulating the problem by 0-1 models (binary programming) will significantly increase the number of variables. This case can not simplify the problem solving.

This enables us:

- to essentially decrease the size of the problem, facilitating a solution and reducing the calculation consumption, the large size problem is reduced to recursive solving of small unique tasks;
- to improve the quality of the solution by more easier trapping into the zone of the global optimum. The number of local extrema is significantly smaller.

---

## The Optimal Circuit Reduction Method

---

The Optimal Circuit Reduction (OCR) method builds the Optimal Reduction Tree  $T^R$  (Figure 2). It is a rooted (generally  $n$ -ary) tree which leaves (level 1) correspond to the set  $P = \{p_1, \dots, p_n\}$  of circuit elements and a root (level  $H$ ) - to all aggregated circuit.

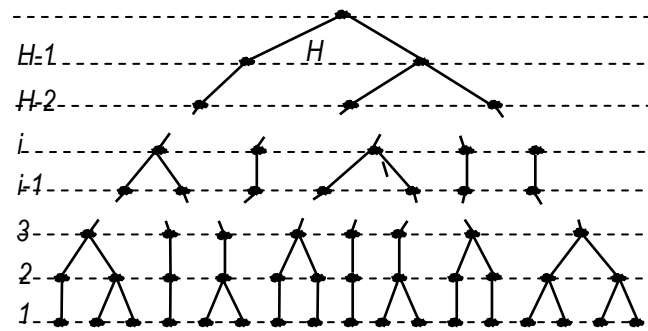


Fig. 2. The Optimal Reduction Tree

The main steps of  $T^R$  generation are:

- consider the set  $C_i$  of all clusters in the level  $i$ . At the first level we consider a set of all initial elements  $P = \{p_1, \dots, p_n\}$ ;
- form a set for all pairs of adjacent clusters for every cluster of the set  $C_i$ ;
- calculate the merging criterion values for all pairs of adjacent clusters;
- create the ordered list  $L(\eta)$  of pairs of adjacent clusters by the chosen merging criterion;
- form the new set  $C_{i+1}$  of clusters of the  $(i+1)$ -th level. There are several possibilities. In the best case - free clustering - we merge only the maximum number of independent pairs of adjacent clusters with the best value of the merging criterion. It could cause a large tree's height and consequently takes a lot of CPU time. The one possible way to reduce the running time is to take all independent pairs with  $\varepsilon$  given decreasing (increasing) of the best criterion value. The second way is to merge the first  $\lambda$  of all possible independent pairs, where  $\lambda$  ( $0 < \lambda \leq 1$ ) - is a reduction parameter. It is partially enforced clustering. In the last case when  $\lambda = 1$  we merge all clusters. It is enforced clustering. Here a height of the ORT is a minimal and therefore it takes the minimal running time but results could be worse. This case corresponds to the enforced circuit reduction that might not generate good natural clusters, because at every level we must merge together some clusters that do not have good criterion's value;
- form the new  $(i+1)$ -th level of the tree  $T^R$  by including a set of the new clusters, defined by merging and the rest clusters from the previous level that are not merged.

We must draw attention that we do not have to build binary Reduction Tree obviously. If, for example, one element creates two or more pairs with the same criterion's value, it is possible to join three or more elements together at one step. It reduces the height of tree and, of course, the CPU time. It is not easy to choose the criteria for clusters merging to receive the best clusters. There are many possible merging criteria. The main of them are:

- maximization the full number of the internal clusters' nets;
- minimization the full number of the external clusters' nets;
- maximization the full number of the subtraction of the of internal and external nets.

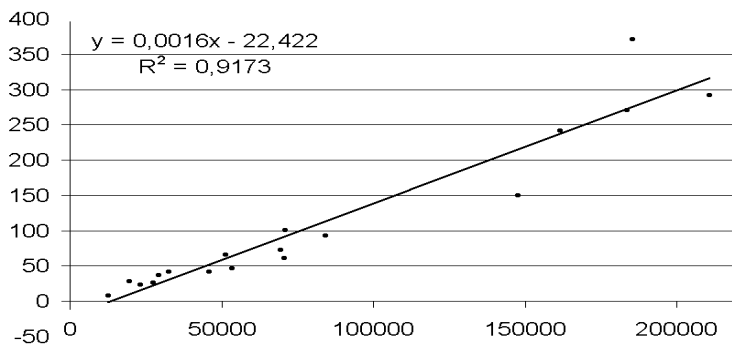


Fig.3. Circuit size vs time for TR building

There are possibilities of exploiting the various mixed merging criteria with the weight coefficients for the individual nets, the element numbers, the sizes of clusters, the time delay, etc. Very important is also the dependency circuit size vs time. As our experiments show, this dependency is close to linear by using the OCR method (Figure 3). Experiments are conducted with the library of the IBM01-IBM18 [Alpert, 1998].

Figure 4 shows the example of the Reduction Tree  $TR$  and the dependency of the cluster external nets' number vs reduction steps starting from element 11 for some circuit with 17 elements. It can help to receive the better dividing the circuit into partitions. The cutting  $\chi_8$  at eighth level shows that partitions (1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10) and (11, 12, 13, 14, 15,16, 17) create the minimal cut with 5 nets. Other cuttings have more external nets. No other partitioning method has such possibility.

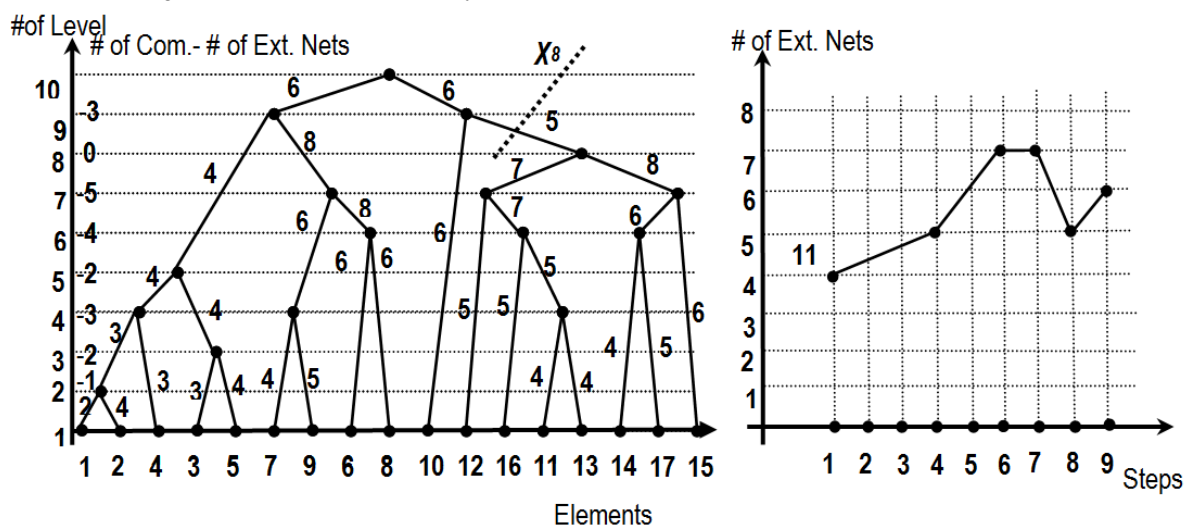


Fig. 4. Example of the Reduction Tree and the diagram dependency of the external nets clusters' number vs reduction steps

**Partitioning**

It is necessary to obtain the partitioning  $P^* = \{P_1, \dots, P_k\}$  for the set of elements  $P = \{p_1, \dots, p_n\}$  so that the quality function is optimized:

$$Q(P^*) \rightarrow \text{opt } Q(\tilde{P}), \quad \tilde{P} \in D,$$

while satisfying such or some other constraints:

$$(\forall P_i, P_j \in P^*) [|P_i| \approx |P_j|].$$

Set  $\tilde{P}$  is the arbitrary partitioning in the feasible region  $D$ ,  $k$  – the number of partitions. The solution should also satisfy the following additional conditions:

$$(\forall P_i \in P^*) [P_i = \{p_{i1}, \dots, p_{ini}\}, p_{ij} \in P; i = 1, \dots, k; j = 1, \dots, n_i];$$

$$(\forall P_i \in P^*) (P_i \neq \emptyset);$$

$$(\forall (P_i, P_j) \in P^*) [P_i \cap P_j = \emptyset].$$

The  $n_i$  is the number of elements of  $i$ -th partition.

By the OCR method we recommend to solve partitioning problem in the two stages: initial partitioning and partitioning optimization.

**Initial partitioning**

Using the constructive method, it is desirable to find an initial solution at the first stage, which must be improved by the iterative method at the second stage. The important peculiarity of the approach developed is that it is recommended to use the hierarchical circuit clustering, obtained by the OCR method at the both stages. For initial partitioning it is possible to use following algorithms: serial, parallel-serial, and dichotomy.

By the serial algorithm on the Reduction Tree  $T^R$  the vertex is found, whose number of elements is equal to or greater than the desired value. If the number of elements is what we desire, we create the first partition and move forward to the next partitions. If this number is greater then desired, the problem is to remove the necessary number of elements. The problem recursively continues to final solution.

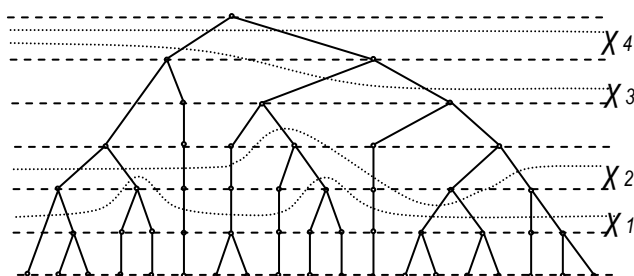


Fig. 5. Cuts in the Reduction Tree

Any cut of the Reduction Tree  $T^R$  by an arbitrary line forms subtrees (forest), the set of initial vertices (leaves) of each subtree can be considered as the set of elements of a single partition. For example, cut  $\chi_1$  at the Figure 5 can be used when it is necessary to split the circuits into the minimum number of partitions with a number of elements not greater than 2. Seven partitions are formed directly. The five elements remain ungrouped. For their assignment it is possible to construct the Reduction Tree in the subsequent repetition, and so on until the completion the problem. Cut  $\chi_2$  creates the three groups for three elements. To form the remaining partitions it is necessary to continue the process as in the previous case. Cut  $\chi_3$  gives good initial solution for three partitions (with 6, 5 and 8 elements). The first subcircuit can be directly incorporated into the solution as one partition, and then it would be necessary to transfer one element from the third to the second subcircuit. We obtain the partitions with 6, 6 and 7 elements. Cut  $\chi_4$  dived the circuit into two partitions with 6 and 13 elements. For receiving two approximately equal partitions we need to transfer three elements from larger partition into smaller one.

Dichotomy algorithm performs the top-down circuit division with constraint on the number of elements that should be equally divided to the desired number of elements at one partition. In the first step we consider the two highest vertices. This determines the number of possible partitions that can be formed from each vertex and the

number of elements in the remainders. The next step is to transfer the remaining elements from one piece to the other in the optimal way. The problem is reduced to the two new problems of the same type but of lesser size. In both cases, we use identical procedures to transfer the small number of elements from one piece to another, procedures which are performed recursively on the sets that decrease from the step to the step.

### Partitioning optimization

For partitioning optimization at the first step we build the separate ORT for two partitions  $T^{R_1}$ ,  $T^{R_2}$  and use the following procedures:

- $P1$ . The exchange: arbitrary element from one partition and arbitrary element (cluster) of other partition.
- $P2$ . The exchange: arbitrary clusters between two partitions.
- $P3$ . The exchange: arbitrary sets of clusters between two partitions.
- $P4$ . The transference: arbitrary element (cluster) from one partition to another.
- $P5$ . The transference: arbitrary set of clusters and elements from one partition to another.

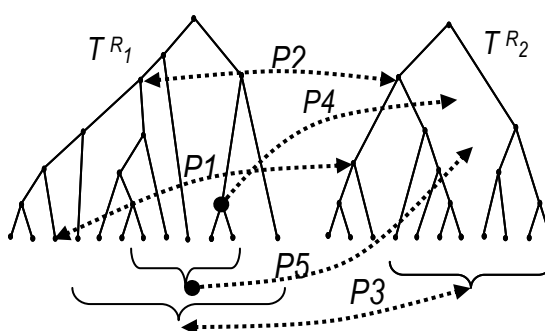


Fig. 6. Procedures for partitioning optimization

Some experimental bipartition results of test-case IBM01 [Alpert, 1998] by using our approach are shown at Table 1. For 100 randomly generated initial solutions we perform optimization. For escaping from the local extreme we use the perturbations by replacing clusters with the smallest value of the solution's worsening. The number of such perturbation is presented at the first column. For all initial solution we received the cut with 180 nets that we think it is an optimal result (our conjecture), as was received from other investigators by using another approaches [Karypis, 1997], [Saab, 2000]. Forth column shows the number of the best solutions that were received for all randomly generated initial solutions. The fifth-eighth columns show the numbers of solutions that have 1, 2, 5 and 10 % deviation from the best solution. Last columns show the average solution, average number of iteration and average runtime.

Table 1. Partitioning results for IBM 01

# steps of perturbations	Maximal solution	Minimal solution	# of optimal solutions	Deviation from the best solution				Average solution	Average # of iterations	Average runtime, s
				$\leq 1\%$	$\leq 2\%$	$\leq 5\%$	$\leq 10\%$			
0	707	180	1	6	27	38	41	307	22	174
1	699	180	1	25	57	61	61	236	33	204
2	699	180	3	43	65	67	67	225	43	233
3	699	180	4	52	68	68	68	222	53	261
5	699	180	8	60	68	68	68	222	73	318

### Packaging



It is necessary to obtain the partitioning  $P^* = \{P_1, \dots, P_k\}$  for the set of elements  $P = \{p_1, \dots, p_n\}$  so that the total number of partitions is minimized:

$$k \rightarrow \min ,$$

while satisfying the given constraints:

$$(\forall P_i \in P^*) [(n_i \leq n_{i \max}) \& (m_i^{ex} \leq m_{i \max}^{ex})].$$

Here  $n_i$  and  $m_i^{ex}$  are the numbers of elements and external nets (IO terminals) in each partition  $P_i$  that can not exceed the upper bounds  $n_{i \max}$  and  $m_{i \max}^{ex}$ . It should also satisfy the same additional constraints as for partitioning.

By the OCR method we also recommend to solve the packaging problem in two stages: initial packaging and packaging optimization.

**Initial packaging**

The algorithm begins to operate on the cluster of the Reduction Tree  $TR$ , which appears first in violation of the constraint on the number of elements. From this cluster we form the first partition with as many as possible elements without violation on constraint on the number of external nets. Two strategies are used: to remove the minimal number of elements and to identify the best cluster without violation on constraints. The next step consists of the addition of the maximum number of elements. The experiments reveal the advantage of simultaneous combination of both strategies that perform iterative removal and addition of elements and clusters. The partitions separated first have a good density; but the final ones – bad. This is caused first of all by the “greedy” partitioning by serial strategy. As a result, the number of partitions can be greater than the optimal.

*Table 2: Packaging Results for FPGAs*

a) with 64 CLB and 58 IO (Xilinx XC2064)							
Circuit	# of CLBs	# of Nets	Numbers of FPGAs				Theoretic optimum
			[Kuznar, 1993]	[Nan-Chi Chou, 1994]	[Bazylevych, 2000]		
					Initial	Opt.	
C499	74	123	2	-	3	2	2
C1355	74	123	2	-	2	2	2
C1908	147	238	3	-	4	3	3
C2670	210	450	6	-	6	6	4
C3540	373	569	6	6	8	6	6
C5315	531	936	11	12	12	10	9
C7552	611	1057	11	11	12	10	10
C6288	833	1472	14	14	14	14	14
b) with 320 CLB and 144 IO (Xilinx XC3090)							
s15850	842	1265	4	3	4	3	3
s13207	915	1377	7	6	6	4	3
s38417	2221	3216	12	10	10	8	7
s38584	2904	3884	17	14	14	10	10

**Packaging optimization**

The partitions with the number of elements lesser than the constraint merge into one or several without violating it. The next step is the optimization on the set of all partitions that allows to increase the number of elements, but not to exceed of constraints on the first group of partitions, which were not subject for merging. Often such optimization substantially decreases the number of external nets of final partitions up to the desired value. If this is

impossible to obtain, then the new final partition is divided into two smaller ones. The first partition should be without violations on constraints; the second may exhibit the violation on the number of external nets, if it is not possible to create it without violation, and so forth, up to the completion of the problem.

The experiments confirm the high efficiency of this approach on the set of some well-known test-cases. We merged only 20% of the better independent pairs at the every level of the Reduction Tree  $TR$  generation. The test results (#FPGAs) are shown at the Tables 2. We used the 64 CLBs and 58 IOs constraints (FPGA Xilinx XC2064) for the tests with Table 2 and 320 CLBs and 144 IOs (FPGA Xilinx XC3090) for the tests with Table 3. As one could see from the tables the obtained results are not worse, and in the 5 cases from 12 they are the best among the known and are optimal. If our results are not being theoretically optimal, they are close to the optimal solutions and differ from them minimally, i.e. only by one partition (circuits c5315, s13207, and s38417) or two partitions (circuit c2670).

### Floorplanning and placement

Combined hierarchical clustering and decomposition can be used for floor planning and placement. Such an approach is especially effective for large and very large-scale problems. The problem is solved in several stages:

- the bottom-up free hierarchical circuit clustering;
- the mathematical description of clusters by macro models at every level of decomposition;
- top-down multilevel placement with global and local optimization at every level of decomposition by using macromodels.

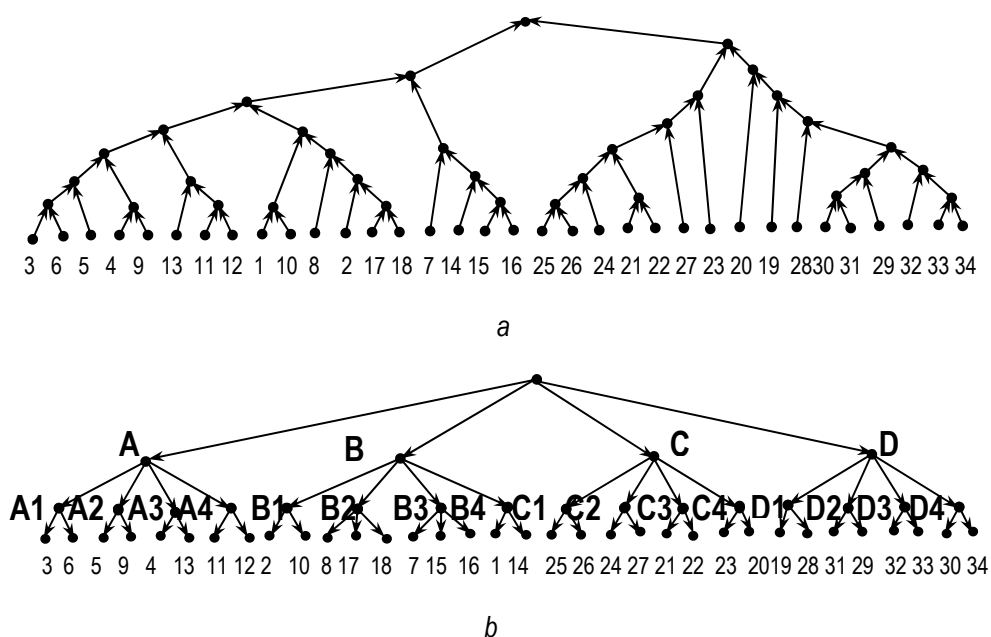


Fig. 7. Bottom-up hierarchical circuit clustering (a) and top-down 3-level of decomposition for Steinberg test-case

Figures 7 and 8 show the results of exploiting the developed approach for the Steinberg placement test-case [Steinberg, 1961]. For the first step (a) we build the ORT, for the next (b) – the tree level of decomposition (with 4, 16 macro models and 34 initial elements at the lowest level). At every level of decomposition we received some initial solution and performed it optimization using macro models (Figure 8 a, b and c) by Scanning-area method [Bazylevych, 1981, 1997]. We got the results of  $L_e = 4119,7$  (the summary length of all connections with the Euclidian metric), which is the best comparatively with the other known solutions.

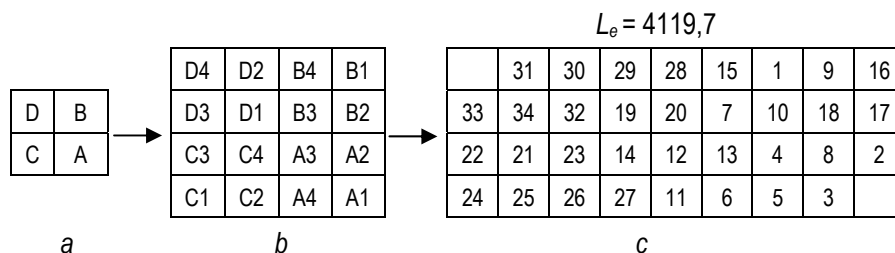


Fig. 8. Multilevel placement for Steinberg test-case

## Conclusions

Hierarchical circuit clustering is a good precondition for solving the physical design problems of large and very large-scale electronic devices - VLSI, SOC and for PCB. For hierarchical clustering we developed the OCR method. Basic algorithms were proposed for partitioning, packaging, floorplanning and placement problems. They were used to obtain the initial solutions with not very large number of macromodels, as well as for their optimization. The proposed algorithms have some new properties, for example, they can be efficient in choosing the most appropriate number of partitions into which it is necessary to divide the circuit; arbitrary division coefficient can be chosen for partitioning; the same procedures can be used for initial solution and their optimization. The suggested algorithms have near linear computational complexity and provide good quality of results. For all test-cases investigated, the results are not worse, and in many cases they are better comparatively with obtained by other known methods. For some cases, the optimal results were received for the first time.

## Bibliography

- [Bazylevych, 1975] R.P. Bazylevych, S.P. Tkachenko. The partitioning problem solving by the parallel reduction method. In: Vychuslitelnaia tehnika: Materialy konferencii po razvitiu technicheskikh nauk po avtomatizirovannomu proektirovaniu, V. 7, Kaunas, 1975, pp. 295-298 (In Russian).
- [Bazylevych, 1981] R.P. Bazylevych. Decomposition and topological methods for electronic devices Physical Design Automation, Lviv: Vyshcha shkola, 1981, 168 P, (In Russian).
- [Bazylevych, 1997] Roman P. Bazylevych, Taras M. Telyuk. VLSI and PCB elements placement optimizing using hierarchical scanning area method. 42 Internationales Wissenschaftliches Kolloquium. Technische Universitat Ilmenau. Ilmenau. 1997, pp. 594-599.
- [Bazylevych, 2000] R.P. Bazylevych, R.A. Melnyk, O.G. Rybak. Circuit partitioning for FPGAs by the optimal circuit reduction method. In: VLSI Design, Vol. 11, No 3, pp. 237-248, 2000.
- [Bazylevych, 2002] Bazylevych R.P. The optimal circuit reduction method as an effective tool to solve large and very large size intractable combinatorial VLSI physical design problems. In: 10-th NASA Symp. on VLSI Design, March 20-21, 2002, Albuquerque, NM, USA, pp. 6.1.1-6.1.14.
- [Bazylevych, 2002] Bazylevych R. P., Podolsky I.V., Bazylevych P.R. Hierarchical clustering – efficient tool to solve nonpolynomial combinatorial problems of high sizes. In: Shtuchnyy intelekt, Ukraine NAN, No. 3, 2002, pp. 447-483, (In Ukrainian).
- [Bazylevych, 2007] R. Bazylevych, I. Podolsky and L. Bazylevych. Partitioning optimization by recursive moves of hierarchically built clusters. In: Proc. of 2007 IEEE Workshop on Design and Diagnostics of Electronic Circuits and Systems. April, 2007, Krakow, Poland, pp. 235-238.
- [Garbers, 1990] Jorn Garbers, Hans Jurgen Promel, Angelika Steger. Finding Clusters in VLSI Circuits. In: Proc. of IEEE/ACM Intern. Conf. on Computer-Aided Design, Santa Clara, 1990, pp. 520-523.

- [Cong, 1993] Jason Cong and M'Lissa Smith. A Parallel Bottom-up Clustering Algorithm with Applications to Circuit Partitioning in VLSI Design. In: Proc. 30th ACM/IEEE DAC, 1993, pp. 755-760.
- [Dutt, 1996] S.Dutt, W.Deng. VLSI circuit partitioning by cluster-removal using iterative improvement techniques. In: ICCAD'96, pp. 194-200.
- [Karypis, 1997] G.Karypis, R.Aggarwal, V.Kumar, and S.Shekhar, Multilevel hypergraph partitioning: Application in VLSI domain. In: DAC 97, pp. 526-529.
- [Saab, 2000] Youssef Saab. A new multi-level partitioning algorithm. In: VLSI Design, Vol 11, No 3, pp. 301-310, 2000.
- [Alpert, 1998] C.J. Alpert. The ISPD-98 Circuit Benchmark Suite. In: Proc. ACM/IEEE Intern. Symposium on Physical Design, April 1998, pp. 80-85.
- [Kuznar, 1993] Kuznar, R., Brglez, F. and Kozminski, K. Cost minimization of partitions into multiple devices. Proc. Of IEEE/ACM 30th Design Automation Conference, 1993, pp. 315 - 320.
- [Nan-Chi Chou, 1994] Nan-Chi Chou, Lung-Tien Liu, Chung-Kuan Cheng, Wei-Jin Dai and Rodney Lindelof. Circuit partitioning for huge logic emulation systems. Proc. 31 ACM/IEEE Design Automation Conference, 1994, pp. 244-249.
- [Steinberg, 1961] Steinberg L. The backboard wiring problem a placement algorithm. SIAM Review, 1961, v.3, '1, pp.37-50.

---

#### Authors' Information

---



**Roman Bazylevych** – Prof. dr.hab., University of Information Technology and Management in Rzeszow; 1/42 Dobjusha str., Lviv, 79008, Ukraine;

email: [rbazylevych@wsiz.rzeszow.pl](mailto:rbazylevych@wsiz.rzeszow.pl);

Major Fields of Scientific Research: combinatorial optimization, computer-aided design, algorithms

**Marek Palasiński** – Prof. nadzw. dr hab., Chair of Mathematics and Computer Science Foundations, University of Information Technology and Management in Rzeszow;

email: [mpalasiniski@wsiz.rzeszow.pl](mailto:mpalasiniski@wsiz.rzeszow.pl);

Major Fields of Scientific Research: theoretical computer science, theory of algorithms, graph theory, data mining and algebraic logic



**Lubov Bazylevych** – senior scientist, Institute of Mechanical and Mathematical Applied Problems of the Ukrainian National Academy of Sciences;

email: [lbaz@iapmm.lviv.ua](mailto:lbaz@iapmm.lviv.ua);

Major Fields of Scientific Research: applied mathematics and mechanics, combinatorial optimization, computer science problems.

## ARTIFICIAL INTELLIGENCE IN MONITORING SYSTEM

Lucjan Pelc, Artur Smaroń, Justyna Stasieńko

**Abstract:** *The article presents the neural network constructed in order to use it for monitoring. Its role is to recognize the events in alarm situations (theft, burglary etc.). The film presenting a real break-in into the car was used while testing this network. The main task of monitoring system based on the neural network is to compile such a network which shows the alarm situations as soon as possible using the available equipment.*

**Keywords:** *neural network, monitoring system, neuron.*

**ACM Classification Keywords:** *1.2 Artificial Intelligence – 1.2.6. Learning*

---

### Introduction

---

The monitoring programs are often used in order to gather the information about the amount and the quality of the observed object. The gathered information makes it easier to make right decisions, especially if this state is hazardous for human being and the surrounding. It allows also to improve or remove the results in the existing situation. The article aims at showing one of these solutions. There are many monitoring systems available on his market. They are different as far as the quality, functionality and price are concerned. If we take into consideration the functionality the following types can be distinguished: systems which record on the continuous basis, systems which record if the movement is detected and those which record before and after the incident. The current monitoring systems are able to: record on the continuous basis, detect the movement and record the image, detect the disappearance of the image, follow and count the objects, control the cameras, one or more visual channels, and inform the operator by an e-mail or sms. The cost of such systems amounts to 300 – 10000zł. Unfortunately, there are not many systems which can define the situation before the burglary or theft etc. The majority of systems posses an insignificant or do not posses any functionality as far as the process of supporting the system operator is concerned especially in case of recognizing important facts and events. The monitoring systems are used rather for recording and gathering the courses of events, even undesirable ones. It requires large disk storage for recording an image. In some systems (e.g. Taiwanese ACTI- APP-2000-32) an intelligent management of memory was used by means of recording the events before they occur which is often called "buffering the image". If the system posses such ability, it buffers the image all the time and in the moment of detection the alarm it has also the images recorded before the accident, for example 5sec. Supporting the system operator while recognizing undesirable situations on the recorded image and responding automatically requires from the system to solve problems connected for example with pattern recognition. One of the authors of this article used probabilistic timed automata which if connected with particular spheres on the observed image allowed to define important actions in reference to the given problem [Pelc, 2008]. This article is a kind of a supplement and expansion of that approach. Instead of using probabilistic timed automata the neural network was used here.

While creating the monitoring system three rules should be taken into consideration: the periodicity of measurement, the unification of the equipment and methodology used for the measurement and observation as well as the unification of the results interpretation. This case study uses a film from the Internet showing the real break-in into the car [<http://www.youtube.com/watch?v=pLKjm2uGrU4&feature=related>]. It resulted

---

in the idea of creating the monitoring system based on one of the Artificial Intelligence applications “neural network”

---

### **The aim of article**

---

The aim of this article is to create a low-cost monitoring system which can be compiled on the basis of computers designed for the general use as well as typical cameras (e.g. network cameras)

The system assumptions:

- one or two cameras
- limited computability of the equipment and disk
- better functionality than in typical monitoring systems
- algorithm of artificial intelligence – the neural network whose computability is not very complex
- the ability to recognize particular actions and reactions depending on their character:
  - ✓ neutral actions – normal work of the system
  - ✓ suspicious actions – generating the warnings by the system, starting recording the sequence of frames
  - ✓ prohibited actions – alarming, calling the operator and recording the sequence of frames

The main feature which distinguishes the monitoring system from many others is its ability to take decisions in order to prevent the prohibited actions and not only to record them. The following method of solving the problem was accepted:

- action was defined as the occurrence of the given movement trajectory in particular period of time
- the movement trajectory is determined on the basis of recognition of the object being observed in the successive spheres of observation
- resignation of the constant recording of the images in aid of the recording which is caused by events
- the event causing the recording is the occurrence of the beginning of the action
- introducing the spheres of observation on the image recorded by the camera limiting at the same time the size of the in-put data.

---

### **The method of defining the spheres of observation**

---

The image coming from the camera is covered by the network of spheres of observation consisting of 125 elements. The size of the network is determined automatically depending on the picture definition. The network density can change adjusting the image from the camera. The network is related to the scene or objects being observed whose position may change. However, the main stress is not put here on fluent change of the position and following the object.

The example of using the network related to the scene can be the observation of the gateway. However, in case of the network connected with the object, it can be presented by means of observation of a car parked in different ways and positions. To sum up, the steps of defining the spheres are as follows:

- (1) determining if the spheres should take the form of a uniform or condensed network
- (2) choosing the object or the scene
- (3) indicating the essential points of the image for which the system should condense the network
- (4) if needed, defining the size of grid's fields which is different than the implicit one (implicitly there are three types of the size and density of grid's fields: large, medium and small)

- (5) in case of the network related to the particular object there is a need of creating the definition of this object. The system defines the object on the basis of its simplified shape, colour or its texture.
- (6) If the system is going to be used in applications whose shape does not change or the changes are small than in (5) the size of the object is also taken into consideration. It is worth remembering that in the systems with cameras there is an apparent change of the object size depending on its distance from the camera.

Fig.1 shows the example of the image seen by means of a camera with a network characterized by reduced density. The biggest density is intended for the observed vehicles and it is the situation when the network is related to the observed objects.



*Fig 1. An example of the definition of the sphere of observation. Source: Own elaboration.*

**The logic of the created system**

As it has been mentioned earlier, the system should inform about the potential risk of the occurrence of the prohibited action, which as a result, allows to take remedial. In consequence, the logic of the system should provide the proper sensitivity as far as the recognition of the action is concerned and it should also be characterized by the ability of limited prediction. It should be remembered that one of the assumptions at the beginning of this article was the system's ability to recognize the neutral, suspicious and prohibited actions. Detection of the prohibited action is connected with alarming and calling the operator and that is why the system cannot do it precipitously. In other words, the system cannot be too sensitive and its reactions should not be exaggerated. It is especially important in case of determining the prediction. Taking into consideration all of these assumptions, the logic of the system should demand the occurrence of one of these three sorts of action on the basis of the information in the defined spheres of observation. The classic neural networks were used to support this solution. It was necessary to decide about the number of layers in the network. It was assumed that the exit layer will consist of at least three neurons indicating the occurrence of a particular kind of action. In the entrance layer, there should be as many neutrons as the spheres of observation, however, this solution would be enough only in case of recognizing the statistical actions. In the presented example the time lag also should be mentioned. Time is measured with the occurrence of successive frames. That is why, it is necessary to add one more neuron informing about the number of frames which have just appeared. It allows for taking into

consideration their movement dynamics, and not only the statistical trajectory. Therefore, it can be assumed that in the “minimal” version the number of neurons in the entrance layer amounts to the number of spheres of observation enlarged by one.

Classic neural network consists of one or few hidden layers. In the experiment the number of the hidden layers and the number of the neurons in each of them were the object of the simulation analysis whose results will be soon discussed in this article. Taking into consideration the previous assumptions referring to computability, two and more hidden layers were analyzed.

### The simulation

The aim of the simulation was to determine the construction of neural network which is the most suitable from the viewpoint of the given problem. Such a network should for example:

- recognize defined actions accurately
- predict the occurrence probability almost without any mistakes
- include the least neurons
- learn quickly

According to the assumptions accepted previously the network with one and two hidden layers were analyzed paying particular attention to the accuracy of recognizing the actions as well as fast learning.

Fig.2, Fig.3, and Fig.4 present hypothetical and simplified movement trajectories referring to three types of actions discussed previously. Fig.2 deals with the forbidden action, Fig.3 is related to the suspicious action and Fig.4 presents the neutral action. The protected sphere is marked with “x” and darkened. The black squares represent the recorded trajectory of movement. The number under each of the pictures informs about the number of frames falling on the particular trajectory.

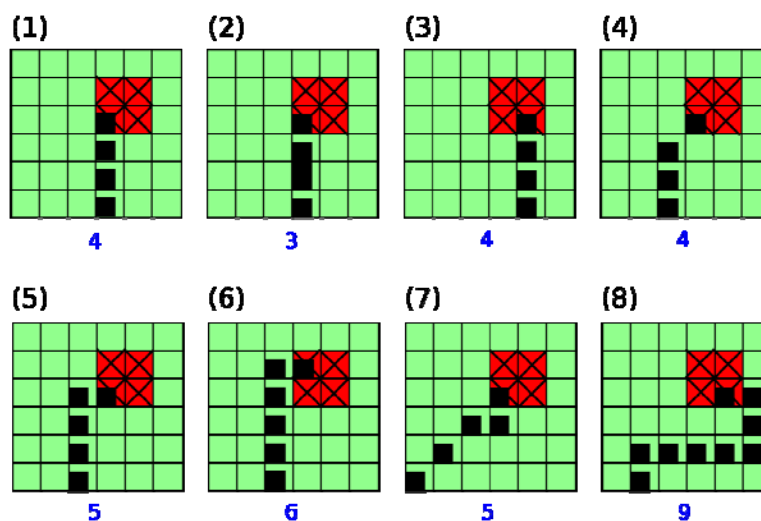


Fig. 2. Examples of movement trajectory for the forbidden-alarm action. Source: Own elaboration.



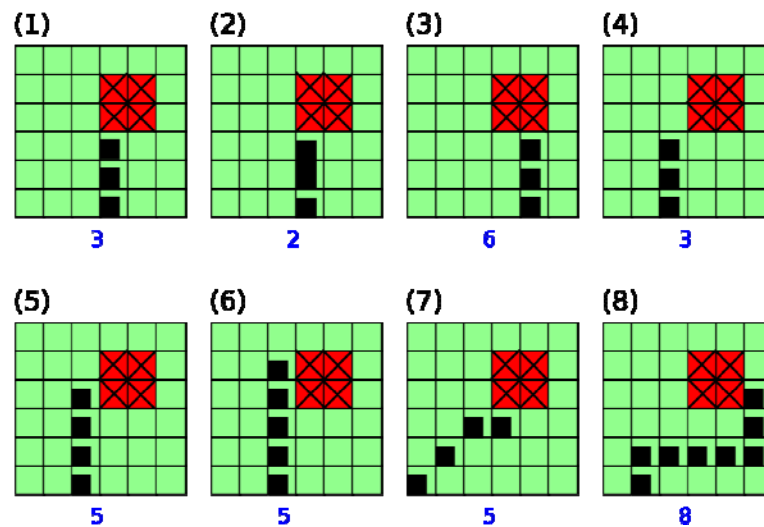


Fig. 3. Examples of movement trajectory for the suspicious-warning action. Source: Own elaboration.

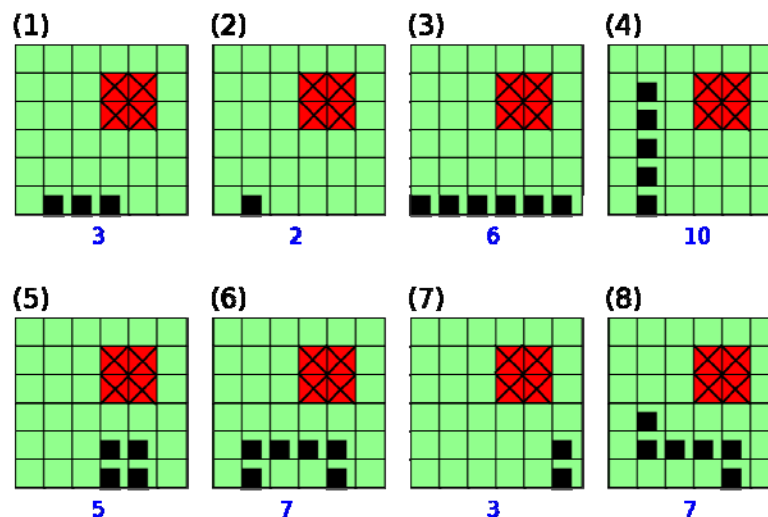


Fig.4. Examples of movement trajectory for the neutral-normal action. Source: Own elaboration.

The elementary picture (1) presented by Fig.2 represents the trajectory going in four steps from the bottom part directly to the marked sphere. Below there are some situations, which could occur for the discussed example of the scene:

- The prohibited action – if the trajectory ends or goes through the marked sphere (Fig.2)
- The prohibited action which is very probable – if the trajectory with great dynamics goes towards the marked sphere (compare the elements (1) and (3) from Fig.2)
- The suspicious action – when the trajectory goes towards the marked sphere but does not reach it, the movement dynamics is not important here
- The suspicious action which is very probable - when the trajectory is near the marked sphere regardless of the dynamics (compare elements (5), (6), (7), (8) from Fig.3)
- The neutral action – the movement trajectory is in a short distance away from the marked sphere regardless the dynamics (Fig.4).

Even after analysing Fig.2 roughly it can be seen that if the object moves quickly on the trajectory with the marked sphere, the system should inform as quickly as possible the occurrence of the prohibited action before

the trajectory reaches the marked sphere. In practice, it would allow to prevent the prohibited action to reach the marked sphere in the given example. The permanent persistence of the suspicious action may indicate the potential occurrence of the prohibited action. The situations presented on Fig.3 and Fig.4 were chosen in such a way that the situations which at the beginning looked like forbidden after the occurrence of the successive frames turned out to be suspicious for example Fig.3 (1) and Fig.2 (1), Fig.3 (2) and Fig.2 (2) etc. Taking into consideration the assumptions that the system should be characterized by the ability to predict the situation, but it should not react precipitously, such kind of situations are interesting.

---

## The simulation results

---

During the simulation it was checked how quickly the network with one or two hidden layers is able to learn as well as the accuracy of recognition if the action is forbidden, suspicious or neutral.

- **The comparison of the capacity of learning for the network with one or two hidden layers**

Below, there are the tests results of the capacity of learning for the network with one hidden layer depending on the number of the neurons in the hidden layer. Fig.5 presents the reduction of mistakes in the process of learning depending on the number of epochs. Other progresses refer to various numbers of neurons in the hidden layer. It is important to emphasize that at the beginning the number of neurons in the hidden layer improves the parameters of learning but after exceeding the certain number of the neurons the situation changes the other way round. On the basis of the conducted analysis the formula was defined. It counts the number of neurons in the single hidden layer as a quotient of the product and the sum of the number of neurons in the input and output layer in the network (1).

$$N_h \approx \frac{N_{in} * N_{out}}{N_{in} + N_{out}} \quad (1)$$

where

$N_{in}$  – the number of neurons in the entrance layer

$N_{out}$  – the number of neurons in the exit layer

$N_h$  – the number of neurons in the hidden layer.

If the number of neurons in the entrance layer is much bigger than the number of neurons in the exit layer (at least of an order of magnitude), the radical dependence, which combines the number of neurons of the hidden layer with these from the entrance and exit layer (2), seems to be more accurate.

$$N_h \approx 1.2 * \sqrt{N_{in} * N_{out}} \quad (2)$$

The analogous simulations were carried out with two hidden layers. The number of the neurons in the first hidden layer is chosen on the basis of relations resulting from formula (2).

The comparison of progresses for various number of neurons in particular hidden layers presented by means of Fig.6 indicates the validity of accepting the correlation that the number of neurons in the second layer should be smaller by half than the number of neurons in the first layer (according to the rule of pyramids) (3).

$$N_{h2} \approx 0.5 * N_{h1} \quad (3)$$

Having analysed the networks which learned faster with one or two hidden layers it was included that the network with only one layer is able to learn faster.

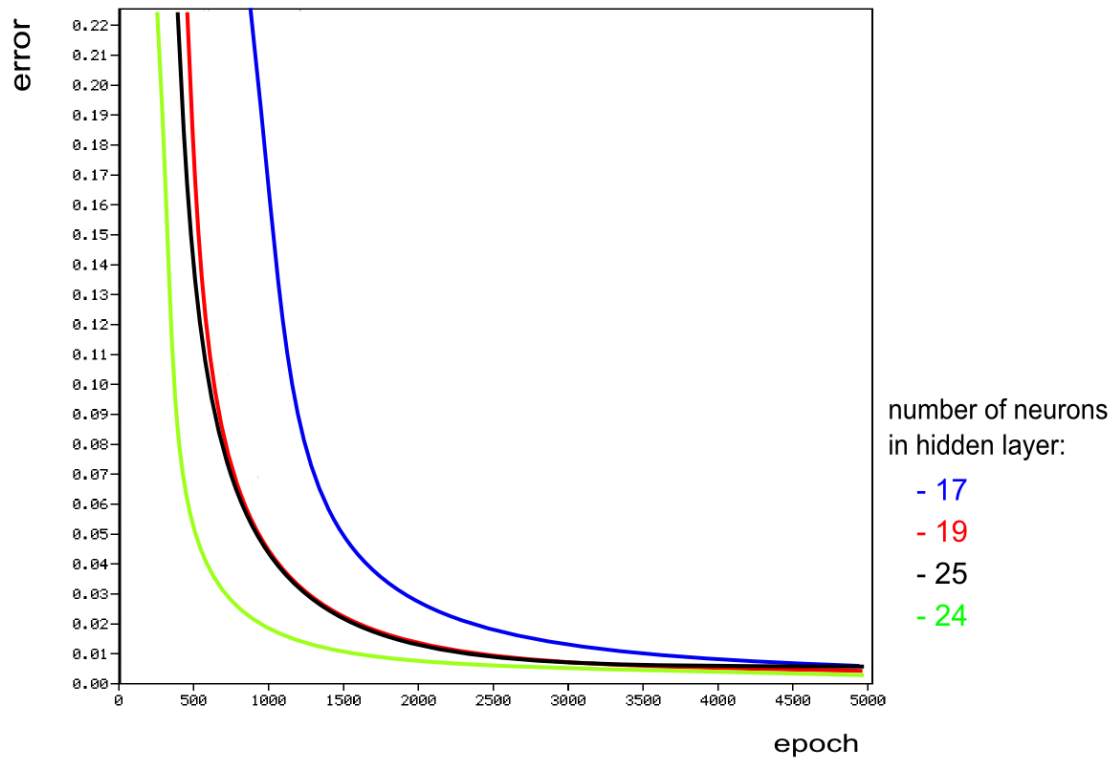


Fig.5. The capacity of learning depending on the number of epochs. Source: Own elaboration

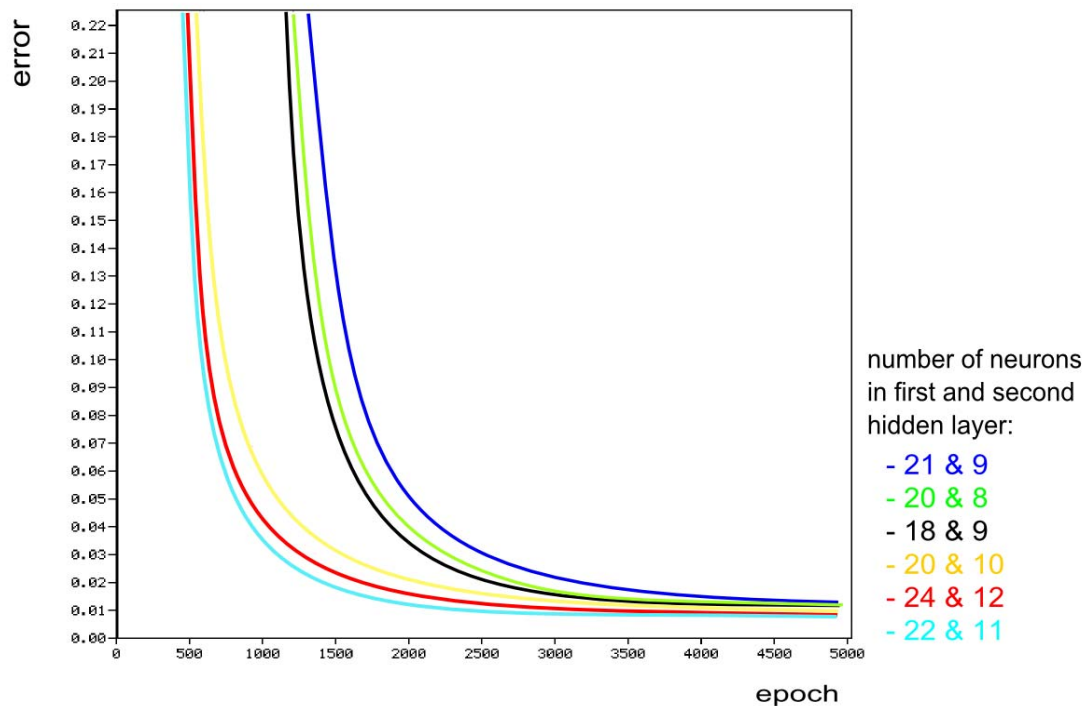


Fig.6. The capacity of learning depending on the number of neurons for two hidden layers. Source: Own elaboration.

- **The comparison of the accuracy in actions recognition**

It turned out that the accuracy of the recognition of the action is bigger for the network with two hidden layers than the network with just one hidden layer. However, in this case the recognition of the forbidden action during the trajectory analysis was precipitate. In consequence, it was the trajectory mostly connected with the suspicious

action. That is why, at the dynamic recognition of the action or rather the prediction of the action, the network with one layer turned out to be more accurate solution. It is illustrated by Table 1 which shows how the neuron network recognized the forbidden-alarm action, and suspicious-warning action in the particular period of time, described in Table as steps. Analyzing for example case, it is easy to notice that the network with two hidden layers reacts too rapidly and notices in the developing warning situation the alarm situation. It can be easily seen in steps 4 and 5 where the network guesses the alarm situation wrongly with the probability bigger than 99%. In the successive steps the network recognizes the warning correctly, but taking into consideration the previous assumptions such behavior of the network is delayed. For the same steps (4 and 5) the network with one layer recognized the alarm with the probability of 1,5% and the proper warning situation with more than 92%. Although the network with one hidden layer recognizes the finished actions less accurately than the network with two hidden layers, for the investigated application the network with one hidden layer is more suitable.

Table 1 Example situations recorded by monitoring system

Example situations recorded by monitoring system		Response of the system in percentage					
		One hidden layer – 19 neurons			Two hidden layers – 22 and 11 neurons		
		alarm	warning	neutral	alarm	warning	neutral
Case 1: warning	Step 1	1,72%	0,19%	98,09%	28,06%	0,08%	71,86%
	Step 2	0,82%	0,41%	98,76%	22,72%	0,08%	77,20%
	Step 3	2,85%	1,93%	95,22%	39,02%	0,08%	60,90%
	Step 4	1,55%	92,43%	6,02%	99,13%	0,79%	0,08%
	Step 5	1,49%	94,62%	3,89%	99,12%	0,80%	0,08%
	Step 6	0,67%	98,83%	0,50%	1,77%	98,23%	0,00%
	Step 7	0,60%	99,03%	0,37%	0,58%	99,42%	0,00%
	Step 8	0,55%	99,08%	0,36%	0,48%	99,52%	0,00%
	Step 9	0,73%	98,97%	0,30%	0,34%	99,66%	0,00%
Case 2: alarm	Step 1	6,73%	0,00%	93,26%	5,16%	0,09%	94,76%
	Step 2	1,61%	0,24%	98,15%	56,15%	0,08%	43,77%
	Step 3	0,87%	0,95%	98,18%	73,90%	0,10%	26,00%
	Step 4	1,44%	7,81%	90,74%	94,39%	0,15%	5,46%
	Step 5	97,20%	0,64%	2,16%	98,70%	0,26%	1,04%
	Step 6	76,49%	23,40%	0,10%	2,76%	97,24%	0,00%
	Step 7	98,65%	1,28%	0,07%	12,84%	87,16%	0,00%

Source: Own elaboration.

### The applied algorithm of learning

In the elaborated program was used the algorithm of error backpropagation based on the example found in the literature (4).

$$\begin{aligned} \delta_{ij} &= O_{ij} * (A_j - O_{ij}) * (1 - O_{ij}) \\ \delta^t_{ij} &= \delta^{t-1}_{ij} + O_{ij} * (1 - O_{ij}) * \delta^{t-1}_{i+1,k} * W_{i+1,k,j} \\ W^t_{ij,k} &= W^{t-1}_{ij,k} + \eta * \delta_{ij} * O_{i-1,k} + \alpha * (W^{t-1}_{ij,k} - W^{t-2}_{ij,k}) \\ O_{ij} &= \frac{1}{1 + e^{\beta * (-I_{i,j} + bias_{i,j})}} \end{aligned} \tag{4}$$

where

- $i,j,k$  – index of: layer, neuron and weight
- $t$  – point of time
- $\alpha, \beta, \eta$  – momentum, beta and learning rate
- $I, O, A$  – input, output and expected value
- $\delta$  – error
- $W$  – weight

Paying attention to the given problem the adjustment of the algorithm and coefficients was conducted. The results of the adjustment of coefficients  $\alpha$  and  $\eta$  is shown by Fig.7. The most proper choice turned out to be  $\alpha = 0.6$  and  $\eta = 0.2$ .

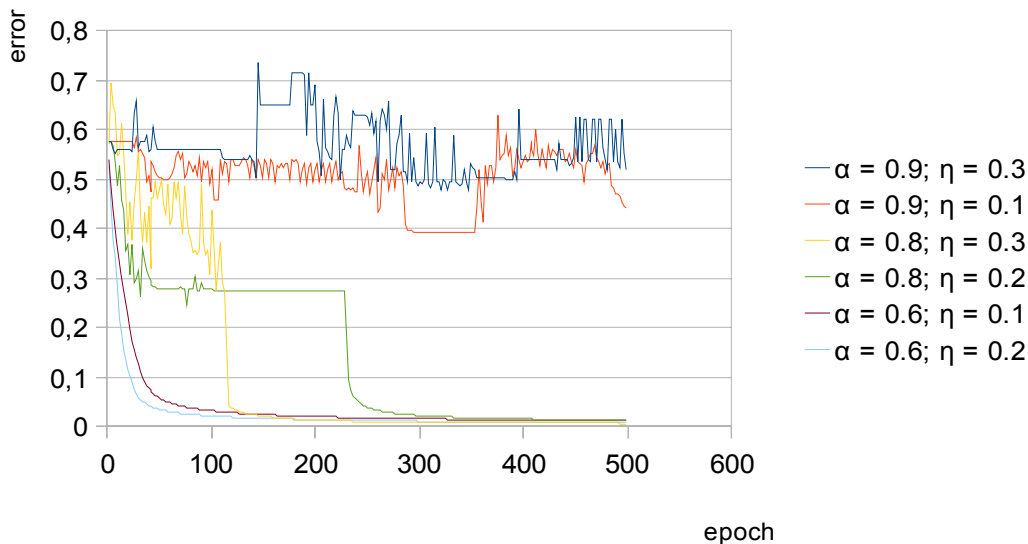


Fig.7. The selection of  $\alpha$  and  $\eta$ . Source: Own elaboration

After choosing  $\alpha$  and  $\eta$  it is possible to estimate the coefficient  $\beta$ . Fig.8 shows the results of experiments for various  $\beta$  coefficients.

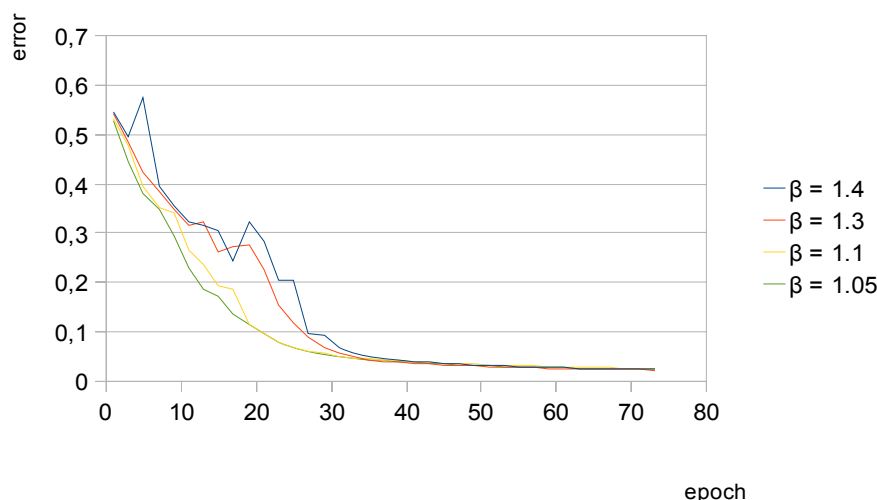


Fig.8. The selection of  $\beta$ . Source: Own elaboration

Analysing the mistake dependence upon the number of epochs it can be noticed from Fig.8 that the most satisfactory results occur for  $\beta$  which is close to 1.

The elaborated program is equipped with the mechanism of measuring out the time needed for learning and recognizing. According to the assumptions from the beginning of this project the computers intended for the general usage were used with Windows 7 Home Premium 64 bit, Dual Core AMD Athlon 64X2 L310 processor and 1.2 GHz RAM 2 GB memory. Satisfactory results of learning were achieved after less than 500 epochs and the time of learning was 5 seconds. The time of recognition fluctuated in single milliseconds. Those two results confirmed the validity of the proposed attitude on the computer equipment intended for the general use which is not very expensive.

### The conducted experiments

The given monitoring conception of the system was implemented according to the policy and used for the image analysis coming from the real monitoring. The image referred to the recorded burglaries into the parked cars and thefts (film). Fig.9a) presents the network connected with the real car (first from the left) whereas Fig.9b) shows the network connected with the silver car- on the right. The conception accepted here related to the network with the variable density.



Fig.9. Image from the camera with the network put by the monitoring system. Source: Own elaboration

Two examples of the neuron network with one hidden layer were considered. The number of neurons in the hidden layer was established according to pattern (2) while in the adjustment algorithm the accepted coefficients were  $\alpha = 0.6$ ,  $\eta = 0.2$  and  $\beta = 1.05$ . The network was taught on the basis of the patterns worked out according to the human intuition patterning on the trajectories analysed on Fig.2, Fig.3 and Fig.4. The expert's knowledge was used for working out these patterns.

---

### **The practical method of adjusting the system**

---

The simulation analysis and the experiments conducted in real situations let to formulate some practical rules of configuring of the elaborated system for other applications.

- **Determining the network**

The network can be chosen as homogeneous or with changeable density. Then, the system operator has to indicate the crucial sphere in the image seen by the camera and establish if the network should be connected with the scene being observed or if it should be the object in that scene.

- **Teaching the network**

In the situation where the current application is very similar to the previous one, the predefined network without teaching can be used. For example, if take into consideration the car park similar to the one presented by Fig.9 etc. If the current situation is totally different from the ones available in the previous study, the proper patterns should be prepared.

It is important that they can be prepared in two ways:

- 1) Recording the mock situation or
- 2) Determine on the basis of the expert knowledge the trajectories by indicating the proper spheres of the network.

Generally, there are even several dozen and that is why it is the task available for the human perception and possible to be released in short time. The system operator has to prepare the trajectories analogous for those presented by Fig.10 b) and c) in the form of the marked spheres of the network.

a)

37	38	39	40				41				42				43
44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	
			59	60	61	62	63	64	65						
			66	67	68	69	70	71	72						
73	74	75	76	77	78	79	80	81	82	83	84	85	86		
			87	88	89	90	91	92	93						
			94	95	96	97	98	99	100						
106	108	107	109				110	111	112	113	114	115	116		
	121		117		118		119		120						
	122	123	124				125								

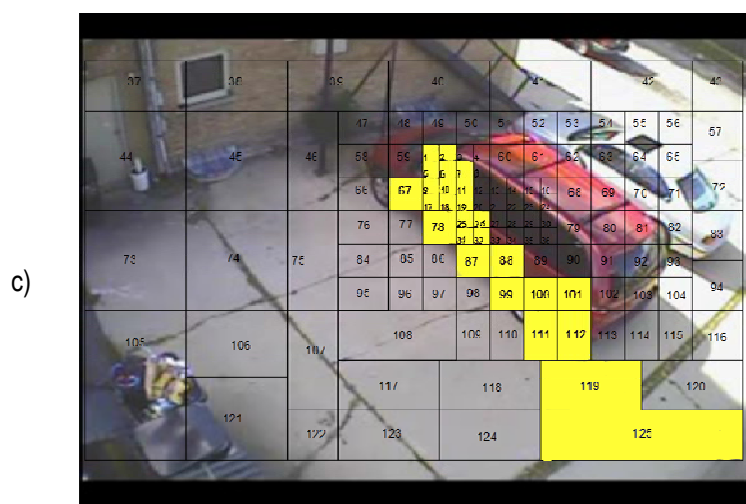
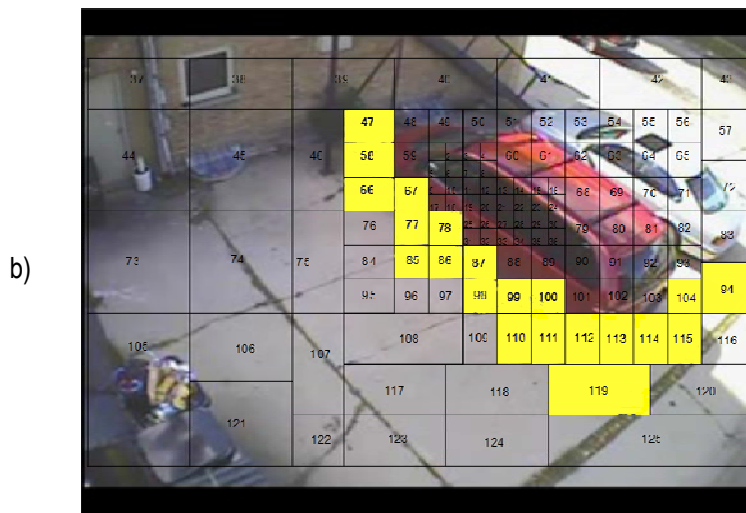


Fig 10. Course of events in the burglary: a) The network put by the system, b) Suspicious action from the point of view of the in-put neurons, c) The prohibited action. Source: Own elaboration

### Conclusions

The presented monitoring system fulfils all the aims established at the beginning of this article which referred to creating low-cost monitoring system which can be combined on the basis of computers intended for general use as well as typical cameras such as network cameras. In addition, the main characteristic feature distinguishing this particular monitoring system from others available on the market was discovered. This feature refers to the



ability of take to prevent the prohibited actions and not only to record them. The use of typical neuron network with the standard configuration allowed to receive satisfactory results.

---

### Bibliography

---

[Pelc, 2008] L.Pelc, B.Kwolek. Activity Recognition Using Probabilistic Timed Automata. In Peng-Yeng Yin: Pattern Recognition. Techniques, Technology and Application. In-Tech, November 2008, pp. 345-360.

[Rutkowski, 2009] L.Rutkowski. Metody techniki sztucznej inteligencji. PWN, Warszawa, 2009.

[Tadeusiewicz, 1993] R.Tadeusiewicz. Sieci Neuronowe. Wydawnictwo RM, Warszawa, 1993.

[Wszolek, 2006] W.Wszolek. Metody sztucznej inteligencji w systemach monitoringu hałasu lotniczego. Zeszyt 3/2006 [patenty.bg.agh.edu.pl/graf/Newsletter\\_AGH\\_Czerwiec\\_2011.pdf](http://patenty.bg.agh.edu.pl/graf/Newsletter_AGH_Czerwiec_2011.pdf) [06.2011]

[Zurada, 1996] J.Żurada, M.Barski, W.Jędruch. Sztuczne sieci neuronowe. PWN, Warszawa, 1996

Netography:

<http://www.youtube.com/watch?v=pLKjm2uGrU4&feature=related>

<http://www.ai.c-labtech.net/sn>

<http://kik.pcz.pl/nn/index.php>

<http://www.neuron.kylos.pl/pliki/start.html>

<http://www.willamette.edu/~gorr/classes/cs449/Backprop/backprop.html>

---

### Authors' Information

---



**Lucjan Pelc** – lecturer, Rzeszow University of Technology; W. Pola Street 2, 35-959 Rzeszów and The Institute of Technical Engineering, The Bronisław Markiewicz Higher State School of Technology and Economics, Czarnieckiego Street 16, 37-500 Jarosław, Poland; e-mail: [lpelc@prz-rzeszow.pl](mailto:lpelc@prz-rzeszow.pl)

Major Fields of Scientific Research: Real-time protocols and systems, Formal description techniques, industrial vision systems



**Artur Smaroń** – computer science's student, The Bronisław Markiewicz Higher State School of Technology and Economics, Czarnieckiego Street 16, 37-500 Jarosław, Poland; e-mail: [arturo.s@op.pl](mailto:arturo.s@op.pl)



**Justyna Stasienko** – lecturer, The Institute of Technical Engineering, The Bronisław Markiewicz Higher State School of Technology and Economics, Czarnieckiego Street 16, 37-500 Jarosław, Poland; e-mail: [justyna.stasienko@pwste.edu.pl](mailto:justyna.stasienko@pwste.edu.pl)

Major Fields of Scientific Research: Management Information Systems, Business information technology

---

---

## DATABASE AND KNOWLEDGE BASE AS INTEGRAL PART OF THE INTELLIGENT DECISION SUPPORT SYSTEM, CREATED FOR MANUFACTURING COMPANIES

**Monika Piróg-Mazur, Galina Setlak**

**Abstract:** *The paper presents the structure of a database and a knowledge base which are the integral part of the Intelligent Decision Support System, being developed for a manufacturing company operating in the glass industry. Both modules will be used in the advisory system, whose functions will be classification of defects of products (here: glass packaging, e.g. bottles, jars) and selection of an appropriate (the most beneficial) method of elimination of defects arising in the manufacturing process.*

**Keywords:** *intelligent decision support systems, knowledge base, knowledge representation, reasoning process.*

**ACM Classification Keywords:** *I. Computing Methodologies, I.2.1 Applications and Expert Systems, J. Computer Applications,*

---

### Introduction

---

In a manufacturing company making decisions in the scope of production preparation processes is a basic and key element of the whole manufacturing process.

It is essential for manufacturing companies to obtain information necessary to make an appropriate decision as soon as possible. They employ a lot of experts and specialists to react appropriately at every moment and to make suitable decisions. Decision-making processes are also supported by integrated computer systems, which continuously gather data and analyse different areas of the manufacturing process.

Decision support systems using artificial intelligence techniques are the systems which combine the potential for gathering and processing enormous amounts of data, using increasingly diversified models and intelligent utilizing data and knowledge gathered. The main purpose of development of an intelligent decision support system is to reflect experts' knowledge and experience, which are indispensable for solving problems by the system. Integration of intelligent methods allows to create better and more precise methods which can be applied in this field. In intelligent systems reasoning plays the most important role. The indicator of a system's "intelligence" is the ability to make decisions (through the reasoning process) and the ability to learn and acquire knowledge [Rojek, 2010].

Intelligent Decision Support Systems (IDSS) are the systems which have the potential for gathering and processing huge amounts of decision-making information, conducting analyses of the information and using diversified models, data and knowledge gathered to solve complex decision-making problems. The essential parts of an intelligent decision support system are a database and a knowledge base [Zieliński, 2000].

A database is one of the most important sources of decision-making information for a knowledge base in an IDSS. A database plays a basic role when developing a knowledge base in order to support the technological process, including the quality control process. The purpose of this paper is to present the issues concerning the development of these two component parts of the advisory system being developed.

---

### Characteristics of the selected object of research.

---

The design of the intelligent decision support system, which is presented in this paper, is being realized for the Glassworks, a company operating in the sector of large companies. In total, the glassworks has 14 production lines, which work in a three-shift system; the capacity of one production line per one shift is 200,000 items of

finished product. To illustrate it better – an automatic machine works with the speed of 275 drops (gobs of molten glass) per minute.

The concept to develop an intelligent decision support system arose after analysing the literature on the subject and numerous visits in the Glassworks. It was found that there was no algorithm to be applied in the case of discovering a defect of a product (here: glass packaging, e.g. bottles, jars) and for selection of an appropriate (the most beneficial) method of elimination of defects which occur during the production process.

The intelligent decision support system being designed should allow to classify product defects and to select an appropriate (the most beneficial) method for their elimination. The system being developed should support a line operator and a production line manager to a degree which is comparable with the support provided by a specialist (an expert) with high qualifications. Effects of the operation of the advisory system operation will allow operators and line managers to make appropriate decisions to eliminate production defects, and in consequence, to improve the technological process [Piróg-Mazur, 2010].

Before setting about working on the expert system being discussed initial assumptions and a method of the system development have been defined:

- the system should suggest solutions within the defined range – supporting a user in solving decision-making problems in the process of finished product quality control, i.e. classification of defects of products (bottles) as well as analysis and selection of an appropriate method of defect elimination, which will also allow to improve the technological process,
- the system should be user-friendly, a user is not required to be an expert in the field, the user interface will be based on questions and answers in the natural language,
- the system should provide texts, drawings and possibly simulations - databases in the form of text and graphic files, which contain additional or more complete information,
- the system can be developed in any programming language.

The technological process in the Glassworks precisely defines the process of converting a raw material (semi-finished product) into a finished product, which is compliant with requirements specified in a project. Development of technological processes is a very important phase of production preparation. However, its automation is very difficult due to large contribution of the experience of process engineers to the designing process. Traditional designing of technological processes is dominated by the activities which to a large degree are based on the experience, skills and intuition of a process engineer. Technological processes and their costs are dependent on a process engineer's experience.

When designing a technological process information from different sources is used. Technological processes are influenced by different kinds of information and limitations: information on a product, limitations related to technological capabilities of a manufacturing company and output, requirements concerning a product manufacturing, competences of a process engineer (professional experience, creativity), methods and resources used in technological planning and data gathered previously (technological databases and knowledge bases) [Rojek, 2010].

According to the definition, a technological process is a quantitatively and qualitatively structured set of actions that change physical properties (shape, size), the form or chemical properties of a specific substance (material). Technological process, together with support actions (transfer of material), constitutes a production process that results in a final product. The process of glass production comprises 9 main actions, connected with transforming raw materials and materials into ready products (for the purpose of an external recipient).

- Glass batch preparation – accurately weighed out and mixed raw materials constitute the so called batch. Glass cullet is a very important raw material. Even 80% of natural resources can be replaced with it.

- 
- 
- Melting – the batch is transported to the foundry furnace, namely the glass melting furnace and melts in the temperature of 1500°C. Such a high temperature is provided by gas fire burners, situated on both sides of the glass melting furnace. Molten glass is pushed by a new portion of batch.
  - Forming process – a stream of molten glass is cut into sections – drops of glass (known as gobs) of a weight that corresponds to the weight of the container being formed. Gobs are transferred to forming machines. The compressed air shapes, in the initial phase, a glass bubble that falls into the moulds where it takes on its final shape.
  - Hot refinement – bottles or jars are transferred to a tunnel-chamber where the compound of tin is sprayed. It penetrates the glass surface and results in higher mechanical resistance of wares and gives them shine.
  - Tempering – Wares are moved slowly on a conveyor belt inside a tunnel-lehr and solidify under control. It protects bottles against future cracking.
  - Cold refinement – cooled glass wares are subject to a process that makes them still more shining and flexible.
  - Quality check and sorting – it is automatically checked whether wares have flaws. If yes, the machine immediately eliminates defective wares.
  - Packaging and lamination – once the wares have been checked, they are conveyed to an automatic machine – palletizes, which arranges them in layers on pallets and protects them with a heat-shrinkable film. Packed wares are transferred to warehouses.
  - Storage and shipping – prepared to client's order wares wait in the warehouse to be shipped [Stowarzyszenie Opakowań Szklanych, 2010].

In the case of such an extended production process the probability of occurrence of defects is very high. Optimization of the technological process is essential for the company.

---

### Characteristics of database

---

A database development process comes down to defining objects in individual objects/tables and their attributes.

When designing the database the following questions have been asked:

- what data are we interested in?
- what format will they have?
- how are they related to each other?

Large production lines, which consist of several dozen machines linked to each other, have measurement points. Currently, in the glassworks data from measurement points are collected by PIC - Production Information Computer. Controlling the software (setting parameters which should be checked: glass wall thickness, profiled body, bottle neck, bottom, setting sensitivity – permissible norm and ideal norm) is indispensable for maintaining desired parameter values on a constant level. These parameters are adjusted every time when a product range (a product) is being changed. Control and measurement apparatus adjusts sensitivity. Sensitivity to critical defects is set to 100%. The higher sensitivity (expressed numbers) the larger number of rejects.

PIC software provides the following information:

- summary of losses on a specific production line,
- summary of losses in the whole glassworks,
- summary of rejects on a measurement point (cold end),
- losses on a selected production line,

- losses on a selected production line (detailed report),
- rejects per specific defects (percentage value),
- rejects per specific defects (number of items),
- machine downtime report,
- summary of results on all production lines,
- switching to another production line.

Dynamic data are related to the defects being monitored - a number of defects in time intervals: in 10 min., in an hour, in one shift and in 24 hours. Table 1 contains real data from one production line. There are 5 measurement points (FP1, ..., FP5) situated along the line, which record quality and quantity of defect occurrences. The table presents numbers of defects, their abbreviated names, percentage values in individual measurement points (FP) and their totals.

Table 1. Data extracted from measurement points one production line. Source:: System PIC

DETECTOR ID	FP1 %	FP2 %	FP3 %	FP4 %	FP5 %	TOTAL %
L1 SPEK.101	0.05	0.11	0.08	0.19	0.16	0.11
L2 SPEK.101	0.09	0.05	0.19	0.35	0.00	0.15
L3 SPEK.119	0.00	0.00	0.00	0.00	0.00	0.00
L4 PECH. W GL.	0.34	0.43	0.24	0.44	0.92	0.39
L5 SPEK.102	0.03	0.01	0.03	0.30	0.05	0.08
L6 SPEK.121	0.40	0.25	0.59	0.27	0.48	0.39
L7 NIEROWNLOGLY	0.83	0.56	0.38	0.34	0.98	0.56
L9 KRZYWY	0.26	0.20	0.44	0.37	0.59	0.33
L10 OOR-CMG	0.02	0.15	0.02	0.14	0.04	0.07
L11 SPEK.119	0.43	0.25	0.15	0.56	0.17	0.32
L12 SWA	0.57	0.48	0.43	0.40	0.58	0.48
L13 OOR-IPS	0.01	0.00	0.00	0.00	0.00	0.00
L14 SPEK.DNO	0.01	0.05	0.20	0.16	0.03	0.10
L16 FTA	1.81	1.98	1.72	1.94	1.71	1.85
L17 CID	0.45	0.52	0.36	0.36	0.67	0.44
L18 ROZDMUCHANA	0.00	0.00	0.01	0.00	0.00	0.00
L19 CIENKI GORA	0.09	0.14	0.08	0.03	0.27	0.10
L20 CIENKI DOL	1.26	0.89	0.95	0.57	0.82	0.93
L21 SSG1	0.00	0.00	0.00	0.00	0.00	0.00
L22 SSG2	0.00	0.00	0.00	0.00	0.00	0.00
L23 SSG3	0.00	0.00	0.00	0.00	0.00	0.00
L24 BHA	0.73	0.58	0.61	0.97	0.73	0.71
% REJECTED	6.22	5.58	5.42	5.86	6.84	5.82
INSPECTED	152638	144079	150891	115036	30026	592670

One of the first phases of the database designing process is development of a conceptual data model, which is of key importance for usefulness and quality of a database being designed. It is created independently of solutions characteristic for any logical models and database management systems. The conceptual model will allow to present the technological process described above in a formalized way. The main purpose of database conceptual modelling is to create a design which reflects the fragment of reality being analyzed, which is free of details that would locate it among models of a specific class (object, relational or others) and which is independent of a programming platform. The final effect of the conceptual designing process is a design containing three kinds of elements [Put, 2009]:

- facts, i.e. objects and events, which are to be stored in a database,
- attributes which describe individual facts,
- types of relationships between facts.

A design is typically presented graphically in a form of an entity-relationship diagram, which is supplemented with a detailed text description of information which it contains. In the diagram, facts are usually denoted with rectangles, attributes are denoted with ellipses and relationships between facts are denoted with lines linking rectangles and with symbols near lines which describe a type of relationship (Fig. 1).

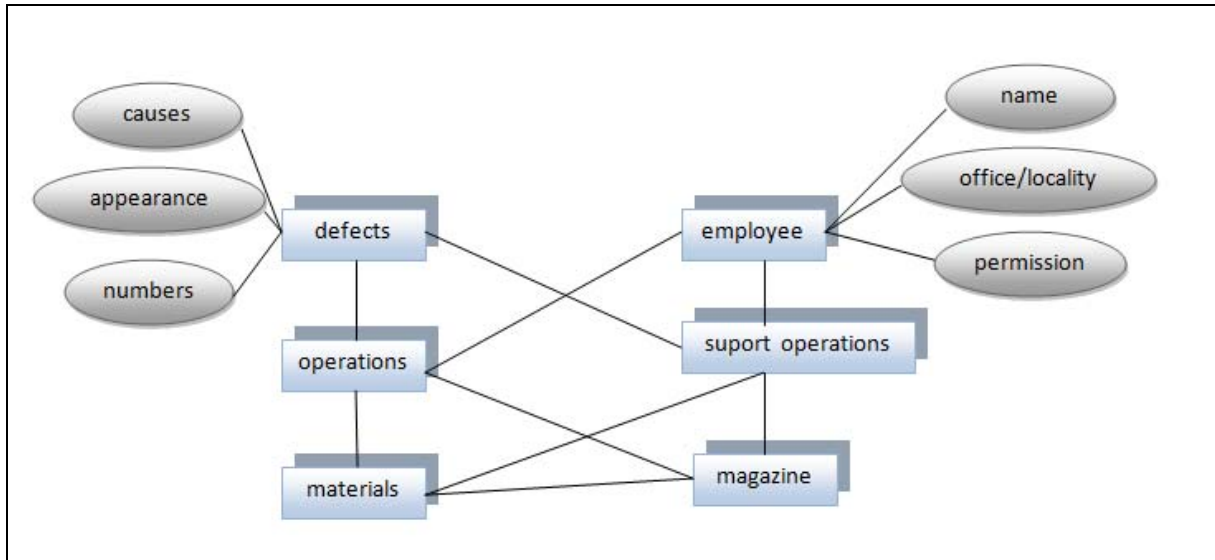


Fig. 1 Example of conceptual design - graphic form (fragment/episode). Source: own work

Selection of the relational model as a data storage method means that it is necessary to translate a universal conceptual design into a design in which data about facts is stored in tables, and attributes - according to the assumptions of the relational model – are atomic, which sometimes means the need to create additional tables and relationships between them. The logical design, presented in the form of a diagram, in the further phase of the process will be a basis for creation of a physical design and its implementation in a selected relational DBMS [Put, 2009].

Fig. 2 presents the relational database design developed on the basis of the conceptual design.

A database conceptual design, which is the final effect of the designing process participated by a future user, is the basis for creation of a logical design, which takes into account the specifics of a system in which the database will be implemented. The universal character of the conceptual model and its independence of the logical model allow to design databases not taking into account details of a particular model type.

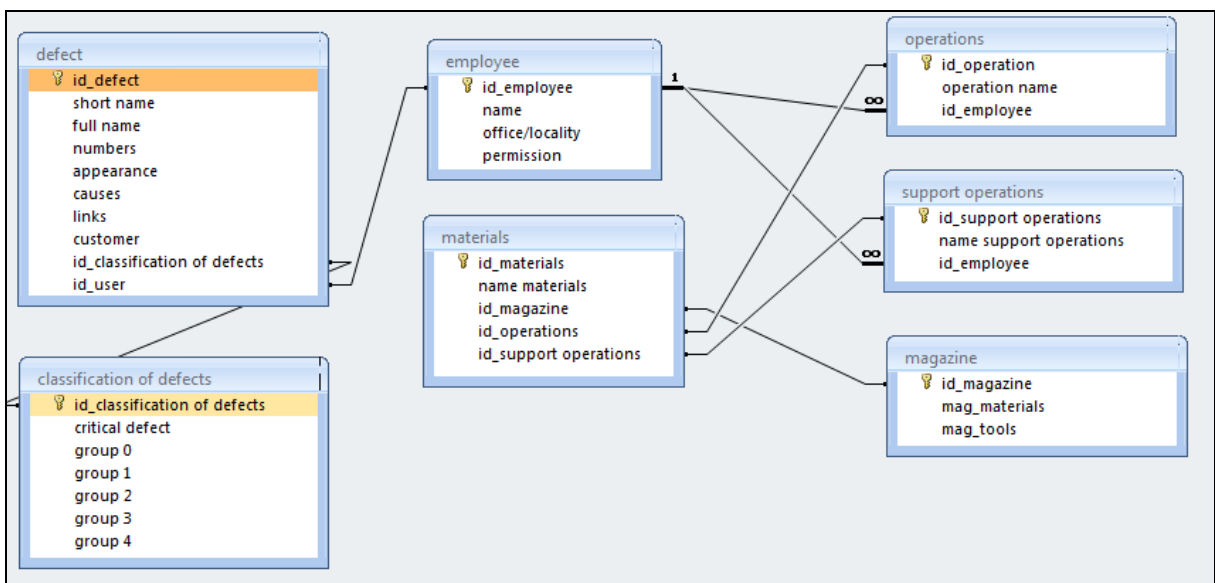


Fig. 2. Relationship database Source: own work

The following tables have been included in the relational database model:

- Products table - contains information on a product manufactured (product card, finished product specification),
- Materials table - contains information on materials (semi-finished products) used,
- Operations table - contains a list of operations carried out subsequently (technological operations),
- Auxiliary Operations table - contains a list of auxiliary operations to be carried out in the case of specific products (additional operations),
- Defects table - contains detailed information on individual defects.

The database contains vocabulary data and dynamic data and operates in real time. Vocabulary data concerns subject-matter objects and it changes much slower; however, it is also updated if objects are changed. As far as dynamic data is concerned, the database is archived periodically due to fast growth of data and a huge number of records. Dynamic data is related to gathering measurements from measurement points (FP), which read instantaneous measurements. This data will be used in the knowledge acquisition system as teaching files. The database structure is still being developed.

In the current phase, the largest collection of data is data containing classification of defects and complementary data containing a list of defects, defect descriptions, photos, reasons for occurrence of defects and methods of their elimination.

Measurement data indicates correlations between defects. A few correlations between defects are presented below:

- If the defect "melted bottom" occurs, the defect "deformed bottle lip" in a similar quantity can also be expected; in this case there is a lack of glass needed to form a bottle lip,
- uneven bottle – uneven bottle lip,
- thin bottom – thin glass walls – thin product.

Defects which are classified as critical are defects which may cause hazardous conditions for a product user (every defect which may result in glass inside a bottle). Products with such defects should not get into the annealing furnace. If they get into the annealing furnace, this fact should be reported to the cold end and these products should be rejected before they reach the end of the annealing furnace. It should be remembered that any of these defects may cause injury to a customer or a consumer.

---

### Technological knowledge base in a company

---

Technological knowledge plays a very important role in a manufacturing company. Systems supporting the design of technological processes, currently being developed, allow to make different methods of data presentation available, transform and exchange data.

Knowledge acquisition is a process of defining knowledge, on the basis of which an expert system will provide answers in the form of an expertise. Defining knowledge consists in acquiring knowledge from an expert in a form which allows formalizing it. An expert in a given field is responsible for the content of knowledge, whereas a knowledge engineer is responsible for its form.

Technological knowledge is a collection of information on a technological process realized in specific conditions of a given enterprise. The contribution of an expert and a knowledge engineer are described in the reference titles [Rojek, 2007]. Technological knowledge is a dynamic collection, i.e. it changes in time along with changes of a technological process. Additionally, it is assumed that technological knowledge may be processed in the way specific to the phases of an advisory system development. The following phases of the process are distinguished:

- acquisition of technological knowledge,
- development of technological knowledge representation models,
- recording knowledge in a system's technological knowledge base.

A knowledge engineer uses the following information for assessment of knowledge sources:

- information necessary to carry out work (materials from non-serial and serial publications),
- information concerning all processes realised in the production system (materials collected in the Glassworks, consultations with the plant manager of O-I Produkcja Polska, consultations with specialists having expertise knowledge on different phases of the technological process - an expert's knowledge),
- methods of finished product quality assessment (ideal norm and permissible norm),
- possible variants of modernisation (purchase of new machines, modernisation of existing machines, new technologies, new materials etc.),
- criteria for assessment of variants of the system development.

As a result of a dialogue, on the basis of data entered by a user and data from measurement points the expert system will perform a process consisting of:

- recognition of a defect of a product (here: bottles) and its classification into one of the groups (e.g. Group 0 - critical defect - leaky bottle lip, overblown bottle lip/collar, scratches in a bottle lip/collar),
- recognition of the cause of a defect (a mechanical defect, a defect of a form etc.),
- determining ways or methods of elimination of a defect,
- selection of an optimal solution out of previously determined methods.

On the basis of the process presented above the system suggests a method of elimination of product defects occurring on the production line.

Artificial intelligence package SPHINX by AITECH will be used for development of the intelligent decision support system. The following software tools will be used for the implementation of the system: PC-Shell – expert system shell – for development of basic modules of the system, CAKE – for presentation of knowledge elements and explanation how they are used and DeTreex – to acquire knowledge, decision-making rules from the database.

PC-Shell shell system is a hybrid system with the blackboard architecture, so it may use different sources of knowledge for solving problems. PC-Shell 4.5 supports the following sources of knowledge: expert knowledge bases, applications based on neural networks and databases with text explanations.

A knowledge base in the PC-Shell system is divided into five blocks: the block of knowledge sources description, blocks of facets, rules, facts description and the block of control. A knowledge base in PC-Shell may contain the following elements:

- descriptions, or in other words, facts, which are indicative sentences. A fact may be represented in the form of a relationship between certain objects and have different features (attributes),
- rules, which are indispensable for solving a problem in a given field,
- relationships,
- procedures.

The general format of description of rules in PC-Shell is presented below:

*[number\_of\_rule :] conclusion1 if  
condition\_1 & condition\_2 &...& condition\_n.*

Example:



*[Rule No.1:] <misadjustment of plunger and guide ring> if  
<uncentred plunger cylinder>&<plunger cylinder is not aligned with invert>  
& <too low plunger cylinder> & <glass reaches plunger>*

All rules are numbered and express logical associations between elements of knowledge in a given field or they contain a description of certain actions. Facts, expressed in the form of indicative sentences, represent elements of knowledge and they are treated as statements or conclusions. There are clear semantic associations between rules and facts [Piróg-Mazur, 2011, Buchalski, 2005].

---

## **Conclusion**

---

Manufacturing companies currently operating in the market collect more and more data on production processes, delivery processes, customers and their requirements, products' susceptibility to failure and control processes. The decision-making process is a process consisting in processing information. Classical methods of acquiring and analyzing information often fail, and additionally, they often refer to legacy data.

The paper presents the characteristics of the selected object of research, the conceptual data model, formalization of knowledge and the information collected during visits in the manufacturing company for which the intelligent decision support system is being developed.

The need for development of an intelligent decision support system arose from the practice and numerous meetings with a production line manager. It was realized that there was a lack of algorithms of action in the case of finding a defect of a product (here: glass packaging, e.g. bottles, jars) and for selection of an appropriate (the most beneficial) method of elimination of defects.

The system being designed, which is based on integration of selected tools of artificial intelligence and a knowledge base will allow to solve complex problems occurring in the production system faster and more effectively, using experience and intuition of a manager as an expert [Piróg-Mazur, 2011].

---

## **Bibliography**

---

- [Buchalski, 2006] Buchalski Z., „Zarządzanie wiedzą w podejmowaniu decyzji przy wykorzystaniu systemu ekspertowego”, Bazy danych: Struktury, algorytmy, metody, WKiŁ, Warszawa, 2006.
- [Buchalski, 2005] Buchalski Z., „Akwizycja wiedzy w systemie ekspertowym wspomagającym podejmowanie decyzji”, Bazy danych: Modele, Technologie, Narzędzia, Kozielski S., Małysiak B., Kasprowski P., Mrozek D. (red), WKiŁ, 2005
- [Piróg-Mazur, 2010] Piróg-Mazur, „Wykorzystanie systemu ekspertowego do wspomaganie podejmowania decyzji w przemyśle szklarskim”, w Bazy Danych: Aplikacje i Systemy, Studia Informatica Volume 31, Number 2B (90) 2010.
- [Piróg-Mazur, 2010] Piróg-Mazur, „Przykład postaci systemu doradczego do zastosowania w przemyśle szklarskim”, w: Technologie informatyczne i ich zastosowania. pod red. Aleksandra Jastriebowa, i inni, Politechnika Radomska, Radom, 2010
- [Piróg-Mazur, Setlak 2011] Piróg-Mazur M., Setlak G., „Budowa bazy danych oraz bazy wiedzy dla przedsiębiorstwa produkcyjnego w przemyśle szklarskim”, w Bazy Danych: Aplikacje i Systemy, Studia Informatica Volume 32, Number 2B (97) 2011.
- [Piróg-Mazur, Setlak 2010] Piróg-Mazur M., Setlak G., „Conceptual model of decision support system in a manufacturing enterprise”, 2nd International Conference on Intelligent Information and Engineering Systems, INFOS 2010, Rzeszów-Krynica, Poland
- [Pondel, 2003] Pondel M, „Wybrane narzędzia informatyczne pozyskiwania wiedzy i zarządzania wiedzą”, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 975. Pozyskiwanie wiedzy i zarządzanie wiedzą, Wrocław, 2003

- 
- [Put, 2009] Put D., „Model konceptualny jako wspólna platforma niejednorodnych modeli logicznych”, Akademia Ekonomiczna w Krakowie, 2009.
- [Rojek, 2009] Rojek I., „Wspomaganie procesów podejmowania decyzji i sterowania w systemach o różnej skali złożoności z udziałem metod sztucznej inteligencji”, Wydawnictwo Uniwersytetu Kazimierza Wielkiego, Bydgoszcz, 2010
- [Rojek, 2009] Rojek I., „Inteligentny system wspomagania decyzji dla sterowania siecią wodociagową”, projekt badawczy Nr R11 001 01.
- [Rojek, 2009] Rojek I., „Baz danych i baza wiedzy dla miejskiego systemu wodno – ściekowego”, w: Bazy danych: Nowe Technologie, Kozielski S., Małysiak B., Kasproski P., Mrozek D. (red), WKŁ, 2007
- [Weiss , 2002] Weiss Z., „Techniki komputerowe w przedsiębiorstwie”, Wydawnictwo Politechniki Poznańskiej, Poznań, 2002
- [Zieliński, 2000] Zieliński S., „Inteligentne systemy w zarządzaniu”, Teoria i praktyka, PWN, Warszawa, 2000.
- [Źródło] Stowarzyszenie forum opakowań szklanych, 2010

---

### Authors' Information

---



**Monika Piróg-Mazur** – Editor in chief, Institute of Technical Engineering, The Bronislaw Markiewicz State School of Technology and Economics in Jaroslaw, Czarniecki Street 16, 37-500 Jaroslaw, Poland; e-mail: [m\\_pirog@pwste.edu.pl](mailto:m_pirog@pwste.edu.pl)

Major Fields of Scientific Research: knowledge representation, hybrid intelligent system.



**Galina Setlak** – D.Sc, Ph.D., Associate Professor, Rzeszow University of Technology, Department of Computer Science, W. Pola 2 Rzeszow 35-959, Poland, and The State Professional High School, Czarnieckiego 16, Jaroslaw, Poland, e-mail: [gsetlak@prz.edu.pl](mailto:gsetlak@prz.edu.pl)

Major Fields of Scientific Research: decision-making in intelligent manufacturing systems, knowledge and process modeling, artificial Intelligence, neural networks, fuzzy logic, evolutionary computing, soft computing.

## STUDY OF INTEGRATION ALGORITHM AND TIME STEP ON MOLECULAR DYNAMIC SIMULATION

Janusz Bytnar, Anna Kucaba-Piętal

**Abstract:** A simulation is reliable when the simulation time is much longer than the relaxation time of the quantities in question. The aim of this work is to address the question when Molecular Dynamics (MD) simulation is reliable and how it depends on the integration algorithms and optimal time step. There were certain problems related to the choice of integration algorithms on Molecular Dynamics simulations. The effect of time step on convergence to equilibrium in Molecular Dynamics simulation has been studied.

**Keywords:** Molecular Dynamics, computer simulations, integration algorithms

**ACM Classification Keywords:** A.0 General Literature - Conference proceedings

---

### Introduction

---

After obtaining the results of a research, each scientist needs to consider verification and validation of those results. Computer simulation of molecular systems is playing an ever growing role in academic and industrial research. In areas ranging from materials science and chemistry to pharmacy and molecular biology, computer simulation is already a part of daily practice. The behavior of a variety of molecular systems can be studied by using the Molecular Dynamics (MD) simulation method. These include liquids, solutions, electrolytes, polymers such as proteins, DNA, and polysaccharides, as well as membranes, liquid crystals, crystals, and zeolites [Allen, 1987], [Bicout, 1996].

Computer simulation of molecular systems requires software to calculate the interatomic interactions and to integrate the equations of motion [Griebel, 2007].

Many models, for example in materials science or in astrophysics, contain large number of interacting bodies (called particles), as for example stars and galaxies or atoms and molecules. In many cases the number of particles can reach several millions or more. For instance, every cubic meter of gas under normal conditions (i.e., at temperature of 273.15 Kelvin and pressure of 101.325 kilopascal) contains  $2.68678 \times 10^{25}$  atoms (Loschmidt constant). 12 grams of the carbon isotope C12 contain  $6.02214 \times 10^{23}$  atoms (Avogadro constant).

These are some of the reasons why computer simulation has recently emerged as a third method in science besides experimental and theoretical approaches. Over the past years, computer simulation has become an indispensable tool for the investigation and prediction of physical and chemical processes. In this context, computer simulation means the mathematical prediction of technical or physical processes on modern computer systems [Griebel, 2007].

The deterministic method of Molecular Dynamics (MD) simulation, although theoretically valid for the whole range of densities, is employed mainly for liquids and solids [Allen, 1987]. The long flight paths between collisions of gas molecules make the method of Molecular Dynamics prohibitively expensive, while other methods, like e.g. Direct Monte-Carlo Simulation, can give satisfactory results at much lower computational cost. Molecules in liquids are densely packed and remain in constant contact with the neighbours. Under such conditions Molecular Dynamics seems to be the most accurate and, at the same time, the most efficient simulation method.

Molecular Dynamics requires the description of the molecules and the forces acting between them. Perhaps the most often, to describe the Van-der-Waals forces, the Lennard-Jones potential is used. It assumes that the molecules are spherically symmetric, repelling one another at close and attracting at far distances.

---

### Procedure of Molecular Dynamic

---

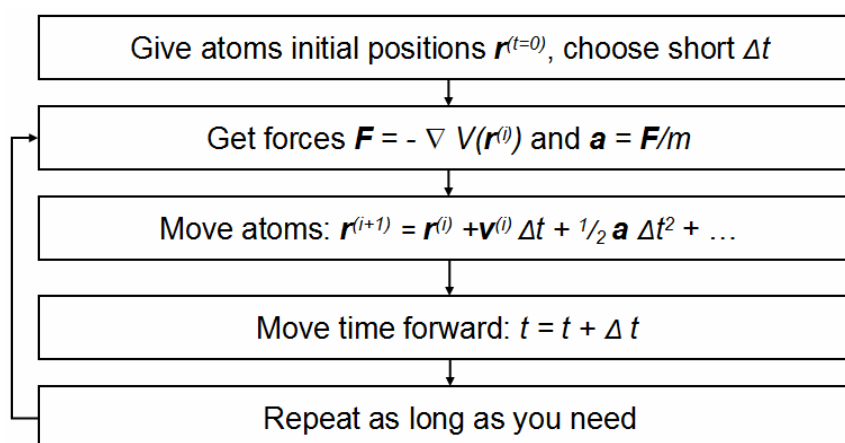
Molecular Dynamics (MD) is a computer simulation technique:

- the time evolution of interacting atoms is followed by integrating their equations of motion.
- the laws of classical mechanics are followed, and most notably Newton's law:

$$F_i = m_i a_i \quad (1)$$

$$a_i = d^2 r_i = dt^2 \quad (2)$$

- The Molecular Dynamics (MD) procedure can be written as follows:



- Given the initial set of coordinates and velocities, the subsequent time evolution is in principle completely determined.
- Atoms and molecules will 'move' in the computer, bumping into each other, vibrating about a mean position (if constrained), or wandering around (if the system is fluid), in a way similar to what real atoms and molecules do.
- The computer calculates a trajectory of the system
- 6N-dimensional phase space (3N positions and 3N moments).

The orientation of the molecules can be represented in several ways, however the use of quaternions [Refson, 2001] seems to be the most advisable. The most important advantage of quaternions is the fact, that they lead to equations of motion free of singularities (which is not the case for e.g. Euler angles). This, in turn, leads to good numerical stability of the simulation.

Integration algorithms used in Molecular Dynamics simulation are based on finite difference methods, with discretized time and the time step equal to  $\Delta t$ . Knowing the positions and some of their time derivatives at time  $t$  (the exact details depend on the type of algorithm), the integration scheme gives the same quantities at a later time  $(t + \Delta t)$ . With such procedure the evolution of the system can be followed for long times [Allen, 1987].

### Stages of simulation:

**Initiation:** placing the molecules of water and the copper atoms in the knots of crystalline mesh. After that the velocities of the molecules are initialized. Their values are sampled at random from the Maxwell – Boltzmann distribution for the assumed temperature.

**Balancing:** after initiation the positions of molecules are far from equilibrium. The whole ensemble is allowed to move freely for some time to attain equilibrium positions. This is always connected with decreasing the potential and increasing the kinetic energy of the molecules, i.e. increasing the temperature of the medium. This excess temperature must be removed by a suitable “thermostat”.

**Actual simulation:** after attaining equilibrium, the simulation starts. The required data (specified in advance) are accumulated in “dump-files” in preselected time intervals. Any property of interest.

---

### Integration Algorithms

---

The engine of a Molecular Dynamics program is the time integration algorithm, required to integrate the equation of motion of the interacting particles and follow their trajectory.

The integration scheme gives the possibility to find particle position at a later time  $t + \Delta t$ . By iterating the procedure, the time evolution of the system can be followed for long times.

Of course, these schemes are approximate and there are errors associated with them. In particular, one can distinguish between:

- Truncation errors, related to the accuracy of the finite difference method with respect to the true solution. Finite difference methods are usually based on a Taylor expansion truncated at some term, hence the name. These errors do not depend on the implementation: they are intrinsic to the algorithm.
- Round – off errors, related to errors associated to a particular implementation of the algorithm. For instance, to the finite number of digits used in computer arithmetic.

Both errors can be reduced by decreasing  $\Delta t$ . For large  $\Delta t$ , truncation errors dominate, but they decrease quickly as  $\Delta t$  is decreased. For instance, the Verlet algorithm has a truncation error proportional to  $\Delta t^4$  for each integration time step. Round – off errors decrease more slowly with decreasing  $\Delta t$ , and dominate in the small  $\Delta t$  limit [Ercolessi, 1997].

Using time integration techniques, it is possible to determine the velocity and position of a particle from its acceleration. There is a variety of different numerical methods available, however the nature of Molecular Dynamics simulations has narrowed down the field to a handful of methods. Methods which require more than one force calculation per time step are considered wasteful and can only be considered if the time step can be proportionally increased, while still maintaining the same accuracy.

Similarly, adaptive methods that change the time step dynamically are useless due to the rapidly changing neighbourhood of each atom. As a result, only two methods have become mainstream in Molecular Dynamics field, that is, the Verlet method and predictor-corrector method [Rapaport , 2004].

Both methods are based on finite difference techniques, derived from the Taylor expansion of the  $r(t)$ .

#### Basic Verlet Method

In Molecular Dynamics, time integration algorithm that is used very common is Verlet algorithm [Verlet, 1967]. The basic idea is to write two third-order Taylor expansions for the positions  $r(t)$ , one forward and one backward in time. The basic form of the Verlet method is defined by the equation:

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + (\Delta t)^2 a(t) + O(\Delta t^4) \quad (3)$$

where  $a(t)$  is the acceleration. Via the combination of the force calculation with Newton's second law of motion, the acceleration is defined as

$$a(t) = -(1/m)\nabla U(r(t)) \quad (4)$$

While not required for computation, the velocity variable can be found by using the equation

$$v(t) = \frac{r(t + \Delta t) - r(t - \Delta t)}{2\Delta t} \quad (5)$$

The Verlet algorithm uses positions and accelerations at time  $t$  and positions from time  $t-\Delta t$  to calculate new positions at time  $t+\Delta t$ . The Verlet algorithm uses no explicit velocities. The advantages of the Verlet algorithm are:

- It is straightforward
- The storage requirements are modest

The disadvantage is that the algorithm is of moderate precision.

A problem with this version of the Verlet algorithm is that velocities are not directly generated. While they are not needed for the time evolution, their knowledge is sometimes necessary. Moreover, they are required to compute the kinetic energy  $E_K$ , whose evaluation is necessary to test conservation of the total energy  $E = E_K + E_P$ . We can also calculate temperature of the simulated molecular system from kinetic energy.

$$E_K = \frac{1}{2}k_B T \quad (6)$$

Where  $k_B$  is the Boltzman constant,  $T$  is the temperature.

This is one of the most important tests to verify that a Molecular Dynamics simulations of real processes is proceeding correctly.

However, the error associated to this expression is of order  $\Delta t^2$  rather than  $\Delta t^4$ . To overcome this difficulty, some variants of the Verlet algorithm have been development.

### Velocity Verlet Method

However more common algorithm is a related one, Velocity Verlet algorithm. Here the velocity, position and accelerations at time  $t+\Delta t$  are obtained from the same quantities at time  $t$  [Verlet, 1967]. This uses a similar approach but explicitly incorporates velocity, solving the first-time step problem in the Basic Verlet algorithm:

$$r(t + \Delta t) = r(t) + v(t)\Delta t + 1/2 a(t)\Delta t^2 \quad (7)$$

$$v(t + \Delta t) = v(t) + 1/2[a(t) + a(t + \Delta t)]\Delta t^2 \quad (8)$$

### Beeman Method

This algorithm is also closely related to the Verlet algorithm

$$r(t + \Delta t) = r(t) + v(t)\Delta t + 2/3 a(t)\Delta t^2 - 1/6 a(t - \Delta t)\Delta t^2 \quad (9)$$

$$v(t + \Delta t) = v(t) + v(t)\Delta t + 1/3 a(t + \Delta t) + 5/6 a(t)\Delta t - 1/6 a(t - \Delta t)\Delta t \quad (10)$$

The advantage of this algorithm is that it provides a more accurate expression for the velocities and better energy conservation. The disadvantage is that the more complex expressions make the calculation more expensive [Beeman, 1976].

A variant of the Verlet method, called the velocity-Verlet method, addresses this problem by directly including the velocity in computation. As a result, particle velocities are known at the same time step as coordinates, and the high-order accuracy of the method is maintained. Additionally, particle velocities are necessary for kinetic energy calculations, which play a critical role in most Molecular Dynamics simulations [Haile, 1997].

### Time Step

Lennard – Jones potential is the most popular interaction potential used in Molecular Dynamics (MD) simulations to describe Van-der-Waal forces [Karniadakis, 2005]. The form of the Lennard – Jones potential is as follows:

$$V(r) = 4\varepsilon \left[ \left( \frac{\delta}{r} \right)^{12} - \left( \frac{\delta}{r} \right)^6 \right] \tag{11}$$

where  $\varepsilon$  and  $\delta$  are the Lennard – Jones parameters that depend on the atoms involved in the interaction. Note that:

- $\varepsilon$  is related to the interaction strength, and a higher  $\varepsilon$  corresponds to a higher interaction energy between the atoms
- $\delta$  corresponds to the distance at which the potential between the two atoms goes to zero, which can be approximately taken as the diameter of a fluid atom.

The term  $\sim 1/r^{12}$ , dominating at shorter distance, models the repulsion between atoms when they are brought very close to each other.

The term  $\sim 1/r^6$ , dominating at large distance, constitute the attractive part. This is the term which gives cohesion to the system. A  $1/r^6$  attraction is originated by van der Waals dispersion forces, originated by dipole – dipole interactions in turn due to fluctuating dipoles. These are rather weak interactions, which however dominate the bonding character of closed – shell systems, that is, rare gases such as Argon. Therefore, these are the materials that Lennard – Jones potential could mimic fairly well [Ercolessi, 1997]. The parameters  $\varepsilon$  and  $\delta$  are chosen to fit the physical properties of the material.

In the Molecular Dynamics (MD) simulation with Lennard – Jones interaction potentials, the time and the other physical quantities are represented and typically computed using reduced units. Table 1 summarizes the units for various quantities used in simulations for instance, length, temperature, and density. In the Table 1, symbols  $\varepsilon$  and  $\delta$  denote constants as defined in equation (10),  $k_B$  is the Boltzman constant, and  $m$  is the mass of a atom.

Table 1. Units for various quantities in Lennard – Jones fluids [Griebel, 2007], [Karniadakis, 2005]

Length	$\delta$	Velocity	$(\varepsilon / m)^{1/2}$
Mass	$m$	Shear rate	$(\varepsilon / m \delta^2)^{1/2}$
Energy	$\varepsilon$	Stress	$\varepsilon / \delta^3$
Time	$(m \delta^2 / \varepsilon)^{1/2}$	Viscosity	$(m \varepsilon)^{1/2} \delta^2$
Number density	$\delta^{-3}$	Diffusivity	$\delta (\varepsilon / m)^{1/2}$
Temperature	$\varepsilon / k_B$		

In many publications [Griebel, 2007], [Karniadakis, 2005] authors calculate time step from the formula:

$$\Delta t = 0,001 * \sqrt{\frac{m\delta^2}{\varepsilon}} \quad (12)$$

We intended to use as large a time step as possible so that we can explore more of the phase space of the system. However, since we truncate the Taylor's series expansions, the time step needs to be small enough so that the expansion can provide a reliable estimate of the atomic positions and velocities at the end of the time step (see Fig. 1). For typical algorithms with a time accuracy of order three, one uses a time step that is a fraction of the period of the highest-frequency motion in the system [Karniadakis, 2005]. For a typical simulation of water transport, where the O – H bond length is fixed, a time step size of 1.0 to 2.0 fs is commonly used.

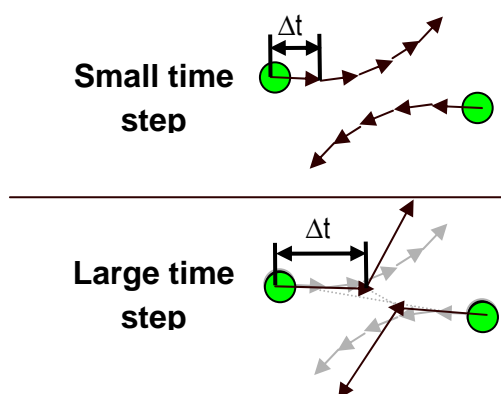


Fig. 1. The effect of the small and large temporary step

## Results

The aim of the research was to analyze the influence of integration algorithm and time step on the computational time and memory complexity. Also influence of the size of molecular systems on efficiency integration algorithms was studied.

The simulations were carried out for one molecular model of water TIP4P and three integration algorithms were applied: Velocity Verlet, Beeman and Beeman algorithm with Predictor – Corrector modifications.

We use for all algorithms different time steps  $\Delta t = 0,00001$ ,  $\Delta t = 0,00002$ ,  $\Delta t = 0,00005$ ,  $\Delta t = 0,0001$  and  $\Delta t = 0,0005$  picosecond long. The calculations were carried out over 100 000 time steps.

The program MOLDY [Refson, 2001], suitably modified, was used for this purpose. Moldy is free software; which may redistribute it and/or modify it under the terms of the GNU.

The physical properties of materials and their electrostatic interactions were taken into account. The number of water molecules was equal to 500 and 20000. The periodic boundary conditions were applied. The Lennard-Jones potential was assumed for interactions between water molecules [Kucaba – Pietal, 2004], [Bytnar, 2008].

Molecular Dynamics is always an approximate science approach, the longer the time step, the less accurate the results. In the worst case scenario, the time step will allow atoms to move too far between single iterations, allowing atoms to get closer together than they ever could in a real liquid. This usually causes an incorrect “chain reaction”, whereby two close particles repel at a much faster speed than normal causing them to bump even closer other atoms, which are repelled at an even greater velocity. This effect compounds until all atoms are moving at unrealistic speeds and eventually arithmetic overflows will occur. When we attempted to use larger time step (1 fs) the program crashed.



If we only consider figures 2-5 then we can deduct that ideally time step should be as small as possible.

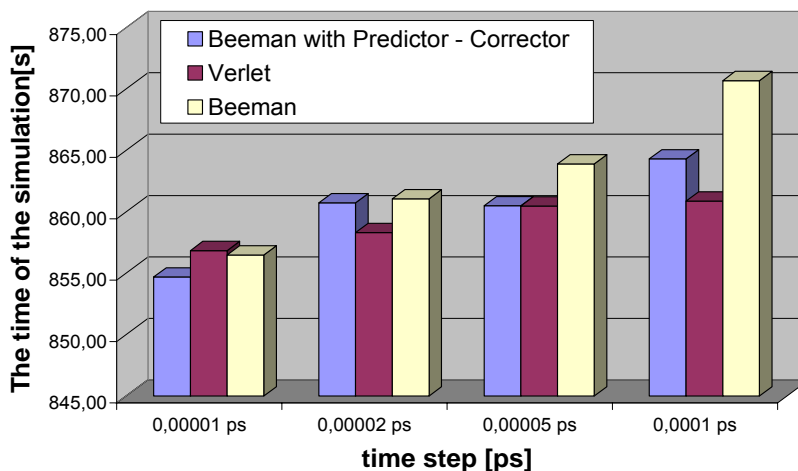


Fig. 2. The time of the whole simulation (500 molecules of water) – Integration algorithms with various time steps

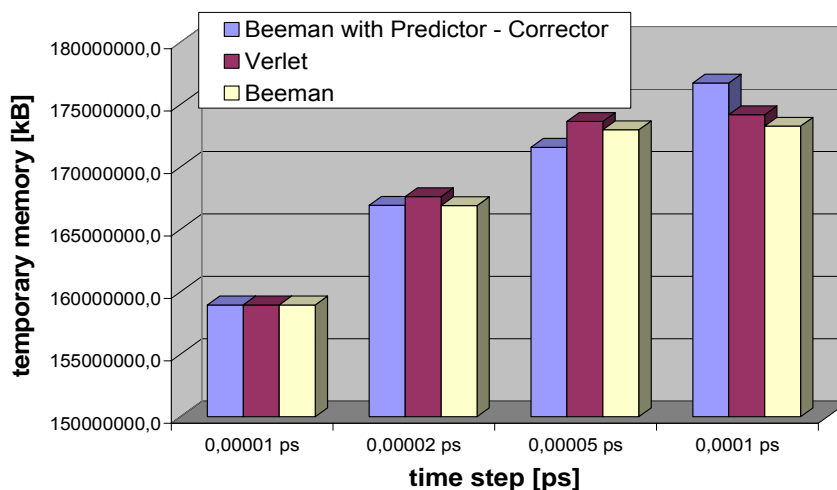


Fig. 3. The reservation of the temporary memory (500 molecules of water) – Integration algorithms with various time steps

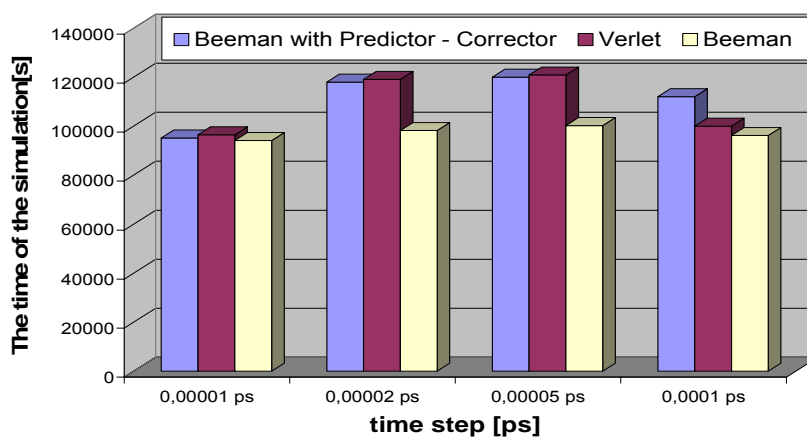


Fig. 4. The time of the whole simulation (20000 molecules of water) – Integration algorithms with various time steps

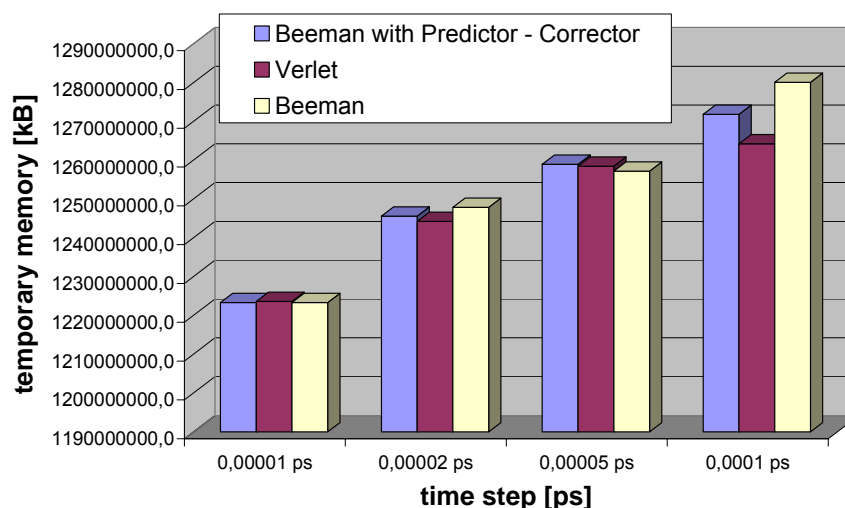


Fig. 5. The reservation of the temporal memory (20000 molecules of water) – Integration algorithms with various time steps

In figures 2-5, several different time step simulations were run in different integration algorithm. Using a small time step of 0,00001 ps gives good results connected with the reservation of the memory and the time of the whole simulation. The increment of the time step forces the use of the greater spaces of the memory and time of the whole simulations.

## Conclusions

In Molecular Dynamics simulation, there exist several algorithms to realize equations of motion of molecular systems.

From the presented diagrams we can see that the choice of these algorithms with correct time steps decides about the time of the whole simulations and also about the necessary reservation of the memory.

The choice of the time step in integration algorithm from equation (11) can be good for abstract molecular models. However, this type of simulations does not take into account many factors, which are very important for simulations of real materials. In Molecular Dynamics simulations of real processes not only speed of calculations and reservation of the memory are important. In this case what is very important is the thermodynamics of the simulated molecular configuration during the whole simulation as also the obtained results of physical properties simulated materials which in confrontation with the results of these materials in the larger scale will be satisfactory.

The influence of the correct choice of the time step was also present in paper [Bytnar, 2010]. From the presented diagrams (considered in paper [Bytnar, 2010]) it is clear, that for the problems considered, i.e. real flows in nanochannels, the Gaussian thermostat is much more efficient if the time step  $\Delta t$  is well chosen.

If we consider only time of the whole simulations and required memory (see Figures 2-5) we can say that the best performance and accuracy is if the time step is as small as possible also for small and large molecular systems.

Additionally we can see that the performance of all integration algorithms (Velocity Verlet, Beeman and Beeman with Predictor – Corrector modifications) is very similar also for small (500 molecules of water) and large (20000 molecules of water) molecular systems. This means that the choice of the integration algorithm does not have large signification for Molecular Dynamics simulations.

The smaller time step effects enlargement of the time of the whole simulation and the required memory, however the larger temporary step can effect generating the inaccurate trajectory of motion. Clearly, the larger time step, the less accurately our solution will follow the correct classical (see Fig 1).

From the presented figures in this paper it can be deduced that a good way of checking whether the time step is satisfactory is to run an equilibrium simulation of small molecular system because results for small (500 molecules of water) and large (20000 molecules of water) molecular system are wery similar.

In summary, the choice of time step has a big impact on accuracy of simulations. It is recommended that, to avoid the incorrect "chain reaction" phenomena, if two atoms get unrealistically close, the user should be warned that the time step should be decreased and be given the option to terminate the program since the results are already effectively useless. A more advanced program might provide warnings if the user enters an unrealistically large time step before the simulation is allowed to start.

The physical explanation of other peculiarities of the presented diagrams, particularly the thermodynamics properties molecular systems, requires further investigation.

The study of Verification and Validation methods gives the possibility of the construction of the correct computer model to the description of the studied phenomenon, how also receipt of the exact and authentic results of computer simulations.

Influence on the results of the simulation in the Molecular Dynamics method has e.g. Integration algorithms and time step, molecular model of water or another material, mechanisms to control the temperature of the system (thermostats). Therefore, the study of the methods of Verification and Validation will be closely connected with the above mentioned factors.

---

## **Acknowledgment**

---

Calculations were made in the Interdisciplinary Centre for Mathematical and Computational Modeling (ICM) University of Warsaw (grant no. G44-9):

<https://granty.icm.edu.pl/lcmGrants/displayGrant/showGrants.jsp>

We also want to thank for the possibility of the realization of calculations in the Institute of Fundamental Technological Research PAN, Department of Mechanics and Physics of Fluids in Warsaw.

---

## **Bibliography**

---

- [Allen, 1987] M. P. Allen and D. J. Tildesley: Computer Simulation of Liquids, Clarendon Press, Oxford, 1987,
- [Bicout , 1996] D. Bicout and M. Field: Quantum Mechanical Simulation Methods for Studying Biological Systems, Springer, Berlin, 1996,
- [Griebel, 2007] M. Griebel S. Knapek, G. Zumbusch: Numerical Simulation in Molecular Dynamics, Numerics, Algorithms, Parallelization, Applications, Springer-Verlag Berlin Heidelberg, 2007,
- [Ercolessi, 1997] F. Ercolessi: A molecular dynamics primer, Spring College in Computational Physics, ICTP, Trieste, Italy, 1997,
- [Rapaport , 2004] D. C. Rapaport: The Art of Molecular Dynamics Simulation. Cambridge University Press, 2004,
- [Verlet, 1967] L. Verlet: Computer `experiments' on classical fluids. I. thermodynamical properties of Lennard-Jones molecules, Phys. Rev. 165, 201-214, 1967,
- [Beeman, 1976] D. Beeman: Some multistep methods for use in molecular dynamics calculations, J. Comp. Phys. 20 130-139, 1976,
- [Haile, 1997] J. Haile: Molecular Dynamics Simulation: Elementary Methods. New York: John Wiley and Sons Inc., 1997,

- 
- 
- [Karniadakis, 2005] G. Karniadakis, A. Beskok, N. Aluru: Microflows and Nanoflows – Fundamentals and Simulation, Interdisciplinary Applied Mathematics, Springer, 2005,
- [Refson, 2001] K. Refson: Moldy User's Manual. Chapter II, <ftp://ftp.earth.ox.ac.uk/pub>,
- [Bytnar, 2010] J. Bytnar, A. Kucaba-Piętal, Z. Walenta: Verification and Validation of Molecular Dynamics Simulation, Publications S.H.F. Second European Conference in Microfluidics, ISBN 978-2-906831-85-8, Session P4-3-Liquid Microflows, Toulouse, France, 2010,
- [Bytnar, 2008] J. Bytnar, A. Kucaba-Piętal, Z. Walenta: Influence of molecular models of water on computer simulations of water nanoflows - Proceedings of International Conference on Computer Science and Information Technology, Volume 3, ISSN 1896-7094 pages 269 – 275, IEEE CS Press, Los Alamitos, CA, 2008
- [Kucaba – Piętal, 2004] A. Kucaba-Piętal: Microflows modelling by use micropolar fluid model, OW RUT, Rzeszow, 2004;

---

### Authors' Information

---

**Janusz Bytnar** –*Technical and Economical State School of Higher Education in Jaroslaw, Institute of Technical Engineering, ul. Czarnieckiego 16, 37-500 Jaroslaw, Poland; e-mail: [janusz.bytnar@pwste.edu.pl](mailto:janusz.bytnar@pwste.edu.pl)*

*Major Fields of Scientific Research: Software technologies, Molecular Dynamics, Computer simulations, Integration algorithms research*

**Anna Kucaba - Piętal** – *Technical and Economical State School of Higher Education in Jaroslaw, Institute of Technical Engineering, ul. Czarnieckiego 16, 37-500 Jaroslaw, Poland;*

*Rzeszow University of Technology, The Faculty of Mechanical Engineering and Aeronautics Powstancow Warszawy 8, 35-959 Rzeszow, Poland; e-mail: [anpietal@prz.edu.pl](mailto:anpietal@prz.edu.pl)*

*Major Fields of Scientific Research: Scientific Calculations, Computational Mechanics,, Molecular Dynamics, Computer simulations, Nano and Micromechanics*

## INFORMATION SYSTEMS FOR METROLOGY

Roman A. Tabisz, Łukasz Walus

**Abstract:** An important application of the information technology (IT), which is the creation and improvement of Information Systems for Metrology (ISM) is discussed. These systems initially had a form of Metrological Database (MD) and Metrological Knowledge Bases (MKB). At present, the advanced versions of these systems have a form of the Metrological Expert Systems (MES) using artificial intelligence methods. Two selected examples of Information Systems for Metrology and their practical use are described. An actual state of the ISM ("CAMPV System") designed and developed in the Department of Metrology and Diagnostic Systems, Faculty of Electrical and Computer Engineering at Rzeszow University of Technology, is presented. The system consist of the portal for communication with the system via the Internet and the Metrological Database (MD) collecting the data on the specification of measurement equipment as well as the data obtained in the process of equipment calibration. It has been planned that the system will be expanded by adding the Analytical Metrological Database (AMD) and Metrological Knowledge Base (MKB). A full version of "CAMPV-Expert System" will be dedicated to the computer-aided design of electrical and electronic measurement channels as well as the computer-aided validation of measurement processes, which include such channels.

**Keywords:** information systems for metrology, metrological databases, metrological knowledge bases, metrological expert systems.

**ACM Classification Keywords:** A.0 General Literature - Conference proceedings

**Conference topic:** Industrial Control and Monitoring

---

### Introduction

---

The development of microelectronics and related development of information technology (IT) has enabled the creation of measurement information systems (MIS), which allow for measuring and collecting large amounts of measurement data. A vital part of the measurement equipment is the software, which carries out important functions related to the performance of measurement activities and therefore determines the accuracy of final results obtained in the measurement process. Such kind of software is called measurement systems software (MSS). The measurement results are now the primary source of our knowledge about the objects and the phenomena of the real world. The metrology therefore has become the basis for the development of many different fields of science, industry and trade. Measurement results are also the foundation on which the scientific claims are created. In the industrial applications, the measurement results are used for the assessment of products' conformity with their technical specification, the evaluation of the quality of manufacturing processes as well as the control and monitoring of their condition. In the trade, the mutual financial settlements are often based on the measurement results. In each of these areas the highest possible quality of measurement results is desired. The basic condition of reaching this goal is the proper design and appropriate execution of the measurement process. Moreover, the quality of measurement results largely depends on the quality of the hardware and software as well as the level of competence of the operators supervising the measurements processes. The management of the measurement equipment and the effective control of the measurement processes require some additional metrological actions such as calibration, repeatability and reproducibility analysis (R&R) as well as inter-laboratory comparison (IC). This increases the need to collect the data concerning the properties of the measurement equipment and the data obtained in the calibration processes. For this reason, the Metrological Databases (MDB) are created with the use of appropriate IT.

The general model of the measurement process, which incorporates the main factors determining the numerical values of the final measurement results is presented on the Fig.1., on which the measuring equipment is specified by a rectangle. The measurement equipment includes both hardware and software. The circle inscribed in the rectangle marks software as an integral part of the measurement equipment. The software is now incorporated in almost all kinds of measuring equipment as it outscored other technologies when it comes to the implementation of important measurement functions such as: measurements process control, correction of the numerical values of the measured quantity, saving the measurement results as well as reading, decoding and displaying them.

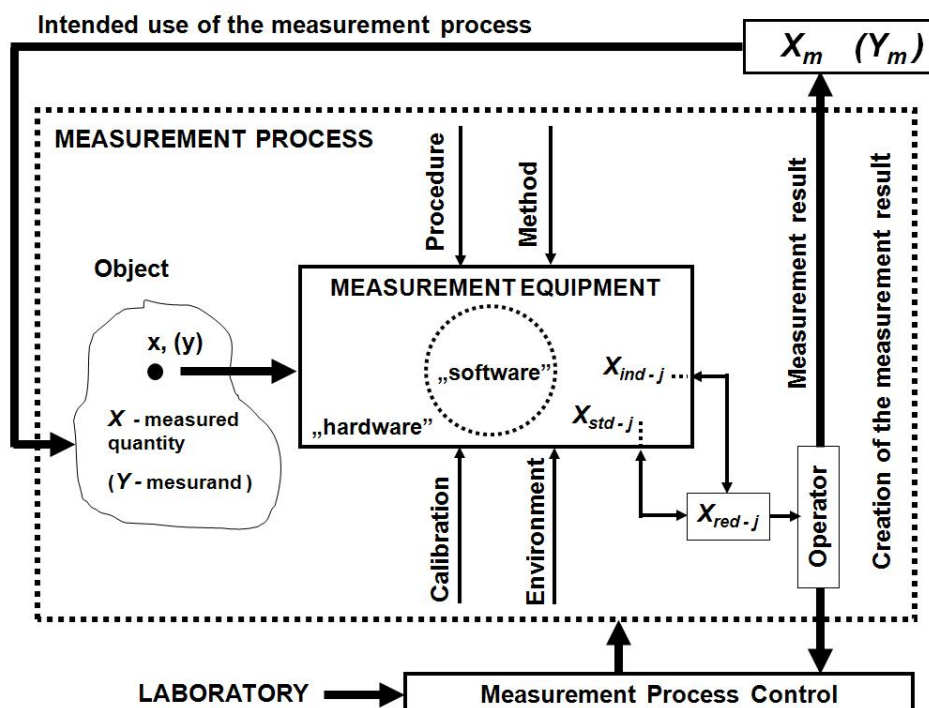


Fig.1. The general model of the measurement process.

$x$  - numerical value of the measured quantity

$y$  - numerical value of the measurand

$X_{ind-j}$  - numerical value indicated on the display of measurement equipment

$X_{std-j}$  - numerical value stored in the memory of the measurement equipment

$X_{red-j}$  - read the value, on which measurement result ( $X_m$ ), is created

$Y_m$  - measurand, calculated based on several measured quantities ( $X_i$ )

The general model of the measurement process, as shown in Fig.1., has been developed taking into account the original definition of the measure proposed in [1], and models of Information Measurement Systems (IMS) proposed in [2]. These models take into account the types of measurement scales. With such approach, this model is universal and can be applied to all situations in which measurements are carried out. It is independent from the measurement scale adopted to implement the measurement procedure as well as from the composition of the hardware and the software.

The types of the measurement scales have been defined in [3]. In every area of possible application - scientific research, industry or trade - the measurement processes are performed in the specific order. The ultimate goal is always to describe the properties of the real object according to its true state. It allows for making right decisions

when it comes to the objects' management and helps the scientists in formulating the accurate and consistent descriptions of the reality of the phenomena. Relevance assessment of the actual state of measured properties of an object, event or phenomenon, depends largely on the accuracy and proper interpretation of obtained results, which, in turn, is strictly associated with the type of the measurement scale. If above is not taken into account, not only the measurements become ruseless but their false results may lead to wrong decisions.

A definition of the measurement proposed in [1] is as follows:

*„The measurement is a process of empirical, objective assignment of symbols to attributes of objects and events in the real world, in order to describe them”. [1]*

However, the proposal formulated in [2] differentiates the MIS due to the measurement scales used and has been developed in the following order:

*“The models proposed in this paper can be used for the development of concepts, principles, and guidelines, which can support decision making in the arrangement of design and application of the Measurement Information Systems (MIS)”. [2]*

*„In order to apply these scientific and technical achievements, the designer or the user of a MIS has a possibility to choose between different measurement procedures and also between different constructive modules of the MIS”. [2]*

Such approach - taking into account the definition proposed in [1] and the proposal of different types of MISs presented in [2] – allows for the correct interpretation of the measurement results and collected information about the measured object, event or phenomenon. Consequently, it makes possible to gain credible knowledge about the real world, manufactured products and also about the measurement science. In order to ensure the desired quality of the measurement results, one should design, manufacture and exploit the MIS using the appropriate hardware and software composition as well as taking into account the type of measurement scale applied in the measurement process. Only properly designed, implemented and applied MIS can be used to effectively collect large amounts of credible data. Credible measurement data can and should be collected in the Operational Metrological Databases (OMD). These databases currently represent the most basic form of Information Systems for Metrology (ISM). The more complex form of the ISM are systems in which the Analytical Metrological Database (AMD), also called data warehouse, is added to the OMD. The extraction of data from OMD to the AMD made through suitably designed ETL operation (Extraction Transaction Loading) provides a grouping of the data according to a well-thought-out strategies and allows for their storage in a useful and inviolable form. This is necessary to carry out various kinds of analysis, which require extracting desired information and knowledge. The concept of extracting information from the measurements and acquiring knowledge from obtained information is described in [4]. It justifies the usefulness of the definition of measurement proposed in [1].

The most complex form of the ISM are Metrological Expert Systems (MES), in which the properly prepared Metrological Knowledge Bases (MKB) are supplemented by automatic inference systems using artificial intelligence methods. Information Systems for Metrology (ISM), including OMD, AMD, MKB and Inference Systems (IS) can be used in metrology for the computer-aided design of measurement processes and measurement equipment (hardware and software). They can also be used to improve the competence of the operators, who supervise the measurement processes and improve the metrological characteristics of MISs, by adjusting the measured numerical values in order to ensure the accuracy of the final results.

In order to present the specific applications of ISM, two selected examples will be described as well as the current state of the third one, which currently has been in the process of design, implementation and development. The first example is the Metrological Knowledge Base - one of major components of the education system operators of Coordinate Measurement Machines (CMM) [5]. The second example is the metrological expert system using

---

---

neural networks to predict corrections, calculate the Polish Universal Time Coordinate - UTC (PL) and propagate the Polish Official Time - OT (PL) [6]. The third example is a computer system supporting the validation of the measurement process, designed, implemented and developed in the Department of Metrology and Diagnostic Systems, Faculty of Electrical Engineering and Computer Science, Rzeszow University of Technology. Each of described examples implements the concept of knowledge acquisition from information extracted from the measurements data collected in the appropriate metrological databases and then stored in the metrological knowledge bases.

The process of extracting knowledge from data described in [4] goes in the order, which can be simply described as: "*measurement-information-knowledge*". This order has been used in all described examples, but in each one for different reason. In the first example, the objective is to improve the competence of the operators of CMM. The operator is the key person responsible for the creation of the final result in the model of measurement process shown in Fig.1. Therefore, the more complex is the measurement equipment, the more qualified the operator should be. In the second example, the metrological expert system analyzes the data published by the Time Laboratory of the International Bureau of Weights and Measures in Paris (BIMP) [7] and predicts the corrections, which allow for calculation of the exact value of the Polish Universal Time Coordinated - UTC (PL). The difference between these corrections and the corrections published by the BIMP should not exceed 10 ns. In the third example, the main goal is to use the "*CAMPV System*" (*Computer Aided Measurement Process Validation System*) for validation of the measurement process, conducted on the basis of historical calibration data of measurement equipment, stored in the metrological database. This system also allows testing and calibration laboratories for cooperating on-line.

---

### **The IT System for Metrology used in the European Education and Training Programme for operators of CMMs**

---

The European education and training programme for operators of CMMs was developed in 2001-2005 within the European Research Project EUKOM [5]. This project was funded by "LEONARDO DA VINCI" programme of European Commission, DG Education and Training. The main objective of the project "European Training for Coordinate Metrology" was to create a common, Europe-wide approach to the training in the field of coordinate measuring technology, which would meet the requirements of continuing education. The results of the project were implemented by the association CMTrain e.V. [8], independent of CMM producers, which currently is being developed and tested of training by a mixed method. This method combines the capabilities of distance learning (e-learning) with verification of acquired knowledge during the practical handling of CMMs. In the original version the distance learning module was created with the ILIAS system [9], which is the open source software. The effectiveness of learning was achieved through a carefully prepared Metrology Knowledge Base developed by experts from the six centers from different countries. The project was coordinated by the Chair of Quality Management and Manufacturing Metrology at University of Erlangen-Nuremberg. [10]. The structure of Metrological Knowledge Base EUKOM system has been adapted into the learning material coherent and common for the whole of Europe.

The starting point for the study was to establish three levels of competence of the operators: level 1 - "CMM-User", level 2 - "CMM-Operator" and level 3 - "CMM-Expert." For each level of competence approx. 15 training modules was developed and made available online. Respective modules include the learning content appropriate for a given level of competence and concerning areas where knowledge is required and needed for the CMMs' operators. The areas include: metrology, geometry, statistics, computer science, quality management, standardization, measurement equipment, technology of production and technical drawing systems used in computer-aided design systems (CADs). Thanks to the cooperation of specialists in the field of construction and



exploitation of Coordinate Measuring Machines, a three-levels metrological knowledge base structure was developed, which is shown in Table.1.

*Table 1. Three-levels structure of the Metrological Knowledge Base of the EUKOM system for education of CMMs' operators*

Types of qualification	Levels	Modules of the Metrological Knowledge Base of the EUKOM System.
CMM- User	I	<div style="display: flex; justify-content: space-around; border: 1px solid black; padding: 2px;"> <span>1</span><span>2</span><span>3</span><span>4</span><span>5</span><span>6</span><span>7</span><span>8</span><span>9</span><span>10</span><span>11</span><span>12</span><span>13</span><span>14</span><span>15</span> </div>
CMM- Operator	II	<div style="display: flex; justify-content: space-around; border: 1px solid black; padding: 2px;"> <span>1</span><span>2</span><span>3</span><span>4</span><span>5</span><span>6</span><span>7</span><span>8</span><span>9</span><span>10</span><span>11</span><span>12</span><span>13</span><span>14</span><span>15</span> </div>
CMM- Expert	III	<div style="display: flex; justify-content: space-around; border: 1px solid black; padding: 2px;"> <span>1</span><span>2</span><span>3</span><span>4</span><span>5</span><span>6</span><span>7</span><span>8</span><span>9</span><span>10</span><span>11</span><span>12</span><span>13</span><span>14</span><span>15</span> </div>

The structure of the Metrological Knowledge Base of the EUKOM system, shown in Table.1., was developed by the eminent experts from seven European universities [10] engaged in the development of CMM metrology and its applications in the industry. The 3-levels structure of education, including about 15 basic training modules at every level of competence, was established after the adoption of the basic assumption that the industries using the CMMs need three categories of operators. Three categories of operators are needed so that the very expensive and complex measurement equipment, such as CMM, could be properly used and the desired accuracy of measurement results was ensured. It was found that the lowest level of competence (1-level) requires an employee (called "CMM-User") able to perform actions such as a simple measurement tasks involving: mounting the measured parts on a test bench and preparing the measurement procedure to start as well as operating the machine-controlled software. An employee called "CMM-Operator", whose qualifications were classified as "Level II" competence, should be able to define the measuring task on the basis of technical drawings, create software that controls the measurements, assess the accuracy of obtained measurements results and perform the correction of measurement results, taking into account deviations caused by various factors. The highest, "Level-3" of qualifications is intended for an employee who has been called "CMM-Expert." He should be able to plan, program and optimize the measurement for any established measurement tasks as well as estimate the uncertainty of the measurement results. Additionally, he should know and use quality management methods.

The training concept described above is presented in detail in [11]. This concept has been used to develop the Metrological Knowledge Base, intended for training of CMMs' operators near their workplace, but through the "e-learning" form. Rich experience and valuable results obtained during the execution and implementation of the EUKOM project are good case study showing how the Information System for Metrology should be designed for distance learning of the operators who supervise measurement processes. The first action should be to define the target group (or groups) of operators, supervising a particular type of measurement process. As in each "e-learning" system an essential part is the knowledge base, the structure of such a base should follow the basic assumption concerning the competence of the operator obtained after completion of training. The levels of these competencies should be determined by the professional research and educational centers, which work closely with the industry that uses a particular type of measurement equipment. The structure of a knowledge base, designed in such a particular way, and integrated into appropriately selected training materials, is the primary factor determining the quality of education systems of operators, who supervise measurement processes.

The information technology used in e-learning system for preparing and making available the knowledge base is a secondary matter. Currently, there are many platforms (open source software) for creating "e-learning" systems, for example: [9], [12], [13]. The choice between one of these or one of commercially offered platforms depends on such factors as an expected number of people simultaneously using the "e-learning" system, a kind of supplementary teaching materials that are to be made available to learners. Among other factors taken into consideration are: the ability of teachers to prepare individual learning modules, the possibility of easy modification of the content of the various education modules and the ability to verify the acquired knowledge. In any case, the decisive factor in the quality of Information System for Metrology (ISM) designed for distance learning of operators will be the structure of the metrological knowledge base. It should be developed by the experts in the field of measurements on the basis of agreed level of competence, which the person benefiting from the training and intended to play certain role in the measurement process should reach.

---

### **The IT System for Metrology designed to predict the corrections needed to calculate of the Polish Universal Time Coordinated UTC(PL)**

---

Another type of Information System for Metrology (ISM) is an expert system designed for predicting corrections necessary to calculate the Polish Universal Time Coordinated - UTC (PL). This predictive information system for metrology has been continuously improved. Prediction of corrections values is implemented in this system by using various types of neural networks. [14]. This system allows for determining the Polish Official Time OT(PL) introduced in [15], [16] and valid in the Republic of Poland since 2004. This time is calculated as increased by one or two hours referring to the UTC (PL) [17], [18], [19]. It should be determined and maintained with an accuracy not exceeding 10 ns in relation to the universal time coordinated UTC determined by the Time Laboratory of the International Bureau of Weights and Measures in Paris (BIMP). BIMP creates UTC by calculating a weighted average based on systematic comparisons of 300 most accurate frequency and time atomic standards appearing in many countries around the world. BIMP also calculates corrections (PNMI) for each of National Metrology Institutes (NMI) including the corrections to the UTC (PL). In Poland, the Central Office of Measures (GUM) in Warsaw responsible for creation of the values of UTC (PL). The role of GUM in Poland is relevant to NMIs in other countries. The corrections for Poland (PPL) are determined by the BIMP and are published in a special bulletin "Circular T" [7]. These corrections are calculated for the day as:

$$P_{PL(BIMP)} = UTC - UTC(PL) \quad (1)$$

The problem is that the values of these corrections are announced only a month after their calculation in Paris. Therefore NMIs around world are forced to use the appropriate method to determine the predicted value of the corrections (PPRED) for the day. Predicting of the values of corrections is based on the historical collection of the corrections for each country published in the "Circular T" bulletin [7]. In different countries various methods of prediction are used. The most common is the analytic prediction method based on the linear regression method extended for stochastic differential equations [20]. Regardless of what prediction method will be applied, it is assumed that the error of such prediction for the day, for the NMI of the each country, should not exceed 10 ns, referring to the value of the corrections that will be published for this country in the bulletin "Circular T" [7]. Hence for Poland this condition can be formulated as follows:

$$\Delta = P_{PRED(PL)} - P_{PL(BIMP)} \ll 10 \text{ ns} \quad (2)$$

Although the analytic prediction method described in [20] is widely applied by NMIs and brings good results, it is certainly time consuming and quite difficult. For this reason, the studies [6], [14], [21] have been undertaken to

develop and implement an IT system for metrology, enabling the automatic prediction of the corrections for Poland PPRED (PL). This system has been implemented and tested in the Time and Frequency Laboratory of GUM, responsible for designating and propagation of Polish Official Time OT (PL). This time is established for winter and summer and is calculated as:

$$OT(PL) = UCT(PL) + 1 \text{ hour.} \quad \text{or} \quad OT(PL) = UCT(PL) + 2 \text{ hours.} \quad (3)$$

UTC (PL) is determined with use of the Polish Atomic Time Scale - TA (PL) created on the basis of mutual comparisons of atomic time standards used by over 20 Polish laboratories under and two foreign (Lithuania and Latvia) [19] collaborating under the agreement [17], [18]. The leading laboratory is the laboratory of the Polish Academy of Sciences, Space Research Centre (AOS) in Borowiec [22], for which BIMP separately calculates corrections PAOS (BIMP) and publishes them in the "Circular T" bulletin. AOS also participates in the creation of the Galileo-European Satellite Navigation System [24]. Studies [21], [23] showed that thanks to TA (PL) it is possible to meet the requirements of a specific inequality (2). Using the TA (PL) and the ISM, which automatically predicts the corrections PPRED (PL), GUM creates UTC (PL), which is one of the most accurate national UTCs in the world. Fig.2. shows the block diagram of the ISM using artificial neural networks [21]. This system combined with GUM's selected atomic time standard ensures the creation of UTC (PL) with an accuracy not exceeding 10 ns with respect to UTC time set by the BIMP in Paris.

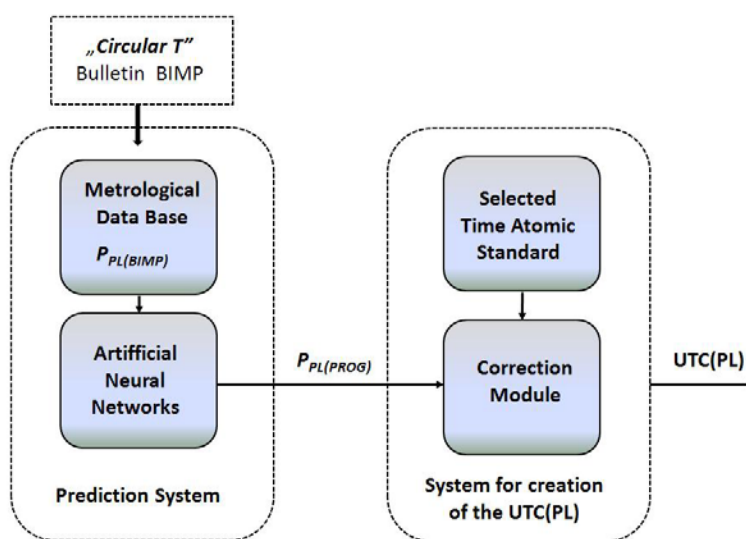


Fig.2. Block diagram of the ISM intended for automatic predicting of corrections PPRED(PL) and for creation of the UTC(PL)

---

### The IT System for Metrology intended for Computer-Aided Measurement Processes Validation - "CAMPV System"

---

Validation is an activity aimed at experimental verification of whether the properties of the method, object, process or software fulfill the requirements for their intended use. The summary of validation must provide the appropriate certificate definitely confirming that the outcome of the validation is positive. In case of the designing process, the goal of validation is to check and confirm that the prototype of the product meets the requirements of the application, for which the product is designed. In case of the measurement process, validation is an experimental verification and confirmation that the measurement characteristics of the measurement process

meet the requirements of its intended use. The full validation cycle of the measurement process - developed and described in [25] - is a set of action, which should result in the issuing of a certificate confirming that the metrological characteristics of the measurement process meet the requirements of its intended use. The full validation cycle consists of 7 specific steps:

1. Identification and characterization of the intended use of the measurement process (IUMP)
2. Determination of metrological requirements for the intended use of measurements (MRIUMP)
3. Choosing the right measurement process (SMP)
4. Determination of metrological characteristics of the selected measurement process (MCMP)
5. Comparison of the metrological requirements of the intended use of measurement process (MRIUMP) with the metrological characteristics of the measurement process (MCMP)
6. Determining result of the comparison and say whether it is positive or negative
7. Execution of one of alternative actions:
  - 7.1. If the result of the step 6 is positive: preparation and generation of validation certificate
  - 7.2. If the result of the step 6 is negative: selection of a different measurement process followed by determination of the measurement characteristics of the new measurement process (step 4) leading steps 5 and step 6.

Fig. 3. shows the flow diagram of the full validation cycle of the measurement process [25].

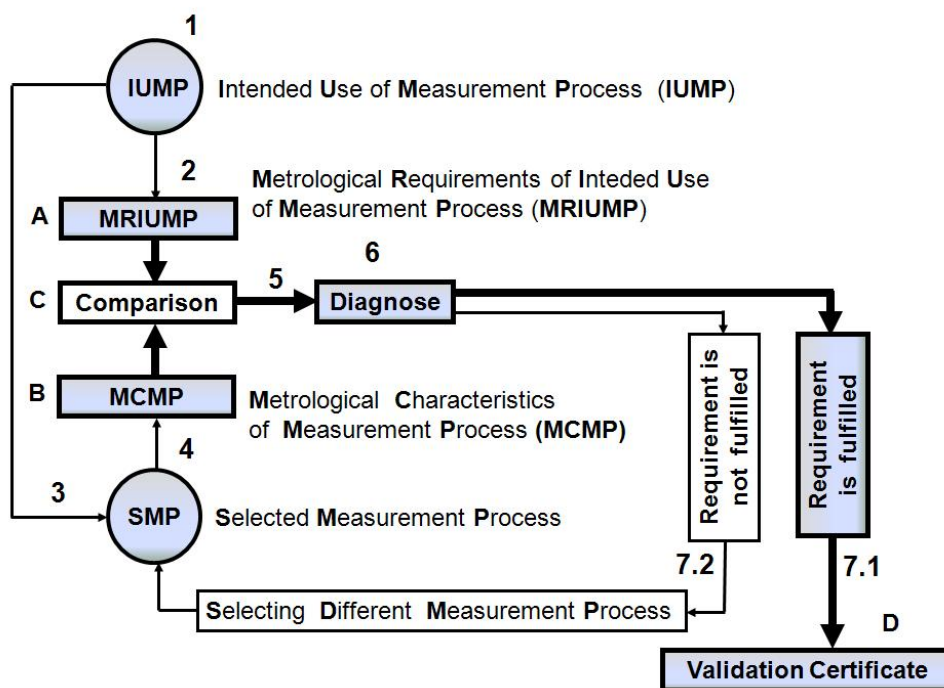


Fig.3. The flow diagram of full validation cycle of measurement processes consisting of 4 phases (A, B, C, D) and 7 specific actions

It should be emphasized, though, that the loop of steps: (7.2) - (4) - (5) - (6) can be repeated as many times as necessary to achieve the positive result of the step 6.

All seven steps are described in details in [25]. They can be grouped into 4 phases:

- A** – determination of the Metrological Requirements of the Intended Use of the Measurement Processes (MRIUMP),
- B** – determination of the Metrological Characteristics of the Measurement Process (MCMP),

- C – comparison of MRIUMP with the MPMP according to assumed criterion of the validation,
- D – generation of the Validation Certificate (VC) confirming that the selected measurement process (SMP) meets the requirements of its intended use.

The measurement process can be accepted and implemented for intended use only if the result of validation is positive. Such rule particularly refers to production quality control, conformity assessment of products or health and environmental protection. In order to execute all steps of the full validation cycle of the measurement process, relevant metrological data must be collected and stored, such as technical specifications of measurement equipment (hardware and software), results of its calibration and results of their statistical analysis, which are the metrological characteristics of evaluated measurement processes.

Having regard to above, the Department of Metrology and Diagnostic Systems, Faculty of Electrical Engineering and Computer Science in Rzeszow University of Technology has been undertaking systematic efforts [25], [26], [27] aiming in development of the most adequate methodology for validation of the measurement processes and creation of the information system, which would support the implementation of this methodology. The outcome of this work is the information system for metrology named “CAMPV-system” (the first part of the name is an abbreviation derived from the full name: Computer Aided Measurement Processes Validation).

Fig.4 shows the structure of the ISM of “CAMPV-system”, including website portal providing the online access to the system and the metrological operations database (MOD) designed to collect technical data of the measurement equipment and the results of its calibration. Already designed and implemented modules of the “CAMPV system” are marked by dark background in Fig.4.

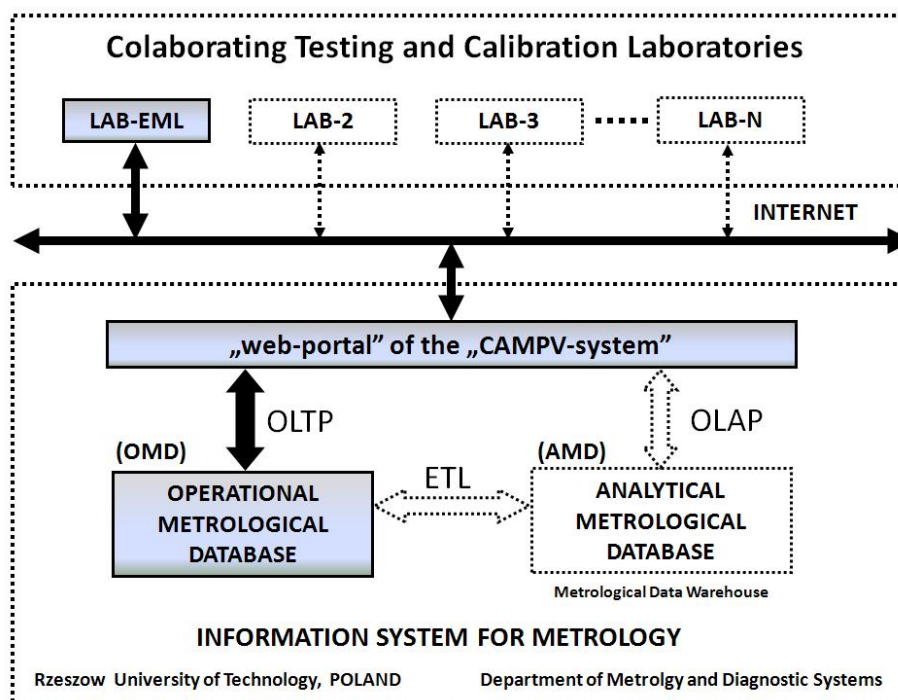


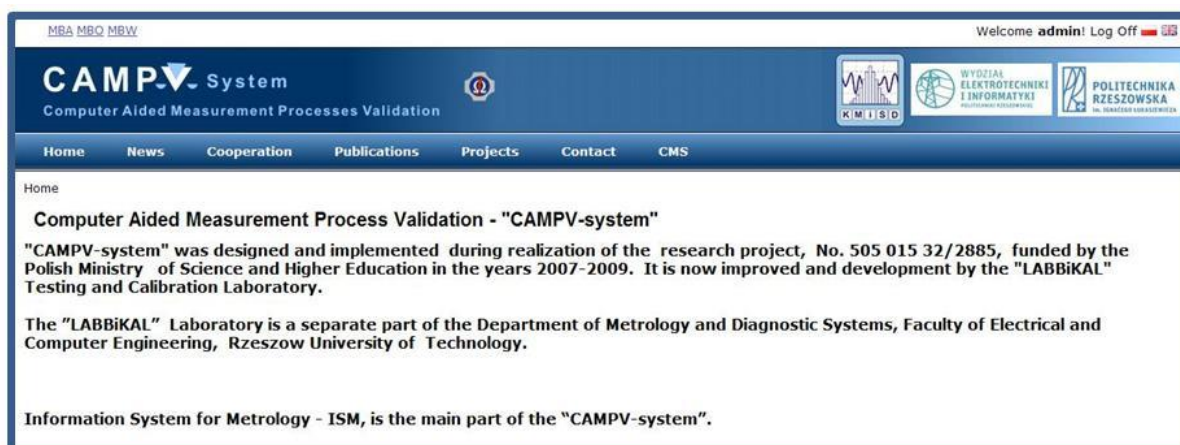
Fig.4. Structure of the CAMPV-system including already developed modules (dark background) and the planned module: Analytical Metrological Database (AMD).

The choice of information technology for respective modules depended on the progress of system’s development and the experience gained during the process. Regardless chosen technology, the supreme and indisputable assumption has been as follows:

*“ensuring authorized online access to the “CAMPV system” for collaborating laboratories and enabling the extension of system’s functionality by adding new modules”.*

The first prototype of the website portal for the “CAMPV-system” was based on PHP technology, whilst the metrological database used MySQL [28]. The next version of the website portal was based on MS ASP.NET technology and metrological database used MS SQL 2005 technology. The choice of these technologies is explained in [29]. The most important reason for choosing the Microsoft’s software environment was to ensure the continuation of the development of the “CAMPV-system”. Moreover, the Microsoft Technologies can help in creating of more complex models of metrological databases in the future. The last but not least, Department of Metrology and Diagnostic Systems has had an access to Microsoft’s MSDN license, including the license for use of the MS Windows Server. This software ensures compatibility between the created information system for metrology (ISM) and the operating system, which manages the work of the servers.

In the near future the CAMPV-system is expected to be completed by adding the Analytical Metrological Database (AMD), indicated by dotted line on the diagram shown in Fig.4. Such component will ensure an appropriate grouping of the collected data and separating the AMD users from On-Line Transaction Process (OLTP) operations. The AMD also allow for using the Extraction Transaction Loading (ETL) operations to group data in the most suitable form for their future analysis. In order to find this form, the On-Line Analytical Processing (OLAP) has to be determined. Having above in mind, the future version of “CAMPV-system” will have an adequate structure, as presented in Fig. 4. CAMPV system has recently been improved by using newer versions of Microsoft technologies. The website portal has been rebuilt with use of Model View Controller technology (MVC v.3.), with the RAZOR engine and MS SQL 2008 environment for creating databases. This technology allows for dynamic changes in the structure of the menu names of the website portal as well as tabs and content published through Content Management System (CMS). It also allows for creating different language versions. Following this improvements, the website portal has become a module that not only can be quickly rebuilt but also easily expanded with use of the Microsoft Silverlight technology and WCF RIA Services. Fig.5. shows one of the portal screens appearing after selecting the “home” button.



*Rys.5. The view of the website “home” of the website portal of the “CAMPV-system”*

Operational Metrological Database (OMD) is currently being reconstructed with use of MS Silverlight technology. The advantage of this solution is that the User Interface (UI) of the webpage is loaded only once, and the data used to fill the forms are loaded later during work with the server. Other innovations are planned to be implemented in the future, including drag&drop technology applied for creating single measurement channels consisting of elementary modules or creating a complex measurement process with use of several single measurement channels.

Fig.6. shows the design of the future OMD website. It provides access to one of the modules of the OMD, which collects technical data of measurement equipment and loads files with calibration results of measurement equipment. This module - named Applied Measurement Processes (AMP) - will also be equipped with features such as creating measurement processes, collecting the results of the repeatability and reproducibility analysis as well as collecting the results of interlaboratory comparisons. The results of calibration loaded into the OMD are saved in the Excel 2007 files. In the future other formats are expected such as .csv and .xml.



*Fig.6. The design of the future Operational Metrological Database (OMD) website accessing Applied Measurement Processes (AMP) module of the "CAMPV-system"*

---

## **Conclusion**

Three examples of different Information Systems for Metrology (ISM) are described. The first one has been designed to train operators of coordinate measuring machines (CMM) with use of the "e-learning" system. A Metrological Knowledge Base (MKB) structure plays an important role in this type of ISM. Such MKB contains elementary teaching modules tailored to the level of competence of the operators, who improve their measurement skills. The second system has been designed to predict the corrections, which allow for creation of the Polish Universal Time Coordinated UTC (PL). The key module in this type of ISM is the structure of the artificial neural network predicting the corrections values. The third system has been designed for the computer-aided measurement processes validation. In this type of ISM the major role plays an open and modular structure, accessible through the website portal and ready for a continuous expansion. This ISM called "CAMPV-system" has been tested and developed in the Department of Metrology and Diagnostic Systems, Faculty of Electrical and Computer Engineering of the Rzeszow University of Technology. The target version of "CAMPV-system" will include two databases and a knowledge base. This form may become the basis for creation of the metrological expert system "CAMPV-EXPERT-system" supporting the validation of the measurement processes using artificial intelligence methods.

Based on above-described examples, one can conclude that the information technology contribute significantly to the development of the metrology and its applications. A key condition of successful creating of the high quality IT systems for the metrology is a close cooperation between scientists and practitioners in the field of computer science and metrology.

---

## **Bibliography**

[1] [Finkelstein-1982] Finkelstein L.: Theory and philosophy of measurement, In: Sydenham P.H. (ed.) Handbook of Measurement Science, Chichester: Wiley, 1982, pp. 1-30.

- 
- [2] [Muravyov-1997] Muravyov S.V., Savolainen v.: Towards describing semantic aspect of measurement. Proceedings of the XIV IMEKO World Congress, Tampere, Finland, 1997.
- [3] [Stevens-1946] Stevens S.S.: On the theory of scales of measurements. SCIENCE.
- [4] [Finkelstein-2003] Finkelstein L.: Analysis of the concept of measurement, information and knowledge. Proceedings of the XVII IMEKO World Congress. Dubrovnik. Croatia.2003. pp. 1043-1047.
- [5] [Keferstein-2007] Keferstein C.P., Marxer M.: EUKOM-European Training for Coordinate Metrology. Proceedings of the 13 International Metrology Congress. Lille. France. 2007.
- [6] [Cepowski-2009] Cepowski M, Miczulski W.: Metody prognozowania państwowej skali czasu. Materiały VII Konferencji Naukowo-Technicznej Podstawowe Problemy Metrologii. Sucha Beskidzka. 2009. ss.12-15
- [7] Circular T". Bulletin. (<ftp://ftp2.bipm.fr/pub/tai/publication/cirt/>)
- [8] [www.cm-train.org](http://www.cm-train.org)
- [9] <http://www.ilias.de>
- [10] Research project: „European Training for Metrology”. University Erlangen-Nuremberg. Chair Quality Management and Manufacturing Metrology. (in polish version: [www.lm.ath.bielsko.pl](http://www.lm.ath.bielsko.pl))
- [11] [Weckenmann-2004] Weckenmann A., Jakubiec W., Płowucha W.: Training concept EUKOM. Leonardo da Vinci pilot Project. European Training for Coordinate Metrology. D/02/B/F/PP 112 662 EUKOM. Erlangen. 2004.
- [12] <http://moodle.org>
- [13] <http://www.openelms.org>
- [14] [Miczulski-2011] Miczulski W., Sobolewski Ł.: Wpływ sposobu przygotowania danych na wynik prognozowania poprawek dla UTC(PL) z zastosowaniem sieci neuronowej GMDH. Materiały z VIII Konferencji Naukowo-Technicznej. Podstawowe Problemy Metrologii. Krynica. 2011.
- [15] Ustawa z dnia 10 grudnia 2003. O czasie urzędowym na obszarze Rzeczypospolitej Polskiej (Dz. U. z 2004. Nr16, poz.144.)
- [16] Rozporządzenie Ministra Gospodarki, Pracy i Polityki Społecznej z dnia 19 marca 2004, w sprawie sposobów rozpowszechniania sygnałów czasu urzędowego i uniwersalnego czasu koordynowanego UTC(PL).
- [17] [Czubla-2006] Czubla A.,Konopka J., Nawrocki J.: Realization of atomic SI second definition In context UTC(PL) and TA(PL). Metrology and Measurement System. No 2, 2006, pp. 149-159.
- [18] [Nawrocki-2006] Nawrocki J., Rau Z., Lewandowski W., Małkowski M., Marszałec M., Nerkowski D.: Steering UTC(AOS) and UTC(PL) by TA(PL), In Proceedings of the Annual Precise Time and Time Interval (PTTI) Systems and Applications Meeting, 7-9 December 2006.
- [19] [Marszałec-2011] Marszałec M., Lusawa M., Nerkowski D.: Wyniki badań algorytmów zespołowych skal czasu z wykorzystaniem bazy danych dla TA(PL). Materiały z VIII Konferencji Naukowo-Technicznej. Podstawowe Problemy Metrologii. Krynica. 2011.
- [20] [Panfilo-2008] Panfilo G., Tavella P.: Atomic clock prediction based on stochastic differentia equations. Metrologia 45, 2008, pp.108-116.
- [21] [Cepowski-2009] Cepowski M., Miczulski W.: Zastosowanie sieci neuronowych do prognozowania państwowej skali czasu. Materiały z XVII Sympozjum Modelowanie i Symulacja Systemów Pomiarowych. Krynica 2009.
- [22] Polska Akademia Nauk. Centrum Badań Kosmicznych. Obserwatorium Astrogeodynamiczne w Borowcu (<http://www.cbk.pl>)
- [23][Miczulski-2010] Miczulski W., Cepowski M.: Wpływ sieci neuronowej i sposobu przygotowania danych na wynik prognozowania poprawek UTC- UTC(PL). Pomiar Automatyka Kontrola vol.56, nr 11, 2010, ss.1330-1332.



- [24] Polski Punkt Informacyjny GALILEO (<http://galileo.kosmos.gov.pl>)
- [25] [Tabisz-2010] Tabisz R.A.: Walidacja procesów pomiarowych. Przegląd Elektrotechniczny (Electrical Review). R86 Nr 11a, 2010, ss. 313-318
- [26] [Tabisz-2009] Tabisz R.A.: Validation of Industrial Measurement Processes. Proceedings of the 13th International Metrology Congress. 18-21 June. Lille. France. In the book: Transverse Disciplines in Metrology. ISTE-Wiley. 2009 pp.791-801.
- [27] [Tabisz-2009] Tabisz R.A.: Computer Aided measurement Process Validation. Proceedings of the 14th International Metrology Congress, 22-25 June, Paris. 2009.
- [28] [Adamczyk-2007] Adamczyk K., Tabisz R.A.: Zastosowanie wybranych technologii informatycznych do tworzenia Metrologicznej Bazy Danych. Pomiary Automatyka Kontrola. Vol.53. nr 12, 2007, ss. 51-54.
- [29] [Świerzowicz-2008] Świerzowicz J., Adamczyk K., Tabisz R.A.: Kluczowe etapy tworzenia Informatycznego Systemu Metrologicznej Bazy Danych przeznaczonego dla sieci laboratoriów badawczych i wzorcujących. Pomiary Automatyka Kontrola. Vol.54, nr 12, 2008, ss. 869-873.

---

### **Authors' Information**

---



**Roman Aleksander Tabisz** – *Department of Metrology and Measurement Systems, Faculty of Electrical and Computer Engineering, Rzeszow University of Technology. W. Pola.2. 35-959 Rzeszow, Poland. e-mail: [rtabisz@prz.edu.pl](mailto:rtabisz@prz.edu.pl)*

*Major Fields of Scientific Research: Industrial metrology, measurement processes validation*



**Łukasz Walus** – *The owner of the IT enterprise: SolvSoft. Łukasz Walus. 36-230 Domaradz 460A, Poland, e-mail: [ukaniow@gmail.com](mailto:ukaniow@gmail.com)*

*Major Fields of Scientific Research: Business applications of the IT Technology.*

## Business Intelligence Systems

---

### BUSINESS DISCOVERY – A NEW DIMENSION OF BUSINESS INTELLIGENCE

**Justyna Stasieńko**

**Abstract.** *This article deals with the issue of Business Intelligence (BI), especially its next generation - Business Discovery (BD). This tool can help to fill the gap between traditional solutions we get from BI and standalone office productivity applications. Its users are able to forget new paths and make new discoveries. Here we want to present BD as being thoroughly complementary to traditional ERP, CRM, BI, and data warehousing systems. BD brings a whole new level of analysis, insight and value to the information stored within these systems. What is more, its users are not burdened with interfaces which are difficult to use and configure.*

**Keywords:** *Business Intelligence, Business Discovery, information, analysis, Qlickview*

**ACM Classification Keywords:** *K.6 Management of Computing and Information Systems - K.6.0 General Economics*

*“Computers are useless. They can only give you answers.”*

Pablo Picasso

---

### Introduction

---

Currently, one can notice a sea-change approach to the role and importance of information. So far, information has often been treated as a by-product or, at best, co-implemented business processes. Now the information is one of the most important organising resources. The information is a factor that increases the knowledge about the reality surrounding a man, or specific intangible asset, which, with economic progress and development of means and forms of social communication, is becoming more and more important, transforming the face of many traditionally organised economies of the world.

The importance of information in the modern world should be analysed in various aspects. One can even attempt to say that in a sense, the information is the engine of progress in every area of human life. Following an increasing pace, technological development to a great extent depends on the speed and the quality of information. Thus, the access to information should be easy, and the way it is used should present its values.

In order to achieve business benefits the strategic decisions are taken. These decisions result from the study of gathered information. For hundreds of years experience had been mainly a source of information. They were mainly mathematics and mathematical models, then the statistical models, econometric, and now the Internet.

A growing number of data, a flow of information, fast communication via the Internet and new market challenges mean that data analysis, broadening knowledge and skilful decision-making in an increasingly complex market are essential to survive and run business. Solutions in an enterprise operate at many levels and play various roles supporting the administration or management, yet they have their own limitations. Especially in the

processing of a large number of diverse data and in the use of information in many fields. In order to meet the demands BI solutions have been created .

Business Intelligence is now the most important and inevitable point of contact between sciences and business. Business Intelligence tools enable an easy access to information, its analysis and sharing across the organisation and its business environment. They give the possibility to integrate data from different sources and their comprehensive analysis in terms of business needs. BI gives a preview of all business organisations. Their goal is to support effective business management and business planning by providing the right information. They support the work of managers in managing key areas of business. Most generally, one can present it as a process of transforming data into information and the information into knowledge that can be used to enhance the competitiveness of enterprises. Among these tools there are both management systems of information resources, reporting and analysis tools, and also solutions enabling to boost managing performance.

Information in the organisation is generated and processed mostly in the transactional software, such as ERP, CRM. These systems have evolved over the years. The first stage of its development was automating tasks and processes, reporting and logging. This has contributed to an increase in the information registration. The second phase was the emphasis on resource planning and the efficiency of business processes. For better reporting and data analysis data warehouses have been introduced— an aggregation of transactional information and multi-dimensional analytical models. Due to the need of a larger amount of information for analysis and its presentation in a friendly manner, tools in the field of decision support proved to be necessary. The solution which came out to meet the problems of modern organisations are Business Intelligence Systems (BI). Traditional BI technology use data warehousing, analytic tools (OLAP and Data Mining), and presentation techniques. They enable to optimise operations, to increase efficiency, to reduce the risk of taking wrong decisions, as well as to reduce costs and to maximise profits.

---

### Analytical tools

---

Analytical tools existing on the market support a decision maker by providing him/her with the necessary knowledge - in the form of reports based on historical and current data - to make decisions. They allow the standard and advanced reports using statistical analysis, forecasting, relationship between the data search, research trends [Nycz, 2008].

The basic analytical tools include query generation tool and reporting (Query & Report - Q & R), spreadsheets, OLAP mining and data visualisation tools [Dudycz, 2004].

Query and reporting tools are the most basic tools for data analysis, in particular, gathered in the data warehouses. There are two types of reporting: the standard and so called 'ad hoc'. Frequently they answer the questions "what happened?", "What level of sales was as in the previous year?" etc. In the second half of the 1990s, it was noted that the data stored in databases, transactional systems used in companies caused a lot of trouble to analysts who carried out the assessment of a business enterprise. This problem was solved by introducing analytical techniques implemented, inter alia, in spreadsheets [Dudycz, 2004], which enable to create models that generate periodic reports automatically [Sierocki, 2007]. Spreadsheet offers flexibility when it comes to the definition of the conditions of analysis and ease of use. The difficulty arises when the basis for analysis are large volumes of data or high complexity of the model. In order to achieve the desired analytical flexibility a wide enough diagram of processing must be built, often based on large amounts of macro-commands and sheets. This solution cause difficulties of managing the data, and at the same time it is prone to user's errors. Sheets have a limited working capacity for the data, which almost eliminates their usefulness for the analysis of large portions of data, reaching hundreds of thousands of transactions.

---

---

Facing new challenges, a concept of multidimensional databases and OLAP technology emerged, which allowed for a dynamic and multidimensional analysis of business data.

OLAP architecture is encountered in data warehouses and tools for analysis such as query languages, data mining, artificial intelligence, as well as report generators. Through proper presentation, visualization and aggregation, it enables to display and view data from different points of view allowing its user to examine them quickly. In addition, it is characterized by the possibility of an interactive reporting without knowledge of programming languages and to obtain answers to complex and often non-standard queries in a current mode. Therefore, OLAP tools are often used to perform analysis of sales trends, financial analysis (data warehouse), or to pre-screen the data set by the analyst in the initial phase of statistical analysis [Sierocki, 2007].

Data visualization tools are designed to increase transparency and legibility of presented information. Most of the analytical tools offer simple dependence images between the data.

Existing tools of data acquisition and processing of analytical reports, such as generators or spreadsheets were not able to fully meet the needs of managers who have to make a relatively fast growing in-depth analysis.

Facing these challenges, a concept of multidimensional databases and OLAP technology was introduced, HOLAP, MOLAP and ROLAP, which allow for a dynamic and multidimensional analysis of business data. Architecture of the OLAP (MOLAP, DOLAP, ROLAP, HOLAP) may be encountered in data warehousing and data analysis tools such as generators, reports, query languages, data mining, artificial intelligence. It enables to display and view the data from different points of view allowing its user to examine them quickly through an appropriate method of presentation, visualization and aggregation.

OLAP systems are characterized by the possibility of:

- perform multidimensional analysis according to complex search criteria,
- interactive reporting without knowledge of programming languages,
- obtaining answers to complex and often non-standard (so called 'ad hoc') queries in a current mode.

Therefore, OLAP tools are often used to perform analysis of sales trends, financial analysis (data warehouse), or to pre-screen the data set by the analyst in the initial phase of statistical analysis.

In order to deepen the analysis and discovery of repetitive behaviours in large data sets through data mining, matching various models and relationships between data analysis, special methods are used for data mining. Data mining is a methodology that refers to a technique derived from mathematical statistics and machine learning algorithms. Information extracted by using these tools can be used in areas such organisations as the support of decision making, forecasting, financial analysis and risk analysis, optimisation.

A frequently used data mining tool is a universal, integrated system for statistical data analysis - STATISTICA. This software not only contains statistical and graphical procedures for general use, but also powerful tools for analysis and visualisation of data, as well as specialised analytical techniques (e.g. social studies, biomedical, or technical) [Dudycz, 2004].

Other tools to cope with the analysis of the large amounts of data processed into information, and then into knowledge are BI systems.

BI systems should improve the management of knowledge in an organisation at the three levels presented in Table 1

*Table 1. Tasks of Business Intelligence Systems*

<b>Management level</b>	<b>BI tasks</b>
<b>Operating</b>	Analysis carried out ad hoc, information on current operations, finances, sales, collaboration with suppliers, customers, clients, etc.
<b>Tactical</b>	Fundamentals of decision making in marketing, sales, finance, capital management. Optimising future actions and modification of financial factors, technology in the implementation of strategic objectives.
<b>Strategic</b>	Precise setting of goals and tracking their implementation, to perform various comparative statements, conducting simulation development, forecasting future performance under certain assumptions.

Source: Own elaboration.

### **The new generation of Business Intelligence systems**

Business Intelligence turns out to be the new quality in the management conception. BI systems are used in order to create and improve the relationship with a customer, yet at the same time to boost management effectiveness. Unfortunately, traditional BI software seems to have failed, as far as delivering on this vision is concerned, as a result of its complexities, time lags, and expensive professional services requirements.

Forrester Research's definition of BI is "a set of methodologies, processes, architectures, and technologies that transform the raw data into the meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making." This definition of BI covers the whole data-to-insight process (including data preparation). It all appears to be time-consuming, especially to plan and implement - from collecting requirements, to building a data warehouse, to populating a metadata layer. Traditional BI software users have problems to learn and use it, and, in consequence, adoption is limited. Moreover, distribution of information and analytic tools is tightly controlled.

Traditional BI systems are high cost and IT driven. They are chosen, installed, and maintained by IT organisations and in most cases not by the business people themselves who will use them later on. Owing to the complexity of the system, not many people (in an organisation) feel skilled enough to form business insights. When business analysts and IT professionals want to be sure that they deliver the right analysis, the back-and-forth questions and answers with their business constituents make it really difficult for them. Traditional BI usually constitutes more or less centralised, pre-packaged reports or predetermined queries which users can run to get updated numbers. It often happens that the information we get is mostly static. Therefore, users having a question which is outside the standard configuration need to log a ticket with IT and expect their assistance (sometimes weeks or even months). The drawback of BI is the fact that it is centralised, possessed by IT, difficult to change or modify, and slow to deliver results. What is more, it is also expensive and highly complex.

In addition, if one wants to deploy a traditional BI solution, it can take him/her up to a year and a half. By and large, it seems to be due to requirements gathering and data modelling and integration efforts. This is definitely not what the business needs as an organisation can live and die within this time. Furthermore, traditional BI requires lots of services and support in order to keep its various components working smoothly.

BI in the company combines finance, manufacturing, warehousing, logistics, purchasing, sales, HR, planning and strategy - in short, all aspects of a company. Therefore, BI uses a common repository of information - Data Warehouse. All facts come from individual branch systems through ETL processes converted into information and

---

---

stored in the DSA - Data Staging Area data warehouse. From this information the system uses the second part of the Business Intelligence which converts this information into knowledge and provides the user through the presentation layer. Thanks to the class Business Intelligence supports managers effectively, and enables, inter alia, building a What-If analysis, budgets and controlling systems.

Traditional BI turns out to be excessively bloated and rigid. Further evolution of these systems will lead to the revival of petrified BI. During this evolution Business Discovery (BD) has emerged. Trying to answer why BI platform displaces BD, one would indicate the four trends that have caused the evolution of the BI software market. The first one is the ability to search the Internet and to obtain a rapid response. The second trend is the community network (social networking) that enables to communicate, to share information and to develop robust, professional and personal networks (with no requirement of technology background). Other trends are the development of mobile and task-specific applications. In conclusion, BD are much faster, open and straightforward at the same time, mobile and addressed for everyone.

Business Discovery platforms (offering new solutions), as opposed to its ancestor, can have a total cost of ownership that is half that of other BI solutions. Business Discovery is a whole new way of doing things for Business Intelligence.

---

### **Concept of Business Discovery**

---

If we decide to store data in-memory, it means we no longer have to deal with a database located somewhere else, and we receive no queries and no retrieval. Thanks to it, there is no delay in returning results, whatsoever.

Business Discovery (BD) is a complement to traditional BI, ERP, CRM, and data storage systems. It also introduces

a new level of analysis, knowledge and value of information that fall within these systems. Additionally, it became a response to the unmet needs of users as it is a new way of doing BI. It is a bottoms-up approach that puts the user in control, fulfilling the promise of BI. The main aim of BD is to help users to solve specific business problems in a timely way to get answers to the most critical questions and also to share knowledge and analysis among individuals, groups, and even organisations.

Most essentially, users are capable of gaining insights that address their individual needs at every level of the organisation. It seems that they are not limited to particular paths they must follow, or questions they are obliged to formulate in advance. The key issue is the fact that they can ask about what they need.

Business Discovery gives an opportunity of a whole new level of analysis, insight, and value to existing data stores with user interfaces that are clean, simple, and straightforward. It is complementary to traditional BI software and other enterprise applications. BD is enriched with new opportunities for BI. The most vital one would include an application interface (insight for everyone), the time to provide results of the analysis, mobility of applications, remixability and reassembly, and finally social and collaborative environment.

Importantly, everyone can create an insight by means of Business Discovery. It's the equivalent of open source computing or peer creation. This is definitely intelligence creation — rather than just information consumption. First of all, Business Discovery is not a large collection of centrally-controlled, pre-packaged, and tightly-distributed data. Secondly, rather, it provides data access and analysis to individuals and groups, and allows them to get what they ask for faster and more accurately than ever before.

BD enables instant analysis. The user receives the results at the bedside, where it lasted weeks with the traditional BI. It has a direct access to all necessary data. Technology can ask any questions to which answers are kept on-line.

At all levels in an organisation business decision makers need data at their fingertips, wherever they are. They want to work how and where they like - whether that be in the warehouse, on the customer site, or on the trade show floor. Tablets and other large-form-factor mobile devices promise to make business data ubiquitous. Unlike traditional BI solutions, Business Discovery platforms provide an intuitive interface and an application infrastructure that is tailor-made to exploit the opportunity of a truly mobile, well-informed workforce.

BD supports mobile applications. This allows the provision of data while using mobile devices such as iPhone, Android.

Business Discovery platforms empower anyone to quickly develop and deploy simple, focused, and intuitive applications that can be easily reused. These applications are easy to modify, mash up, and share, allowing innovation to flourish at the edges of the organisation and spread inward. The new opportunity is leveraging a model that lets any user quickly develop and deploy task-specific, purpose-built BI applications. BD platform enables its users to quickly create and implement among others their own applications. These applications are able to quickly solve specific problems.

Nobody can predict what questions business users will have when they start exploring data — not even the users themselves. Traditional BI solutions require IT or power users to get involved whenever new questions arise. In contrast, Business Discovery platforms make it easy for business users to remix and reassemble data in new views and create new visualisations for deeper understanding. With BD, users generate insights like never before.

BD makes it easy to "remix" and reassemble data to the new views (previews) and a fast way to create visualisations.

BD is a social and collaborative environment. It enables its users to share and collaborate on insight and analysis. They can share insights within Business Discovery apps or through the integration with collaboration platforms. Business Discovery is about creating a community of users who engage in wiki-like decision-making to drive knowledge that can cascade across an organisation.

---

### **QlikView BI in-memory**

---

QlikView is a modern and innovative approach to Business Intelligence. What adds value to the existing QlikView BI applications include: making the process of assembling, associating, and preparing data for analysis simple and straightforward, allowing users to interact with data in the way they think-associatively. A great advantage of QlikView is that data is collected in memory, which improves analysis and convenience to use.

QlikView was built with a simple architectural premise. All data should be held in memory, and all calculations should be performed when requested and not prior. The goal was to deliver the powerful analytic and reporting solutions in a quarter of the time, at half the cost and with twice the value of competing OLAP cube based product.

QlikView is the world's first associative, in-memory BI platform. It manages associations among data sets at the engine level, not the application level, by storing individual tables in its in-memory associative engine. Every data in the analytic dataset is associated with every other data point in the dataset.

Associative cheese means finding answers to questions, but also the questions that have not yet been started. What is meant by a simple application is creating questions that do not require knowledge of creating queries in SQL. This associative experience gives decision makers a better overview of their business.

Visualisation is of dual significance. Firstly, it refers to the visual display of summarised forms of information. Secondly, the ability to see those displays change as the selected date is changed. QlikView offers various ways

of data presentation: graphs, charts, tables and others. It also enables its users to create different types of measures that enhance the analysis process. It provides flexible, intuitive and powerful data visualizations.

Using the QlikView application is pleasant and not too complicated (Fig.1)

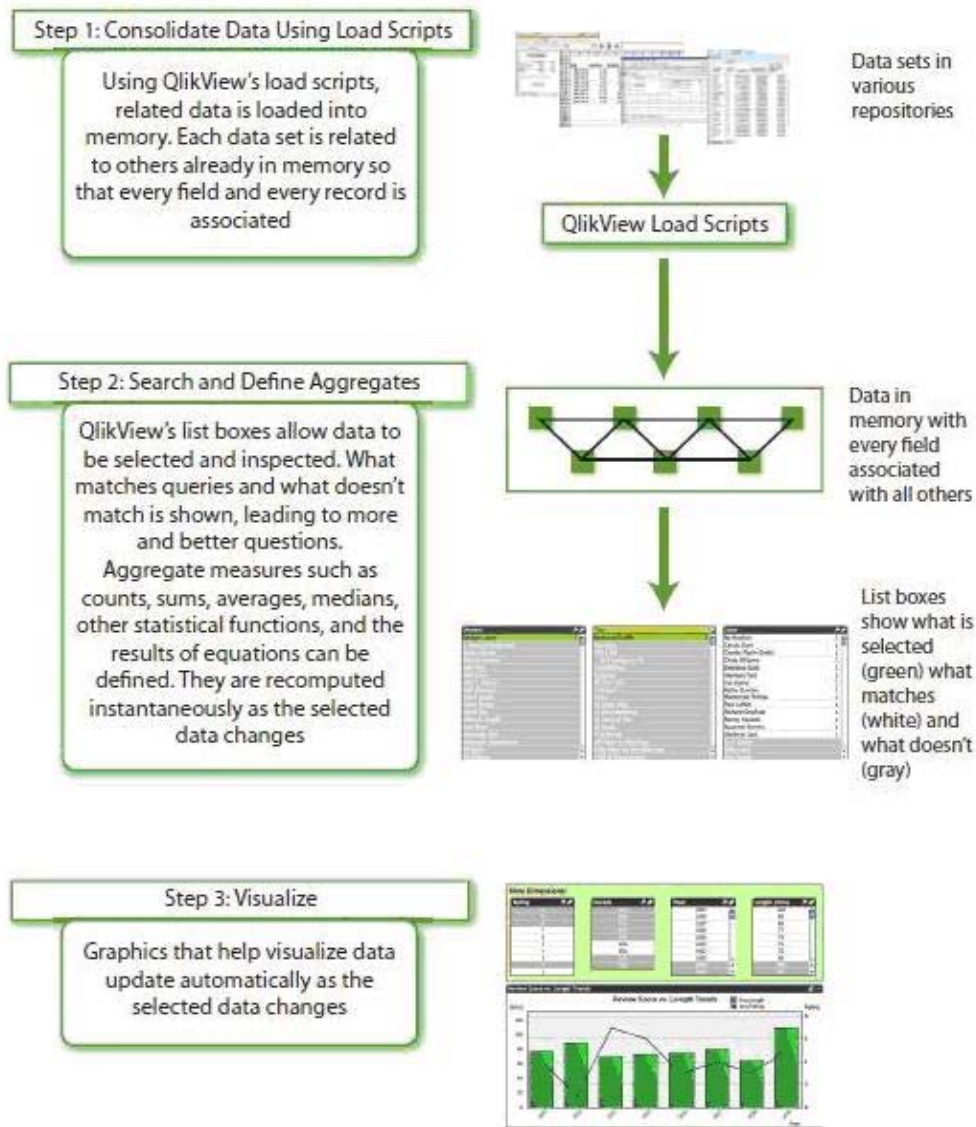


Fig. 1. A Step-by-Step Tour Through QlikView. Source: [The Art of Business Discovery p.8]

QlikView stores data in memory instead of retrieving it from databases and OLAP cubes, it cannot only display what is included in your search, but also what is excluded as well. This is certainly a real departure from traditional BI (i.e. for the first time, unknowns become known).

Compared to other analytical tools or BI based on the OLAP 'cube', an application of QlikView present the analysed data to the user in a faster and simpler way. QlikView gives a chance to impose criteria onto the data. In a straightforward and quick way, QlikView gives the opportunity to return to the previous data or to add further ones upon existing criteria. Due to its simplicity and ease to use, a user-friendly and attractive QlikView interface is a modern and highly efficient application. It has also, not available for other solutions, the time of submission of new studies, computing power and flexibility. Software flexibility leads to the lack of restrictions on the number



of dimensions and measures, and its power – to virtually immediate response to inquiries from the system, even with databases of up to five hundred million records. It also provides the possibility of an immediate transition to a single transaction.

QlikView enables to view data from different perspectives (Fig. 2). Indeed, some analysis can also be done, for example in a spreadsheet, but QlikView provides instant viewing of data by analysing different criteria.

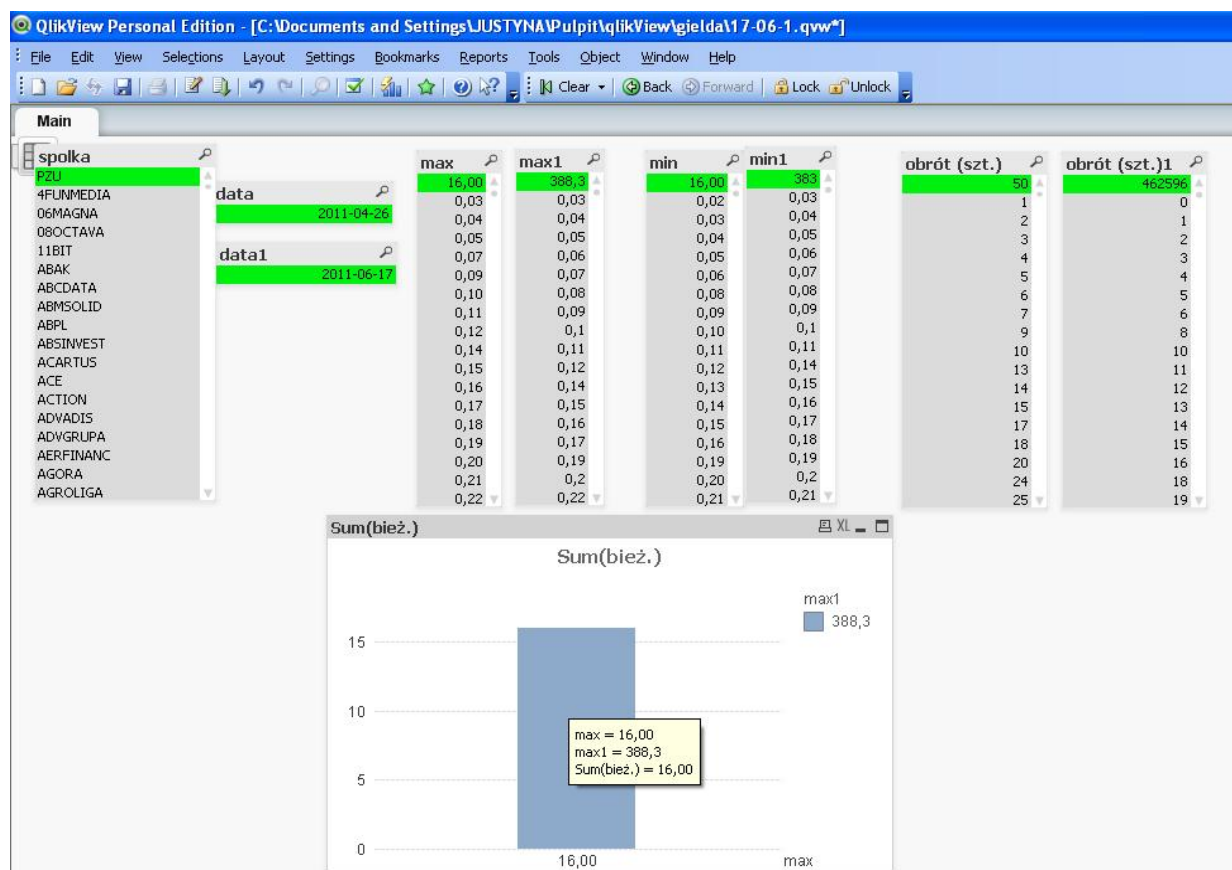


Fig. 2. Comparison of speed, start and end of a listed company PZU session held on 26.04.2011 and 17.06.2011

Source: Own elaboration

It is not necessary to switch from a previous analysis (Fig. 2), all one needs to do is to simply click or select the appropriate criteria and get a completely different analysis (Fig. 3) that no longer applies to the company, PZU, but only to the same level of stock prices such as 0.19 PLN during the trading session on 17.06.2011. This is a great help because another analysis from the beginning is not required. While working on some data one can analyse them in many ways without switching between windows, sheets, etc.

QlikView software enables to transfer analysis results, which remain still just as functional, onto hardware. The application allows printing of the results in the form of reports, exporting them to MS Excel or saving as PDF. It is practically capable of integrating all data formats - from standard relational data into text reports, the data from Excel and XML streams. QlikView is quickly and easily deployed and integrated with existing enterprise systems.

Table 2 shows the advantages of QlikView applications compared to traditional BI systems.

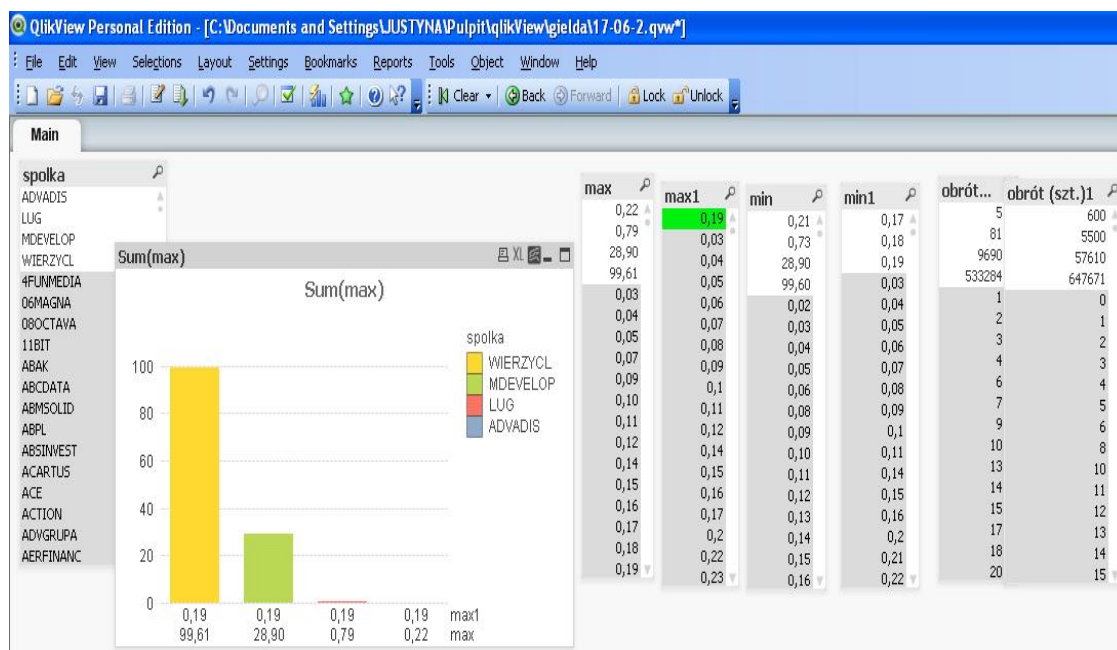


Fig. 3. Maximum price for shares at the 0,19 PLN level on June 17th, 2011. Source: Own elaboration

Table 2. The differences between traditional BI and in-memory BI systems, such as QlikView

Traditional BI	QlikView
Lots of tools: data warehouse and data marts, OLAP, query and reporting tools, data mining	Simple architectural premise - all data should be held in memory
Presentation techniques of analysis (dashboards, scorecards, reports)	Presentation techniques of analysis (charts)
Many users	One user
Longer time of implementation (approximately 18 months)	Short time of implementation (several weeks)
High costs of implementation	Low costs of implementation
Less convenient and flexible for user	Easy to use, flexible
Time-consuming and complex process of information processing	Fast query and on demand calculation engine

Source: Own calculation.

At the end of 2010, there is another version (QlikView 10). A new feature is the ability to deploy the software on any platform: local, cloud computing and mobile devices. The release of the platform in the cloud is via Amazon's Elastic Compute Cloud (EC2). It has also enormous power to allow flexible analytical processing of large data sets while maintaining access to the details. It is not limited by any number of dimensions. Changes in the designed applications can be performed quickly and without possessing advanced knowledge of programming. The application offers new opportunities to present the results of studies using AJAX. New ways to visualise data facilitate the understanding of the data presented. Searching for information through associative search capabilities has been improved as well.

Summarising QlikView in-memory analysis and reporting is simplifying analysis for everyone and has clearly demonstrated its affordability and value to organizations across industries for solving their performance and information challenges.

---

### Conclusion

BI systems have existed / operated in the market for about two hundred years. Transforming data into information and information into knowledge enabled business decisions, allowing users to make effective and informed choices based on data analysis. Technology development in this field has also brought change. It turned out that traditional BI systems are too complicated, delayed and requires professional services, which are costly. Therefore, the direction of change went toward cheaper and faster applications. Thus, there arose the Discovery Business Systems which are a whole new way of doing things for Business Intelligence.

BD bridges the gap between traditional BI solutions and standalone office productivity applications, enabling users to forge new paths and make new discoveries. BD works with what you have and infuses new capabilities into BI: insight for everyone, zero-wait analysis, mobility, an app-like model, remix ability and reassembly, and a social and collaborative experience.

---

### Bibliography

- [Dudycz, 2004] Dudycz H. Przetwarzanie analityczne podstawą rozwiązań informatycznych klasy Business Intelligence. Materiały konferencyjne SWO, Katowice, 2004. [http://www.swo.ae.katowice.pl/\\_pdf/138.pdf](http://www.swo.ae.katowice.pl/_pdf/138.pdf)
- [Januszewski, 2008] Januszewski A. Funkcjonalność Informatycznych Systemów Zarządzania, t.2 Systemy Business Intelligence, PWN, Warszawa, 2008.
- [Nycz, 2008] Nycz M., Smok B. Business Intelligence w zarządzaniu. Materiały konferencyjne SWO, Katowice, 2008. [http://www.swo.ae.katowice.pl/\\_pdf/421.pdf](http://www.swo.ae.katowice.pl/_pdf/421.pdf)
- [Olszak, 2005] Olszak C. Wspomaganie decyzji w erze informacji i wiedzy. W: Systemy wspomaganie organizacji. Praca zbiorowa pod redakcją H. Sroki i T. Porębskiej-Miąc. AE, Katowice, 2005, s.346-353.
- [Sierocki, 2007] Sierocki R.: OLAP to efektywna technologia przetwarzania danych analitycznych. Controlling i Rachunkowość Zarządcza nr 1/2007. <http://www.infosynergia.eu/pliki/publikacje%20-%20Robert%20Sierocki%20-20OLAP%20to%20efektywna%20technologia.pdf>
- [Surma, 2007] Surma J. Wsparcie strategicznych decyzji zarządczych z wykorzystaniem Business Intelligence. Innowacyjne Aspekty Strategii Przedsiębiorstwa pod red. Lidii Nowak. Politechnika Częstochowska, Częstochowa, 2007. <http://www.surma.edu.pl/wp-content/czestochowa-jerzy-surma-2007.pdf> .
- [Surma 2009] Surma J. Business Intelligence – Systemy Wspomaganie Decyzji Biznesowych, PWN, Warszawa, 2009.

### RELATED WHITE PAPERS

*The QlikView Product Roadmap*, December, 2010

<http://www.qlikview.com/us/explore/resources/whitepapers/qlikview-product-roadmap>

*The Associative Experience: QlikView's Overwhelming Advantage*, October, 2010

<http://www.qlikview.com/us/explore/resources/whitepapers/the-associative-experience>

*QlikView Architectural Overview*, October, 2010

<http://www.qlikview.com/us/explore/resources/whitepapers/qlikview-architectural-overview>

<http://community.qlikview.com/blogs/theqlikviewblog/archive/2010/09/21/software-should-fit-the-business-not-the-other-way-around.aspx>

A QlikView White Paper; *Business Discovery: The Next Generation of BI*; Published January, 2011;  
<http://www.qlikview.com/us/explore/resources/whitepapers/business-discovery-the-next-generation-of-bi>

A CITO Research Explainer: *The Art of Business Discovery*; May 2010

<http://www.qlikview.com/us/explore/resources/whitepapers/the-art-of-business-discovery> [March, 2011]

A QlikView Technology White Paper: *QlikView Architectural Overview* Published October, 2010  
<http://www.qlikview.com/us/explore/resources/whitepapers/qlikview-architectural-overview> [March, 2011]

A White Paper from QlikTech International AB: *QlikView 7. The In Memory Business Intelligence Revolution*.  
[http://www.businessintelligence.bg/files/presentations/Intro\\_to\\_QlikView\\_061206.pdf](http://www.businessintelligence.bg/files/presentations/Intro_to_QlikView_061206.pdf) [February, 2011]

---

### Authors' Information

---



**Justyna Stasienko** – lecturer, *The Institute of Technical Engineering, The Bronisław Markiewicz Higher State School of Technology and Economics, Czarnieckiego Street 16, 37-500 Jarosław, Poland*; e-mail: [justyna.stasienko@pwste.edu.pl](mailto:justyna.stasienko@pwste.edu.pl)

*Major Fields of Scientific Research: Management Information Systems, Business information technology*

---

## Intelligent Applications: Medical and Diagnostic System

---

---

### PERFORMANCE OF COMPUTER-AIDED DIAGNOSIS TECHNIQUES IN INTERPRETATION OF BREAST LESION DATA

Anatoli Nachev, Mairead Hogan, Borislav Stoyanov

**Abstract:** *This study explores and compares predictive abilities of six types of neural networks used as tools for computer-aided breast cancer diagnosis, namely, multilayer perceptron, cascade-correlation neural network, and four ART-based neural networks. Our experimental dataset consists of 803 patterns of 39 BI-RADS, mammographic, sonographic, and other descriptors. Using such a combination of features is not traditional in the field and we find it is better than traditional ones. The study also focuses on exploring how various feature selection techniques influence predictive abilities of the models. We found that certain feature subsets show themselves as top candidates for all the models, but each model performs differently with them. We estimated models performance by ROC analysis and metrics, such as max accuracy, area under the ROC curve, area under the convex hull, partial area under the ROC curve with sensitivity above 90%, and specificity at 98% sensitivity. We paid particular attention to the metrics with higher specificity as it reduces false positive predictions, which would allow decreasing unnecessary benign breast biopsies while minimizing the number of delayed breast cancer diagnoses. In order to validate our experiments we used 5-fold cross validation. In conclusion, our results show that among the neural networks considered here, best overall performer is the Default ARTMAP neural network.*

**Keywords:** *data mining, neural networks, heterogeneous data; breast cancer diagnosis, computer aided diagnosis.*

**ACM Classification Keywords:** *I.5.1- Computing Methodologies - Pattern Recognition – Models - Neural Nets*

---

#### Introduction

---

Breast cancer is one of the leading causes of death for women in many countries. Mammography is currently the most widely used screening method for early detection of the disease, but it has a low negative predictive value. Many investigators have found that more than 60% of masses referred for breast biopsy on the basis of mammographic findings are actually benign [Jemal et al., 2005], [Lacey et al., 2002]. One goal of the application of computer-aided diagnosis (CAD) to mammography is to reduce the false-positive rate. Avoiding benign biopsies spares women unnecessary discomfort, anxiety, and expense. The problem is nontrivial and difficult to solve. Breast cancer diagnosis is a typical machine learning problem. It has been dealt with using various data mining techniques and tools such as linear discriminant analysis (LDA), logistic regression analysis (LRA), multilayer perceptions (MLP), support vector machines (SVM), etc. [Chen et al., 2009], [Jesneck et al., 2006].

Current CAD implementations tend to use only one information source, usually mammographic data in the form of data descriptors defined by the Breast Imaging Reporting and Data System (BI-RADS) lexicon [BI-RADS, 2003]. Recently, Jesneck et al. [2007] proposed a novel combination of BI-RADS mammographic and sonographic descriptors and some suggested by Stavros et al. [1995], which in combination with MLP show promising results. The MLP have been largely applied in the data mining tasks, but one of their major drawbacks is unclear optimal architecture, which includes number of hidden nodes, activation functions, and training algorithm to learn to predict. Another major problem is to specify optimal set of descriptors used for data mining, which effectively reduces the training and testing datasets to a dimensionality which provides best performance for the application domain. Our study was motivated by addressing those problems and particularly focusing on how reduction of dimensionality of that new combination of descriptors affects performance of not only MLPs, but also other neural network models, such as cascade-correlation nets and those based on the adaptive resonance theory (ART), introduced by Grossberg [1976].

The paper is organized as follows: Section 2 provides a brief overview of the neural networks used in this study: multilayer perceptron (MLP), cascade-correlation neural networks, fuzzy ARTMAP, distributed ARTMAP, default ARTMAP, and ic ARTMAP; Section 3 introduces the dataset and its preprocessing; Section 4 presents and discusses results from experiments; and Section 5 gives the conclusions.

---

## Neural Networks for Data Mining

---

A variety of neural network models are used by practitioners and researchers for clustering and classification, ranging from very general architectures applicable to most of the learning problems, to highly specialized networks that address specific problems. Each model has a specific topology that determines the layout of the neurons (nodes) and a specific algorithm to train the network or to recall stored information.

### Multilayer Perceptions (MLP)

Among the neural network models, the most common is the multilayer perceptron, which has a feed-forward topology and error-backpropagation learning algorithm [Rumelhart & McClelland, 1986]. Typically, an MLP consists of a set of input nodes that constitute the input layer, an output layer, and one or more layers sandwiched between them, called hidden layers. Nodes between subsequent layers are fully connected by weighted connections so that each signal travelling along a link is multiplied by its weight  $w_{ij}$ . Hidden and output nodes receive an extra bias signal with value 1 and weight  $\theta$ . The input layer, being the first layer, has input nodes that distribute the inputs to nodes in the first hidden layer. Each hidden and output node computes its activation level by

$$s_i = \sum_j w_{ij} x_j + \theta \quad , \quad (1)$$

and then transform it to output by an activation function. The MLP we use in this study has one hidden layer with two hidden nodes and log-sigmoid activation function

$$O_i(s_i) = \frac{1}{1 + e^{-\beta s_i}} \quad , \quad (2)$$

We trained the MLP by adaptive learning rate algorithm developed by Jacob [1988], also called delta-bar-delta, or TurboProp. The adaptive learning rate method proposes more flexibility and a higher speed of convergence, compared to the classic backpropagation algorithm.

### Cascade-Correlation Neural Networks (CCNN)

CCNN [Fahlman & Libiere, 1990] are supervised self-organizing networks with structure similar to backpropagation networks. Instead of adjusting the weights in a network of fixed topology, a CCNN begins with a minimal number of nodes, then automatically trains and adds new hidden nodes one by one and do not change them over the time. It creates a multi-layer structure called a 'cascade' because the output from all input and hidden nodes already in the network feed into new nodes.

A CCNN has three layers: input, hidden and output. Initially, the network begins with only input and output nodes. The output layer consists of a single neuron if the network is used for regression problems, or contains several neurons for classification problems, one per class label. The hidden layer is empty in the beginning – every input is connected to every output neuron by a connection with an adjustable weight. Such a simple cascade-correlation network has considerable predictive power and for a number of applications it provides excellent predictions. If not, however, the network adds new hidden nodes one by one as illustrated in Figure 1, until the residual error gets acceptably small or the user interrupts this process.

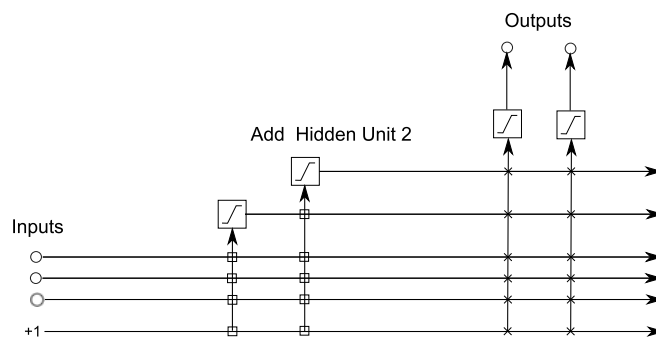


Fig.1. Cascade architecture after adding two hidden nodes (adapted from [Fahlman & Libier, 1991]). The vertical lines sum all incoming activation. Boxed connections are frozen, 'x' connections are trained repeatedly

The cascade-correlation architecture has several advantages over the traditional backpropagation neural nets. First, as the network is self-organizing and determines its own size and topology during the growth of the hidden layer during training, there is no need to decide how many layers and neurons to use in the network. This is a major problem of the backpropagation networks which implies use of predefined network architecture and designing a near optimal network architecture is a search problem that still remains open. Secondly, cascade-correlation nets learn very quickly (often 100 times as fast as a backpropagation network) and retain the structures they have built even if the training set changes. Finally, they have less chance to get trapped in local minima compared to the backpropagation nets.

**Fuzzy, Distributed, Default, and IC ARTMAP Neural Networks**

The adaptive resonance theory (ART) introduced by Grossberg [1975] led to the creation of a family of self-organizing neural networks, such as the unsupervised ART1, ART2, ART2-A, ART3, fuzzy ART, distributed ART and the supervised ARTMAP, instance counting ARTMAP, fuzzy ARTMAP (FAM), distributed ARTMAP, and default ARTMAP. ARTMAP is a family of neural network that consists of two unsupervised ART modules, *ARTa* and *ARTb*, and an *inter-ART* module called map-field as shown in Figure 2. An ART module has three layers of nodes: input layer *F0*, comparison layer *F1*, and recognition layer *F2*. A set of real-valued weights  $W_j$  is associated with the *F1-to-F2* layer connections between nodes. Each *F2* node represents a recognition category that learns a binary prototype vector  $w_j$ . The *F2* layer is connected through weighted associative links to a map field  $F^{ab}$ .

The ARTMAP learning can be described by the following algorithm [Carpenter et al., 1991]:

1. *Initialization*: All *F2* nodes are uncommitted, and all weight values and network parameters are initialized.

2. *Input pattern coding*: When a training pattern is presented to the network, a process called complement coding takes place. It transforms the pattern into a form suited to the network. A network parameter called vigilance parameter ( $\rho$ ) is set to its initial value. This parameter controls the network 'vigilance', that is, the level of details used by the system when it compares the input pattern with the memorized categories.

3. *Prototype selection*. The input pattern activates layer  $F1$  and propagates to layer  $F2$ , which produces a binary pattern of activity such that only the  $F2$  node with the greatest activation value remains active, that is, 'winner-takes-all'. If such a node does not exist, an uncommitted  $F2$  node becomes active and undergoes learning.

4. *Class prediction*. The class label  $t$  activates the  $F^{ab}$  layer in which the most active node yields the class prediction. If that node constitutes an incorrect class prediction, then another search among  $F2$  nodes in Step 3 takes place. This search continues until an uncommitted  $F2$  node becomes active (and learning directly ensues in Step 5), or a node that has previously learned the correct class prediction becomes active.

5. *Learning*. The neural network gradually updates its adaptive weights towards the presented training patterns until a convergence occur. The learning dynamic can be described by a system of ordinary differential equations.

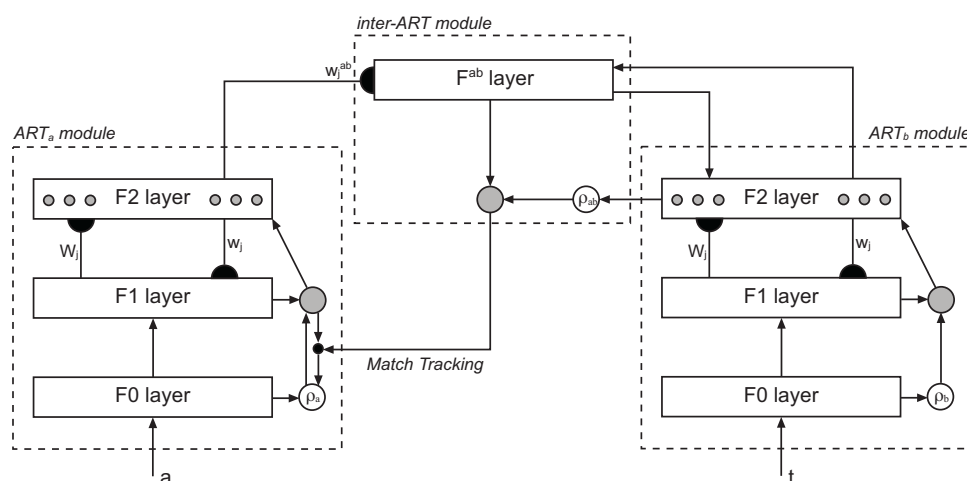


Fig. 2. Block diagram of an ARTMAP neural network adapted from [Carpenter et al, 1991]

**Fuzzy ARTMAP** was developed as a natural extension to the ARTMAP architecture. This is accomplished by using fuzzy ART modules instead of ART1, which in fact replaces the crisp (binary) logic embedded in the ART1 module with a fuzzy one. In fact, the intersection operator ( $\cap$ ) that describes the ART1 dynamics is replaced by the fuzzy AND operator ( $\wedge$ ) from the fuzzy set theory ( $(p \wedge q)_i \equiv \min(p_i, q_i)$ ) [Carpenter et al., 1992]. This allows the fuzzy ARTMAP to learn stable categories in response to either analog or binary patterns in contrast with the basic ARTMAP, which operates with binary patterns only.

ART1 modules in an ARTMAP net map the categories into  $F2$  nodes according to the winner-takes-all rule, as discussed above, but this way of functioning can cause category proliferation in a noisy input environment. An explanation of that is that the system adds more and more  $F2$  category nodes to meet the demands of predictive accuracy, believing that the noisy patterns are samples of new categories. To address this drawback, a new distributed ART module was introduced. If the ART1 module of the basic ARTMAP is replaced by a distributed ART module, the resulting network is called **Distributed ARTMAP** [Carpenter, 1997].

**Instance Counting (IC) ARTMAP** adds to the basic fuzzy ARTMAP system new capabilities designed to solve computational problems that frequently arise in prediction. One such problem is inconsistent cases, where identical input vectors correspond to cases with different outcomes. A small modification of the fuzzy ARTMAP



match-tracking search algorithm allows the IC ARTMAP to encode inconsistent cases and make distributed probability estimates during testing even when training employs fast learning [Carpenter & Markuzon, 1998].

A comparative analysis of the ARTMAP modifications, including Fuzzy ARTMAP, IC ARTMAP, and Distributed ARTMAP, has led to the identification of the **Default ARTMAP** network, which combines the winner-takes-all category node activation during training, distributed activation during testing, and a set of default network parameter values that define a ready-to-use, general-purpose neural network for supervised learning and recognition [Carpenter, 2003]. The Default ARTMAP features simplicity of design and robust performance in many application domains.

---

## Data and Preprocessing

---

Our tests used a dataset that contains data from physical examination of patients, including mammographic and sonographic examinations, family history of breast cancer, and personal history of breast malignancy, all collected from 2000 to 2005 at Duke University Medical Centre [Jesneck et al., 2007]. Samples included in the dataset are those selected for biopsy only if the lesions corresponded to solid masses on sonograms and if both mammographic and sonographic images taken before the biopsy were available for review. Data contain 803 samples, 296 of which are malignant and 507 benign. Out of 39 descriptors, 13 are mammographic BI-RADS, 13 sonographic BI-RADS, 6 sonographic suggested by Stavros et al. [1995], 4 sonographic mass descriptors, and 3 patient history features [BI-RADS, 2003], [Jesneck et al., 2007], [Nachev & Stoyanov, 2010]. There are also class label that indicates if a sample is malignant or benign.

We preprocessed the dataset in order to addresses the problem of large amplitude of variable values caused by their different nature and different units of measurements. Consistency we achieved by mapping all data values into the unit hypercube (i.e. all values between 0 and 1), using a linear transformation

$$x_i^{new} = \frac{x_i^{old} - \min_i}{\max_i - \min_i} \quad (3)$$

applied to each variable (data column) separately. This scaling down of values is essential requirement for certain types of neural networks, and particularly for the ARTMAP models we used in our study.

Another preprocessing step was feature selection. In many cases and application domains removing redundant features from the data can help to alleviate effect of curse of dimensionality, avoid overfitting, and speed up learning process. Exhaustive search approach among all possible subsets of features is not applicable in our case as the dataset has cardinality 39. Alternative approaches could be using subset selection algorithms or feature ranking techniques. The former one is preferable as it usually provides good results. We tested genetic search, best first search, subset size forward selection, race search, and scatter search. Our tests showed that the subset size forward selection, proposed by Guetlien et al. [2009] gives good results with all types neural networks we experimented with. This method output a set of 17 descriptors (s17): patient age, indication for sonography, mass margin, calcification number of particles, architectural distortion, anteroposterior diameter, mass shape, mass orientation, lesion boundary, special cases, mass shape, mass margin, thin echo pseudocapsule, mass echogenicity, edge shadow, cystic component, and mass margin. Two of these are general descriptors; three - mammographic BI-RADS; five - sonographic BI-RADS; four - Stavros'; and three - sonographic mass descriptors. The feature set is relatively balanced in representing different categories of data. In our experiments we also used a set of 14 descriptors (s14) proposed by Jesneck et al. [2007] and obtained by stepwise feature selection. We also used the original full set of 39 descriptors (s39).

## Empirical Results and Discussion

We used simulators of the neural network models explored here. Series of test showed that best architecture of the multilayer perceptron is one hidden layer with two nodes. We also used a cascade-correlation neural network with Turboprop2 learning based on the Fahlman's work [Fahlman & Libiere, 1990]. Each of the four types ARTMAP neural networks were tested with 41 vigilance parameter values from 0 to 1 and step of increment 0.025. In order to avoid bias in training due to the specific order of training samples, we applied 5-fold cross validation and summarized results in four categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Prediction accuracy was calculated by  $Acc=(TP+TN)/(TP+TN+FP+FN)$ . For the purposes of ROC analysis, we also calculated true positive rate  $TPR=TP/(TP+FN)$ , and false positive rate  $FPR=FP/(TP+FN)$ .

No doubts, accuracy is the most common performance estimator of a model, which is used in a vast amount of studies and applications, but in many cases and problem domains it is not sufficient, even can be misleading where important classes are underrepresented in datasets (i.e. class distribution is skewed), or if errors of type I and type II can produce different consequences and have different cost. Secondly, the accuracy depends on the classifier's operating threshold, such as threshold values of MLP or vigilance parameter of ARTMAP NN, and choosing optimal threshold can be challenging. The deficiencies of accuracy can be addressed by the Receiver Operating Characteristics (ROC) analysis [Fawcett, 2006], which plots curves between two indices: TPR and FPR.

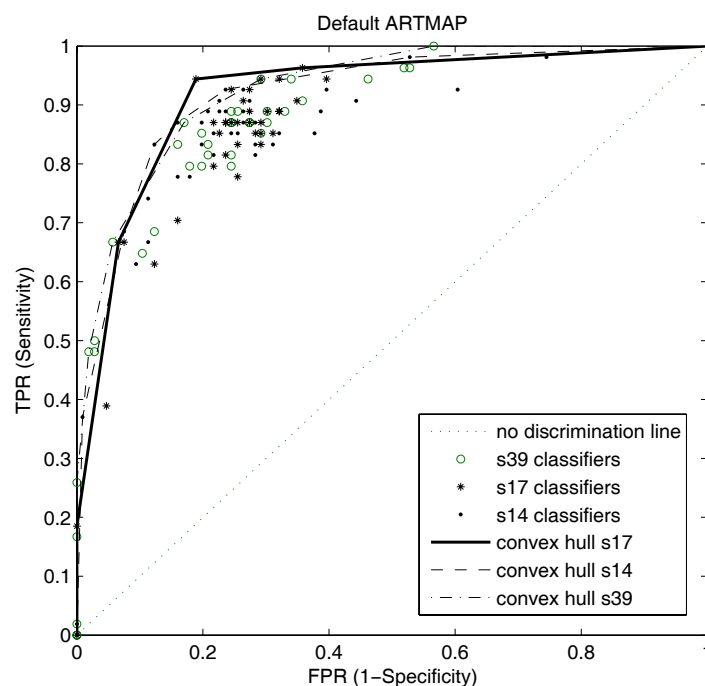


Fig. 3. ROC analysis of Default ARTMAP neural network tested with three sets of descriptors: all available (s39), a selection proposed by Jesneck et al. [2007] (s14), and a set of 17 descriptors (s17) proposed by authors. The model was tested with 41 vigilance parameter values from 0 to 1 with step of increment 0.025

ROC curves are step functions which can be used to select the optimal decision threshold by maximizing any pre-selected measure of efficacy. In general, the best possible prediction method would yield a point in the upper left corner or coordinates (0, 1) of the ROC space, representing 100% sensitivity (all true positives are found) and 100% specificity (no false positives are found). A completely random guess would give a point along a diagonal

line, also known as line of no-discrimination, which is from the left bottom to the top right corners. The optimal classifier would be represented by the most 'northwest' point on the curve, which is the most distant one from the no-discrimination line. Plotting discrete classifiers, such as ARTMAP, require additional processing. Since they do not use operational threshold to be varied in order to generate the ROC curve, the vigilance parameter variance plays the same role. Each parameter value plots one point on the ROC space, and the curve that connects the most northwest points along with the two trivial classifiers (0,0) and (1,1), called also ROC Convex Hull (ROCCH) represents the model as a whole. The ROCCH lines that link points of 'real' classifiers define a continuum of possible classifiers that can be obtained by linear combinations of the plotted ones. ROC analysis also provides additional metrics for estimation of model performance, such as Area Under the ROC curve (AUC) and partial Area Under the ROC curve (pAUC) where sensitivity is above certain value (p). The bigger the AUC / pAUC, the better the model is.

*Table 1. Performance of MLP, CCNN, Fuzzy ARTMAP, Distributed ARTMAP, Default ARTMAP, and IC ARTMAP. Metrics for comparison include: area under the ROC curve (AUC), partial AUC at sensitivity above 90% ( $0.90AUC$ ), specificity at 98% sensitivity, and maximal accuracy ( $ACC_{max}$ ). Models have been tested with three variable selections: s39, s17, and s14. Typical radiologist assessment values are also included.*

<b>MLP</b>	s39	s17	s14	Radiologist	<b>CCNN</b>	s39	s17	s14	Radiologist
AUC	0.89	0.91	0.86	0.92	AUC	0.896	0.911	0.907	0.92
$0.90AUC$	0.62	0.68	0.55	0.52	$0.90AUC$	0.648	0.68	0.731	0.52
Spec /98% sens	0.37	0.49	0.27	0.52	Spec /98% sens	0.427	0.5	0.49	0.52
$ACC_{max}$	0.89	0.91	0.87	n/a	$ACC_{max}$	0.828	0.848	0.838	n/a
<b>Fuzzy ARTMAP</b>	s39	s17	s14	Radiologist	<b>Distributed ARTMAP</b>	s39	s17	s14	Radiologist
AUC	0.851	0.815	0.838	0.92	AUC	0.786	0.819	0.744	0.92
$0.90AUC$	0.586	0.393	0.413	0.52	$0.90AUC$	0.226	0.272	0.187	0.52
Spec /98% sens	0.099	0.082	0.091	0.52	Spec /98% sens	0.047	0.057	0.039	0.52
$ACC_{max}$	0.838	0.813	0.819	n/a	$ACC_{max}$	0.831	0.856	0.819	n/a
<b>Default ARTMAP</b>	s39	s17	s14	Radiologist	<b>IC ARTMAP</b>	s39	s17	s14	Radiologist
AUC	0.927	0.931	0.918	0.92	AUC	0.776	0.821	0.860	0.92
$0.90AUC$	0.725	0.809	0.778	0.52	$0.90AUC$	0.215	0.282	0.384	0.52
Spec /98% sens	0.686	0.649	0.690	0.52	Spec /98% sens	0.045	0.059	0.081	0.52
$ACC_{max}$	0.85	0.856	0.863	n/a	$ACC_{max}$	0.831	0.850	0.856	n/a

As long as AUC provides an overall estimation of the model, the partial area is more relevant to the domain of computer-aided diagnosis, and particularly where  $p=0.9$ . Another clinically relevant metric used in the application domain is sensitivity at a very high level of specificity (98%).

Table 1 summarizes results from numerous experiments where networks were trained and tested with three different sets of descriptors: s39 that contains the original 39 variables; a selection of 14 variables (s14) proposed

by Jesneck et al. [2007] as a result from stepwise feature selection technique, and our set of 17 variables (s17) we obtained by using the subset size forward selection of Guetlien et al. [2009].

Figure 4 illustrates the results. We obtained highest prediction accuracy of 91.1% by using CCNN with s17. The Default ARTMAP outperforms all other models in terms of overall performance measured by AUC. Here again, s17 is best performer. The figure also illustrates, that Default ARTMAP is the only model (among the studied here) that outperforms the average radiologist performance [Jesneck et al. 2007] in terms of AUC, but more important from a clinical viewpoint are the metrics pAUC and sensitivity at very high specificity. Figures show that again Default ARTMAP beats the others with s17 in terms of pAUC, and again is best performer in terms of sensitivity at high specificity, no matter which subset of descriptors is used.

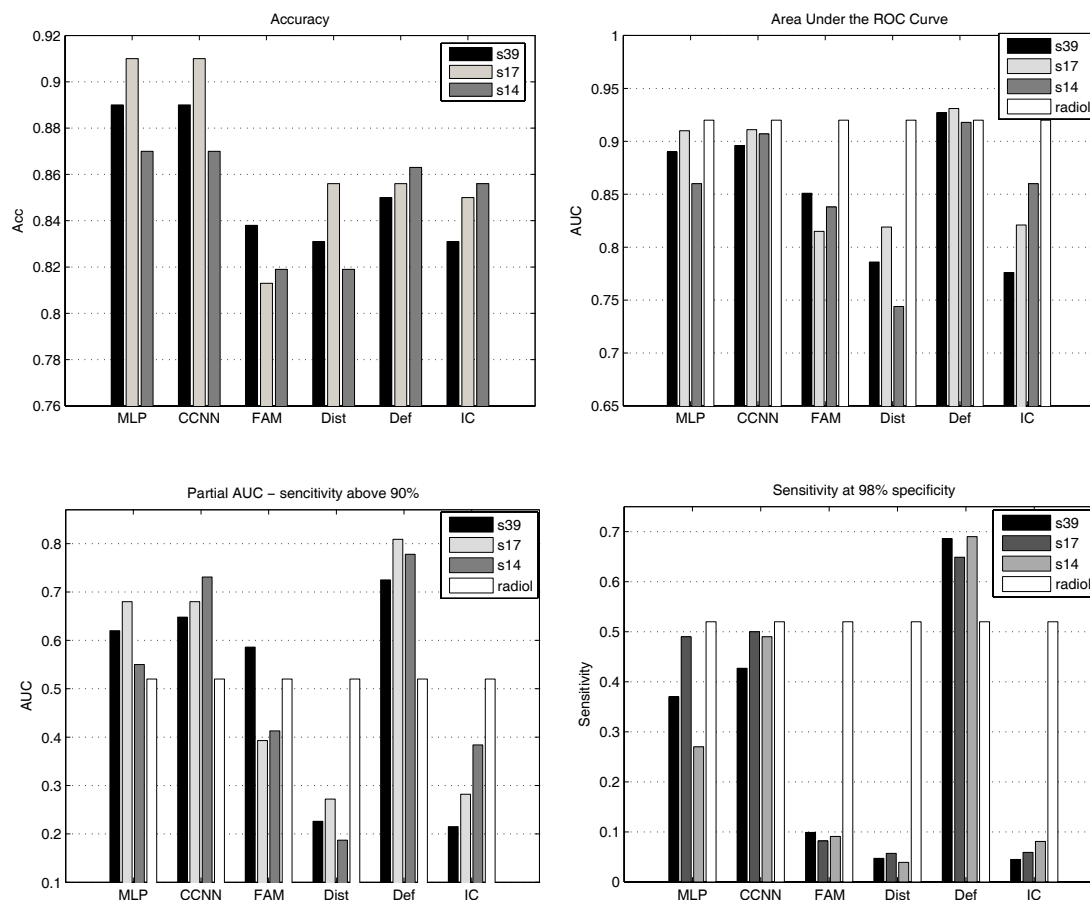


Fig. 4. Performance of three-layer perceptron (MLP), cascade-correlation NN (CCNN), Fuzzy ARTMAP (FAM), Distributer ARTMAP (Dist), Default ARTMAP (Def), and IC ATRMAP (IC), measured by max accuracy (Acc), area under the ROC curve (AUC), partial area under the ROC curve where sensitivity is above 90%, and sensitivity at 98% specificity. Models were tested with all available descriptors (s39), a selection proposed by Jesneck et al. [2007] (s14), and a new proposed set of 17 descriptors (s17). Results were also compared with a typical radiologist performance (radiol)

Figure 3 gives further details on the Default ARTMAP neural network performance obtained by ROC analysis. Bold line represents the best descriptor set. Finally, we find that Default ARTMAP is appropriate for solving the task as its clinically relevant characteristics are good, however a limitation of the model is that it requires a very careful tuning.

---

## Conclusion

---

Many CAD systems for breast cancer screening improve lesion detection sensitivity, but improving specificity is still challenging. This study explores and compares predictive abilities of six types of neural networks: MLP, CCNN, Fuzzy, Distributed, Default, and IC ARTMAP by using a recently proposed combination of BI-RADS mammographic, sonographic, and other descriptors. We also focused our study on how various feature selection techniques influence predictive abilities of those models and found that a subset obtained by subset size forward selection provides best overall results. Our performance estimations were based on ROC analysis and metrics, such as max accuracy, area under the ROC curve and convex hull. We paid particular attention on clinically relevant metrics, such as partial area under the ROC curve with sensitivity above 90%, and specificity at 98% sensitivity, as a higher specificity reduces false positive predictions, which would allow decreasing unnecessary benign breast biopsies while minimizing the number of delayed breast cancer diagnoses. In conclusion, our results show that among all neural networks we explored for this application domain, highest prediction accuracy of 91.1% can be obtained by cascade-correlation neural network, but Default ARTMAP outperforms all other models in terms of overall performance and clinically relevant metrics. All the results confirm that the set of descriptors we propose outperforms the ones used in previous studies.

---

## Bibliography

---

- [BI-RADS, 2003] American College of Radiology. BI-RADS: ultrasound, 1st ed. In: Breast imaging reporting and data system: BI-RADS atlas, 4th ed. Reston, VA: American College of Radiology, 2003
- [Carpenter et al, 1991] Carpenter, G., Grossberg, S., & Reynolds, J.: ARTMAP: Supervised Real-Time Learning and Classification of Non-stationary Data by a Self-Organizing Neural Network. *Neural Networks*, vol. 6, pp. 565-588 (1991)
- [Carpenter et al., 1992] Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H. and Rosen, D.B. "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps", *IEEE Transaction on Neural Networks*, 3(5), pp. 698-713, 1992.
- [Carpenter, 1997] Carpenter, G.. Distributed Learning, Recognition, and Prediction by ART and ARTMAP Neural Networks. *Neural Networks*, vol. 10:8, 1473-1494, (1997)
- [Carpenter & Markuzon, 1998] Carpenter, G. & Markuzon, N.. ARTMAP-IC and Medical Diagnosis: Instance Counting and Inconsistent Cases. *Neural Networks*, vol. 11:2, pp. 323-336 (1998)
- [Carpenter, 2003] Carpenter, G.: Default ARTMAP. *Proceedings of the International Joint Conference on Neural Networks (IJCNN'03)*, Portland, Oregon, pp. 1396-1401, (2003)
- [Grossberg, 1976] S. Grossberg, "Adaptive pattern classification and universal recoding. II: Feedback, expectation, olfaction, and illusions." *Biological Cybernetics*, 23, 1976.
- [Fahlman & Lebiere, 1990] Fahlman, S. and Lebiere C. "The Cascade-Correlation Learning Architecture" in D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2*, Morgan Kaufmann, 1990.
- [Fawcett, 2006] Fawcett, T. "An introduction to ROC analysis"; *Pattern Recognition Letters*, Vol. 27 Issue 8, pp. 861-874, 2006.
- [Guettlin et al., 2009] Guettlin, M., Frank, E., Hall, M., Karwath, A. "Large Scale Attribute Selection Using Wrappers"; In *Proc. IEEE Symposium on CIDM*, pp.332-339, 2009.
- [Jacob, 1988] Jacob, R. "Increased Rates of Convergence Through Learning Rate Adaptation"; *Neural Networks*, Vol. 1 Issue 4, pp. 295-307, 1988.
- [Jemal et al., 2005] Jemal, A., Murray, T, Ward, E., Samuels, A., Tiwari, R., Ghafoor, A., Feuer, E., Thun, M.. "Cancer statistics", *Ca-Cancer J. Clin* 2005;55:10-30, 2005.
- [Jesneck et al., 2006] Jesneck, J., Nolte, L., Baker, J., Floyd, C., Lo, J. "Optimized Approach to Decision Fusion of Heterogeneous Data for Breast Cancer Diagnosis."; *Medical Physics*, Vol. 33, Issue 8, pp.2945-2954, 2006

- 
- [Jesneck et al., 2007] Jesneck, J., Lo, J., Baker, J. "Breast Mass Lesions: Computer-Aided Diagnosis Models with Mamographic and Sonographic Descriptors"; *Radiology*, vol.244, Issue 2, pp 390-398, 2007.
- [Lacey et al., 2002] Lacey, J., Devesa, S., Brinton, L. "Recent Trends in Breast Cancer Incidence and Mortality."; *Environmental and Molecular Mutagenesis*, Vol. 39, pp. 82–88, 2002.
- [Nachev & Stoyanov, 2010] Nachev, A. and Stoyanov, B., "An Approach to Computer Aided Diagnosis by Multi-Layer Preceptrons", In *Proceedings of International Conference Artificial Intelligence (IC-AI'10)*, Las Vegas, 2010.
- [Rumelhart & McClelland, 1986] Rumelhart, D., McClelland, J. "Parallel Distributed Processing"; *Explorations in the Microstructure of Cognition*, Cambridge, MA: MIT Press, 1986.
- [Stavros et al., 1995] Stavros, A., Thickman, D., Rapp, C., Dennis, M., Parker, S., Sisney, G. "Solid Breast Modules: Use of Sonography to Distinguish between Benign and Malignant Lesions"; *Radiology*, Vol. 196, pp. 123-134, 1995.
- 

### Authors' Information

---



**Anatoli Nachev** – *Business Information Systems, Cairnes Business School, National University of Ireland, Galway, Ireland; e-mail: [anatoli.nachev@nuigalway.ie](mailto:anatoli.nachev@nuigalway.ie)*

*Major Fields of Scientific Research: data mining, neural networks, support vector machines, adaptive resonance theory.*

**Mairead Hogan** – *Business Information Systems, Cairnes Business School, National University of Ireland, Galway, Ireland; e-mail: [mairead.hogan@nuigalway.ie](mailto:mairead.hogan@nuigalway.ie)*

*Major Fields of Scientific Research: HCI, usability and accessibility in information systems, data mining.*



**Borislav Stoyanov** – *Department of Computer Science, Shumen University, Shumen, Bulgaria; e-mail: [bpstoyanov@abv.bg](mailto:bpstoyanov@abv.bg)*

*Major Fields of Scientific Research: artificial intelligence, cryptography, data mining.*

---

## Mechanical Engineering

---

### DESCRIPTION OF SURFACES HAVING STRATIFIED FUNCTIONAL PROPERTIES

**Wiesław Graboń**

**Abstract:** *The article presents the characteristic of surfaces having functional properties. It discusses the parameters used to describe the roughness of this type of surfaces. It proposed the method of constructing software which calculates probability parameters.*

**Keywords:** *two-process surfaces, roughness parameters.*

**ACM Classification Keywords:** *Algorithm, Measurement.*

---

#### Introduction

---

During friction in the presence of a lubricant too smooth surfaces hold lubricating oil poorly (it can lead to erosion of a coupling), whereas too rough surfaces are worn away intensively. The opposing influence of small and large heights of surface roughness of a cylinder liner on the functional properties of the piston-rings-cylinder assembly caused that the researches began to conduct studies in order to find surfaces combining sliding properties of smooth surfaces with the oil storage capacity which is typical of porous surfaces. Thanks to those works, in the 1980s structures of cylinder liner surface achieved after two processes (plateau honing surfaces) came into being. They should be similar to the geometrical structure of cylinder liner surfaces which is created during running-in period, then the time of running-in process and wear should be smaller. The example of these kinds of surfaces is the surface of a cylinder liner after plateau honing. The basic tasks of this surface are: to ensure leak-tightness as well as to provide the piston-rings-cylinder assembly with optimal greasing of gear. The most difficult functioning conditions among all of tribological systems of an internal combustion engine are precisely in piston-rings-cylinder assembly [Niewiarowski, 1983]. In this system the particularly difficult conditions are found in the area of the top dead centre position of the first piston ring, where the thickness of an oil film between a packing ring and smooth surface of a cylinder comprises 0-3  $\mu\text{m}$ . The coefficient of friction between the packing ring and a cylinder comes to 0.1-0.15. The piston-rings-cylinder assembly should assure mileage up to 500000 km with reference to personal cars and 1500000 km to trucks. Between the smooth surface of a cylinder and the surface of piston rings there are various greasing conditions from the boundary lubrication up to hydrodynamic lubrication in the middle of the piston's distance line [Shin, 1983], [Sudarshan, 1983]. To the dominant types of wearing-out of cylinders in an internal combustion engine the researches include: abrasive wear, corrosive wear, adhesive wear and from time to time fatigue wear.

According to the author [Kozaczewski,1986] of the research articles, the geometrical structure of cylinders surface influences engine properties, mainly in the initial stage of its functioning (the period of running-in). It is considered that the rough surface of cylinders causes little tendency to erosion, whereas smooth surface ensures their little wear in the period of running-in. At first, the researches stated that little wear in the period of running-in

was found in cylinders characterised by great smoothness. As a result, this kind of cylinders was recommended. However, as progression in engine construction was marked (most of all in the load growth), erosions of cylinders' smooth surface happened. The author of publication [Wiemann, 1971] claimed that the bigger height of roughness of cylindrical liner which does not have additional surface treatment is, the greater its erosion resistance. It appeared that considerable increase of roughness height is also unfavourable because it causes acceleration of chromium plated piston rings wear. The researches [Sreenath, 1976] proved that above the optimal roughness height of  $R_a=0.8 \mu\text{m}$  parameter, the linear wear of cylinder rises. Duck [Duck, 1974] determined the advisable value range of parameter  $R_t$  for spark-ignition engines:  $2\text{-}5\mu\text{m}$ , whereas for diesel engine  $4.7 \mu\text{m}$ . The authors of the review work [Day, 1986] came to the similar conclusions. The advisable greater roughness in diesel engine results probably from the greater loads in this kind of engine. The difference in functioning conditions of various types of engine is the cause of discrepancy with regard to honing cross-hatch angle  $\alpha$  (fig.1), usually smaller in spark-ignition engines in comparison to diesel ones.



*Fig.1 Schematic diagram of a cylinder after honing (a) [Zwierzycki, 1990] and a photograph of the surface of a cylinder liner (b)*

The smaller honing cross-hatch angle affects the decrease in consumption of oil which is particularly aimed at in spark ignition engines, but in the case of compression-ignition engines more important issue is to eliminate the galling. Decreasing of the honing cross-hatch angle causes oil film thickness reduction which leads to greater loss of energy and increasing of wear.

The authors of the publications [Pawlus, 1994 ], [Willis, 1986] confirmed that plateau honing provides lesser linear wear during running-in period and the time of this process might be shortened in comparison to one process honing. In his work [Campbell, 1972], Campbell affirmed that the achievement of linear wear corresponding to 30% of bearing ratio requires two times less of volumetric consumption in the case of plateau honing cylinders in comparison to the similar surface of the same roughness height after one process honing. The usage of plateau honing caused significant reduction of running-in time [Willis, 1986]. Santochi and Vignale [Santochi, 1982] employed plateau honing with reference to air cooled motorcycle engines. They reached geometrical structure of surface characterized by  $R_a=1\mu\text{m}$ ,  $R_z=12 \mu\text{m}$ ,  $R_v/R_p=2$  parameters, however traditional structure was characterised by the following parameters:  $R_a=2.4\mu\text{m}$ ,  $R_z=18 \mu\text{m}$ ,  $R_v/R_p=1.1$ . One obtained faster stabilization as well as improvement of functional parameters of engine after running-in by replacing the traditional honing with the plateau honing (fig. 2). Dolecki et al. in their research [Dolecki, 1983] proved smaller oil consumption (at about 20%) by engine in the case of equipping it with plateau honing cylinders. The research conducted on Polonez engine shows that the plateau honing has a positive influence on the linear wear of cylinders during running-in. It allowed to conclude that the linear wear in the period of running-in is proportional to roughness height as well as to emptiness coefficient  $R_p/R_t$  [Pawlus, 1999]. The evidence for the supremacy of plateau honing over one-process honing are also works [Essig, 1990], [Barber, 1987]. One of a few comprehensive works concerning the influence of two processes surface on frictional tribological properties is Jeng's work [Jeng, 1996]. He highlighted



that the research described in literature was not able to grasp only the influence of microgeometry of surface on the wear, therefore he was not sure about the need of using additional process of honing.

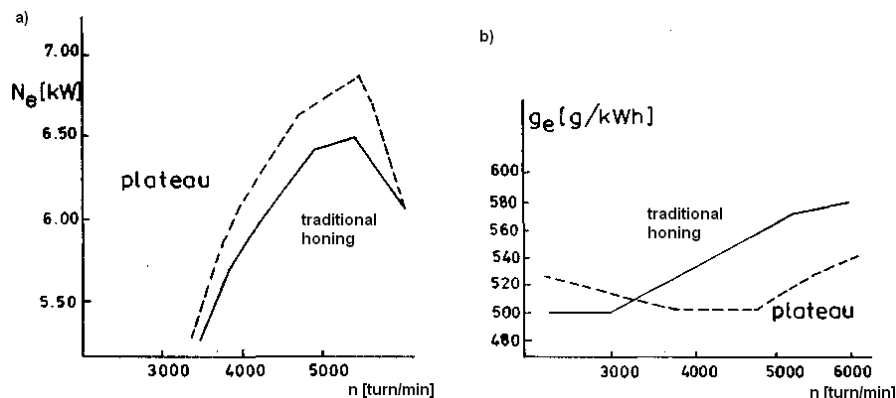


Fig.2 The influence of geometrical structure of surface on power output  $N_e$  (a) and individual fuel consumption  $g_e$  (b) [Santochi, 1982]

He studied the one process honing surfaces and plateau honing surfaces, both kinds of the surfaces had the same values of the  $R_q$  parameter. On the basis of experiments carried out on the special simulation stand he found that the two processes surfaces have a shorter running-in period and less galling resistance in relation to the one process surfaces (the resistance may be increased by the use of oil additives).

The initial wear of these surfaces during the running-in period was greater, however they quickly attained a constant intensity of wear. Therefore, during the work of piston-rings-cylinder assembly they ensure less wear in comparison with surfaces after one process honing. He also performed friction coefficient test and found that in the course of the work in a homogenous hydrodynamic lubrication conditions, the coefficient of friction two-process surfaces is the same as the one-process surfaces. However, the two-process structures of surface are more advantageous from the point of view of the influence on the mixed friction coefficient.

Nosal considers [Nosal, 1998] that the increase in resistance of galling of the plateau honing structure is caused by the increase of oil surface capacity (which causes lubricating layer thickness increase, the reduction of friction resistance and temperature in the contact zone), and frequent interruptions of two surfaces being in contact caused by valleys (which reduces the probability of a galling centre).

The authors of the publications [Sudarshan, 1983] paid attention to the possibility of accumulation of abrasive particles in valleys created during honing process. It should lead to reduction of intensity of abrasive wear. The considerable deterioration in lubrication conditions since the disappearance of valleys was observed by the authors of the publications [Stout, 1990]. This kind of situation causes possibility of intensification of the abrasive wear or generation of galling danger. The specific danger appears in the high loaded diesel engines, in this case on the cylinder liner a very smooth texture comes into existence, it resembles bore polishing surface of a significant tendency to galling. This is the case when together with the increase of smoothness friction increases as well. The authors of the publications [Michalski, 1994] studied the influence of roughness of cylinders liner surfaces on the value of abrasive wear which significantly exceeds the initial roughness height. As a result of the conducted studies, they claimed that the abrasive wear of the cylinder liner is proportional to the distance between the honing valley and to the value of the emptiness coefficient  $R_p/R_t$ . Excessive increase of height roughness causes oil consumption increase in the engine [Kozaczewski, 1986]. Figure 3 shows an example of the impact of roughness height of the cylinder liner on the oil consumption.

During recent years there have been a few changes in requirements relating to cylinder liner roughness: from very smooth, mirror-surface ( $R_a = 0.20$  mm), by changing the roughness to higher one ( $R_a$  in range 0.8-1.2 mm) caused by the increase in power of an engine, up to plateau structure which is now almost universally applicable.

The increased roughness diminishes the ability to galling, however, it causes the increased consumption of oil and increased toxicity of exhaust gases.

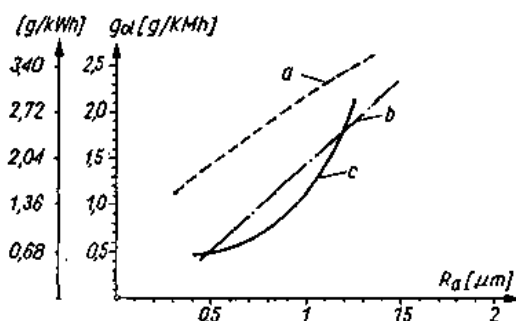


Fig. 3. Influence of cylinder liner roughness on the oil consumption in internal combustion engine [Kozaczewski, 1986]: a) compression-ignition engine,  $V_s$  (engine cubic capacity) = 6 dm<sup>3</sup>, b) spark ignition  $V_s$  = 1.3 dm<sup>3</sup>, c) compression-ignition,  $V_s$  = 4 dm<sup>3</sup>

We can also find the opposite statement. For example, Gruszka [Gruszka, 1983] in his doctoral thesis maintains that too low roughness increases the consumption of oil.

During the construction process of new engines manufacturers of trucks must take into consideration very stringent requirements which are constituted by the European Union standards included in the series EU III/IV/V relating particularly to toxicity of exhaust gases. Therefore, they strive to reduce the roughness of cylinder liner, which leads to reduction of oil film thickness and consequently to smaller oil consumption and exhaust gases emissions. Currently, there are the following ways of honing development of cylinder liner surfaces:

- glide or slide honing,
- honing using laser beams to cracks cutting,
- manufacturing of lubricating pockets on the smooth honing surface.

There are two types of glide honing due to the honing cross-hatch angle:

- measuring 60° angle,
- measuring 140° angle, so called – spiral [Cieślak, 2008].

In currently manufactured (according to plateau standard) cylinder liners, cylinder bearing surface is described by the following roughness parameters:  $R_{pk} < 0.3 \mu\text{m}$ ,  $R_k = 0.8\text{-}1.4 \mu\text{m}$ ,  $R_{vk} = 1.7\text{-}3.2 \mu\text{m}$ . In order to decrease the thickness of oil film and reduce oil consumption as well as exhaust gases emissions it has been proposed to reduce the roughness to the following parameter levels:  $R_{pk} < 0.2 \mu\text{m}$ ,  $R_k = 0.2\text{-}0.5 \mu\text{m}$ ,  $R_{vk} = 1.4$  to  $3.0 \mu\text{m}$ . New requirements focus on even greater reduction of parameter  $R_k$  value.

The researchers from the Volvo and the University of Halmstad (Sweden) followed the program Piston Simulation for the analysis of impact of the geometrical structure of cylinder liner surface on the oil film thickness and friction force. It was found that the increase of the value of the  $R_k$  parameter has an impact on increasing both analysed values in a top dead centre position of the piston [Johansson, 2008].

The authors of the article [Ohlsson, 2003] explored the correlation between roughness parameters of a cylinder liner and consumption of oil by the engine. Oil consumption is proportional to the value of many parameters of geometrical structure of cylinders liner surface, but only  $R_q$ ,  $R_{vk}$  and  $R_k$  parameters in 2D and 3D correlation coefficients are included within the limits of 0.9-1.

The Authors of the articles [Hassis, 1999], [Schmid, 2006], [Schmid, 1999] from Nagel company think that producers should aim at minimizing the value of the parameters  $R_k$  and  $R_{pk}$  in order to reduce oil consumption.

They presented the results of the studies in accordance with which the change of the classical plateau honing ( $R_{pk} < 0.2 \mu\text{m}$ ,  $R_{vk} = 1.4$  to  $2.0 \mu\text{m}$ ,  $R_k = 0.6$ - $0.8 \mu\text{m}$ ) to the glide honing ( $R_{pk} < 0.1 \mu\text{m}$ ,  $R_{vk} = 0.8$ - $1.2 \mu\text{m}$ ,  $R_k < \frac{1}{4} R_{vk}$ ) leads to the reduction in consumption of oil by the engine (over 60%). The reduction of oil consumption is even more affected by the application of spiral honing.

It is essential to modify the honing cross-hatch angle by the application of spiral honing. Application of angle less than 90 degrees counteracted too easy blow-by of oil into the combustion chamber. However, the increase of the honing cross-hatch angle (fig. 1) leads to the decrease of abrasive wear of cylinder liner. The applied honing cross-hatch angle less than 180 degrees allows rotations of the piston rings. The laser honing process is a perspective technology, since it allows to reach the geometrical surface structure with the guarantee of small oil consumption and adequate lubrication surface of cylinder liner at the most loaded place - in piston's top dead centre position. The remaining surface of cylinder liner can have small height of roughness. For example, Klink [Klink, 1997] received the cracks with a width of 40-80  $\mu\text{m}$ , depth 5-25  $\mu\text{m}$ , and the mutual distance of 300  $\mu\text{m}$ . Figure 4 shows examples of cylinder liner surfaces made by different methods.

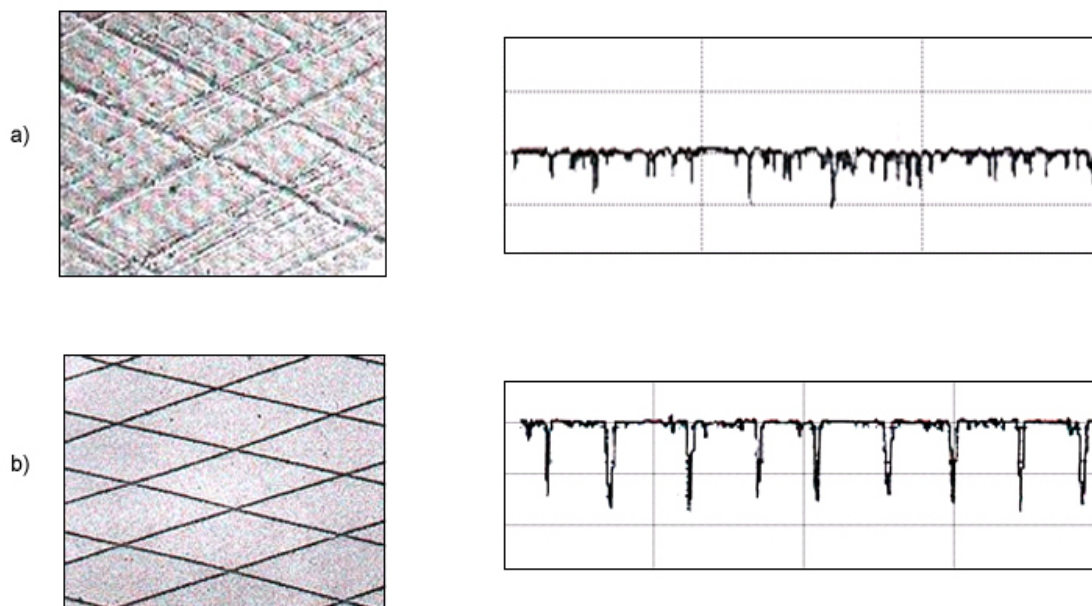


Fig.4. The surface of cylinder liner after glide honing (a) and laser honing (b) [Cieślak, 2008]

As a result of literature analysis it can be concluded that the topography of the cylinder liners' surfaces formed by honing process has a significant impact on the operating parameters of the combustion engines, particularly in the early period of their work.

### Description of two-process surfaces

In accordance with the information contained in the works of [Chusu, 1975], [Nowicki, 1991] the profiles of cylinder liner surfaces after plateau honing are irregular, they usually have random character. When on this kind of profiles periodical irregularity appears then we can qualify them to the mixed profiles. Periodic irregularities depending on oscillation and kinematics of machining process may be a consequence of the earlier boring process of cylinders. Due to the explicit directionality of cylinders' patterns the authors [Wieczorowski, 1996] rated the cylinders' surfaces after honing to the mixed surfaces. According to the author of the work [Michalski, 1998] waviness of surface of honing cylinders has the random characteristics.

In connection with an important impact of the geometrical structure of surface of cylinder liner on the combustion engines exploitation properties, their producers have high requirements concerning the roughness parameters of plateau honing surfaces.

These requirements are mainly related to the proper determination of material ratio curve (so called bearing curve or Abbot Firestone curve) of roughness profile, the profile height and the distance between the deep valleys. The requirements of Berliett company from the 1970s (see fig. 5.) are typical example.

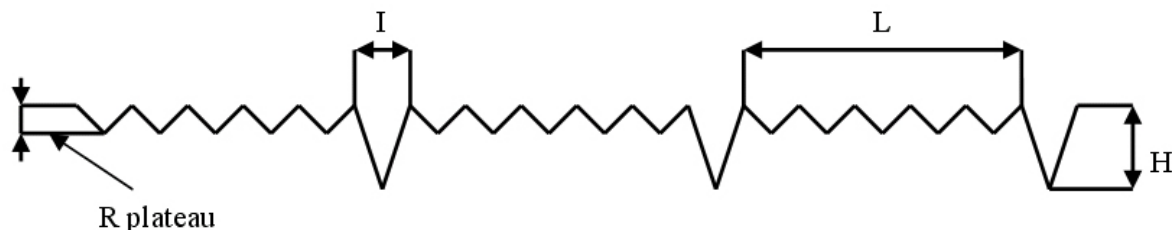


Fig 5. Roughness parameters describing the cylinder liner surface profile [Cieślak, 2008]

Parameters should take the following values:

- average roughness depth:  $2.5 \mu\text{m} \leq R \leq 6.5 \mu\text{m}$
- "plateau" roughness depth:  $1 \mu\text{m}_{\text{plateau}} \leq R \leq 3 \mu\text{m}$
- depth of valleys:  $6 \mu\text{m} \leq H \leq 16 \mu\text{m}$
- width of the plateau:  $125 \mu\text{m} \leq L \leq 600 \mu\text{m}$
- width of valleys:  $10 \mu\text{m} \leq I \leq 65 \mu\text{m}$

Important parameters are the horizontal ones, it is demonstrated by the fact that the average distance between valleys and their dimensions are often included in the requirements of engine manufacturers.

The method of determining the width of the valleys previously required by the GOETZE company was subjective. More objective manner, similar to the method described in the article [Michalski, 1994], is shown in the work [Lenhof, 1997]. It is based on the number of intersections of profile with the line described in DIN 4776 standard. Other parameters, used by researchers because of their statistical significance are the statistical moments of the third and fourth order:  $R_{sk}$  (skewness or asymmetry) and  $R_{ku}$  (eksces or kurtosis). Willn [Willn, 1972] said, however, that these parameters are correlated with each other, therefore, he proposed additional parameters based on analysis of distribution of the number of peaks or distribution of cross sections of the profile lines parallel to the geometric mean line.

Many researchers associated with automotive companies use bearing area curve to cylinder liner surface analysis. In this curve we can distinguish 3 basic parts: peak, central and valley area, responsible for the different properties of surface. Abbott and Firestone [Abbott, 1993] considered that part of the peak corresponds to 2-5% of the bearing ratio, the central 25-75%, valley 75-98%.

German researchers proposed the profile roughness description method described in DIN 4776 (and later ISO 13565-2). Nielsen thought that honing process can be controlled by changing of  $R_k$  parameter value [Nielsen, 1988]. Authors of the work [King, 1994] tried to determine the value of the five parameters from the group "Rk" on the basis of the value of the parameters  $R_{sk}$ ,  $R_{ku}$ ,  $R_q$ . It is possible only for certain types of ordinate distributions, which are characterized by small asymmetry.

Criticizing the method defined in DIN 4776, Zipin in his work [Zipin, 1983], discredited its usage for analysis of surface's profiles having Gaussian's ordinates distribution. Also the authors of the work [Malburg, 1993] express doubts as to the correctness of the determination of  $Mr_2$  parameter. It depends on the slope of the material

bearing ratio curve in its middle area. The parameters contained in ISO 13565-2 are used by most European manufacturers of internal combustion engines.

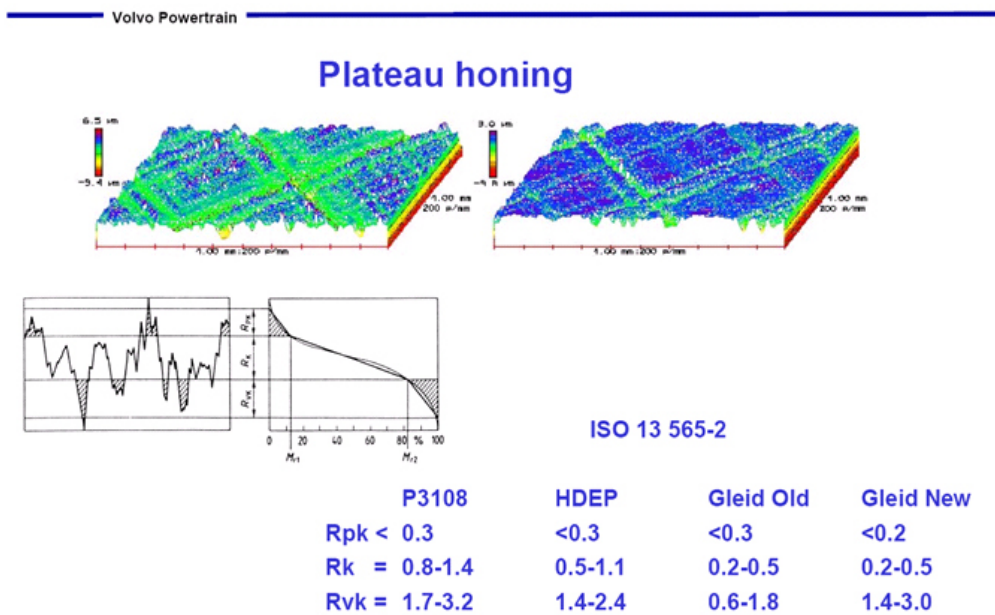


Fig 6. The values of cylinder liner surface parameters used by the Volvo company [Cieślak, 2008]

Figure 6 shows the requirements applied by the VOLVO company for compression ignition engines used in trucks. Authors in the work [Michalski, 1992] suggested a method based on the analysis of approximated material ratio curve. They applied the following equation (1):

$$R = 0.35[1 + (2/\pi) \arctg(A\{tg[(\pi/2)(2tp - 1)] - tg[(\pi/2)(2X0 - 1)]\})] + 0.3tg(Btp)/tg(B) \quad (1)$$

where A, X, B-independent parameters, R - standardized height of the roughness, tp - bearing ratio.

This approach allows you to specify the minimum and maximum curvature of the coordinates (xrk1, yrk1, xrk, yrk), coordinates of a point of inflexion bearing curve (xpp, ypp) and tangent of slope of this curve at that point (del). These parameters and the ones resulting from ISO 13565-2 are shown in Figure 7. The parameter yrk1 is analogous to yr1, yrk-yr2, ypp-pp, xrk1-Mr1, xrk- Mr2, del-Rk/Rt, Vok-Vo2.

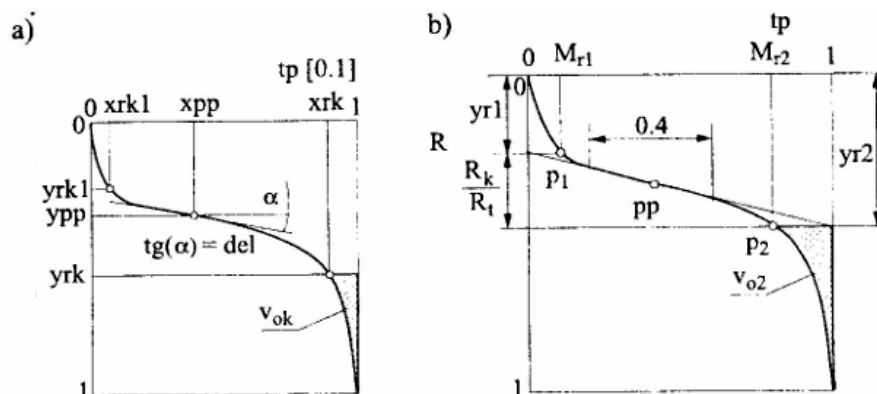


Fig 7. Some parameters of roughness resulting from the approximation of the Abbott curve, (a) and ISO 13565-2 standard (b) [Pawlus, 1999]

The authors of the publications [Malburg, 1993], [Sanna-Reddy, 1997] proposed a method of analysis the surface obtained after many processes. It was used in the American company Cummins producing internal combustion engines. This method is described in ISO 13565-3 standard (fig. 8).

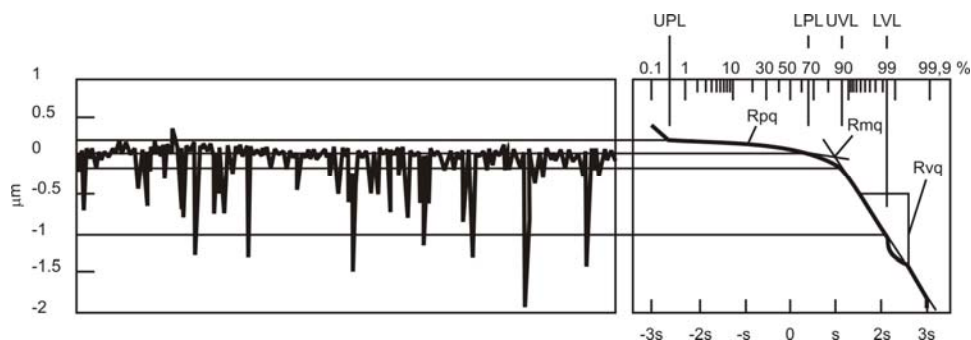


Fig. 8. Graphical interpretation of parameters contained in PN-EN ISO 13565-3: 2002 standard [Jakubiec, 2004]

This method called the probabilistic one, is based on an analysis of the data becomes from the bearing area curve plotted on normal probability paper. In this co-ordinate system, surface which has the Gaussian ordinate distribution is described by a straight line, but the surface after two-process honing, both of them having Gaussian ordinate distribution too, is described by two straight lines of different slopes. The intersection point of those lines in this graph according to the authors of the work [Malburg, 1993] separates the plateau and base texture.

In this graph the abscissa of this intersection point is defined as  $R_{mq}$  parameter (see fig.8) which is important feature of the model because it depends on the honing time.  $R_{pq}$  parameter is the slop of a regression line drawn through the plateau region, but  $R_{vq}$  – through the walleyes region.  $R_{pq}$ ,  $R_{vq}$  and  $R_{mq}$  are three parameters independently characterizing each area of plateau honed surfaces, therefore the honing process controlled with their use should be precise.

This method was recommended in the work: [Whitehouse, 1985] [Zipin, 1983], it was also used for the analysis of the zero wear process and to study running-in process. It is conceptually simpler and more elegant than the method described previously (ISO 13565-2), but the practical difficulty of the parameters  $R_{pq}$ ,  $R_{vq}$  and  $R_{mq}$  computations exists [Pawlus, 2009].

The method has been used to model the geometrical surface structure after two processes. The authors of the works: [Rosen, 2004], [Anderberg, 2009] analyzed the connections between parameters from ISO 13565-2 ( $R_k$ ) and ISO 13565-3 ( $R_q$ ) standards with honing process parameters. However, the conclusive answer to the question which group of parameters is more associated with the manufacturing process was not found. Parameters contained in ISO 13565 standard make a major contribution to an analysis of profiles after several processes. These parameters can be used also in 3D system. The analysis of the relative differences between parameters contained in the standards ISO 13565-2 and ISO 13565-3 and their three-dimensional counterparts is interesting. It can give the answer to the question whether these parameters describe the statistics (average) properties of the surface or they are susceptible to the presence of accidental valleys and peaks. In order to find the intersection point between plateau and valley areas, methodology described in ISO 13565-3 standard and other methods (for example [Sannareddy, 1998]) used different curves to the approximation probability plot of cumulative distribution. Those methods have some imprecisions which were noticed by the authors of the works [Jakubiec, 2004], therefore, the author proposes a different way of solving the problem. This method will be presented in the following subsection.

**Implementation**

To automatize the determination process of parameters  $R_q$  in plateau and valley area, computer program was created. This program was partly based on algorithm described in ISO 13565-3 standard. The main problem is to determine transition point between two random regions. To find this point material probability curve graph was rotated by  $\psi$  angle anticlockwise according to the following equation:

$$x' = x \cos \psi - y \sin \psi$$

$$y' = x \sin \psi + y \cos \psi$$

$\psi$  angle is the slope of straight line passing by the first and the finishing point of the material ratio curve (see Figure 9).

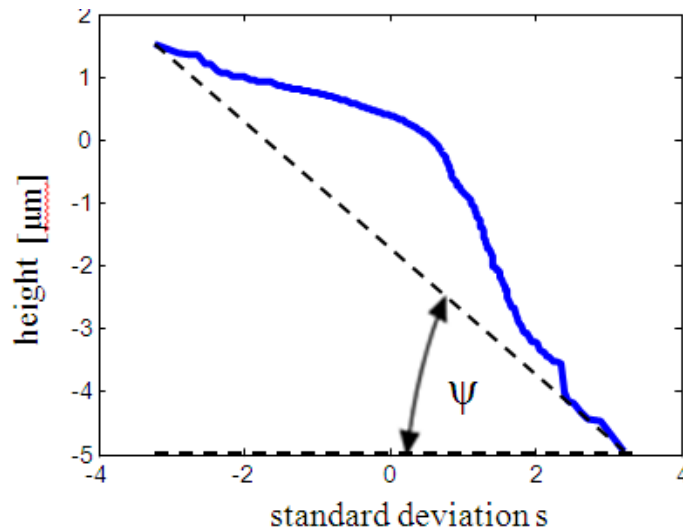


Fig.9. Material probability plot with straight line passing by the first and the finishing point and  $\psi$  angle. In rotated diagram C point of the highest ordinate was determined (see Figure 10). This point is treated as transition between two random regions.

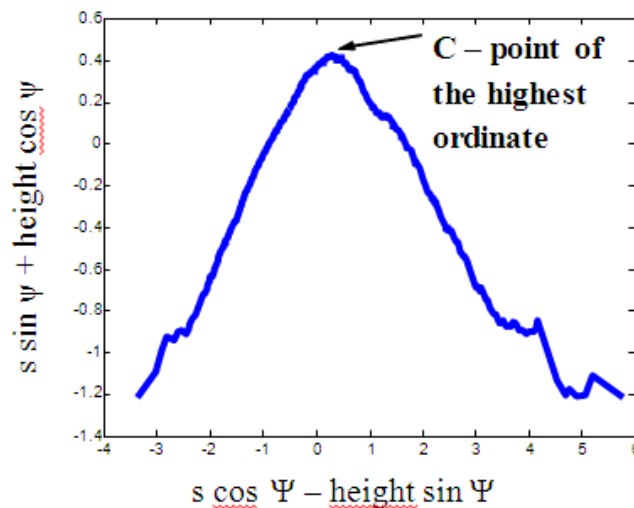


Fig.10. Material probability plot rotated by  $\psi$  angle

According to methodology which is recommended in ISO 13565-3 standard, nonlinear material probability curve graph regions were eliminated and Upper Plateau Limit (UPL), and Lower Valley Limit (LVL) points were assigned.

In material probability curve graph the lower boundary of the region plateau (LPL) and upper region valley (UVL) were determined by elimination of a few points which are situated partly right and partly left from transition point. The number of eliminated points was determined from the value of curvature material ratio curve in transition area. Afterwards, linear regression lines between points UPL and LPL and between points UVL and LVL were determined. The values of directional coefficients of these lines were assigned as values  $R_{pq}$  and  $R_{vq}$ . Parameter  $R_{mq}$  was assigned as value of abscissa in intersection of regressions lines drawn in plateau and valley region - Fig.11. Areal  $S_{pq}$ ,  $S_{vq}$  or  $S_{mq}$  parameters can be also determined in this way.

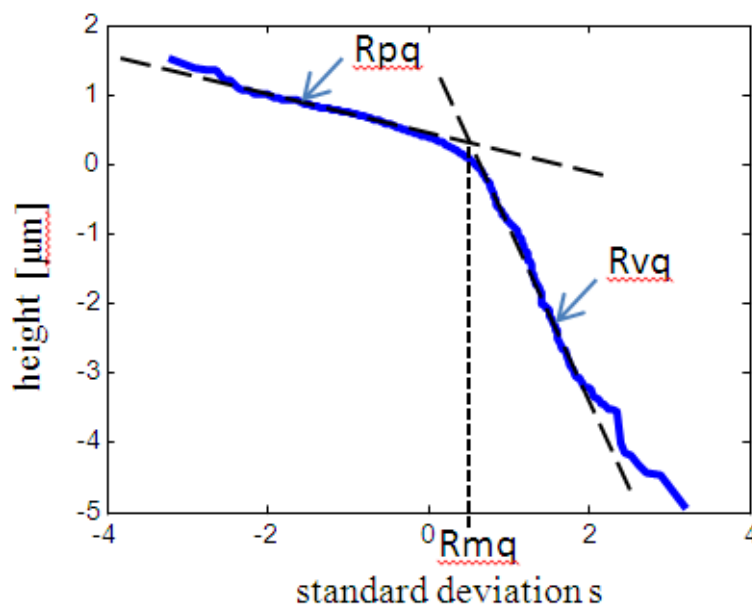


Fig.11. Material probability plot with regression lines passing through random regions and transition point between those regions

## Conclusion

Parameters describing roughness profiles of surfaces having stratified functional properties were proposed in the standards ISO 13565-2 ( $R_k$ ) and ISO 13565-3 ( $R_q$ ). Parameters from  $R_q$  group can be used to profile simulation and zero-wear evaluation or to control of manufacturing process. The number of those parameters is lesser than parameters from  $R_k$  group and they are not determined arbitrarily as in ISO 13565-2 standard. However, in European industry the parameters from ISO 13565-2 are used. Methodology of determining the parameters included in ISO 13565-3 standard is imprecise, therefore its incorrect usage can lead to significant mistakes. That is why different methodology to determine those parameters was suggested. After the analysis of many surfaces it was found that this method is useful.

## Bibliography

- [Abbott, 1993] E.J.Abbott, F.A.Firestone: Specifying surface quality. Mech. Eng., 55, 1993, 556-572.
- [Anderberg, 2009] C.Anderberg, P.Pawlus, B.G.Rosen, T.R.Thomas: Alternative descriptions of roughness for cylinder liner production. Journal of Materials Processing Technology 209, 2009, 1936-1942.
- [Barber, 1987] G.C.Barber, J.Lee, K.C.Ludema: Materials and surface finish effects in the breaking-in process of engines. ASME Journ. of Eng. for Gas Turbines and Power, 109, 1987, 380-387.
- [Campbell, 1972] J.C.Campbell: Cylinder bore surface roughness in internal combustion engines: its appreciation and control. Wear, 19, 1972, 163-168.



- [Cieślak, 2008] T.Cieślak: Wpływ wybranych parametrów procesu gładzenia na strukturę geometryczną powierzchni cylindrów. Rozprawa doktorska, Politechnika Rzeszowska, Rzeszów 2008.
- [Chusu, 1975] A.P.Chusu, Ju.R.Witenberg, W.A.Palmow: Szerochowatość powierzchni. Moskwa, Nauka 1975.
- [Day, 1986] R.A.Day, T.J.Reid, D.C. Evans: Developments in liner technology. AE Technical Symp, paper 2, New York, USA, 1986.
- [Duck, 1974] G. Duck: Zur Laufflachengestaltung von Zylindern und Zylinderlaufbahnen. MTZ, 35 (10), 1974, 339-340.
- [Dolecki, 1983] W.A.Dolecki, W.N.Buntów, A.K.Argusow, B.A.Opoczyński, M.A.Grigorjev: Zwiększenie trwałości maszyn metodami technologicznymi. WNT, Warszawa 1983.
- [Essig, 1990] G.Essig, H.Kamp, E.Wacker: Diesel engine emissions reductions - the benefit of low oil consumption design. SAE, paper 900591, 1990.
- [Gruszka, 1983] J.Gruszka: Badanie wybranych cech warstwy wierzchniej żeliwnych tulei w oparciu o próby zużyciowe. Rozprawa doktorska, Politechnika Poznańska, Poznań 1983.
- [Hassis, 1999] G.Hassis, U-P.Weigmann: New honing technique reduces oil consumption. Industrial Diamond Review, 3/99, 205-211.
- [Jakubiec, 2004] W.Jakubiec, J.Malinowski: Metrologia wielkości geometrycznych. PWN, Warszawa 2004.
- [Jakubiec, 2004] W. Jakubiec, M. Brylki: Validation of software for calculation the surface roughness parameters according to ISO 13565-3. Proceedings of 9th International Conference on Metrology and Properties of engineering Surfaces, Halmstad University, Sweden 2004, 77-84.
- [Jeng, 1996] Y.Jeng: Impact of plateaued surfaces on tribological performance. Tribology Transactions, 39/2, 1996, 354-361.
- [Johansson, 2008] S.Johansson, P.H.Nilsson, R.Ohlsson, C.Anderberg, B.-G.Rosen: New cylinder liner surfaces for low oil consumption. Tribology International, 41, 2008, 854-860.
- [King, 1994] T.G.King, N.E.Houghton: Describing distribution shape: Rk and central moment approaches compared. Proceedings of the 6-th Conference on Metrology and Properties of Engineering Surfaces, Birmingham, UK, 1994, 110-118.
- [Klink, 1997] L.Klink: Laserhonen für Zylinderlaufbahnen. MTZ, 58/9, 1997, 554-556.
- [Kozaczewski, 1986] W.Kozaczewski: Konstrukcja złożeń: tłok-cylinder silników spalinowych. WKŁ, Warszawa 1986.
- [Lenhof, 1997] U.Lenhof, A.Robota: Assessment of honed cylinder bore surfaces for IC engines. AE GOETZE GmbH, Braunschweig 1997.
- [Malburg, 1993] M.C.Malburg, J.Raja: Characterization of surface texture generated by plateau-honing process. CIRP Annals, 42/1, 1993, 637-640.
- [Michalski, 1992] J.Michalski, P.Pawlus: Description of the bearing length curve of the inner surface of piston engine cylinders. Wear, 157, 1992, 207-214.
- [Michalski, 1994] J.Michalski, P.Pawlus: Effects of metallurgical structure and cylinder surface topography on the wear of piston ring-cylinder assemblies under artificially increased dustiness conditions. Wear, 179, 1994, 109-115.
- [Michalski, 1998] J.Michalski: Badania poprawności odwzorowania nierówności powierzchni cylindrów po gładzeniu płasko-wierzchołkowym. Materiały konferencji: „Mechanika '98”, część 2, Rzeszów 1998, 149-158.
- [Michalski, 1994] J.Michalski, P.Pawlus: Description of honed cylinders surface topography. Int. J. Mach. Tools Manufact., 34/2, 1994, 199-210.
- [Nielsen, 1988] H.S.Nielsen: New approaches to surface roughness evaluation of special surfaces. Precision Engineering, 10, 1988, 209-213.
- [Niewiarowski, 1983] K. Niewiarowski: Tłokowe silniki spalinowe. WKŁ, Warszawa 1983.
- [Nosal, 1998] S.Nosal: Tribologiczne aspekty zacierania się węzłów ślizgowych. Rozprawy nr 328, Wydawnictwo Politechniki Poznańskiej, Poznań 1998.
- [Nowicki, 1991] B.Nowicki: Struktura geometryczna. Chropowatość i falistość powierzchni. WNT, Warszawa 1991.
- [Ohlsson, 2003] R. R.Ohlsson, B.-G.Rosen: Surface texture knowledge – ISM. In: Advanced Techniques for Assessment Surface Topography. L. Blunt, X. Jiang (Eds.), Kogan Page, London and Sterling 2003, 325-336.
- [Pawlus, 1994] P.Pawlus: A study on the functional properties of honed cylinder surface during running-in. Wear, 176, 1994, 247-254.
- [Pawlus, 2009] P. Pawlus, T. Cieslak, T. Mathia: The study of cylinder liner plateau honing process Journal of Materials Processing Technology 209.

- [Pawlus, 1999] P.Pawlus: Struktura geometryczna powierzchni cylindrów podczas eksploatacji silnika spalinowego. Oficyna Wydawnicza Politechniki Rzeszowskiej, Rzeszów 1999.
- [Rosen, 2004] B.G.Rosen, C.Anderberg, R.Ohlsson, Characterisation of cylinder roughness for manufacturing control and quality assurance. Addendum to Proceedings of XI International Colloquium on Surfaces, Chemnitz (Germany) 2004.
- [Sannareddy, 1998] H. Sannareddy, J. Raja, K.Chen: Characterization of surface texture generated by multi-process manufacture. Int. J. Mach. Tools Manufact. Vol 38. Nos 5-6, pp.529-536, 1998.
- [Sanna-Reddy, 1997] H.Sanna-Reddy, J.Raja, K.Chen: Characterization of surface texture generated by multi-process manufacture. Transactions of the 7th International Conference on Metrology and Properties of Engineering Surfaces, Gothenburg, Sweden, 1997, 111-118.
- [Santochi, 1982] M.Santochi, M.Vignale: A study on the functional properties of a honed surface. CIRP Annals, 31, 1982, 431-434.
- [Schmid, 1999] J.Schmid: Honing technology for optimal cast-iron cylinder liners. Engine Technology International, 1(2), 1999, 108-109.
- [Shin, 1983] K.Shin, Y.Tateishi, S.Furuhamu: Measurement of oil film thickness. SAE, paper 830068, 1983.
- [Sudarshan, 1983] T.S.Sudarshan, S.B.Bhaduri: Wear in cylinder liner. Wear, 91, 1983, 269-277.
- [Sreenath, 1976] A.V. Sreenath, N.Raman: Running-in wear of a compression ignition engine: factors influencing the conformance between cylinder liner and piston ring. Wear, 38, 1976, 271-289.
- [Sudarshan, 1983] T.S.Sudarshan, S.B.Bhaduri: Wear in cylinder liner. Wear, 91, 1983, 269-277.
- [Stout, 1990] K.J.Stout, E.J.Davis, P. I.Sullivan: Atlas of machined surfaces. Chapman and Hall, London 1990.
- [Schmid, 2006] J.Schmid: Optimized honing process for cast iron running surfaces. VDI Symposium "Piston running surfaces, pistons and conrods". Boblingen, Germany, Schmid.
- [Willis, 1986] E.Willis: Surface finish in relation to cylinder liners. Wear, 109, 1986, 351-366.
- [Wieczorowski, 1996] M.Wieczorowski: Stereometryczna ocena porównawcza powierzchni przy zastosowaniu różnych układów odniesienia profilometrów. Rozprawa doktorska Politechnika Poznańska, Poznań 1996.
- [Willn, 1972] J.E.Willn: Characterisation of cylinder bore surface finish: a review of profile analysis. Wear, 19, 1972, 143-162.
- [Whitehouse, 1985] D.J.Whitehouse: Assessment of surface finish profiles produced by multi-process manufacture. Proceeding of the Inst. Mech. Engrs, 199/4, 1985, 263-270.
- [Wiemann, 1971] L.Wiemann: Die Bildung von Brandspuren auf den Laufflächen der Paarung Kolbenring - Zylinder in Verbrennungsmotoren. MTZ, 32/2, 1971, 43-49.
- [Zipin, 1983] D.J.Zipin: The analysis of profile strata for surface texture specification. Appl. Surface Science, 15, 1983, 334-360.
- [Zwierzycki, 1990] W.Zwierzycki: Wybrane zagadnienia zużywania się materiałów w ślizgowych węzłach tarcia. PWN, Warszawa 1990.

---

## Authors' Information

---



**Wiesław Graboń** – Department of Computer Science, Rzeszow University of Technology; W. Pola 2, 35-959 Rzeszów, Poland ; e-mail: [wgrabon@prz.edu.pl](mailto:wgrabon@prz.edu.pl)

Major Fields of Scientific Research: Software engineering information systems, Modeling of manufacturing processes.